JMIR Perioperative Medicine

Technologies for pre- and post-operative education, preventative interventions, and clinical care for surgery and anaesthesiology patients, as well as informatics applications in anesthesia, surgery, critical care, and pain medicine

Volume 9 (2026) ISSN 2561-9128 Editor in Chief: Nidhi Rohatgi, MS, MD, SFHM

Contents

Reviews

AI Trustworthiness in the Perioperative Period for Patients with Serious Illness: Scoping Narrative Review (e89936)

Bhagvat Maheta, Rachel Ross, Isabella Raspi, Christina Keny, Marti Okech, Selen Bozkurt, Karleen Giannitrapani. 4

Patient Satisfaction With Perioperative Services and Associated Factors in Ethiopia: Systematic Review and Meta-Analysis (e84457)

Melesse Munie, Tegene Kitaw, Aregash Terefe, Alemu Zemariam, Molalign Adisu, Molla Azmeraw, Tesfaye Habtie, Betelhem Dagnaw, Abebe Faris, Tilahun Wodaynew, Amsalu Taye, Gebrie Yirga, Yabibal Desro. 21

Multidimensional Assessment of Recovery After Total Knee Arthroplasty in Clinical Practice: Critical Narrative Review (e84011)

Abderrahmane Boukabache, Nimalan Maruthainar, Vikrant Manhas, Darren Player. 36

Corrigenda and Addenda

Correction: Forced-Air Warming Temperature Settings for Treating Postoperative Hypothermia in the Postanesthesia Care Unit: Randomized Controlled Trial (e110801)

Koravee Pasutharnchat, Rattaphol Seangrung, Sirikarn Sirisophaphong, Wilailuck Wongkum. 52

Original Papers

Radiology-to-Surgical Consultation Timing for Suspected Appendicitis in the Emergency Department: Retrospective Observational Study (e91385)

Phillip Jenkins, Obert Xu, Shelby Willis, Ethan Kimball, Ruchi Thanawala, Beech Burns, Julie Doberne. 56

Forced-Air Warming Temperature Settings for Treating Postoperative Hypothermia in the Postanesthesia Care Unit: Randomized Controlled Trial (e85045)

Koravee Pasutharnchat, Rattaphol Seangrung, Sirikarn Sirisophaphong, Wilailuck Wongkum. 67

Virtual Reality for the Management of Postoperative Pain and Anxiety in Children and Adolescents Undergoing Nuss Repair of Pectus Excavatum: Randomized Controlled Trial (e80902)

Charlotte Walter, Dillon Froass, Nora Bell, Lauren Haack, Chloe Boehmer, Claudia Bruguera Torres, Rachel Spivak, Max Chou, Kristie Geisler, Keith O'Conor, Sara Williams, Lili Ding, Christopher King, Vanessa Olbrecht. 80

XSL•FO
RenderX

Evaluating the Impact of Virtual Reality on Orthopedic Trauma Skills Acquisition Among Surgical Residents: Randomized Crossover Study (e79343)
Sandipan Chatterjee, Khairul Mohammad, Monica Ghidinelli. 94

A Novel Customizable Datamart and Tableau Dashboard to Monitor Multiple Enhanced Recovery After Surgery Programs: Development and Validation Study (e82472)
Sunitha Singh, Susannah Oster, Efrat Bolze, Aaron Sasson, James Nicholson, Elliott Bennett-Guerrero. 104

Comparison Between Ultrasound and Magnetic Resonance Imaging Measurements of the Optic Nerve Sheath Diameter in Patients Undergoing Intracranial Surgery: Prospective Observational Single-Center Study (e67480)
Mauricio Giraldo, Luz Lopera, Aly Balbaa, Nelson Gonzalez, Raffael Cezar Zamper, Michael Mayich, Mel Boulton. 117

Survival Prediction in Patients With Bladder Cancer Undergoing Radical Cystectomy Using a Machine Learning Algorithm: Retrospective Single-Center Study (e86666)
Francesco Causio, Vittorio De Vita, Andrea Nappi, Melissa Sawaya, Bernardo Rocco, Nazario Foschi, Giuseppe Maioriello, Pierluigi Russo. .
1 2 5

Physician Perspectives on ChatGPT-4o as a Patient Resource for Abdominal Cancer Surgeries: Cross-Sectional Survey (e81374)
Christina Lindsay, Devika Shenoy, Allison Martin, Christie Clipper, Kevin Shah, Michael Lidsky, Daniel Nussbaum, Ralph Snyderman. 144

AI-Generated Avatar Videos for Postoperative Patient Education Among Health Care Workers: Pilot Randomized Controlled Trial (e89277)
Syed Haider, Srinivasagam Prabha, Cesar Gomez-Cabello, Ariana Genovese, Bernardo Collaco, Nadia Wood, James London, Sanjay Bagaria, Mark Lifson, Cui Tao, Antonio Forte. 158

Evaluating the Performance of Large Language Models in Vascular Surgery Cases: Cross-Sectional Study (e91973)
Asanka Wijetunga, Yunyi Wang, Chantell Yaghi, Mauro Vicaretti. 174

Machine Learning–Based Prediction of Major Postoperative Complications After Noncardiac Surgery Under General Anesthesia Using Routinely Available Electronic Health Record Data: Retrospective Multicenter Development and Validation Study (e91315)
Takuya Nishino, Hiroshi Mase, Chol Kim, Shinji Sugita, Yukihiro Kondo, Masashi Ishikawa. 192

Comparing the Quality of Patient Recovery After Open Appendectomy Under General Versus Spinal Anesthesia: Prospective Cohort Study (e86273)
Negesse Gobezie, Temesgen Asmare, Kumlachew Belete, Gezahagn Gedefaw, Biruk Demissie, Abere Abuhay, Belayneh Abate, Birhanu Munie, Wubet Wondie, Habtie Wubet. 211

Exploring Provider Perspectives on Gaps in Trauma-Informed Care Delivery Within the Acute Care Surgery Timeline at a Federally Qualified Health Center: Human-Centered Design Study (e81004)
Angelica McDaniel, Willow Frye, Neema Rashidi, Christiana von Hippel, Marianna Salvatori, Adrienne Greer, Manami Diaz Tsuzuki, Amanda Sammann. 221

Patient Blood Management Adoption and Transfusion-Threshold Concordance Among Romanian Anesthesiologists: National Survey Analysis (e106731)
Mihai Stefan, Daniela Filipescu, Dan Corneci, Cornelia Predoi, Silvius Negoita, Dorel Sandesc, Liana Valeanu, Serban Bubenek Turconi, Stefan Andrei, Dana Tomescu. 234

Enhancing the User Experience of a Perioperative Digital Health Tool for Information Exchange Using a Human-Centered Design Thinking Approach: Qualitative Observational Study (e79349)
Charlé Steyl, Carljohan Orre, Greg Foster, Hanel Duvenage, Michelle Chew, Hyla Kluyts. 247

XSL•FO
RenderX

Immersive Virtual Reality for Pain and Relaxation in Older Adults Following Elective Inpatient Abdominal Surgery: Single-Arm Study Examining Feasibility and Acceptability (e81791)
Christina Keny, Ujala Shafiq, Karl Lorenz, Marcia Russell, Heather Leutwyler, Laura Wagner, Victoria Tang, Linda Park. 268

Clinical Feasibility and Outcomes of Surgeon-Performed Laparoscopic-Guided Subcostal Transversus Abdominis Plane Block in Laparoscopic Cholecystectomy: Prospective Observational Study (e87622)
Sarun Mahasupachai, Thawatchai Tullavardhana. 283

Preoperative Mortality Risk Prediction and Equity Using Preoperative Score to Predict Postoperative Mortality in a Diverse US Surgical Population: Retrospective Cohort Study (e99623)
Jackson Hall, Haley Warzecha, Chase Penhallurick, Venkatsai Bellala, Shyamal Asher. 292

Association Between Complications and Death Within 30 Days After Orthopedic Surgery: Vascular Events in Noncardiac Surgery Patients Cohort Evaluation (VISION) Substudy (e90823)
Lily Park, P Devereaux, Ameen Patel, Vikas Tandon, Diane Heels-Ansdell, Lehana Thabane, Pablo Serrano, Matthew Chan, Wojciech Szczeklik, Sadeesh Srinathan, Ignacio Garutti, Gerard Urrutia, Ernesto Guerra-Farfan, Hassaan Abdel Khalik, Emmanuelle Duceppe, Sandra Ofori, Maura Marcucci, David Conen, Michael Wang, Jessica Spence, Daniel Tushinski, Kamal Bali, Anthony Adili, Vickas Khanna, Ana Tonelli, Francesca Mulazzani, Wenjun Jiang, Olufemi Ayeni, Gerard Slobegean, Theodore Miclau, Mohit Bhandari, Flavia Borges. 302

Assessing the Effects of eHealth Literacy and the Area Deprivation Index on Barriers to Electronic Patient Portal Use for Orthopedic Surgery: Cross-Sectional Observational Study (e72035)
Audrey Litvak, Nicholas Lin, Kelly Hynes, Jason Strelzow, Jeffrey Stepan. 316

Automated Identification of Surgical Site Infections From Electronic Medical Records: Retrospective Observational Predictive Modeling Study (e87896)
Arjun Chakraborty, Peter Tarczy-Hornoch, Dustin Long, Meliha Yetisgen. 327

In-House 3D Planning for Corrective Osteotomy of the Distal Radius: Prospective Cohort Study With 1-Year Radiological and Clinical Follow-Up (e101764)
Charlotte Stor Swinkels, Peter Axelsson, Katleen Libberecht, Per Fredrikson, Emilia Gryska, Anders Björkman. 340

Benchmark Development for Fundamental Arthroscopic Skills Using a Simulation-Based Training Program: Observational Study (e82723)
Eric Davis, Brianna Caraet, Robert Pedowitz, Gregg Nicandri. 352

Machine Learning for Predicting Critical Postoperative Interventions: Proof-of-Concept Study Using the INSPIRE Dataset (e65327)
Manan Shukla, Paul Fodor, Suresh Yelika, Nicholas Ahn. 360

Editorial

Advancing Perioperative Medicine With AI: Current Landscape and Research Priorities (e107155)
Nidhi Rohatgi, Piyush Mathur. 185

XSL•FO
RenderX

AI Trustworthiness in the Perioperative Period for Patients with Serious Illness: Scoping Narrative Review

Bhagvat J Maheta[1,2], MD; Rachel L Ross[3], PhD, MPH; Isabella Raspi[2], BA; Christina Keny[4,5], MHA, PhD, RN; Marti N Okech[3]; Selen Bozkurt[6], PhD; Karleen F Giannitrapani[2,3], MPH, PhD

[1]Department of Surgery, Northwestern University, 420 E Superior St, Chicago, IL, United States

[2]VA Palo Alto Health Care System, Palo Alto, CA, United States

[3]Department of Primary Care and Population Health, Stanford University, 3180 Porter Drive, Palo Alto, CA, United States

[4]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, United States

[5]School of Nursing, University of California, San Francisco, CA

[6]Department of Biomedical Informatics, Emory University School of Medicine, GA, United States

Corresponding Author:
Karleen F Giannitrapani, MPH, PhD
VA Palo Alto Health Care System, Palo Alto, CA, United States

Abstract

Background: Despite the promising potential of AI in the perioperative context, the rapid pace of development and diverse implementation warrant a thorough review to consolidate existing knowledge, identify gaps, and assess the use of trustworthiness principles in AI integration into the perioperative period for patients with serious illness.

Objective: The purpose of this study was to address deficiencies in the perioperative AI literature by elucidating the extent to which discussions of equity, ethics, and safety are incorporated, thereby establishing a foundation for the development of robust ethical guidelines for the safe and effective integration of AI in health care.

Methods: We searched PubMed, Embase, CENTRAL (Cochrane Central Register of Controlled Trials), and Scopus for studies published from 2010 to July 2024. We included studies that reported patient functional outcomes, occurred in the perioperative period (30 d before and up to 90 d after surgery), incorporated AI integration, and included patients with serious illness (defined as malignancy, advanced organ failure, frailty, dementia or neurodegenerative disease, or stroke). To ensure reliability and minimize bias, 2 independent reviewers screened all studies at the title or abstract and full-text stages; conflicts were resolved through team consensus. The abstraction form was developed iteratively and was tested through pilot abstractions. Any discrepancies identified during data extraction were resolved through discussion and consensus among the reviewers. The ROBINS-I (Risk of Bias Tool in Nonrandomized Studies of Interventions) tool was used to assess quality. Abstraction and risk assessment were conducted through a blinded, independent dual-review process. A narrative review was compiled from the identified studies.

Results: Of the 10,980 papers identified through the database searches, this review yielded 81 papers that met the inclusion criteria. Analysis of AI implementation strategies revealed foundational efforts toward equitable access, with 6 studies providing open-access tools and several more designing models with simple inputs suitable for low-resource settings (17 studies). Seven studies mentioned their commitment to transparency (eg, publishing code) to enhance safety and trust. However, significant ethical deficiencies persist, particularly regarding input data, as only 2 studies explicitly addressed racial or ethnic disparities, and concerns about lack of sample diversity (16 studies) and the omission of socially relevant features (5 studies) were frequently noted as limitations.

Conclusions: Machine learning for predictive analytics and other types of AI tools for surgical outcomes offer significant potential but require adherence to trustworthiness and safety principles to be clinically viable. Future research should prioritize adherence to guidelines for equity, ethics, and safety, conduct prospective studies, incorporate more external validation of AI models, and facilitate transparent monitoring and reporting of model performance to build clinician and patient trust and encourage broader health care system adoption.

(JMIR Perioper Med 2026;9:e89936) doi:10.2196/89936

KEYWORDS

AI; surgery; perioperative; systematic review

XSL•FO
RenderX

Introduction

Rationale

In the perioperative period, patients with serious illnesses such as malignancies, organ failure, or frailty face a diminished quality of life, and this can be exacerbated during the perioperative period, when patients face additional morbidity and mortality [1-3]. However, AI offers a transformative approach to address these challenges in the perioperative period [4]. By analyzing complex patient data, AI—specifically through methods like machine learning (ML), natural language processing (NLP), and artificial neural networks (ANNs)—can predict risks, optimize treatment plans, and assist in critical surgical decision-making and risk stratification [5]. This technology not only aids clinicians but also holds potential for providing personalized patient support and education [6]. Although these AI tools are already becoming integral to perioperative care, the integration of generative AI is still in its nascent stages.

Although the potential benefits of AI are significant, its integration into health care, particularly for vulnerable patients in the perioperative period, introduces crucial ethical considerations. The core ethical issues include data privacy—ensuring that sensitive patient information remains secure—and algorithmic bias, which can perpetuate and even amplify existing health disparities if the data used to train the AI are not representative [7]. There are also concerns about overreliance on technology, where clinicians may defer too much to AI recommendations without exercising their own critical judgment. The Department of Health and Human Services (HHS) characterizes AI trustworthiness into 3 categories: ethics, equity, and safety [8]. The American Medical Association's (AMA) 2025 governance framework expands trustworthiness by mandating explainable AI tools whose decision logic clinicians can interpret and communicate to patients more transparently [9,10]. The World Health Organization (WHO) emphasizes that trustworthy AI in health care must protect human autonomy, promote safety and well-being, ensure transparency and explainability, foster accountability, advance inclusiveness and equity, and uphold sustainability across the entire AI lifecycle [11]. This is especially important in the perioperative period for seriously ill patients, where the stakes are high, and decisions may need to be made rapidly [12]. A standard for AI ethics in the perioperative period, when patients are highly vulnerable, does not yet exist. It is imperative to establish a robust and appropriate standard to ensure the well-being and safety of patients and to uphold ethical integrity.

The rapid development and diverse implementation of AI in perioperative care have created a compelling need for a thorough review to consolidate existing knowledge, identify critical gaps, and assess the effectiveness and limitations of integrating perioperative AI for patients with serious illness. Although a growing body of literature validates the performance of various AI models, a significant gap remains in understanding the practical strategies for integrating these validated models into clinical workflows to tangibly enhance efficiency and patient outcomes [13,14]. Although there has been work conducted on perioperative AI integration, most of it has not focused on serious illness, and there is still more to uncover related to the ethical implementation of AI in these spaces [14]. Furthermore, there is a distinct lack of research dedicated to the crucial ethical and safety considerations of AI, as well as an exploration of how AI-enabled team augmentation is currently incorporated into the care environment. Anchoring the concept of perioperative AI trustworthiness within standard guidelines, such as HHS and WHO, reinforces the notion that surgical AI systems must not only achieve technical excellence but also demonstrate compliance with internationally accepted principles of ethical design, governance, and continuous quality assurance.

Objectives

The purpose of this study is to address these deficiencies by systematically reviewing the literature to better elucidate the extent to which current publications incorporate equity, ethics, and safety discussions. This research also aims to provide a comprehensive understanding of the current landscape of AI in perioperative care, establish a foundation for future research, and contribute to informing the development of robust, ethical guidelines to ensure the safe and effective integration of AI in health care.

Methods

Protocol and Registration

Our review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis Protocols (PRISMA) Statement and EQUATOR (Enhancing the Quality and Transparency of Health Research) guidelines [15,16]. We registered our protocol with the PROSPERO (International Prospective Register of Systematic Reviews) database under the registration number CRD42024608387 [17].

Eligibility Criteria

This review included peer-reviewed publications from 2010 to July 2024 that met the following criteria, organized by the population, intervention, comparator, outcomes, timing, and setting (PICOTS) framework (Table 1). *Inclusion criteria* were patients with serious illness over the age of 18, integration of AI in the perioperative period, reporting any patient quality outcomes, and being peer-reviewed and published in English (for feasibility). Studies that solely focused on radiology or pathology data were excluded from this study, as AI in these fields has been well characterized [18,19]. Studies that only used logistic regression, which is a traditional statistical modeling approach, were also uniformly excluded from this study for feasibility. Serious illness, as operationalized in this research, included the following categories: malignancy, advanced organ failure (eg, end-stage cardiac, pulmonary, renal, or hepatic disease), frailty with dementia or neurodegenerative disease (progressive, incurable conditions with functional decline, such as dementia or Parkinson disease), and stroke with significant residual impairment [1,2]. AI is defined as the field encompassing machine learning and generative AI, enabling systems to learn from data and generate new content, respectively [20-22]. The perioperative period was defined as

XSL•FO
RenderX

the period 30 days before and up to 90 days after surgery [23]. Patient quality outcomes were defined as both patient-reported measures of quality of life and surgical factors such as mortality, length of hospital stay, and postoperative complications [24,25]. Conversely, biological intermediaries, such as radiology scans and blood biomarker values, were excluded from the definition of patient quality outcomes [26].

Table . Population, intervention, comparator, outcomes, timing, and setting (PICOTS) eligibility criteria.

PICOTS	Eligibility criteria

Population	Inclusion
	• Patients with serious illness
	• Adults (18 y and older)
	Serious illness:
	• Malignancy
	  • Any type of malignant cancer. Benign cancers will not be included
	• Advanced organ failure
	  • Must be irreversible (unless patient gets an organ transplant) and chronic (no acute processes such as infection)
	  • Common examples: end-stage heart failure, COPD[a], cirrhosis, end-stage renal disease
	• Frailty and dementia or neurodegenerative disease
	  • Chronic, irreversible, progressive, and deteriorating conditions with no cure
	  • Multiple chronic conditions, significant functional impairments, and an increased risk of mortality
	  • Common examples: dementia, Parkinson disease
	• Stroke with significant residual impairment
	  • Major stroke and have experienced significant and persistent neurological deficits

Intervention	Inclusion • Describe the integration of AI in the perioperative period Exclusion • AI is only used for diagnostic radiology
Comparator	• N/A[b]

Outcomes	Inclusion
	• Patient quality outcomes reported in the study
	  • Patient-reported outcomes relating to quality of life
	  • Patient-demonstrated outcomes (eg, arm curl, chair stand, back scratch, chair sit, reach, walk test, overall cognition, memory, etc)

Timing	• Interventions with any follow-up period
Study characteristics	Inclusion • Peer-reviewed, published manuscript • English only (for feasibility)

[a]COPD: chronic obstructive pulmonary disease.

[b]N/A: not applicable.

Information Sources and Search

Our comprehensive literature search encompassed terms related to AI, the perioperative period, and serious illness, as detailed in Multimedia Appendix 1. The search was created through an interactive process using prior published search terms, as well as collaboration with experts in AI (CK, SB). We conducted searches across PubMed, Embase, Scopus, and the Cochrane Central Register of Controlled Trials (CENTRAL) for publications from 2010 to July 2024, yielding an initial 10,980 records, which were reduced to 8942 after removing duplicates.

Selection of Sources for Evidence

Our study team advisors included a PhD expert in team science, health care systems design, and health services research (KG), a PhD expert in patient engagement, organizational behavior,

XSL•FO
RenderX

and health services research (RR), and a PhD and registered nurse expert in AI and the perioperative period (CK). Screening involved 4 co-authors (BM, RR, IR, and MO), each independently reviewing studies at the title or abstract and full-text stages, with reviewers blinded to each other's assessments. Discrepancies during title or abstract screening were primarily resolved by a designated "gold standard" reviewer (BM or KG), with occasional recourse to majority consensus. At the full-text screening stage, exclusion reasons were systematically recorded, and all conflicts were adjudicated solely by the "gold standard" reviewer (BM or KG). If a study did not specify what percentage of their sample size was diagnosed with malignancy compared to benign pathology, the corresponding author was contacted. If the information was not given within 3 weeks, the study was rejected. Three studies were rejected because of this. Furthermore, any systematic reviews fulfilling the inclusion criteria were used to identify additional relevant studies, which were then incorporated into the title or abstract screening phase. The study selection process, including the number of studies at each stage, was documented using Covidence software to generate a PRISMA flow diagram [27].

Data Charting Process

The abstraction form was created through an iterative process (Multimedia Appendix 2). This form collected data on participant characteristics, including demographics such as gender, age, and race or ethnicity, as well as the category of serious illness (malignancy, advanced organ failure, frailty or dementia, or stroke). Information regarding the AI intervention was abstracted, including the AI's purpose in the study (prognostication, screening, decision support, etc), the type of AI used (generative AI or ML), and the specific AI interface or tool. Information related to the validity of the AI model was not abstracted, since the primary objective of this review was to better understand equity, ethics, and safety discussions within the studies, and the focus was on implementation of the AI model rather than its validity. Additionally, the form captured whether the study addressed AI equity, ethics, and safety, and how AI was used within the health care setting, such as a team member or a prompting tool. Data abstraction for the included studies was conducted using Google Sheets, with 2 reviewers per paper. Reviewers (either BM, RR, IR, or CK) independently abstracted each paper, were blinded to each other's work, and resolved all abstraction conflicts through a consensus discussion.

Risk of Bias and Quality Assessment

The ROBINS-I tool was used during the data extraction phase to evaluate the quality of each included paper [28], since it allows for risk of bias stratification for nonrandomized studies that examine the effect of an intervention on an outcome. The research team assessed potential biases across the following domains: confounding, selection of participants into the study, classification of interventions, deviations from intended interventions, missing data, measurement of outcomes, and selection of the reported result. This systematic assessment involved 2 independent reviewers (among BM, RR, IR, and CK), who evaluated the risk of bias across multiple domains. Disagreements between the 2 reviewers were resolved through discussion and consensus.

Synthesis of Results

Given the substantial heterogeneity observed across included studies, stemming from variations in AI implementation, use, and reported patient quality-of-life outcomes, a narrative synthesis of the abstracted data was conducted using a grouping method. We synthesized studies based on malignancy type, AI type, as well as the incorporation of AI equity, ethics, and safety. Studies that included equity, ethics, and/or safety related to AI, as defined in the sections below, were included in this review, and details on these topics were systematically extracted from the manuscripts. The type of AI used was categorized based on whether the individual tools were supervised ML, unsupervised ML, supervised deep learning, unsupervised deep learning, NLP, or linear programming. The definitions used are listed in the sections below. The validity of AI was not considered as part of this analysis. For studies that included AI equity, ethics, and safety, the quotes were extracted and open-coded with dual review until saturation of codes was reached (RR, BM) [29,30]. Then the quotes were consolidated into appropriate themes through open coding and dual review. AI equity, ethics, and safety were defined in the following section.

Data Items

The following definitions were used to guide the identification and extraction of data related to AI equity, ethics, and safety:

- AI equity: The fair and just distribution of the benefits and opportunities afforded by AI technologies across all populations, regardless of race, ethnicity, gender, socioeconomic status, geographic location, or other social determinants of health, while also mitigating any potential harms or biases that could exacerbate existing health disparities [31]
- AI ethics: Ensuring that the development, implementation, and use of AI adhere to the fundamental ethical principles of beneficence, nonmaleficence, autonomy, justice, and explainability, while mitigating potential risks, such as biased screening data, to promote equitable and trustworthy health care outcomes [31]
- AI safety: The development and implementation of measures and practices that ensure that AI systems used in medical settings are reliable and robust and function as intended, minimizing the risk of harm to patients, clinicians, or the health care system [32]

Definitions of AI

The following definitions were used to categorize the types of AI identified in the included studies:

- Generative AI focuses on creating new content, such as text, images, or synthetic patient data. It learns the underlying patterns and structures of the data it is trained on and then uses that knowledge to generate entirely new content that is similar in style and characteristics to the training data [33].

XSL•FO
RenderX

- ML algorithms learn patterns from data without explicit programming, enabling them to make predictions or decisions [34].
- NLP focuses on enabling computers to understand, interpret, and generate human language [35].

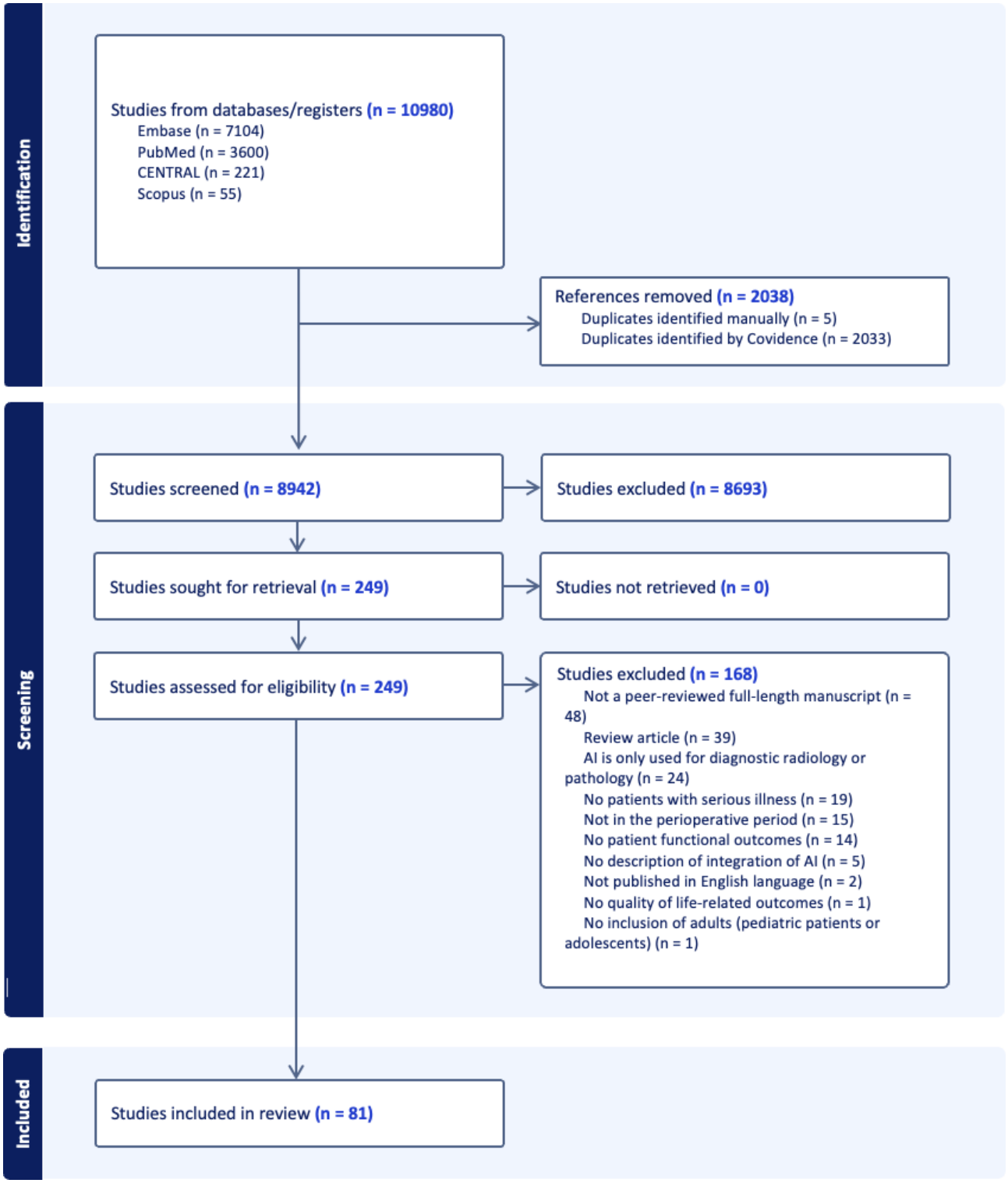
Results

Selection of Sources of Evidence

A total of 10,980 records were considered for review by the research teams. After 2038 (18.6%) papers were removed as duplicates, 8942 were screened at the title or abstract stage. Of 8942 papers, 8687 (97.1%) of those were excluded, and 249 were screened at the full-text stage. Ultimately, 81 studies were eligible for review, as presented in Figure 1 [36-116]. Overall, the included studies were assessed as having low-to-moderate risk when evaluated with the ROBINS-I tool, suggesting that the overall quality of the included papers was high.

XSL•FO
RenderX

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analysis Protocols (PRISMA) flowchart.



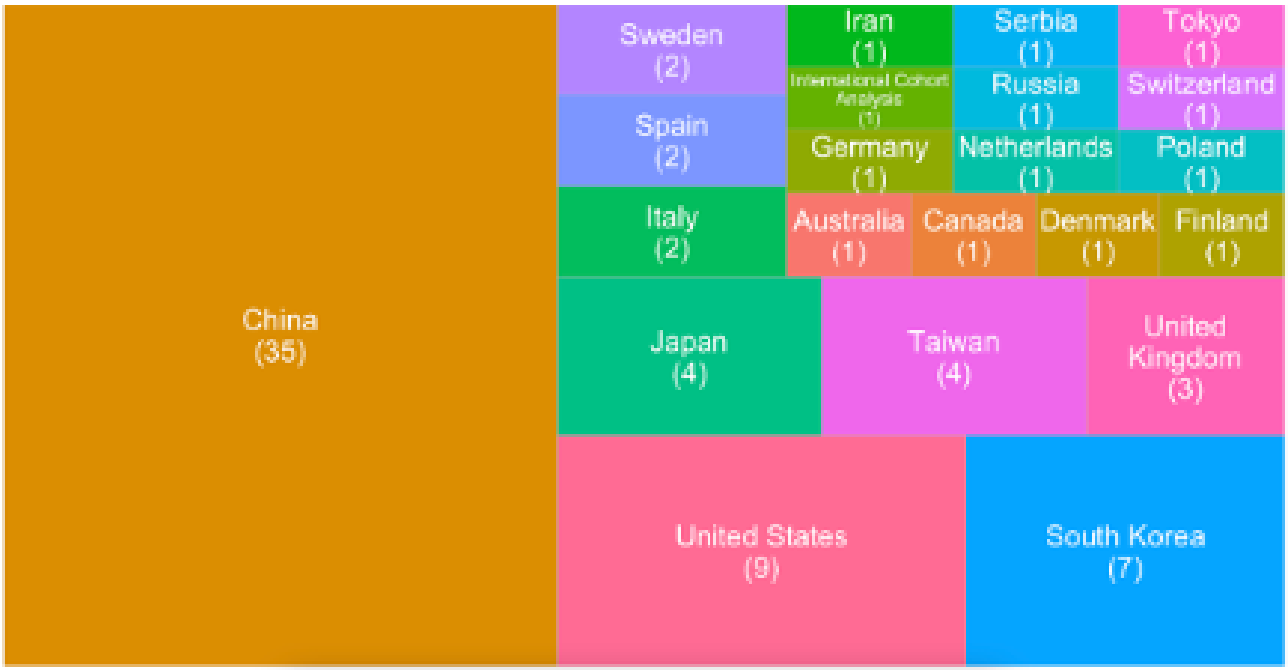
Characteristics of Sources of Evidence

Across all studies, there was an overall total of 714,253 participants. A majority of the studies were published after 2022, with 59 studies published during or after 2022 and 22 studies published prior to 2022. A majority of the studies were published in China (35 studies), with the United States (9 studies) and South Korea (7 studies) having the next highest numbers of publications, as shown in Figure 2. Of the 81 included studies, 80 (98.8%) focused on patients with malignancy, suggesting a growing interest and importance of AI integration within oncology specifically. The most common forms of cancer in the included papers were colorectal cancer (15 studies), hepatic cancer (11 studies), and head and neck cancer (9 studies). One study focused on patients with liver failure requiring surgery, without a primary malignancy in all patients [109]. A majority of studies used AI for prognostication (73 studies), with 4 using it for risk scores, 3 for treatment recommendations, and 1 using AI for decision support. All studies used ML as their form of AI and included different models, including support vector

XSL•FO
RenderX

machine (SVM), least absolute shrinkage and selection operator (LASSO) regression, and extreme gradient boosting (XGBoost). Of the included papers, 75 were retrospective studies, and 6 were prospective cohort studies. A comprehensive overview of relevant study characteristics can be found in Multimedia Appendices 3 and 4.

Figure 2. Country representation of included studies by frequency.



Results of Individual Sources of Evidence and Synthesis of Results

AI Equity

Consideration of Model Use in Low-Resource Settings (n=17)

Seven study teams addressed the potential for their models to be deployed in low-resource settings. Several study teams explained that model inputs were intentionally selected for their ease of acquisition in standard patient care settings (as opposed to data points that may require costly equipment to obtain or advanced analytical software to calculate) [46,59,84,91]. Two study teams specifically called out the low-cost nature of their model development [84,109], and one referenced their model's convenience and potential for widespread distribution relative to comparable existing options [103]. One study team originally stated that their ideal model application would be an electronic health record (EHR)–integrated tool, but included a statement about the potential for the creation of a standalone mobile application for use in low-resource settings that may not have robust information technology infrastructure [45]. These discussions reflect an effort to mitigate structural inequities in technological capacity and underscore the importance of context-aware model design for advancing AI equity in diverse care environments.

Distribution of an Open-Access Tool (n=6)

Six of the studies included in our sample described the creation of a web-based tool that would allow clinicians outside their organizations to implement the model and generate insights for their own patients [41,61,6278, 86,91], reinforcing a commitment to equitable access to these emerging technologies.

Design Considerations for Non–AI-Expert Clinicians (n=4)

Usability for frontline clinicians, regardless of their expertise in AI model development, was explicitly discussed in 4 studies. These authors described human-centered design choices, such as clear instructions, intuitive interfaces, and simplified input-output mechanisms to facilitate real-world implementation of these tools and to ease communication between patients and their providers [41,84,116]. One study focused on how their tool could be embedded easily into clinical workflows by creating publicly available software that could be incorporated into an EHR for individualized risk prediction [45]. These features reflect a growing recognition that equitable deployment of AI depends not only on tool accuracy but also on ensuring that all clinicians, regardless of AI fluency, can effectively incorporate these models into patient care.

Consideration of the Cost Burden to the Patient Resulting From AI Decision Output (n=1)

One study explicitly addressed the potential downstream cost implications of AI-enabled decision-making for patients: Qiao et al [89] cautioned clinicians to interpret model outputs in a manner that avoids unnecessary or overly burdensome interventions, particularly in cases where the AI-generated risk assessments might influence decisions around costly diagnostic or treatment options. This finding highlights an emerging but underdeveloped area of AI equity: namely, the affordability of care pathways shaped by algorithmic outputs. Without attention to the financial consequences of model-informed decisions, there is a risk of exacerbating disparities among patients already facing economic hardship.

XSL•FO
RenderX

AI Ethics

Lack of Patient Sample Diversity (n=16)

Many studies were conducted at a single site and/or involved a retrospective study design, inherently introducing the potential for selection bias and limiting generalizability. This was frequently noted via a brief acknowledgment in the limitations sections of the papers included in our review. However, 16 study teams discussed concerns about selection bias and lack of generalizability more robustly. Our findings indicate that authors primarily relied on 3 techniques to reflect on concerns about lack of patient sample diversity: calling out the specific methodological issues leading to these worries, explaining specific techniques used to address limitations in diversity, and identifying the need for further exploration beyond the scope of the existing work to further validate their models and thereby address these ethical considerations.

Six study teams mentioned specific considerations driving their concerns regarding a lack of representativeness of the study population, ranging from a lack of geographic and demographic diversity in the sample [43,77,89], limited sample sizes [44], and lack of external validation in additional populations [66], to the overrepresentation of patient data from highly resourced settings (eg, hospitals with infrastructure to uphold NSQIP (National Surgical Quality Improvement Program) reporting requirements [61] or from exclusively tertiary care settings [68]).

Other authors mentioned intentional steps taken to enhance the diversity of their analytic sample, including the intentional inclusion of patients from multiple institutions [41,80], strict study recruitment screening guidelines designed to address concerns about population homogeneity [58], and the intentional selection of a patient population that is currently underrepresented in clinical trials [90]. In the manuscript, 1 study team included an explanation of the technique they used to address class-imbalance problems, leading to improved data quality and representativeness of real-world settings [88]. A handful of other studies mentioned concerns about lack of generalizability more generally but also included a call for future work with more diverse populations, inherently implying the need to address ethical considerations of model dissemination [45,50,53,92].

Omission of Socially and Clinically Relevant Features Limiting Model Accuracy (n=5)

Concerns about missing or incomplete data elements were discussed in a small subset of studies, primarily in the context of limitations in model performance. For example, several study teams highlighted the absence of key variables, such as potentially relevant clinical characteristics, smoking status, insurance status, and other socioeconomic factors, which have the potential to meaningfully impact model performance [47,57,74].

One study team mentioned that various methodological options are available to impute missing values, which could help avoid the loss of important information [86]. Relatedly, Kadomatsu et al [60] noted that the inclusion of subjective variables (ie, those requiring clinician judgment) may introduce variability or distortion in the dataset, further complicating model accuracy and reliability.

Discussion of Health Disparities (n=2)

Only 2 studies included in our review explicitly discussed racial and ethnic disparities in the context of model development or evaluation, despite the clear relevance of these considerations. Osman et al [86] specifically mentioned differences in mortality rates across racial and ethnic groups that are observed in national trends; they highlighted that their model performed consistently across datasets with differing baseline characteristics, thereby improving faith in the model's robustness. Verma [101] acknowledged that their model's output identified differences in risks across racial and ethnic groups and called out the need to both understand the mechanism of action behind this finding and to prevent the model from exacerbating existing disparities.

AI Safety

Removing the AI Black Box for Transparency (n=7)

AI safety included clarity and transparency about how the algorithm used the data to generate output and provided insight into what is sometimes thought of as the "black box." Certain studies emphasized the importance of sharing how the AI algorithm generated results through what they termed a "white box," since they felt it provides highly interpretable results [55]. This has been echoed by the AMA's emphasis on explainability and clinician interpretability. Others have emphasized transparency by publishing their code publicly so that others can easily see how the AI model was trained to produce results, and so that the results can be reproduced in other settings [78]. Other studies have argued that increased insight into how AI created the results was important for transparency and to assist surgeon decision-making [79].

AI Considers Additional Factors That May Influence Patient Care (n=6)

Although AI can be a powerful tool, it cannot be used in every setting without clinician input, and the final decision regarding patient care should not be made solely by AI. In addition to clinical factors that affect patient outcomes, there are certain socioeconomic factors that can significantly impact a patient's care that are not fully accounted for in current iterations of AI [60]. Multiple studies have emphasized that although there is great benefit to incorporating AI into serious illness perioperative care, the provider should still make the final decision regarding the patient's care [43,65].

Physical Safety From Additional Testing

Harnessing AI algorithms may also pose a risk of physical harm to patients, threatening AI safety. Liu et al [77] showed that, although there were benefits to incorporating AI, patients were exposed to high radiation from positron emission tomography (PET) scans and incurred increased costs without significant benefit.

AI-Enabled Team Augmentation

Despite growing attention toward the consideration of AI as true "team members," none of the included studies used this lens in the narrative discussion of how the tools were being

XSL•FO
RenderX

included in workflows [117]. However, the majority of papers described how the AI tool under consideration could be used to improve perioperative care by prompting a provider, team member, or patient/caregiver to take a tangible action.

There were 2 studies in which the AI model specifically prompted the provider in terms of patient care. Wang et al [104] used their model to provide personalized recommendations and solutions that the physicians could then act on as part of the study. Zhang et al [113] crafted a study design where their AI model assisted in distinguishing gallbladder carcinoma from xanthogranulomatous cholecystitis in the preoperative setting, thus directly impacting surgical decision-making in the perioperative period. Seventy-seven of the other included studies did not explicitly use their AI model to prompt providers, team members, or patients/caregivers, but described how their model could be incorporated in health care settings to prompt others and lead to real-time action contributing positively to patient care in the perioperative setting.

Discussion

Summary of Evidence

Our study systematically examined the extent to which scholarly work on AI in perioperative care for patients with serious illness explicitly incorporates discussions of equity, ethics, and safety. Only low-to-moderate risk studies, per the ROBINS-I risk of bias tool, were included in this review. However, the ultimate goal of these applications is to use AI-enabled team augmentation that contributes positively to clinician workflow, improves decision-making speed, and enhances overall patient safety. Although numerous studies have focused on the statistical and methodological validation of AI tools in this setting, our findings reveal a significant gap: key ethical considerations—particularly concerning vulnerable patient populations—remain largely unaddressed [14]. Specifically, we noted that limitations in sample diversity and generalizability, often framed by authors in purely methodological terms, carry profound ethical implications for justice and beneficence. By risking unequal performance across diverse patient subgroups and obscuring systemic disparities through the omission of socially and clinically relevant features, these limitations directly challenge the principles required for fair and trustworthy decision-making in real-world clinical applications. Although this study focused primarily on oncology, many core tenets related to the ethics of integrating AI can be extrapolated to other types of serious illnesses, such as advanced organ failure, frailty, dementia, neurodegenerative diseases, and stroke. Therefore, our analysis underscores the critical need for a paradigm shift, moving the conversation beyond technical statistical performance to demand the integration of rigorous, equity-focused analyses into both the validation and the broader implementation planning of AI tools used in the sensitive domain of perioperative patient care by explicitly including equity, ethics, and safety within the implementation of perioperative AI.

The integration of AI into the perioperative care of seriously ill patients presents unique and critical ethical and safety challenges. As a highly vulnerable population, these patients stand to benefit from AI's ability to analyze complex data for refined decision-making and personalized care, but they are also at increased risk of the potential for algorithmic bias and data privacy breaches [118]. The application of AI must be carefully considered, distinguishing between a patient-facing chatbot and other leveraged forms of AI that may also improve patient care [119]. Ensuring the safe deployment of AI is paramount and necessitates a multifaceted approach, including clearly defined safety parameters, robust risk mitigation strategies, and integrated workflows with essential human oversight [120,121]. Given the transformative potential of AI to fundamentally reshape perioperative care, ensuring robust ethical standards is paramount to safeguard patient safety and maintain trust in these technologies [122]. Currently, a gold standard for addressing the equity, ethics, and safety of AI in the perioperative period is absent from most research articles, a gap that must be addressed in all future manuscripts to align with emerging guidelines such as the US Department of Health and Human Services' Trustworthy AI Playbook or the AMA's 8-Step AI Governance Toolkit [8,10,123]. This review helps define the aspects of AI trustworthiness through the definition mentioned in the *Methods* section and encourages all future papers using AI in the perioperative period to explicitly report these criteria to ensure all AI models are held to a rigorous standard for equity, ethics, and safety.

While no existing studies actively use AI as a direct team member, the opportunity to integrate AI in this capacity remains a burgeoning field. The development of trust within human-AI teams is a complex process, and recent research shows that while smaller human-AI teams are initially seen as less trustworthy than human-only groups, this trust deficit diminishes as the team grows, suggesting that team complexity plays a key role in developing trust [124]. AI-powered systems, including chatbots, large language models (LLMs), and agentic AI, are poised to function as effective members of the health care team, either by partnering with providers to deliver clinical decision support or by engaging with patients to facilitate communication, education, and guidance [6,125]. Although minimal literature exists on integrating AI-enabled team augmentation in health care, and none specifically in the context of surgery, AI has the potential to prompt a team member, whether a clinician or a patient. This capacity to prompt offers significant potential to enhance surgical care for seriously ill patients through improved risk stratification, personalized education, and optimized monitoring. However, its successful implementation hinges on addressing critical concerns related to algorithmic bias, data privacy, and patient trust, while also tackling practical challenges related to integration, overreliance, and ensuring patient comprehension, particularly within this vulnerable population. In addition, trust calibration between humans and AI improves after exposure and iterative feedback, suggesting that increased engagement with AI by health care teams may lead to more AI-enabled team augmentation [126].

Moving beyond traditional machine learning models, generative AI presents both unique opportunities and significant pitfalls in the perioperative period. Generative AI offers the potential to create synthetic patient data for enhanced preoperative risk modeling and simulation, allowing for more personalized

XSL•FO
RenderX

surgical planning in complex cases of serious illness [127]. However, this necessitates careful validation to ensure these synthetic datasets accurately reflect real-world patient variability and avoid perpetuating biases [127]. Although generative AI can produce tailored patient education materials and postoperative instructions, improving patient understanding and adherence, there is a risk of generating inaccurate or misleading information, requiring rigorous oversight and quality control to maintain patient safety—a risk not as prominent as with more static ML models [128]. Furthermore, generative AI can streamline administrative tasks, such as generating discharge summaries and patient follow-up plans—a benefit not yet realized in ML—thereby freeing clinicians to focus on direct patient care. Despite these opportunities, critical ethical concerns remain regarding data privacy and the potential for algorithmic bias to impact resource allocation and patient prioritization, underscoring the need for careful and deliberate implementation.

Limitations

This review should be considered with the following limitations. Our broad definitions, used to capture a comprehensive range of AI applications, may have inadvertently led to the exclusion of relevant studies, as the terminology in this constantly evolving field lacks standardization. There were few studies focusing on the workflow integration of AI, which represents a lack of evidence in this space and demonstrates an opportunity to develop this field of research. Due to feasibility constraints, we were unable to conduct a continuous search, and thus, our findings represent a snapshot of a rapidly developing landscape, potentially missing the most recent publications. A significant limitation in the current literature is the scarcity of data on the practical integration of AI models into clinical workflows and a notable absence of discussion regarding ethics, equity, and safety. Furthermore, our study did not analyze the validity or performance of the AI models themselves, focusing instead on their application and integration. Finally, a significant number of included studies were published in China (35 studies), and additional studies published in languages other than English may have been excluded from this study. This limited geographic and demographic diversity may affect generalizability and inherently create a risk of algorithmic bias. Future research should prioritize the exploration of AI-enabled team augmentation and analyze clinical workflows to better understand where and how AI is being integrated into the perioperative period across different regions of the world and in different health care settings. Since a significant majority of papers were in the realm of oncology, further research on AI integration should focus on populations with other serious illnesses. There should also be longer-term follow-up of AI models to truly understand their impact and safety concerns, along with studies that focus on external validation, prospective implementation, and real-world workflow evaluation so that stronger conclusions can be drawn about clinical utility [129]. It is also crucial to explore the perspectives of patients with serious illnesses and their clinicians regarding the ethical implications of AI in perioperative care, examining their concerns, expectations, and preferred levels of AI integration to ensure that future implementation plans are patient-centered and ethically sound.

Conclusions

This investigation highlights the considerable potential of AI tools for predictive analytics and other AI tools in surgical outcomes while simultaneously exposing a significant deficit in the explicit reporting of ethical considerations, equity, and safety in current publications in the perioperative period. Among the 81 included papers, 80 of which focused on malignancy, there were significant ethical deficiencies, as very few studies mentioned key factors such as racial or ethnic disparities, sample diversity, or socially relevant features. Achieving the clinical viability of these tools requires moving beyond initial statistical reporting to incorporate rigorous external validation, continuous performance monitoring, and fundamental adherence to principles of trustworthiness and safety. The evidence suggests that unless researchers prioritize equity-focused design and data transparency, the risk of perpetuating structural health disparities within this vulnerable perioperative population remains high. To address these critical issues, it would be beneficial for future research to uphold the gold standard of including validation sets, a practice currently lacking in many published studies, and to demand comprehensive data transparency, including information on patient race and ethnicity. By prioritizing prospective studies and transparent reporting, the research community can build the trust necessary for adoption, aligning AI innovation in perioperative care with the broader HHS goals for improving health outcomes through innovative, fair, safe, and ethical technology.

Acknowledgments

This paper was presented at AcademyHealth's 2025 Annual Research Meeting in Minneapolis, MN, in June 2025. This study would not have been possible without the support of the Stanford University School of Medicine and the VA Center for Innovation to Implementation. Generative AI was not used in any portion of the manuscript writing.

Funding

KFG is supported by a VA Career Development Award (19-075).

Data Availability

The data sets analyzed during this study are publicly available as detailed in the Methods section.

XSL•FO
RenderX

Authors' Contributions

Data analysis: BJM, RLR, IR, CK, MNO, SB, KFG
Data collection: BJM, RLR, IR, MNO, KFG
Project ideation: BJM, RLR, CK, KFG
Overall project oversight: KFG
Writing – original draft: BJM, RLR
Writing – review & editing: BJM, RLR, CK, MNO, SB, KFG
All authors read and approved the final version of this manuscript.

Generative AI was not used in any portion of the manuscript writing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 7 KB - periop_v9i1e89936_app1.docx]

Multimedia Appendix 2
Abstraction form and ROBINS-I tool.
[DOCX File, 4846 KB - periop_v9i1e89936_app2.docx]

Multimedia Appendix 3
Included study characteristics.
[DOCX File, 20 KB - periop_v9i1e89936_app3.docx]

Multimedia Appendix 4
Structured synthesis table of included studies.
[DOCX File, 20 KB - periop_v9i1e89936_app4.docx]

References

1. Sanders JJ, Curtis JR, Tulsky JA. Achieving goal-concordant care: a conceptual model and approach to measuring serious illness communication and its impact. J Palliat Med 2018 Mar;21(S2):S17-S27. [doi: 10.1089/jpm.2017.0459] [Medline: 29091522]
2. Murray SA, Kendall M, Boyd K, Sheikh A. Illness trajectories and palliative care. BMJ 2005 Apr 30;330(7498):1007-1011. [doi: 10.1136/bmj.330.7498.1007] [Medline: 15860828]
3. Makary MA, Segev DL, Pronovost PJ, et al. Frailty as a predictor of surgical outcomes in older patients. J Am Coll Surg 2010 Jun;210(6):901-908. [doi: 10.1016/j.jamcollsurg.2010.01.028] [Medline: 20510798]
4. Hassan AM, Rajesh A, Asaad M, et al. Artificial intelligence and machine learning in prediction of surgical complications: current state, applications, and implications. Am Surg 2023 Jan;89(1):25-30. [doi: 10.1177/00031348221101488] [Medline: 35562124]
5. Balch JA, Shickel B, Bihorac A, Upchurch GR, Loftus TJ. Integration of AI in surgical decision support: improving clinical judgment. Global Surg Educ 2024;3(1):56. [doi: 10.1007/s44186-024-00257-2]
6. Lee JC, Hamill CS, Shnayder Y, Buczek E, Kakarala K, Bur AM. Exploring the role of artificial intelligence chatbots in preoperative counseling for head and neck cancer surgery. Laryngoscope 2024 Jun;134(6):2757-2761. [doi: 10.1002/lary.31243] [Medline: 38126511]
7. Morley J, Machado CCV, Burr C, et al. The ethics of AI in health care: a mapping review. Soc Sci Med 2020 Sep;260:113172. [doi: 10.1016/j.socscimed.2020.113172] [Medline: 32702587]
8. Trustworthy AI (TAI) playbook executive summary. : U.S. Department of Health & Human Services; 2021 URL: https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook-executive-summary.pdf? [accessed 2026-09-05]
9. Johannssen A, Chukhrova N. The crucial role of explainable artificial intelligence (XAI) in improving health care management. Health Care Manag Sci 2025 Sep;28(3):565-570. [doi: 10.1007/s10729-025-09720-y] [Medline: 41026402]
10. Henry TA. 8 steps to position your health system for AI success. American Medical Association (AMA). 2025. URL: https://www.ama-assn.org/practice-management/digital-health/8-steps-position-your-health-system-ai-success? [accessed 2026-09-05]
11. Ethics and governance of artificial intelligence for health: WHO guidance 1st. : World Health Organization; 2021 URL: https://iris.who.int/server/api/core/bitstreams/f780d926-4ae3-42ce-a6d6-e898a5562621/content [accessed 2026-09-05]

XSL•FO
RenderX

12. Bozkurt S, Fereydooni S, Kar I, et al. Investigating data diversity and model robustness of AI applications in palliative care and hospice: protocol for scoping review. JMIR Res Protoc 2024 Oct 8;13:e56353. [doi: 10.2196/56353] [Medline: 39378420]
13. Zhang H, Wang AY, Wu S, et al. Artificial intelligence for the prediction of acute kidney injury during the perioperative period: systematic review and meta-analysis of diagnostic test accuracy. BMC Nephrol 2022 Dec 19;23(1):405. [doi: 10.1186/s12882-022-03025-w] [Medline: 36536317]
14. Yoon HK, Yang HL, Jung CW, Lee HC. Artificial intelligence in perioperative medicine: a narrative review. Korean J Anesthesiol 2022 Jun;75(3):202-215. [doi: 10.4097/kja.22157] [Medline: 35345305]
15. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
16. Shamseer L, Moher D, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ 2015 Jan 2;350:g7647. [doi: 10.1136/bmj.g7647] [Medline: 25555855]
17. Maheta B, Keny C, Okech M, Raspi I, Ross R, Giannitrapani K. Understanding processes of integration of artificial intelligence into the perioperative period for patients with serious illness: a systematic review [Poster]. Presented at: AcademyHealth 2025 Annual Research Meeting (ARM); Jun 7-10, 2025 URL: https://academyhealth.confex.com/academyhealth/2025arm/meetingapp.cgi/Paper/72971 [accessed 2026-09-05]
18. Jiang S, Bukhari SMA, Krishnan A, et al. Deployment of artificial intelligence in radiology: strategies for success. AJR Am J Roentgenol 2025 Feb;224(2):e2431898. [doi: 10.2214/AJR.24.31898] [Medline: 39475198]
19. McGenity C, Clarke EL, Jennings C, et al. Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. NPJ Digit Med 2024 May 4;7(1):114. [doi: 10.1038/s41746-024-01106-8] [Medline: 38704465]
20. Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. NPJ Digit Med 2020;3(1):126. [doi: 10.1038/s41746-020-00333-z] [Medline: 33043150]
21. Jiang Y, Li X, Luo H, Yin S, Kaynak O. Quo vadis artificial intelligence? Discov Artif Intell 2022;2(1):4. [doi: 10.1007/s44163-022-00022-8]
22. Helm JM, Swiergosz AM, Haeberle HS, et al. Machine learning and artificial intelligence: definitions, applications, and future directions. Curr Rev Musculoskelet Med 2020 Feb;13(1):69-76. [doi: 10.1007/s12178-020-09600-8] [Medline: 31983042]
23. Yefimova M, Aslakson RA, Yang L, et al. Palliative care and end-of-life outcomes following high-risk surgery. JAMA Surg 2020 Feb 1;155(2):138-146. [doi: 10.1001/jamasurg.2019.5083] [Medline: 31895424]
24. Deyo RA. Measuring functional outcomes in therapeutic trials for chronic disease. Control Clin Trials 1984 Sep;5(3):223-240. [doi: 10.1016/0197-2456(84)90026-6] [Medline: 6488807]
25. Deshpande PR, Rajan S, Sudeepthi BL, Abdul Nazir CP. Patient-reported outcomes: a new era in clinical research. Perspect Clin Res 2011 Oct;2(4):137-144. [doi: 10.4103/2229-3485.86879] [Medline: 22145124]
26. Fung CH, Hays RD. Prospects and challenges in using patient-reported outcomes in clinical practice. Qual Life Res 2008 Dec;17(10):1297-1302. [doi: 10.1007/s11136-008-9379-5] [Medline: 18709564]
27. Van der Mierden S, Tsaioun K, Bleich A, Leenaars CHC. Software tools for literature screening in systematic reviews in biomedical research. ALTEX 2019;36(3):508-517. [doi: 10.14573/altex.1902131] [Medline: 31113000]
28. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919. [doi: 10.1136/bmj.i4919] [Medline: 27733354]
29. Williams M, Moser T. The art of coding and thematic exploration in qualitative research. Int Manag Rev 2019;15(1):45-72 [FREE Full text]
30. Glaser BG. Open Coding Descriptions. Grounded Theory Rev 2016 Dec 19;15(2):108-110 [FREE Full text]
31. Abràmoff MD, Tarver ME, Loyo-Berrios N, et al. Considerations for addressing bias in artificial intelligence for health equity. NPJ Digit Med 2023 Sep 12;6(1):170. [doi: 10.1038/s41746-023-00913-9] [Medline: 37700029]
32. Ratwani RM, Bates DW, Classen DC. Patient safety and artificial intelligence in clinical care. JAMA Health Forum 2024 Feb 2;5(2):e235514. [doi: 10.1001/jamahealthforum.2023.5514] [Medline: 38393719]
33. Oluwagbenro MB. Generative AI: definition, concepts, applications, and future prospects. Techrxiv. Preprint posted online on Jun 4, 2024. [doi: 10.36227/techrxiv.171746875.59016695/v1]
34. Badillo S, Banfai B, Birzele F, et al. An introduction to machine learning. Clin Pharma Therapeutics 2020 Apr;107(4):871-885. [doi: 10.1002/cpt.1796] [Medline: 32128792]
35. Martinez AR. Natural language processing. WIREs Computational Stats 2010 May;2(3):352-357. [doi: 10.1002/wics.76]
36. Amparore D, De Cillis S, Alladio E, et al. Development of machine learning algorithm to predict the risk of incontinence after robot-assisted radical prostatectomy. J Endourol 2024 Aug;38(8):871-878. [doi: 10.1089/end.2024.0057] [Medline: 38512711]
37. Cai LQ, Yang DQ, Wang RJ, Huang H, Shi YX. Establishing and clinically validating a machine learning model for predicting unplanned reoperation risk in colorectal cancer. World J Gastroenterol 2024 Jun 21;30(23):2991-3004. [doi: 10.3748/wjg.v30.i23.2991] [Medline: 38946868]
38. Chen D, Afzal N, Sohn S, et al. Postoperative bleeding risk prediction for patients undergoing colorectal surgery. Surgery 2018 Dec;164(6):1209-1216. [doi: 10.1016/j.surg.2018.05.043] [Medline: 30033185]

XSL•FO
RenderX

39. Choi N, Kim Z, Song BH, et al. Prediction of risk factors for pharyngo-cutaneous fistula after total laryngectomy using artificial intelligence. Oral Oncol 2021 Aug;119:105357. [doi: 10.1016/j.oraloncology.2021.105357] [Medline: 34044316]
40. Costantino A, Sampieri C, Pace GM, et al. Development of machine learning models for the prediction of long-term feeding tube dependence after oral and oropharyngeal cancer surgery. Oral Oncol 2024 Jan;148:106643. [doi: 10.1016/j.oraloncology.2023.106643] [Medline: 38006688]
41. Cui Y, Shi X, Qin Y, et al. Establishment and validation of an interactive artificial intelligence platform to predict postoperative ambulatory status for patients with metastatic spinal disease: a multicenter analysis. Int J Surg 2024 May 1;110(5):2738-2756. [doi: 10.1097/JS9.0000000000001169] [Medline: 38376838]
42. Du J, Yang J, Yang Q, Zhang X, Yuan L, Fu B. Comparison of machine learning models to predict the risk of breast cancer-related lymphedema among breast cancer survivors: a cross-sectional study in China. Front Oncol 2024;14:1334082. [doi: 10.3389/fonc.2024.1334082] [Medline: 38410115]
43. Fourman MS, Siraj L, Duvall J, et al. Can we use artificial intelligence cluster analysis to identify patients with metastatic breast cancer to the spine at highest risk of postoperative adverse events? World Neurosurg 2023 Jun;174:e26-e34. [doi: 10.1016/j.wneu.2023.02.064] [Medline: 36805503]
44. Fu Y, Shi W, Zhao J, et al. Prediction of postoperative health-related quality of life among patients with metastatic spinal cord compression secondary to lung cancer. Front Endocrinol (Lausanne) 2023;14:1206840. [doi: 10.3389/fendo.2023.1206840] [Medline: 37720536]
45. Ganguli R, Franklin J, Yu X, Lin A, Heffernan DS. Machine learning methods to predict presence of residual cancer following hysterectomy. Sci Rep 2022 Feb 17;12(1):2738. [doi: 10.1038/s41598-022-06585-x] [Medline: 35177700]
46. Ganguli R, Franklin J, Yu X, Lin A, Lad R, Heffernan DS. Machine learning models to prognose 30-day mortality in postoperative disseminated cancer patients. Surg Oncol 2022 Sep;44:101810. [doi: 10.1016/j.suronc.2022.101810] [Medline: 36088867]
47. Ghaith AK, Ghanem M, Zamanian C, et al. Using machine learning to predict 30-day readmission and reoperation following resection of supratentorial high-grade gliomas: an ACS NSQIP study involving 9418 patients. Neurosurg Focus 2023 Jun;54(6):E12. [doi: 10.3171/2023.3.FOCUS22652] [Medline: 37552633]
48. Guo SW, Shen J, Gao JH, et al. A preoperative risk model for early recurrence after radical resection may facilitate initial treatment decisions concerning the use of neoadjuvant therapy for patients with pancreatic ductal adenocarcinoma. Surgery 2020 Dec;168(6):1003-1014. [doi: 10.1016/j.surg.2020.02.013] [Medline: 32321665]
49. Han IW, Cho K, Ryu Y, et al. Risk prediction platform for pancreatic fistula after pancreatoduodenectomy using artificial intelligence. World J Gastroenterol 2020 Aug 14;26(30):4453-4464. [doi: 10.3748/wjg.v26.i30.4453] [Medline: 32874057]
50. He Y, Luo L, Shan R, et al. Development and validation of a nomogram for predicting postoperative early relapse and survival in hepatocellular carcinoma. J Natl Compr Canc Netw 2023 Dec 20;22(1D):e237069. [doi: 10.6004/jnccn.2023.7069] [Medline: 38118280]
51. Hoek AG, van Oort S, Mukamal KJ, Beulens JWJ. Alcohol consumption and cardiovascular disease risk: placing new data in context. Curr Atheroscler Rep 2022 Jan;24(1):51-59. [doi: 10.1007/s11883-022-00992-1] [Medline: 35129737]
52. Hong Y, Li Y, Ye M, Yan S, Yang W, Jiang C. Identifying an optimal machine learning model generated circulating biomarker to predict chronic postoperative pain in patients undergoing hepatectomy. Front Surg 2023;9:1068321. [doi: 10.3389/fsurg.2022.1068321] [Medline: 36684250]
53. Huang J, Xie X, Wu H, et al. Development and validation of a combined nomogram model based on deep learning contrast-enhanced ultrasound and clinical factors to predict preoperative aggressiveness in pancreatic neuroendocrine neoplasms. Eur Radiol 2022 Nov;32(11):7965-7975. [doi: 10.1007/s00330-022-08703-9] [Medline: 35389050]
54. Huang Y, Chen H, Zeng Y, Liu Z, Ma H, Liu J. Development and validation of a machine learning prognostic model for hepatocellular carcinoma recurrence after surgical resection. Front Oncol 2020;10:593741. [doi: 10.3389/fonc.2020.593741] [Medline: 33598425]
55. Ishizaki T, Mazaki J, Enomoto M, et al. Predictive modelling for high-risk stage II colon cancer using auto-artificial intelligence. Tech Coloproctol 2023 Mar;27(3):183-188. [doi: 10.1007/s10151-022-02685-y] [Medline: 36031650]
56. Ivanics T, Nelson W, Patel MS, et al. The Toronto postliver transplantation hepatocellular carcinoma recurrence calculator: a machine learning approach. Liver Transpl 2022 Apr;28(4):593-602. [doi: 10.1002/lt.26332] [Medline: 34626159]
57. Jeon Y, Kim YJ, Jeon J, et al. Machine learning based prediction of recurrence after curative resection for rectal cancer. PLoS One 2023;18(12):e0290141. [doi: 10.1371/journal.pone.0290141] [Medline: 38100485]
58. Jin F, Liu W, Qiao X, Shi J, Xin R, Jia HQ. Nomogram prediction model of postoperative pneumonia in patients with lung cancer: a retrospective cohort study. Front Oncol 2023;13:1114302. [doi: 10.3389/fonc.2023.1114302] [Medline: 36910602]
59. Jung JO, Crnovrsanin N, Wirsik NM, et al. Machine learning for optimized individual survival prediction in resectable upper gastrointestinal cancer. J Cancer Res Clin Oncol 2023 May;149(5):1691-1702. [doi: 10.1007/s00432-022-04063-5] [Medline: 35616729]
60. Kadomatsu Y, Emoto R, Kubo Y, et al. Development of a machine learning-based risk model for postoperative complications of lung cancer surgery. Surg Today 2024 Dec;54(12):1482-1489. [doi: 10.1007/s00595-024-02878-y] [Medline: 38896280]

XSL•FO
RenderX

61. Karabacak M, Margetis K. A machine learning-based online prediction tool for predicting short-term postoperative outcomes following spinal tumor resections. Cancers (Basel) 2023 Jan 28;15(3):812. [doi: 10.3390/cancers15030812] [Medline: 36765771]
62. Karhade AV, Thio QCBS, Ogink PT, et al. Development of machine learning algorithms for prediction of 30-day mortality after surgery for spinal metastasis. Neurosurgery 2019 Jul 1;85(1):E83-E91. [doi: 10.1093/neuros/nyy469] [Medline: 30476188]
63. Kim K, Han Y, Jeong S, et al. Prediction of postoperative length of hospital stay based on differences in nursing narratives in elderly patients with epithelial ovarian cancer. Methods Inf Med 2019 Dec;58(6):222-228. [doi: 10.1055/s-0040-1705122] [Medline: 32349156]
64. Kinoshita F, Takenaka T, Yamashita T, et al. Development of artificial intelligence prognostic model for surgically resected non-small cell lung cancer. Sci Rep 2023 Sep 21;13(1):15683. [doi: 10.1038/s41598-023-42964-8] [Medline: 37735585]
65. Klén R, Salminen AP, Mahmoudian M, Syvänen KT, Elo LL, Boström PJ. Prediction of complication related death after radical cystectomy for bladder cancer with machine learning methodology. Scand J Urol 2019 Oct;53(5):325-331. [doi: 10.1080/21681805.2019.1665579] [Medline: 31552774]
66. Kuo PJ, Wu SC, Chien PC, et al. Artificial neural network approach to predict surgical site infection after free-flap reconstruction in patients receiving surgery for head and neck cancer. Oncotarget 2018;9(17):13768-13782. [doi: 10.18632/oncotarget.24468] [Medline: 29568393]
67. Kuwayama N, Hoshino I, Mori Y, Yokota H, Iwatate Y, Uno T. Applying artificial intelligence using routine clinical data for preoperative diagnosis and prognosis evaluation of gastric cancer. Oncol Lett 2023;26(5):499. [doi: 10.3892/ol.2023.14087] [Medline: 37854867]
68. Laios A, De Freitas DLD, Saalmink G, et al. Stratification of length of stay prediction following surgical cytoreduction in advanced high-grade serous ovarian cancer patients using artificial intelligence; the Leeds L-AI-OS score. Curr Oncol 2022 Nov 23;29(12):9088-9104. [doi: 10.3390/curroncol29120711] [Medline: 36547125]
69. Laios A, De Oliveira Silva RV, Dantas De Freitas DL, et al. Machine learning-based risk prediction of critical care unit admission for advanced stage high grade serous ovarian cancer patients undergoing cytoreductive surgery: the Leeds-Natal score. J Clin Med 2021 Dec 24;11(1):87. [doi: 10.3390/jcm11010087] [Medline: 35011828]
70. Lee IC, Huang JY, Chen TC, et al. Evolutionary learning-derived clinical-radiomic models for predicting early recurrence of hepatocellular carcinoma after resection. Liver Cancer 2021;10(6):572-582. [doi: 10.1159/000518728] [Medline: 34950180]
71. Lee KS, Jang JY, Yu YD, et al. Usefulness of artificial intelligence for predicting recurrence following surgery for pancreatic cancer: retrospective cohort study. Int J Surg 2021 Sep;93:106050. [doi: 10.1016/j.ijsu.2021.106050] [Medline: 34388677]
72. Lee W, Park HJ, Lee HJ, et al. Preoperative data-based deep learning model for predicting postoperative survival in pancreatic cancer patients. Int J Surg 2022 Sep;105:106851. [doi: 10.1016/j.ijsu.2022.106851] [Medline: 36049618]
73. Li R, Zheng Z, Yang L, et al. Development of a machine learning algorithm to forecast the likelihood of postoperative neurological complications in patients with parotid tumors. Ear Nose Throat J 2024 May 28:1455613241258648. [doi: 10.1177/01455613241258648] [Medline: 38804648]
74. Li Y, Wu JH, Li CP, et al. Multidimensional characteristics, prognostic role, and preoperative prediction of peritoneal sarcomatosis in retroperitoneal sarcoma. Front Oncol 2022;12:950418. [doi: 10.3389/fonc.2022.950418] [Medline: 36387243]
75. Lin B, Chen F, Wu M, Li C, Lin L. Machine learning models for prediction of postoperative venous thromboembolism in gynecological malignant tumor patients. J Obstet Gynaecol Res 2024 Jul;50(7):1175-1181. [doi: 10.1111/jog.15960] [Medline: 38689519]
76. Lin V, Tsouchnika A, Allakhverdiiev E, et al. Training prediction models for individual risk assessment of postoperative complications after surgery for colorectal cancer. Tech Coloproctol 2022 Aug;26(8):665-675. [doi: 10.1007/s10151-022-02624-x] [Medline: 35593971]
77. Liu CT, Peng YH, Hong CQ, et al. A nomogram based on nutrition-related indicators and computed tomography imaging features for predicting preoperative lymph node metastasis in curatively resected esophagogastric junction adenocarcinoma. Ann Surg Oncol 2023 Aug;30(8):5185-5194. [doi: 10.1245/s10434-023-13378-7] [Medline: 37010663]
78. Liu R, Wu S, Yu HY, et al. Prediction model for hepatocellular carcinoma recurrence after hepatectomy: machine learning-based development and interpretation study. Heliyon 2023 Nov;9(11):e22458. [doi: 10.1016/j.heliyon.2023.e22458] [Medline: 38034691]
79. Lopez-Lopez V, Morise Z, Albaladejo-González M, et al. Explainable artificial intelligence prediction-based model in laparoscopic liver surgery for segments 7 and 8: an international multicenter study. Surg Endosc 2024 May;38(5):2411-2422. [doi: 10.1007/s00464-024-10681-6] [Medline: 38315197]
80. Lou SJ, Hou MF, Chang HT, et al. Machine learning algorithms to predict recurrence within 10 years after breast cancer surgery: a prospective cohort study. Cancers (Basel) 2020 Dec 17;12(12):3817. [doi: 10.3390/cancers12123817] [Medline: 33348826]
81. Ma Y, Tan B, Wang S, Ren C, Zhang J, Gao Y. Influencing factors and predictive model of postoperative infection in patients with primary hepatic carcinoma. BMC Gastroenterol 2023 Apr 12;23(1):123. [doi: 10.1186/s12876-023-02713-7] [Medline: 37046206]

XSL•FO
RenderX

82. Mai RY, Lu HZ, Bai T, et al. Artificial neural network model for preoperative prediction of severe liver failure after hemihepatectomy in patients with hepatocellular carcinoma. Surgery 2020 Oct;168(4):643-652. [doi: 10.1016/j.surg.2020.06.031] [Medline: 32792098]
83. Masum S, Hopgood A, Stefan S, Flashman K, Khan J. Data analytics and artificial intelligence in predicting length of stay, readmission, and mortality: a population-based study of surgical management of colorectal cancer. Discov Oncol 2022;13(1):11. [doi: 10.1007/s12672-022-00472-7] [Medline: 35226196]
84. Mazaki J, Katsumata K, Ohno Y, et al. A novel prediction model for colon cancer recurrence using auto-artificial intelligence. Anticancer Res 2021 Sep;41(9):4629-4636. [doi: 10.21873/anticanres.15276] [Medline: 34475091]
85. Obrzut B, Kusy M, Semczuk A, Obrzut M, Kluska J. Prediction of 10-year overall survival in patients with operable cervical cancer using a probabilistic neural network. J Cancer 2019;10(18):4189-4195. [doi: 10.7150/jca.33945] [Medline: 31413737]
86. Osman MH, Mohamed RH, Sarhan HM, et al. Machine learning model for predicting postoperative survival of patients with colorectal cancer. Cancer Res Treat 2022 Apr;54(2):517-524. [doi: 10.4143/crt.2021.206] [Medline: 34126702]
87. Paro A, Hyer MJ, Tsilimigras DI, et al. Machine learning approach to stratifying prognosis relative to tumor burden after resection of colorectal liver metastases: an international cohort analysis. J Am Coll Surg 2022 Apr 1;234(4):504-513. [doi: 10.1097/XCS.0000000000000094] [Medline: 35290269]
88. Pera M, Gibert J, Gimeno M, et al. Machine learning risk prediction model of 90-day mortality after gastrectomy for cancer. Ann Surg 2022 Nov 1;276(5):776-783. [doi: 10.1097/SLA.0000000000005616] [Medline: 35866643]
89. Qiao W, Wang Y, Luo C, et al. Development of preoperative and postoperative models to predict recurrence in postoperative glioma patients: a longitudinal cohort study. BMC Cancer 2024 Feb 28;24(1):274. [doi: 10.1186/s12885-024-11996-2] [Medline: 38418976]
90. Qin L, Liang Z, Xie J, et al. Development and validation of machine learning models for postoperative venous thromboembolism prediction in colorectal cancer inpatients: a retrospective study. J Gastrointest Oncol 2023 Feb 28;14(1):220-232. [doi: 10.21037/jgo-23-18] [Medline: 36915444]
91. Sammour T, Cohen L, Karunatillake AI, et al. Validation of an online risk calculator for the prediction of anastomotic leak after colon cancer surgery and preliminary exploration of artificial intelligence-based analytics. Tech Coloproctol 2017 Nov;21(11):869-877. [doi: 10.1007/s10151-017-1701-1] [Medline: 29080956]
92. Shahriarirad R, Meshkati Yazd SM, Fathian R, Fallahi M, Ghadiani Z, Nafissi N. Prediction of sentinel lymph node metastasis in breast cancer patients based on preoperative features: a deep machine learning approach. Sci Rep 2024 Jan 16;14(1):1351. [doi: 10.1038/s41598-024-51244-y] [Medline: 38228684]
93. Shao S, Liu L, Zhao Y, Mu L, Lu Q, Qin J. Application of machine learning for predicting anastomotic leakage in patients with gastric adenocarcinoma who received total or proximal gastrectomy. J Pers Med 2021 Jul 29;11(8):748. [doi: 10.3390/jpm11080748] [Medline: 34442391]
94. Shen Y, Huang LB, Lu A, Yang T, Chen HN, Wang Z. Prediction of symptomatic anastomotic leak after rectal cancer surgery: a machine learning approach. J Surg Oncol 2024 Feb;129(2):264-272. [doi: 10.1002/jso.27470] [Medline: 37795583]
95. Shi HY, Tsai JT, Chen YM, Culbertson R, Chang HT, Hou MF. Predicting two-year quality of life after breast cancer surgery using artificial neural network and linear regression models. Breast Cancer Res Treat 2012 Aug;135(1):221-229. [doi: 10.1007/s10549-012-2174-6] [Medline: 22836876]
96. Shi X, Cui Y, Wang S, Pan Y, Wang B, Lei M. Development and validation of a web-based artificial intelligence prediction model to assess massive intraoperative blood loss for metastatic spinal disease using machine learning techniques. Spine J 2024 Jan;24(1):146-160. [doi: 10.1016/j.spinee.2023.09.001] [Medline: 37704048]
97. Spelt L, Nilsson J, Andersson R, Andersson B. Artificial neural networks—a method for prediction of survival following liver resection for colorectal cancer metastases. Eur J Surg Oncol 2013 Jun;39(6):648-654. [doi: 10.1016/j.ejso.2013.02.024] [Medline: 23514791]
98. Sun LY, Ouyang Q, Cen WJ, Wang F, Tang WT, Shao JY. A model based on artificial intelligence algorithm for monitoring recurrence of HCC after hepatectomy. Am Surg 2023 May;89(5):1468-1478. [doi: 10.1177/00031348211063549] [Medline: 34894786]
99. van de Beld JJ, Crull D, Mikhal J, et al. Complication prediction after esophagectomy with machine learning. Diagnostics (Basel) 2024 Feb 17;14(4):439. [doi: 10.3390/diagnostics14040439] [Medline: 38396478]
100. van Niftrik CHB, van der Wouden F, Staartjes VE, et al. Machine learning algorithm identifies patients at high risk for early complications after intracranial tumor surgery: registry-based cohort study. Neurosurgery 2019 Oct 1;85(4):E756-E764. [doi: 10.1093/neuros/nyz145] [Medline: 31149726]
101. Verma A, Balian J, Hadaya J, et al. Machine learning-based prediction of postoperative pancreatic fistula following pancreaticoduodenectomy. Ann Surg 2024 Aug 1;280(2):325-331. [doi: 10.1097/SLA.0000000000006123] [Medline: 37947154]
102. Vesovic R, Milosavljevic M, Punt M, et al. The role of the diaphragm in prediction of respiratory function in the immediate postoperative period in lung cancer patients using a machine learning model. World J Surg Oncol 2023 Dec 22;21(1):393. [doi: 10.1186/s12957-023-03278-1] [Medline: 38135875]

XSL•FO
RenderX

103. Wang K, Tang Y, Zhang F, Guo X, Gao L. Combined application of inflammation-related biomarkers to predict postoperative complications of rectal cancer patients: a retrospective study by machine learning analysis. Langenbecks Arch Surg 2023 Oct 13;408(1):400. [doi: 10.1007/s00423-023-03127-5] [Medline: 37831218]
104. Wang L, Song D, Wang W, et al. Data-driven assisted decision making for surgical procedure of hepatocellular carcinoma resection and prognostic prediction: development and validation of machine learning models. Cancers (Basel) 2023;15(6):1784. [doi: 10.3390/cancers15061784] [Medline: 36980670]
105. Wen R, Zheng K, Zhang Q, et al. Machine learning-based random forest predicts anastomotic leakage after anterior resection for rectal cancer. J Gastrointest Oncol 2021 Jun;12(3):921-932. [doi: 10.21037/jgo-20-436] [Medline: 34295545]
106. Wu X, Guan Q, Cheng ASK, et al. Comparison of machine learning models for predicting the risk of breast cancer-related lymphedema in Chinese women. Asia Pac J Oncol Nurs 2022;9(12):100101. [doi: 10.1016/j.apjon.2022.100101] [Medline: 36276882]
107. Xu Z, Xie Y, Wu L, et al. Using machine learning methods to assess lymphovascular invasion and survival in breast cancer: performance of combining preoperative clinical and MRI characteristics. J Magn Reson Imaging 2023 Nov;58(5):1580-1589. [doi: 10.1002/jmri.28647] [Medline: 36797654]
108. Ying T, Borrelli P, Edenbrandt L, et al. Automated artificial intelligence-based analysis of skeletal muscle volume predicts overall survival after cystectomy for urinary bladder cancer. Eur Radiol Exp 2021 Nov 19;5(1):50. [doi: 10.1186/s41747-021-00248-8] [Medline: 34796422]
109. Zaver HB, Mzaik O, Thomas J, et al. Utility of an artificial intelligence enabled electrocardiogram for risk assessment in liver transplant candidates. Dig Dis Sci 2023 Jun;68(6):2379-2388. [doi: 10.1007/s10620-023-07928-y] [Medline: 37022601]
110. Zeng J, Song D, Li K, Cao F, Zheng Y. Deep learning model for predicting postoperative survival of patients with gastric cancer. Front Oncol 2024;14:1329983. [doi: 10.3389/fonc.2024.1329983] [Medline: 38628668]
111. Zeng L, Liu L, Chen D, et al. The innovative model based on artificial intelligence algorithms to predict recurrence risk of patients with postoperative breast cancer. Front Oncol 2023;13:1117420. [doi: 10.3389/fonc.2023.1117420] [Medline: 36959794]
112. Zhang G, Liu X, Hu Y, et al. Development and comparison of machine-learning models for predicting prolonged postoperative length of stay in lung cancer patients following video-assisted thoracoscopic surgery. Asia Pac J Oncol Nurs 2024;11(6):100493. [doi: 10.1016/j.apjon.2024.100493] [Medline: 38808011]
113. Zhang W, Wang Q, Liang K, et al. Deep learning nomogram for preoperative distinction between xanthogranulomatous cholecystitis and gallbladder carcinoma: a novel approach for surgical decision. Comput Biol Med 2024 Jan;168:107786. [doi: 10.1016/j.compbiomed.2023.107786] [Medline: 38048662]
114. Zhang Y, Zhou Q, Chen G, Xue S. Early postoperative prediction of the risk of distant metastases in medullary thyroid cancer. Front Endocrinol (Lausanne) 2023;14:1209978. [doi: 10.3389/fendo.2023.1209978] [Medline: 38075078]
115. Zhao F, Wang P, Yu C, et al. A LASSO-based model to predict central lymph node metastasis in preoperative patients with cN0 papillary thyroid cancer. Front Oncol 2023;13:1034047. [doi: 10.3389/fonc.2023.1034047] [Medline: 36761950]
116. Zheng CY, Wu J, Chen CS, et al. A scoring model for predicting early recurrence of gastric cancer with normal preoperative tumor markers: a multicenter study. Eur J Surg Oncol 2023 Nov;49(11):107094. [doi: 10.1016/j.ejso.2023.107094] [Medline: 37797381]
117. Schmutz JB, Outland N, Kerstan S, Georganta E, Ulfert AS. AI-teaming: redefining collaboration in the digital era. Curr Opin Psychol 2024 Aug;58:101837. [doi: 10.1016/j.copsyc.2024.101837] [Medline: 39024969]
118. Oh O, Demiris G, Ulrich CM. The ethical dimensions of utilizing artificial intelligence in palliative care. Nurs Ethics 2025 Jun;32(4):1285-1296. [doi: 10.1177/09697330241296874] [Medline: 39551621]
119. Burry N, Nakagawa S, Blinderman CD. “You are not alone”: the allure and limitations of artificial intelligence in serious illness communication. J Palliat Med 2024 Jan;27(1):7-9. [doi: 10.1089/jpm.2023.0471] [Medline: 38011011]
120. Sittig DF, Singh H. Recommendations to ensure safety of AI in real-world clinical care. JAMA 2025 Feb 11;333(6):457-458. [doi: 10.1001/jama.2024.24598] [Medline: 39602298]
121. Maliha G, Gerke S, Cohen IG, Parikh RB. Artificial intelligence and liability in medicine: balancing safety and innovation. Milbank Q 2021 Sep;99(3):629-647. [doi: 10.1111/1468-0009.12504] [Medline: 33822422]
122. Guni A, Varma P, Zhang J, Fehervari M, Ashrafian H. Artificial intelligence in surgery: the future is now. Eur Surg Res 2024 Jan 22. [doi: 10.1159/000536393] [Medline: 38253041]
123. Weiner EB, Dankwa-Mullan I, Nelson WA, Hassanpour S. Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice. PLOS Digit Health 2025 Apr;4(4):e0000810. [doi: 10.1371/journal.pdig.0000810] [Medline: 40198594]
124. Georganta E, Ulfert AS. Would you trust an AI team member? Team trust in human–AI teams. J Occupat Organ Psyc 2024;97(3):1212-1241. [doi: 10.1111/joop.12504]
125. Abi-Rafeh J, Henry N, Xu HH, et al. Utility and comparative performance of current artificial intelligence large language models as postoperative medical support chatbots in aesthetic surgery. Aesthet Surg J 2024 Jul 15;44(8):889-896. [doi: 10.1093/asj/sjae025] [Medline: 38318684]
126. Almutairi M. Teaming in the AI era: AI-augmented frameworks for forming, simulating, and optimizing human teams. Proc ACM Conf User Model Adapt Pers 2025 Jun 16:414-418. [doi: 10.1145/3699682.3727574]

XSL•FO
RenderX

127. Rodler S, Ganjavi C, De Backer P, et al. Generative artificial intelligence in surgery. Surgery 2024 Jun;175(6):1496-1502. [doi: 10.1016/j.surg.2024.02.019] [Medline: 38582732]
128. Qin S, Chislett B, Ischia J, et al. ChatGPT and generative AI in urology and surgery—a narrative review. BJUI Compass 2024 Sep;5(9):813-821. [doi: 10.1002/bco2.390] [Medline: 39323919]
129. Jacob C, Brasier N, Laurenzi E, et al. AI for IMPACTS framework for evaluating the long-term real-world IMPACTS of AI-powered clinician tools: systematic review and narrative synthesis. J Med Internet Res 2025 Feb 5;27:e67485. [doi: 10.2196/67485] [Medline: 39909417]

Abbreviations

AMA: American Medical Association
ANN: artificial neural network
CENTRAL: Cochrane Central Register of Controlled Trials
EHR: electronic health record
HHS: Health and Human Services
LASSO: least absolute shrinkage and selection operator
LLM: large language model
ML: machine learning
NLP: natural language processing
PET: positron emission tomography
PICOTS: population, intervention, comparator, outcomes, timing, and setting
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis Protocols
PROSPERO: International Prospective Register of Systematic Reviews
ROBINS-I: risk of bias tool in nonrandomized studies of interventions
SVM: support vector machine
WHO: World Health Organization
XGBoost: extreme gradient boosting

Edited by N Rohatgi; submitted 20.Dec.2025; peer-reviewed by H Lv, L Zhou, M Chakit, Z Liu; revised version received 05.Jul.2026; accepted 21.Aug.2026; published 01.Oct.2026.

Please cite as:
Maheta BJ, Ross RL, Raspi I, Keny C, Okech MN, Bozkurt S, Giannitrapani KF
AI Trustworthiness in the Perioperative Period for Patients with Serious Illness: Scoping Narrative Review
JMIR Perioper Med 2026;9:e89936
URL: https://periop.jmir.org/2026/1/e89936
doi:10.2196/89936

© Bhagvat J Maheta, Rachel L Ross, Isabella Raspi, Christina Keny, Marti N Okech, Selen Bozkurt, Karleen F Giannitrapani. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 1.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patient Satisfaction With Perioperative Services and Associated Factors in Ethiopia: Systematic Review and Meta-Analysis

Melesse Abiye Munie[1*], MSc; Tegene Atamenta Kitaw[1*], MSc; Aregash Birhane Terefe[1*], MSc; Alemu Birara Zemariam[1*], MSc; Molalign Aligaz Adisu[2*], MSc; Molla Azmeraw[2*], MSc; Tesfaye Engdaw Habtie[1*], MSc; Betelhem Walelgn Dagnaw[1*], MSc; Abebe Merchaw Faris[3*], MSc; Tilahun Wodaynew[1*], MSc; Amsalu Baylie Taye[1*], MSc; Gebrie Kassaw Yirga[1*], MSc; Yabibal Asfaw Desro[4*], MSc

[1]Department of Nursing, College of Health Sciences, Woldia University, PO Box 400, Woldia, Ethiopia

[2]Department of Pediatrics and Child Health Nursing, College of Health Sciences, Woldia University, Woldia, Ethiopia

[3]Department of Emergency and Critical Care Nursing, College of Health Sciences, Woldia University, Woldia, Ethiopia

[4]Department of Surgical Nursing, College of Health Sciences, Woldia University, Woldia, Ethiopia

[*]all authors contributed equally

Corresponding Author:
Melesse Abiye Munie, MSc
Department of Nursing, College of Health Sciences, Woldia University, PO Box 400, Woldia, Ethiopia

Abstract

Background: Patient satisfaction is a key indicator of health care quality, and it guides improvement efforts. Although many local studies have examined perioperative patient satisfaction in Ethiopia, there is no comprehensive national synthesis. This gap limits the development of targeted strategies to enhance patient care.

Objective: The aim of this systematic review and meta-analysis is to determine the pooled prevalence of patient satisfaction with perioperative services in Ethiopia and identify associated factors.

Methods: This study included all observational research articles on patient satisfaction with perioperative services in Ethiopia. A multidatabase search strategy, incorporating PubMed/MEDLINE, HINARI, Web of Science, Cochrane Library, African Journals Online, and Scopus, was used alongside a gray literature search to identify all Ethiopian studies on perioperative satisfaction available before January 1, 2024. The Newcastle-Ottawa Scale was used to assess the quality of the studies. To assess heterogeneity, subgroup analyses were conducted, and I^2 statistics were calculated. This study used funnel plots, the Egger test, and a nonparametric trim-and-fill analysis to assess publication bias. A sensitivity analysis was also used to identify any influential studies. Univariate meta-regression examined the association between study-level covariates and perioperative satisfaction.

Results: This review included 21 studies comprising 6858 participants. Overall satisfaction with perioperative services was expressed by 5072 participants (73.96%, 95% CI 68.84%-79.08%; I^2=96.56%). Factors significantly associated with higher satisfaction included effective postoperative pain management (adjusted odds ratio [AOR] 2.23, 95% CI 1.56-2.90), illiteracy (AOR 3.18, 95% CI 1.23-5.13), primary school education (AOR 6.55, 95% CI 3.61-9.49), local anesthesia use (AOR 2.80, 95% CI 2.03-3.57), and history of prior surgery or anesthesia (AOR 2.76, 95% CI 1.51-4.01).

Conclusions: This study found that the pooled prevalence of patient satisfaction with perioperative services in Ethiopia was 73.96% (5072/6858 participants). Postoperative pain management, illiteracy, primary school, local anesthesia, and a history of surgery or anesthesia were significantly associated with patient satisfaction with perioperative services. Health care facilities should focus on providing effective postoperative pain management, clear information about perioperative services, and training for surgical and anesthesia teams to boost patient satisfaction with perioperative services in Ethiopia.

Trial Registration: PROSPERO CRD42024614957; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024614957

(JMIR Perioper Med 2026;9:e84457) doi:10.2196/84457

KEYWORDS

patient satisfaction; perioperative care; surgical patients; Ethiopia; systematic review; meta-analysis; associated factors; prevalence

Introduction

Patient satisfaction is a subjective measure that reflects how well patients' health care experiences align with their expectations and needs. It is an essential indicator of the quality of health care services, particularly in the perioperative setting, where patients undergo significant physical and emotional changes [1]. Patient satisfaction is a vital goal in health care,

XSL•FO
RenderX

acting as both an indicator of care quality and a catalyst for improvement. By assessing satisfaction, providers can pinpoint areas needing enhancement, which leads to better patient outcomes and adherence to treatment protocols [2]. High patient satisfaction is correlated with improved quality of perioperative care and a reduction in patient readmissions, mortality rates, and hospital stays [3].

Perioperative patient satisfaction is a critical outcome that enhances recovery and trust. Perioperative care integrates management across the preoperative, intraoperative, and postoperative phases. This multidisciplinary approach faces challenges, including patient anxiety, pain control, and communication issues. Addressing these challenges through holistic strategies, which commonly prioritize communication, pain management, and emotional support, can enhance perioperative patient satisfaction [4]. Effective perioperative care requires a multidisciplinary approach involving collaboration among surgeons, anesthetists, nurses, and other health care professionals. This teamwork is essential for addressing the diverse needs of patients, particularly those with greater surgical risk [5]. Engaging patients in their care, providing them with clear information, and addressing their concerns can significantly improve their satisfaction and overall experience.

Patient satisfaction with perioperative services is significantly influenced by different factors. Therefore, understanding these factors can help health care providers enhance the quality of perioperative care and improve patient experiences. The determining factors affecting patient satisfaction with perioperative services can be health care provider–related or patient-related. Health care provider–related factors include professional and interpersonal communication skills, perioperative care, patient involvement in care, staff skills and behaviors, multidisciplinary patient care approaches, staff attentiveness and support, access to care, infrastructure, and basic facilities. Patient-related factors include sociodemographic characteristics, such as age, baseline health status, overall experience, and stage of disease; perception of a relationship of trust with health care providers; and experience of intraoperative awareness, postoperative care, nausea and vomiting, postoperative complications, and feeling involved in decisions about their care [1,4,6-9].

Several studies have assessed patient satisfaction with perioperative services. These studies reported different levels of satisfaction in various areas. The overall level of satisfaction varied in studies conducted in the countries of Australia (96.8%) [7], India (73%) [10], and Eritrea (68.8%) [11] and at Sohag University Hospital (61.9%) [12]. In Ethiopia, there is a gap in understanding patient satisfaction with perioperative services at the national level. Evidence is fragmented and inconclusive, with no comprehensive national analysis available. A systematic appraisal of the literature confirms this significant knowledge gap, highlighting both the inconsistency of existing study results and the complete absence of any systematic review or meta-analysis synthesizing findings within the Ethiopian context [13-33].

The findings of this study will help policymakers, stakeholders, and other concerned bodies identify gaps and develop strategies to increase patient satisfaction with perioperative services. Moreover, it will help improve the quality of the health care delivery system by filling the gap accordingly. Understanding patient satisfaction in Ethiopia requires a national lens because the experience of care is shaped by different challenges, including scarce resources, cultural diversity, and geographic inequality. Therefore, the aim of this systematic review and meta-analysis is to determine overall patient satisfaction with perioperative services and identify associated factors in Ethiopia.

Methods

Study Design

In this study, a systematic review and meta-analysis were performed to determine overall patient satisfaction with perioperative services and their significant associated factors in Ethiopia. This study was conducted in line with the recommendations of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [34] (Checklist 1). This systematic review and meta-analysis has been registered on PROSPERO (CRD42024614957).

Inclusion and Exclusion Criteria

Inclusion Criteria

This study considered all research articles conducted in Ethiopia that met the inclusion criteria. Articles focused on patient satisfaction with perioperative services and their associated factors in Ethiopia were included. We included primary research studies that were published in the English language (researchers are encouraged or required to publish in English to gain international recognition), used observational and quantitative methods, and were available in electronic databases up to January 1, 2024. We included gray literature from preprints and institutional repositories to capture timely and locally relevant information. This approach minimized publication bias, time-lag bias, and geographic bias while maintaining methodological rigor through consistent quality appraisal and transparent reporting.

Exclusion Criteria

The systematic review and meta-analysis excluded studies conducted outside surgical departments, low-quality studies, those without clearly defined outcome variables or patient satisfaction information, and those lacking full-text access for data extraction.

Search Strategy and Sources of Information

The search strategy was adapted from and developed according to the population, exposure, outcomes, study design, and setting framework for creating MeSH terms to retrieve potential studies in databases. The population included patients who underwent surgical therapy. Exposure included descriptions of perioperative services, any specific interventions aimed at enhancing patient satisfaction, and factors influencing satisfaction, such as predictors, barriers, and determinants. The outcome included patient satisfaction with perioperative services. Study design

XSL•FO
RenderX

included observational quantitative studies. The setting, or context, was Ethiopia.

To identify relevant primary studies, we created the following review questions based on the previously mentioned framework: (1) What is the pooled prevalence of patient satisfaction with perioperative services in Ethiopia? (2) What factors influence patient satisfaction in Ethiopia?

Electronic Databases

Primary studies were subsequently found by searching PubMed/MEDLINE, HINARI, Web of Science, African Journals Online, Cochrane Library, and Scopus, as well as other Ethiopian research repositories for additional literature and sources. Data retrieval took place from January 1 to 30, 2024. In addition to our search strategy terms, we searched electronic databases using the following free-text terms: "patient satisfaction," "perioperative service," "perioperative care," "preoperative care," "postoperative care," "associated factors," "determinant factors," "operating room," "surgery," "recovery room," "post-anesthesia care unit (PACU)," and "Ethiopia." Articles were searched using Boolean operator strings with "AND" and "OR." All available articles were organized after the data were gathered and saved. We removed duplicate articles and incorporated studies focused on patient satisfaction with perioperative services and its associated factors in Ethiopia (for the full search strategy document, see Multimedia Appendix 1).

Study Selection Process

All included studies were imported into EndNote software (version 20; Clarivate PLC), and unnecessary files were removed from the analysis. Two independent reviewers (MAM and YAD) selected articles based on this study's objectives and eligibility criteria. They identified potentially eligible studies according to the inclusion criteria by title, abstract, and full text and, finally, screened and compiled the articles. Any disagreements between the reviewers during study selection and data extraction were resolved through discussion and consensus. If consensus could not be reached, a third reviewer (TAK) was consulted for a final decision.

Assessment of Study Quality

The 2 independent reviewers (MAM and YAD) independently reviewed each article that fulfilled the inclusion criteria and met the objectives by using a modified Newcastle-Ottawa Scale to confirm the quality of each study [35]. Any disagreements that occurred among the reviewers were resolved by consulting a senior researcher (TAK). The assessment scale evaluates each article's sample representativeness, adequacy, measurement tools, response rates, groups in outcomes, control of confounding variables, outcome evaluation, and statistical testing, yielding a total possible score of 10. A total score of >5 out of 10 indicated a low risk of bias, and all studies included in this systematic review and meta-analysis scored >7 out of 10 (Multimedia Appendix 2).

Data Extraction Tools

The data extraction format was designed and implemented in Microsoft Excel. The data extraction format included author names, study area, year of publication, working unit, study population, study design, sampling technique, sample size, assessment tools, and types of services. The format also included the prevalence for each primary outcome and the adjusted odds ratio for each factor (second outcome) with the corresponding 95% CI.

Measurement of Outcomes

This meta-analysis included 2 outcome variables. The primary outcome was overall patient satisfaction with perioperative services. The outcome was measured using validated assessment tools, such as the Leiden Perioperative Care Patient Satisfaction Questionnaire, or questionnaires adapted from validated tools. Overall patient satisfaction was categorized as satisfied or unsatisfied according to the assessment tool used in each study, based on the median or mean score. The secondary outcome was factors affecting patient satisfaction with perioperative services. Significant factors were determined by analyzing adjusted odds ratios from the included studies. For the meta-analysis, only combined results from studies that controlled for conceptually similar confounders, such as age, sex, and type of surgery, were used.

Statistical Analysis

This meta-analysis used Stata software (version 17; StataCorp) to analyze the retrieved data. Overall prevalence was estimated using SEs, which were calculated as $p = r/n$ and $SE = \sqrt{p(1-p)/n}$, where p is the proportion, r is the total population size, and n is the sample size. The results are reported as the overall prevalence of patient satisfaction with perioperative services with 95% CI, and $P<.05$ was considered statistically significant. This study also used Stata software to identify factors significantly associated with perioperative patient satisfaction. The analysis used a random-effects model with the restricted maximum likelihood method to account for each study's variability [36]. We assessed heterogeneity using the I^2 statistic and the Cochran Q test. I^2 measures the proportion of variation across studies due to heterogeneity, estimating the difference observed and assigning a value from 0% to 100%. I^2 values below 25% imply low heterogeneity, values between 25% and 50% indicate moderate heterogeneity, values between 51% and 74% indicate substantial heterogeneity, and values greater than 75% suggest high heterogeneity. The Cochran Q test was used to estimate whether the observed differences in effect size were statistically significant, which would suggest the presence of heterogeneity [37].

To determine the heterogeneity level between studies, a subgroup analysis was conducted. Publication bias was determined by checking funnel plots for asymmetry and using the Egger test and a nonparametric trim-and-fill analysis. Each of these methods can be used to identify publication bias, but the Egger test is the most objective and effective in providing precise values [38]. A sensitivity analysis was used to evaluate the robustness of each study's results. To identify sources of heterogeneity, study-level covariates were analyzed for their effect on perioperative satisfaction via univariate meta-regression.

XSL•FO
RenderX

Ethical Considerations

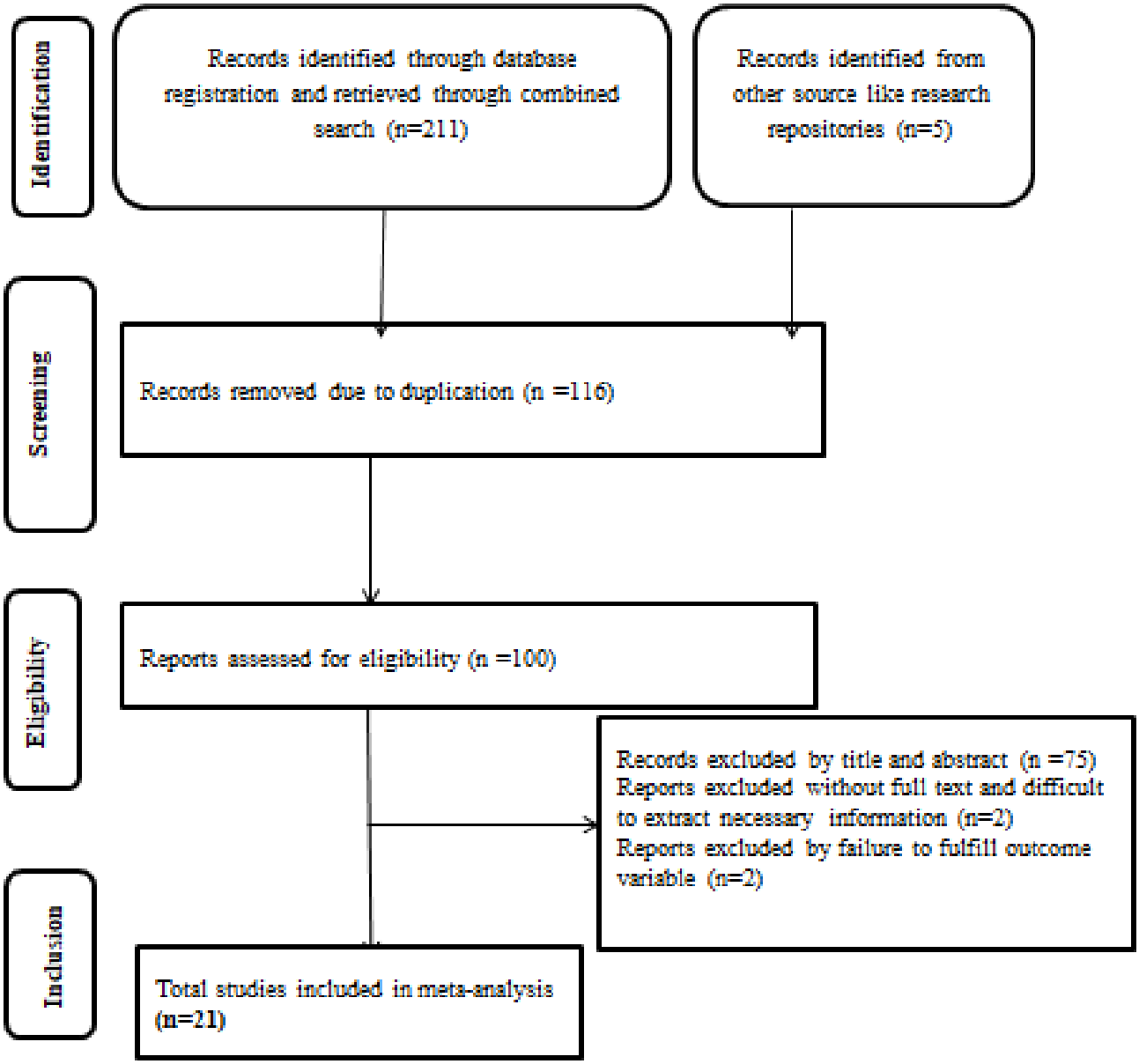
This study did not require ethics approval. The research constitutes a systematic review of existing, publicly available scientific literature. All data analyzed were extracted from previously published studies that had already obtained their own necessary ethics approvals and participant consent. Therefore, as no new human or animal subjects were involved, ethics approval for this analysis was not compulsory.

Results

Literature Search

To conduct this systematic review and meta-analysis, we performed a search as outlined above, and 216 published records were retrieved. From the retrieved studies, a total of 116 published articles were excluded due to duplication or ineligibility by the inclusion criteria. Following a meticulous review of the titles and abstracts of the 100 remaining records, 79 articles were excluded. Finally, a meta-analysis was conducted on the remaining 21 studies (Figure 1).

Figure 1. Flowchart of selection for systematic review and meta-analysis to assess patient satisfaction with perioperative services and associated factors in Ethiopia.



Characteristics of the Included Studies

All cross-sectional studies included in this systematic review and meta-analysis assessed patient satisfaction with perioperative services. The included studies examined 5 areas of Ethiopia: Addis Ababa [15,19,20,24,25], Amhara Region [13,14,16,17,21,27,28,30-33], Oromia Region [23,29], Tigray Region [22], and the Southern Nations, Nationalities, and Peoples (SNNP) Region [18,26]. The minimum and maximum sample sizes of the included studies were 120 and 468, respectively (Table 1).

XSL•FO
RenderX

Table . Characteristics of the 21 studies included in the meta-analysis and systematic review of patient satisfaction with perioperative services and associated factors in Ethiopia [13-33].

Author name and publication year	Area of study	Sampling technique	Assessment tools	Service provided	Sample size, n	Prevalence of patient satisfaction, n (%)	NOS[a] score of quality assessment out of 10
Fetene et al [13] (2022)	Amhara Region	Random	Adopted validated tools	All perioperative	411	263 (63.99)	8
Endale Simegn et al [14] (2021)	Amhara Region	Nonrandom/consecutive	Adopted validated tools	Perioperative anesthetic service	398	295 (74.12)	8
Shamil [15] (2021)	Addis Ababa	Random	Adopted validated tools	Nursing care	299	196 (65.55)	9
Teshome et al [16] (2022)	Amhara Region	Nonrandom/consecutive	Adopted validated tools	Perioperative anesthetic service	387	242 (62.53)	9
Alemu et al [17] (2023)	Amhara Region	Random	Adopted validated tools	All perioperative	422	290 (68.72)	8
Ataroe et al [18] (2024)	SNNPE[b] Region	Nonrandom/consecutive	Adapted questionnaire	Nursing care	468	372 (79.49)	9
Deressa et al [19] (2022)	Addis Ababa	Random	Adapted questionnaire	Nursing care	414	346 (83.57)	9
Kibru et al [20] (2023)	Addis Ababa	Random	Adapted questionnaire	All perioperative	287	276 (96.17)	8
Derso et al [21] (2024)	Amhara Region	Random	Adapted questionnaire	Nursing care	383	238 (62.14)	7
Benwu and Gebremedhin [22] (2019)	Tigray Region	Nonrandom/consecutive	Adapted questionnaire	Perioperative anesthetic service	120	106 (88.33)	9
Alemu [23] (2015)	Addis Ababa	Nonrandom/consecutive	Adapted questionnaire	Preoperative care	224	162 (72.32)	9
Obsa et al [24] (2017)	Oromia Region	Nonrandom/consecutive	Adapted questionnaire	Perioperative anesthetic service	184	144 (78.26)	8
Abdissa et al [25] (2023)	Addis Ababa	Random	Adapted questionnaire	All perioperative	346	286 (82.66)	9
Siraneh et al [26] (2020)	SNNPE Region	Nonrandom/consecutive	Adapted questionnaire	Perioperative anesthetic service	200	120 (60.00)	10
Ayele et al [27] (2022)	Amhara Region	Random	Adapted questionnaire	Preoperative care	404	284 (70.30)	7
Belay Bizuneh et al [28] (2020)	Amhara Region	Random	Adapted questionnaire	Postoperative pain management	418	302 (72.25)	7
Biyazin et al [29] (2022)	Oromia Region	Random	Adapted questionnaire	Preoperative care	372	160 (43.01)	8
Admass et al [30] (2024)	Amhara Region	Random	Adapted questionnaire	Postoperative pain management	424	340 (80.19)	9
Ashebir et al [31] (2024)	Amhara Region	Nonrandom/consecutive	Adapted questionnaires	All perioperative	195	168 (86.15)	8
Bayable et al [32] (2020)	Amhara Region	Nonrandom/consecutive	Adapted questionnaire	All perioperative	382	315 (82.46)	7
Demilew et al [33] (2021)	Amhara Region	Nonrandom/consecutive	Adapted Questionnaire	Perioperative anesthetic service	120	96 (80.00)	8

XSL•FO
RenderX

[a]NOS: Newcastle-Ottawa Scale.

[b]SNNPE: Southern Nations, Nationalities, and Peoples of Ethiopia.

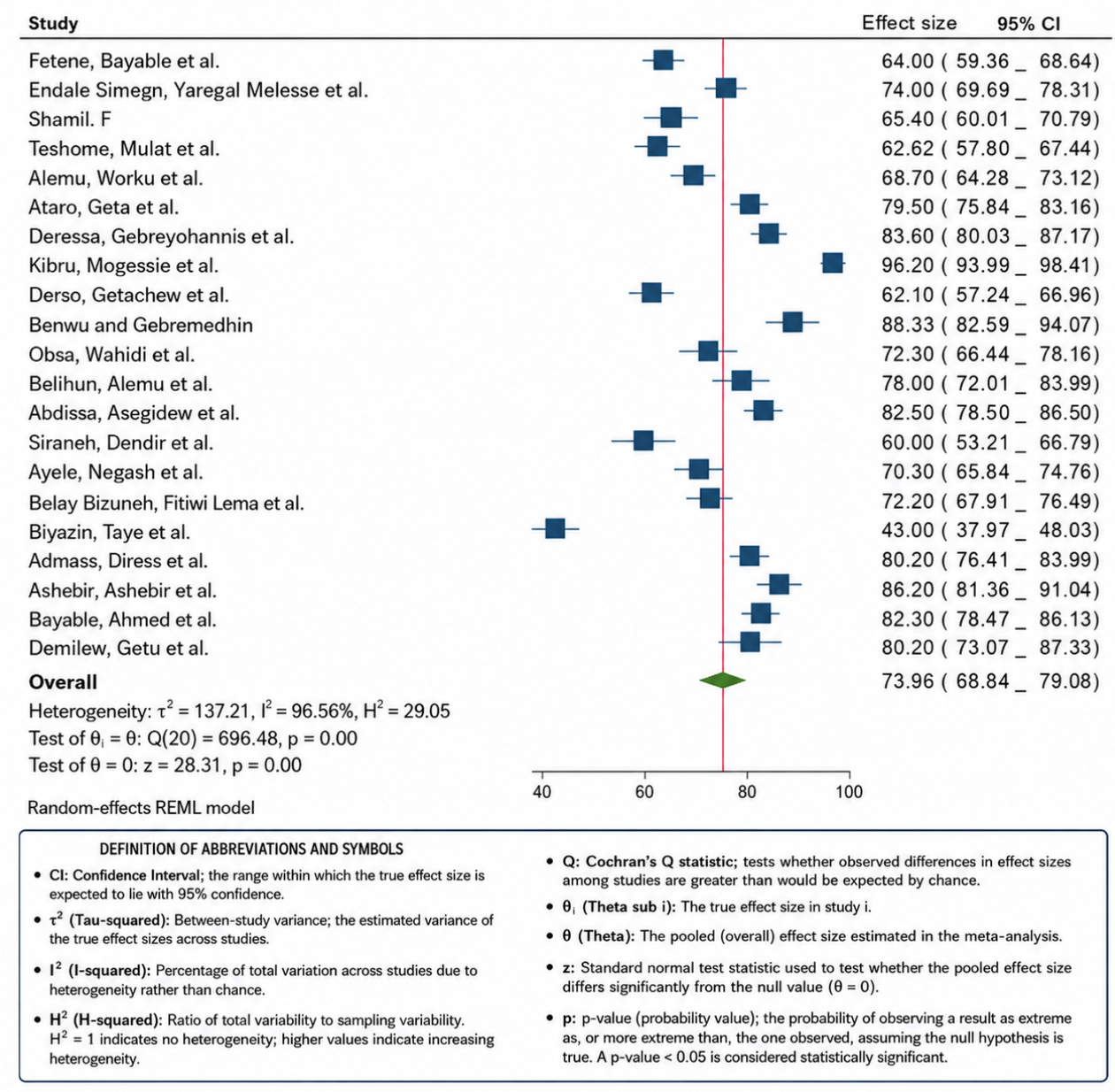
Publication Bias

In this meta-analysis and systematic review, publication bias was assessed using a funnel plot and the Egger test. The funnel plot showed an asymmetrical distribution across the included studies (Multimedia Appendix 3). The Egger test was not statistically significant (P=.07), which suggests no publication bias in the included studies. In addition, a nonparametric trim-and-fill analysis was conducted to estimate the overall effect of the imputed studies. The trim-and-fill analysis revealed that no imputed studies affected the overall prevalence of perioperative patient satisfaction (5072/6858, 73.96%, 95% CI 68.84%-79.08%).

Pooled Prevalence of Perioperative Patient Satisfaction

In total, 21 studies and 6858 participants were included in this analysis. The pooled prevalence of perioperative patient satisfaction was 73.96% (5072/6858, 95% CI 68.84-79.08) using a random-effects model (I^2=96.56%; Q_{20}=696.48; P<.001) (Figure 2).

Figure 2. The forest plot was used to assess the pooled prevalence of patient satisfaction with perioperative services [13-33]. REML: restricted maximum likelihood.



Sensitivity Analysis

A sensitivity analysis was conducted using a random-effects model to identify any outliers or influential studies that affected the overall prevalence of perioperative patient satisfaction. In the analysis, no influential studies were identified, which was confirmed by the fact that all the point estimates remained within the 95% CI (Multimedia Appendix 4).

XSL•FO
RenderX

Subgroup Analysis

Subgroup analyses revealed significant variation in the prevalence of perioperative patient satisfaction. The overall effect was significant (P<.001), with substantial heterogeneity (I^2=96.56%). Prevalence varied notably by region (Tigray: 106/120, 88.33% vs Oromia: 286/473, 60.46%), publication year (pre-2020: 814/1023, 79.56% vs post-2020: 6099/9519, 73.04%), sample size (<300: 706/899, 78.50% vs ≥300: 6207/9643, 71.75%), sampling technique (nonrandom: 2830/4844, 76.42% vs random: 4083/5698, 71.75%), population (all adult surgical patients: 4191/5591, 75.00% vs elective: 2722/4951, 71.85%), and service type (all perioperative: 3618/4519, 80.07% vs preoperative: 598/967, 61.85%). The overall heterogeneity of this subgroup analysis was I^2=96.56%, indicating substantial heterogeneity among the subgroup analyses. The meta-regression analysis confirmed that there was no statistically significant association between the pooled effect size and any of the predefined study-level covariates, including publication year, study area, sample size, type of service provided, study population, sampling technique, or assessment tool (Multimedia Appendix 5).

Meta-Regression Analysis

Meta-regression is used to explore and explain the heterogeneity between study results. It is used to test hypotheses about what factors cause an effect size to be larger in some studies and smaller in others. This systematic review utilized meta-regression to investigate the sources of the considerable heterogeneity (I^2=96.56%) observed among the included studies. The meta-regression analysis showed no statistically significant associations between the pooled effect size and any of the predefined study-level covariates, including publication year, study area, sample size, types of service provided, study population, sampling technique, or assessment tool (Table 2).

Table . Univariate meta-regression analysis results for prevalence of patient satisfaction with perioperative services in Ethiopia.

Specific variable	Coefficient (β)	SE	*P* value	95% CI
Year of publication	−2.93	8.62	.73	−19.82 to 13.96
Sampling technique	−7.10	6.25	.26	−19.35 to 5.15
Assessment tool	−4.32	3.16	.17	−10.50 to 1.87
Sample size category	−9.21	5.36	.09	−19.71 to 1.30
Study population	−0.29	3.58	.94	−7.31 to 6.73
Provided service	−3.55	1.93	.07	−7.32 to 0.23
Study area	−3.60	2.82	.20	−9.14 to 1.93

Factors Affecting Patient Satisfaction With Perioperative Services in Ethiopia

Patient satisfaction with perioperative services was 2.23 times higher among patients satisfied with effective postoperative pain management than patients who had no postoperative pain management (adjusted odds ratio [AOR] 2.23, 95% CI 1.56-2.90; I^2=0.00%; Q_5=1.10; Figure 3).

Satisfaction was 3.18 times greater among illiterate patients than among more educated individuals (AOR 3.18, 95% CI 1.23-5.13; I^2=0.00%; Q_3=0.53; Figure 4).

Patients with a primary school education had 6.55 times higher odds of being satisfied with perioperative services than those with a secondary education or higher (AOR 6.55, 95% CI 3.61-9.49; I^2=49.21%; Q_3=5; Figure 5).

Patients who received local anesthesia were 2.80 times more likely to be satisfied with perioperative services than those who received general anesthesia (AOR 2.80, 95% CI 2.03-3.57; I^2=0.00%; Q_5=1.19; Figure 6).

Patients with a history of surgery or anesthesia were 2.76 times more likely to be satisfied with perioperative services than those without previous surgery or anesthesia (AOR 2.76, 95% CI 1.51-4.01; I^2=0.00%; Q_2=1.03; Figure 7).

XSL•FO
RenderX

Figure 3. The forest plot was used to assess the association between postoperative pain management and patient satisfaction with perioperative services [13-15,26,28,30]. REML: restricted maximum likelihood.

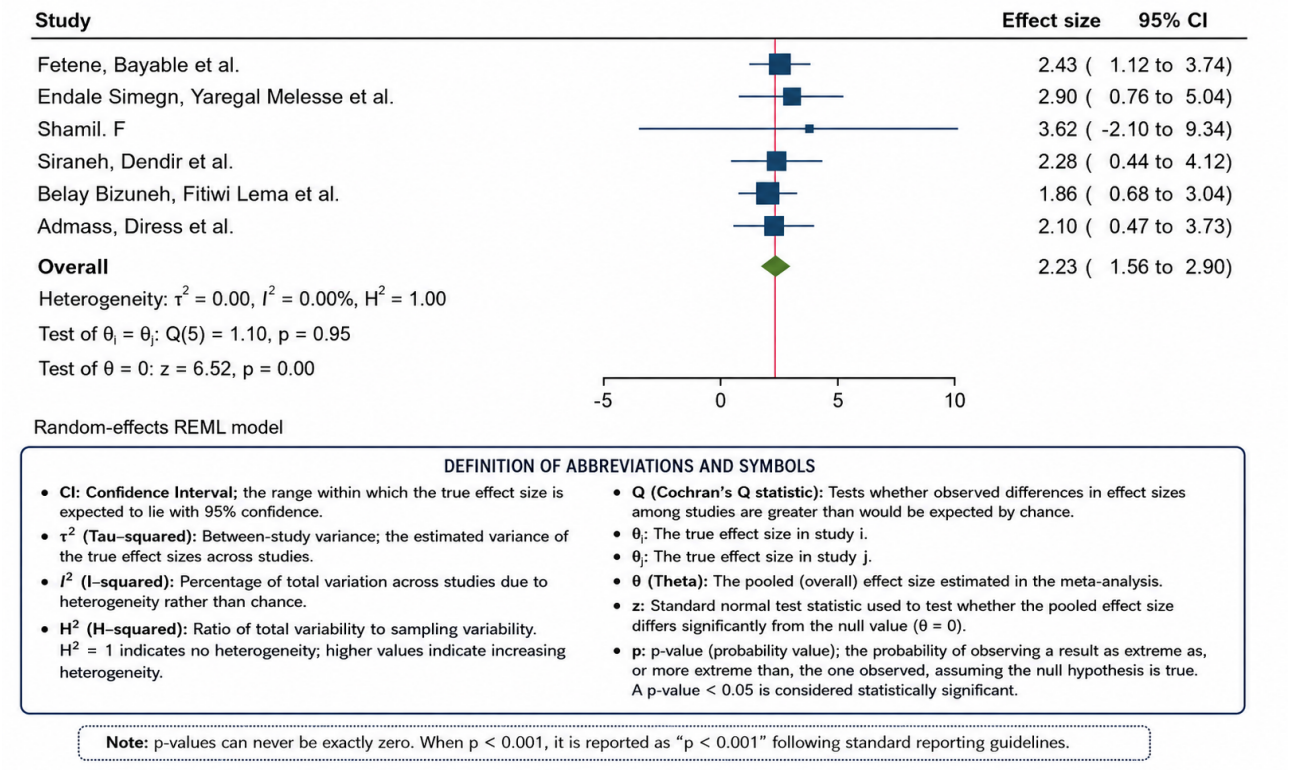
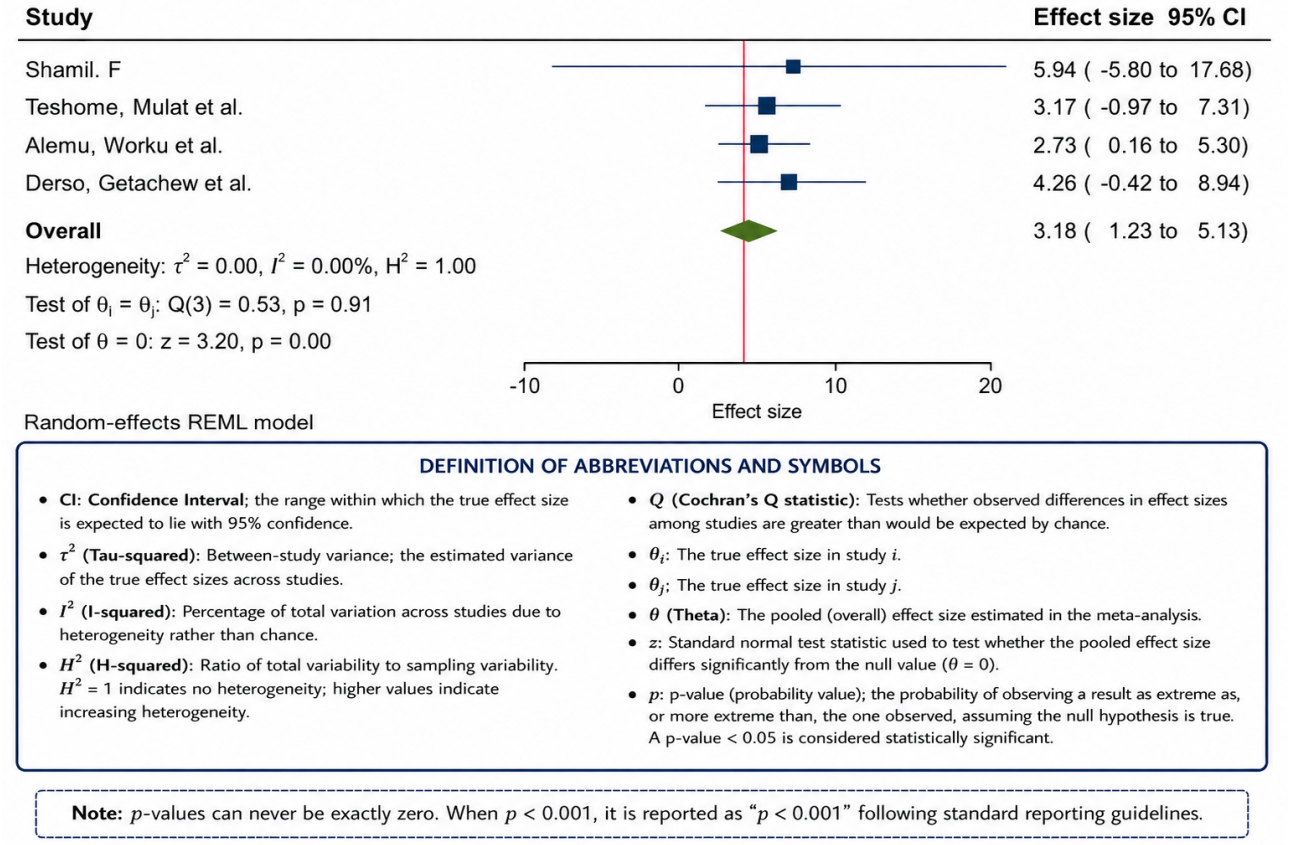


Figure 4. The forest plot was used to assess the association between illiteracy and patient satisfaction with perioperative services [15-17,21]. REML: restricted maximum likelihood.



XSL•FO
RenderX

Figure 5. The forest plot was used to assess the association between primary school and patient satisfaction with perioperative services [15,17,21,27]. REML: restricted maximum likelihood.

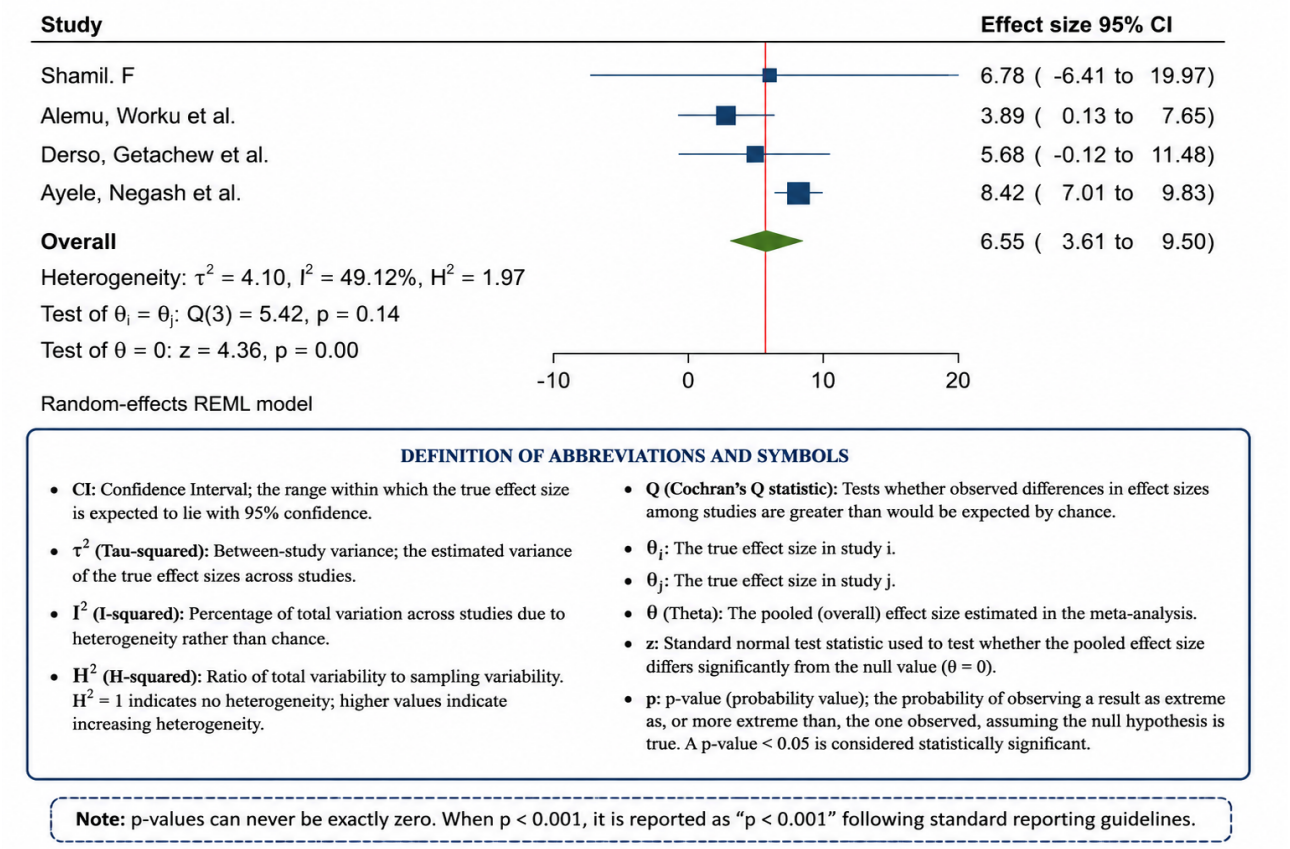
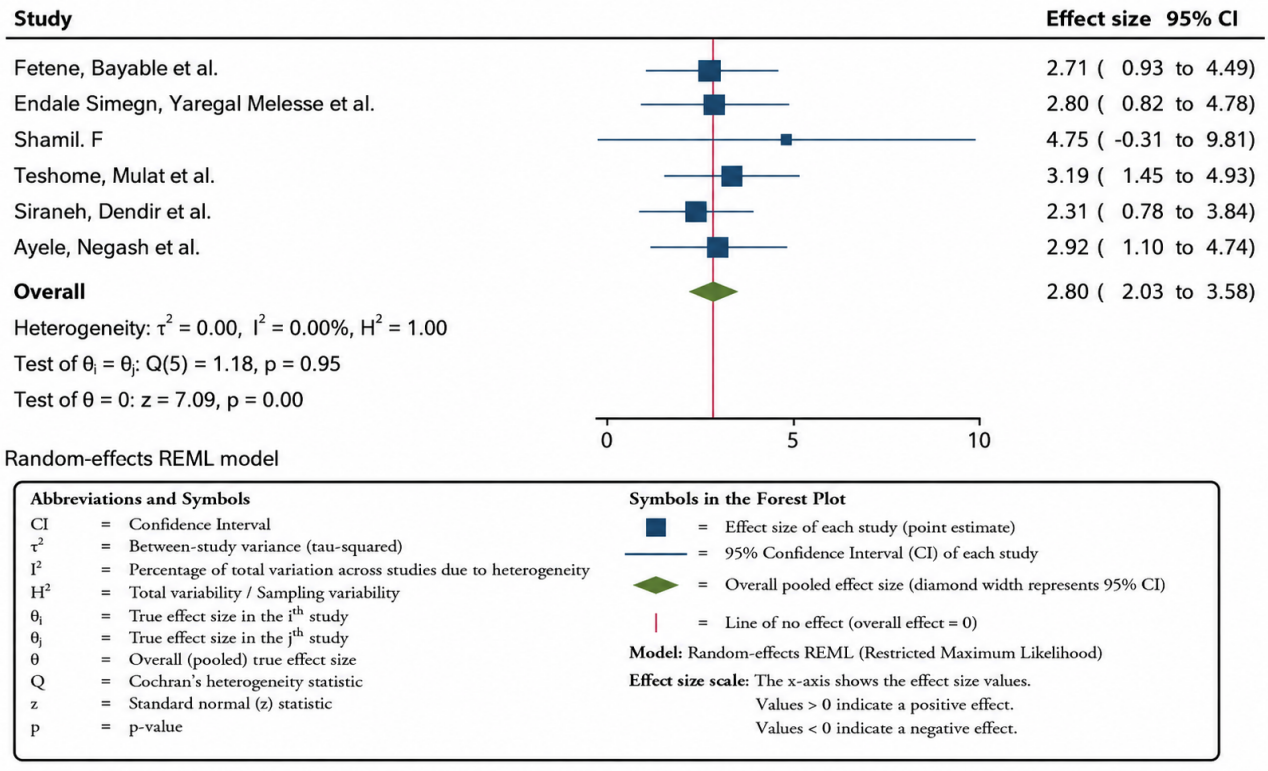
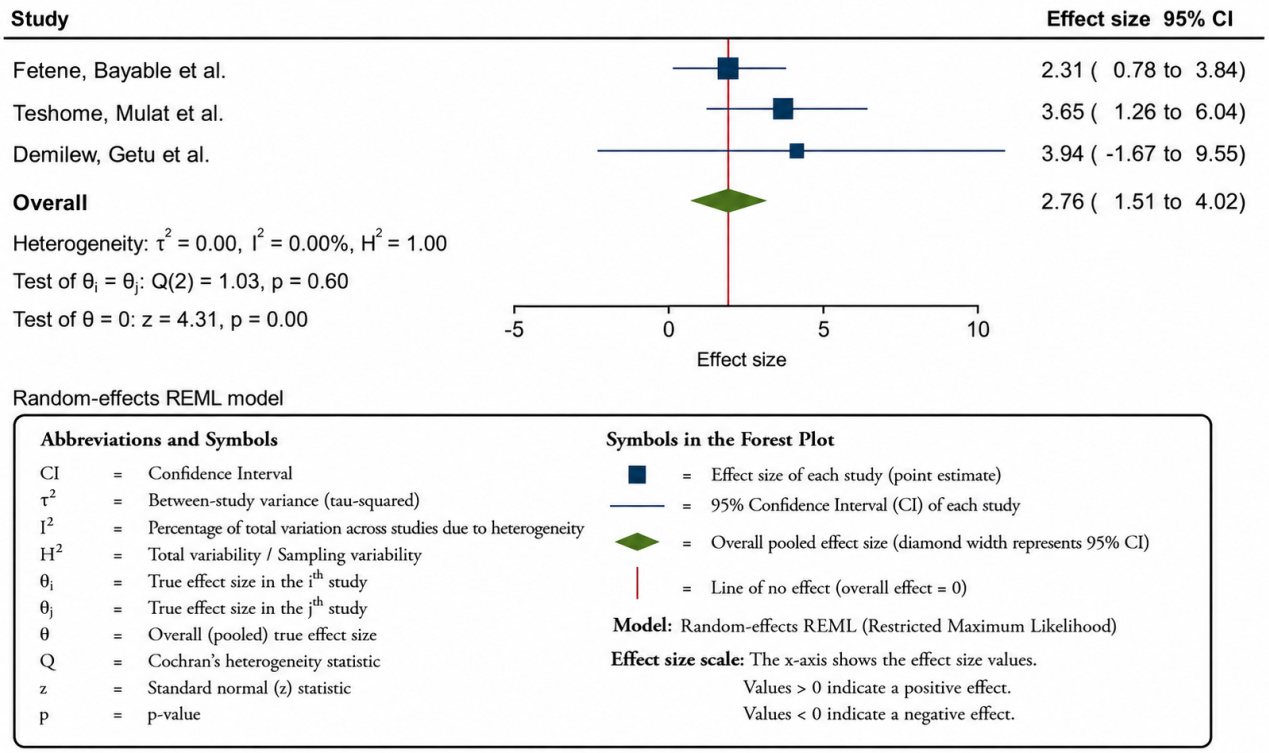


Figure 6. The forest plot was used to assess the association between local anesthesia and patient satisfaction with perioperative services [13-16,26,27]. REML: restricted maximum likelihood.



XSL•FO
RenderX

Figure 7. The forest plot was used to assess the association between a previous history of anesthesia or surgery and patient satisfaction with perioperative services [13,16,33]. REML: restricted maximum likelihood.



Discussion

Principal Findings

This systematic review and meta-analysis synthesized data from 21 studies comprising 6858 participants to estimate the pooled prevalence of patient satisfaction with perioperative services in Ethiopia and identify key determinants of satisfaction. This study is significant because it provides valuable insights into patient satisfaction and informs future approaches to enhance patient well-being and recovery.

The analysis revealed that the pooled prevalence of patient satisfaction was 73.96% (5072/6858, 95% CI 68.84%-79.08%), indicating that more than one-quarter of patients were not fully satisfied with their care. This finding was associated with substantial heterogeneity (I^2=96.56%; Q_{20}=696.48; P<.001). This extremely high heterogeneity indicates that the estimated prevalence should not be interpreted as a single, precise national figure but rather as an overall exploratory estimate. The meta-regression analysis was used to assess the effects of the measured covariates: publication year, study area, and sample size. It suggested that unmeasured factors such as the quality of patient care, patient demographics, or hospital type might be the primary drivers of the variation between studies. Despite this heterogeneity, the pooled estimate is remarkably consistent with findings from other nations, including Nigeria (71%) [39], Portugal (74.5%) [40], and Saudi Arabia (73.5%) [41]. Possible reasons for this consistency could be comparable quality of perioperative services, similar health care challenges and systems, and ongoing health care improvement initiatives in these countries.

Patient satisfaction was greater in the included studies than in studies conducted at Sohag University Hospital (61.9%) [12], in Rwanda (67.43%) [42], at Mount Sinai Hospital (60%) [43], and in the United Kingdom (86.7%) [44]. This discrepancy might be related to disparities in health care systems, including resource access and quality of care; differences in patients' demographics and health status; methodological variations including differences in study design, patient recruitment methods, and assessment tools; cultural perceptions of health care and patient-provider relationships; and focus on patient-centered care. Patient satisfaction was lower in the included studies than in studies conducted in tertiary hospitals in Cameroon (39/45, 86.7%) [45], Nigeria (94/108, 86.85%) [46], the United Kingdom (4595/4709, 97%) [47], and Australia (11067/11,410, 96.8%) [7]. This discrepancy could be due to variations in the quality and availability of infrastructure, resource allocation (staffing levels, medical equipment, and facilities), cultural expectations and experiences, methodological differences, and health care quality initiatives.

The subgroup analysis revealed several factors contributing to the heterogeneity observed across studies. The higher prevalence of patient satisfaction observed in the Tigray Region may be attributed to the limited number of studies conducted there, which potentially introduced biases into the pooled estimate. Studies published before 2020 reported higher satisfaction levels compared to those published afterward. This shift may be explained by greater access to health-related information in recent years, which has raised patient expectations. When these heightened expectations are not met, satisfaction may decline. Studies with sample sizes below 300 participants showed higher satisfaction than those with 300 or more participants. Smaller samples are often more prone to selection and publication biases

XSL•FO
RenderX

and tend to involve more homogeneous populations, which may yield inflated satisfaction estimates. In contrast, larger samples capture greater clinical diversity, often resulting in lower reported satisfaction.

Regarding sampling methodology, nonrandom sampling was associated with higher satisfaction than random sampling. This discrepancy likely reflects selection bias, as nonrandom methods allow researchers to preferentially include individuals with positive experiences, thereby inflating satisfaction rates. Satisfaction was higher among all adult surgical patients compared to those undergoing elective surgery alone. This difference may be explained by variations in surgical urgency, patient expectations, and the overall perioperative experience. Differences in study population also contributed to variation in satisfaction estimates. Patient satisfaction was greater when all perioperative services were considered together than when preoperative services were examined alone. This may be due to the comprehensive nature of full perioperative care, including timely support, effective postoperative pain management, and better alignment with patient expectations.

The findings showed that perioperative patient satisfaction was 2.23 times greater among patients with effective postoperative pain management than those without. This finding was supported by studies conducted in Portugal [40], the United States [48], and Eritrea [11]. This could be due to effective postoperative pain management leading to smoother recovery. Patients frequently associate effective pain relief with high-quality care, which results in more positive experiences and significantly impacts patients' emotional and psychological well-being.

Patient satisfaction with perioperative services was 3.18 times greater among illiterate patients than among those with higher education. Patient perioperative satisfaction was 6.55 times higher among those in primary school than among those in secondary school and above. A study conducted in Saudi Arabia supported this finding [49]. It is likely that less education is not itself a cause of satisfaction but that it is a proxy for other factors (eg, lower expectations, different communication styles, or fear of authority leading to reported satisfaction). Illiterate patients may have lower expectations about their surgical experience, are less likely to evaluate their care critically, and may benefit more from emotional support and reassurance during the process. This finding frames illiteracy not as a positive factor but as a potential indicator of health literacy gaps and a need for better patient communication and empowerment to ensure that satisfaction reports reflect true care quality rather than perceived power dynamics.

Patients who received local anesthesia were 2.80 times more satisfied with perioperative services than those who received general anesthesia. This is congruent with the results of a study carried out in Eritrea [11]. A possible explanation is that local anesthesia tends to reduce anxiety in patients, as they remain conscious during the procedure. It also allows for quicker recovery, enabling them to return to normal activities sooner. In addition, it has fewer side effects, such as grogginess or nausea, and involves less invasive methods, all contributing to the safety and effectiveness of the procedure. Patients with a history of surgery or anesthesia were 2.76 times more satisfied with perioperative services than those without such a history. This result was consistent with a study carried out in Saudi Arabia [49]. The possible reason is that patients with a history of surgery or anesthesia may have a clearer understanding of the perioperative process, greater trust in health care providers, more active participation in care discussions, familiarity with procedures, and realistic expectations, as their past experiences shape their perceptions of quality and care.

Strengths and Limitations

In this study, we performed a systematic review and meta-analysis, along with subgroup, meta-regression, and sensitivity analyses. However, this work included only cross-sectional studies, which may limit the ability to establish causal relationships, and the included studies exhibited a high level of heterogeneity. This study also did not include a trial sequential analysis. These limitations should be considered when interpreting the findings.

Conclusions and Recommendations

Overall patient satisfaction with perioperative services in Ethiopia was good. Significant factors influencing perioperative patient satisfaction include effective postoperative pain management, patients' educational background, use of local anesthesia, and patients' previous experiences with surgery or anesthesia. To further improve patient satisfaction, health care facilities should prioritize comprehensive pain management strategies, ensure the delivery of clear and accessible information about perioperative services, and invest in ongoing training for surgical and anesthesia teams. These steps are essential for enhancing perioperative patient satisfaction and quality of patient care in Ethiopia.

Study Implications for Policymakers

Focusing on comprehensive postoperative pain management strategies, increasing patient awareness and access to information about perioperative services, and investing in ongoing training for surgical and anesthesia teams are crucial steps. Policymakers can significantly advance the quality of perioperative services and enhance overall patient satisfaction with the health care system by adopting these measures.

Acknowledgments

All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

XSL•FO
RenderX

Funding

The authors declared no financial support was received for this work.

Data Availability and Materials

All relevant data required to conduct this meta-analysis are presented within the manuscript. The dataset supporting the conclusions of this study is available in the manuscript/supplementary materials.

Authors' Contributions

MAM and YAD conceived the study, conducted the literature search, performed the statistical analyses, and prepared the study report. All authors contributed to the conceptualization, design, data procurement, analysis and interpretation, software application, and critical revision of the manuscript. All authors have reviewed and approved the final version and submission of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full electronic search strategy.
[DOCX File, 20 KB - periop_v9i1e84457_app1.docx]

Multimedia Appendix 2
Methodological quality assessment of cross-sectional studies using a modified Newcastle-Ottawa Scale.
[DOCX File, 23 KB - periop_v9i1e84457_app2.docx]

Multimedia Appendix 3
Funnel plot for the assessment of publication bias for the pooled prevalence of patient satisfaction with perioperative service.
[PNG File, 13 KB - periop_v9i1e84457_app3.PNG]

Multimedia Appendix 4
The result of the sensitivity analysis conducted on patient satisfaction toward perioperative services and associated factors in Ethiopia.
[PNG File, 55 KB - periop_v9i1e84457_app4.PNG]

Multimedia Appendix 5
Subgroup analysis to assess heterogeneity in the study of patient satisfaction toward perioperative service and associated factors in Ethiopia. K is the number of studies.
[PNG File, 58 KB - periop_v9i1e84457_app5.PNG]

Checklist 1
PRISMA 2020 checklist.
[PDF File, 236 KB - periop_v9i1e84457_app6.pdf]

References

1. Gupta M, Koul M, Mehmood K. Patient satisfaction with OPD services through different healthcare determinants: a study at a newly constructed tertiary care teaching hospital in Northern India. Int J Med Sci Cur Res 2023;6(4):510-518 [FREE Full text]
2. Chow A, Mayer EK, Darzi AW, Athanasiou T. Patient-reported outcome measures: the importance of patient satisfaction in surgery. Surgery 2009 Sep;146(3):435-443. [doi: 10.1016/j.surg.2009.03.019] [Medline: 19715800]
3. Tsai TC, Orav EJ, Jha AK. Patient satisfaction and quality of surgical care in US hospitals. Ann Surg 2015 Jan;261(1):2-8. [doi: 10.1097/SLA.0000000000000765] [Medline: 24887985]
4. Trinh LN, Fortier MA, Kain ZN. Primer on adult patient satisfaction in perioperative settings. Perioper Med (Lond) 2019;8(1):11. [doi: 10.1186/s13741-019-0122-2] [Medline: 31548883]
5. Woodhead K, Fudge L. The context of perioperative care. In: Woodhead K, Fudge L, editors. Manual of Perioperative Care: John Wiley & Sons, Ltd; 2012:1-10 URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118702734 [accessed 2026-07-17] [doi: 10.1002/9781118702734]

XSL•FO
RenderX

6. Lubis* B, Ginting R, Novalinda C, et al. Analysis of factors affecting the quality of service toward inpatient patient satisfaction BPJS DIRSU Bandung. Int J Manag Humanit 2021;5(5):14-17 [FREE Full text] [doi: 10.35940/ijmh.E1217.015521]
7. Myles PS, Williams DL, Hendrata M, Anderson H, Weeks AM. Patient satisfaction after anaesthesia and surgery: results of a prospective survey of 10,811 patients. Br J Anaesth 2000 Jan;84(1):6-10. [doi: 10.1093/oxfordjournals.bja.a013383] [Medline: 10740539]
8. Kouki P, Matsota P, Christodoulaki K, et al. Greek surgical patients' satisfaction related to perioperative anesthetic services in an academic institute. Patient Prefer Adherence 2012;6:569-578. [doi: 10.2147/PPA.S34244] [Medline: 22927750]
9. Lehrich BM, Goshtasbi K, Brown NJ, et al. Predictors of patient satisfaction in spine surgery: a systematic review. World Neurosurg 2021 Feb;146:e1160-e1170. [doi: 10.1016/j.wneu.2020.11.125] [Medline: 33253954]
10. Sharma A, Kasar P, Sharma R. Patient satisfaction about hospital services: a study from the outpatient department of tertiary care hospital, Jabalpur, Madhya Pradesh, India. Natl J Community Med 2014 Jun 30;5(2):199-203 [FREE Full text]
11. Andemeskel YM, Elsholz T, Gebreyohannes G, Tesfamariam EH. Patient satisfaction with peri-operative anesthesia care and associated factors at two national referral hospitals: a cross sectional study in Eritrea. BMC Health Serv Res 2019 Sep 18;19(1):669. [doi: 10.1186/s12913-019-4499-x] [Medline: 31533708]
12. Abd El-Nasser G, Mohamed N. Patient satisfaction with perioperative care and its relationship with patient characteristics. Med J Cairo Univ 2013 Mar;81(2) [FREE Full text]
13. Fetene MB, Bayable SD, Wendimu ES, et al. Perioperative patient satisfaction and its predictors following surgery and anesthesia services in North Shewa, Ethiopia. A multicenter prospective cross-sectional study. Ann Med Surg (Lond) 2022 Apr;76:103478. [doi: 10.1016/j.amsu.2022.103478] [Medline: 35308431]
14. Endale Simegn A, Yaregal Melesse D, Belay Bizuneh Y, Mekonnen Alemu W. Patient satisfaction survey on perioperative anesthesia service in University of Gondar Comprehensive Specialized Hospital, northwest Ethiopia, 2021. Anesthesiol Res Pract 2021;2021(1):3379850. [doi: 10.1155/2021/3379850] [Medline: 34721572]
15. Shamil F. Surgical patients' satisfaction with perioperative nursing care and its associated factors at public teaching hospitals, Addis Ababa, Ethiopia, 2021 [Master's thesis]. : Addis Ababa University; 2021 URL: https://www.scribd.com/document/670476555/Preoperative [accessed 2026-07-20]
16. Teshome D, Mulat Y, Fenta E, et al. Patient satisfaction and its associated factors towards perioperative anesthesia service among surgical patients: a cross-sectional study. Heliyon 2022 Mar;8(3):e09063. [doi: 10.1016/j.heliyon.2022.e09063] [Medline: 35287330]
17. Alemu ME, Worku WZ, Berhie AY. Patient satisfaction and associated factors towards surgical service among patients undergoing surgery at referral hospitals in western Amhara Regional State, Ethiopia. Heliyon 2023 Mar;9(3):e14266. [doi: 10.1016/j.heliyon.2023.e14266] [Medline: 36938460]
18. Ataro BA, Geta T, Endirias EE, Gadabo CK, Bolado GN. Patient satisfaction with preoperative nursing care and its associated factors in surgical procedures, 2023: a cross-sectional study. BMC Nurs 2024 Apr 8;23(1):235. [doi: 10.1186/s12912-024-01881-5] [Medline: 38584268]
19. Deressa D, Gebreyohannis T, Boka A, Geneti Y. Satisfaction with preoperative education and surgical services among adults elective surgical patients at selected public hospitals in Addis Ababa, Ethiopia. SAGE Open Med 2022;10:20503121221143219. [doi: 10.1177/20503121221143219] [Medline: 36532951]
20. Kibru EA, Mogessie YG, Regassa AA, Hailu KT. Patient satisfaction with post-operative surgical services and associated factors at Addis Ababa City government tertiary hospitals' surgical ward, cross-sectional study, 2022. Pan Afr Med J 2023;45(1):189. [doi: 10.11604/pamj.2023.45.189.38416] [Medline: 38020351]
21. Derso YA, Getachew D, Adisu MA, Adu A. Elective surgery patients satisfaction towards perioperative service and its associated factors at Amhara Region comprehensive specialized hospitals of Ethiopia. BMC J Med Sci 2024;5(1):3-11. [doi: 10.70905/bmcj.05.01.0258]
22. Benwu KM, Gebremedhin HG. A prospective study on elective surgical inpatient satisfaction with perioperative anaesthesia service at Ayder comprehensive specialized hospital, Mekelle, Ethiopia. BMC Anesthesiol 2019 Apr 1;19(1):46. [doi: 10.1186/s12871-019-0696-8] [Medline: 30935376]
23. Mengistu Alemu AB. A prospective study on surgical inpatient satisfaction with perioperative anaesthetic service in Jimma University Specialized Hospital, Jimma, South West Ethiopia. J Anesth Clin Res 2015;06(3):1-9. [doi: 10.4172/2155-6148.1000514]
24. Obsa MS, Melekamayhu A, Admas W, Woji BG. Assessment of patient satisfaction with preoperative anesthetic evaluation and associated factors at Menelik II Referral Hospital Addis Ababa, Ethiopia. J Med Physiol Biophys 2017;35 [FREE Full text]
25. Abdissa G, Asegidew A, Bekele S. Assessment of postoperative patient satisfaction towards services provision in Zewditu Memorial Hospital Addis Ababa Ethiopia. A cross-sectional study. Gen Surg Clin Med 2023;1(1):37-47 [FREE Full text]
26. Siraneh AG, Dendir G, Paulos K. Factors associated with patient satisfaction in perioperative anesthesia care at Hawassa university comprehensive specialized hospital, Ethiopia. Cross-sectional study design. Int J Surg Open 2020;26:114-118. [doi: 10.1016/j.ijso.2020.08.016]

XSL•FO
RenderX

27. Ayele TT, Negash TT, Oumer KE, et al. Patients' satisfaction and associated factors towards preoperative informed consent process: A cross-sectional study. Ann Med Surg (Lond) 2022 Jul;79:104104. [doi: 10.1016/j.amsu.2022.104104] [Medline: 35860089]
28. Belay Bizuneh Y, Fitiwi Lema G, Yilkal Fentie D, Woldegerima Berhe Y, Enyew Ashagrie H. Assessment of patient's satisfaction and associated factors regarding postoperative pain management at the University of Gondar compressive specialized hospital, northwest Ethiopia. Pain Res Manag 2020;2020(1):8834807. [doi: 10.1155/2020/8834807] [Medline: 33273994]
29. Biyazin T, Taye A, Belay Y. Patient satisfaction with surgical informed consent at Jimma Medical Center, Ethiopia. BMC Med Ethics 2022 Oct 25;23(1):103. [doi: 10.1186/s12910-022-00841-5] [Medline: 36284338]
30. Admass BA, Diress FT, Fentie DY, Endalew NS. Maternal satisfaction and associated factors with postcesarean section pain management: a cross-sectional study. Anesthesiol Res Pract 2024;2024(1):4885678. [doi: 10.1155/2024/4885678] [Medline: 39263186]
31. Ashebir YG, Ashebir YG, Birhanu D. Assessment of maternal satisfaction and its associated factor after spinal anesthesia for cesarean section in Debere Markose comprehensive specialized hospital, East Gojjam Zone, an Amhara Region, Ethiopia, from January to June 2023 G.C. J Anesth Pain Med 2024;9(1):01-10 [FREE Full text] [doi: 10.33140/JAPM.09.01.03]
32. Bayable SD, Ahmed SA, Lema GF, Yaregal Melesse D. Assessment of maternal satisfaction and associated factors among parturients who underwent cesarean delivery under spinal anesthesia at University of Gondar Comprehensive Specialized Hospital, northwest Ethiopia, 2019. Anesthesiol Res Pract 2020;2020(1):8697651. [doi: 10.1155/2020/8697651] [Medline: 33101405]
33. Demilew BC, Getu D, Tesfaw D, Taye MG. Assessment of satisfaction and associated factors of parturients underwent cesarean section with spinal anesthesia at the General Hospital, Ethiopia; 2019. Ann Med Surg (Lond) 2021 May;65:102282. [doi: 10.1016/j.amsu.2021.102282] [Medline: 33948165]
34. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Int J Surg 2021 Apr;88:105906. [doi: 10.1016/j.ijsu.2021.105906] [Medline: 33789826]
35. Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa Hospital Research Institute. URL: https://ohri.ca/en/who-we-are/core-facilities-and-platforms/ottawa-methods-centre/newcastle-ottawa-scale [accessed 2026-07-24]
36. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis: John Wiley & Sons; 2021. [doi: 10.1002/9780470743386]
37. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002 Jun 15;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
38. Egger M, Smith GD, Altman D, editors. Systematic Reviews in Health Care: Meta-Analysis in Context, 2nd edition: BMJ Books; 2013. URL: https://www.wiley.com/en-us/shop/general-introductory-medical-science/systematic-reviews-in-health-care-meta-analysis-in-context-2nd-edition-p-9781118700044 [accessed 2026-07-24]
39. Olawoye O, Ashaye A, Bekibele C, Ajuwon AJ. A comparative evaluation of patients satisfaction with cataract surgical services in a public tertiary and a private secondary eye care facilities in Nigeria. Ann Afr Med 2012;11(3):157-162. [doi: 10.4103/1596-3519.96877] [Medline: 22684134]
40. Lemos P, Pinto A, Morais G, et al. Patient satisfaction following day surgery. J Clin Anesth 2009 May;21(3):200-205. [doi: 10.1016/j.jclinane.2008.08.016] [Medline: 19464614]
41. Alnashri YM, Alfaqih OY, Buhaliyqh MA, et al. Patient satisfaction and its predictors with perioperative anesthesia care at two general hospitals in southwestern Saudi Arabia. Cureus 2023 Jan;15(1):e33824. [doi: 10.7759/cureus.33824] [Medline: 36819326]
42. Ingabire L. Patients satisfaction with perioperative care at Oshen Fing Feisal Hospital [Master's thesis]. : University of Rwanda; 2017 Jun URL: http://dr.ur.ac.rw/bitstream/handle/123456789/281/INGABIRE%20Leontine.pdf [accessed 2026-07-31]
43. Lovatsis D, José JB, Tufman A, Drutz HP, Murphy K. Assessment of patient satisfaction with postoperative pain management after ambulatory gynaecologic laparoscopy. J Obstet Gynaecol Can 2007 Aug;29(8):664-667. [doi: 10.1016/s1701-2163(16)32552-x] [Medline: 17714620]
44. Jlala HA, Caljouw MA, Bedforth NM, Hardman JG. Patient satisfaction with perioperative care among patients having orthopedic surgery in a university hospital. Local Reg Anesth 2010;3:49-55. [doi: 10.2147/lra.s11381] [Medline: 22915869]
45. Fouogue JT, Tchounzou R, Fouelifack FY, et al. Evaluation of patients' satisfaction after laparoscopic surgery in a tertiary hospital in Cameroon (Africa). Pan Afr Med J 2017;28(216):216. [doi: 10.11604/pamj.2017.28.216.11441] [Medline: 29629002]
46. Ekpe E, Peter A. Surgical patient's satisfaction with services at a tertiary hospital in south -south state of Nigeria. J Med Res 2016;2(5):157-162. [doi: 10.31254/jmr.2016.2508]
47. Hamilton DF, Lane JV, Gaston P, et al. What determines patient satisfaction with surgery? A prospective cohort study of 4709 patients following total joint replacement. BMJ Open 2013;3(4):e002525. [doi: 10.1136/bmjopen-2012-002525] [Medline: 23575998]

XSL•FO
RenderX

48. Schwenkglenks M, Gerbershagen HJ, Taylor RS, et al. Correlates of satisfaction with pain treatment in the acute postoperative period: results from the international PAIN OUT registry. Pain 2014 Jul;155(7):1401-1411. [doi: 10.1016/j.pain.2014.04.021] [Medline: 24785269]
49. Baroudi DN, Nofal WH, Ahmad NA. Patient satisfaction in anesthesia: A modified Iowa Satisfaction in Anesthesia Scale. Anesth Essays Res 2010;4(2):85-90. [doi: 10.4103/0259-1162.73513] [Medline: 25885236]

Abbreviations

AOR: adjusted odds ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by K Cummings; submitted 20.Sep.2025; peer-reviewed by H Karim; revised version received 05.May.2026; accepted 05.Jun.2026; published 07.Aug.2026.

Please cite as:
Munie MA, Kitaw TA, Terefe AB, Zemariam AB, Adisu MA, Azmeraw M, Habtie TE, Dagnaw BW, Faris AM, Wodaynew T, Taye AB, Yirga GK, Desro YA
Patient Satisfaction With Perioperative Services and Associated Factors in Ethiopia: Systematic Review and Meta-Analysis
JMIR Perioper Med 2026;9:e84457
URL: https://periop.jmir.org/2026/1/e84457
doi:10.2196/84457

© Melesse Abiye Munie, Tegene Atamenta Kitaw, Aregash Birhane Terefe, Alemu Birara Zemariam, Molalign Aligaz Adisu, Molla Azmeraw, Tesfaye Engdaw Habtie, Betelhem Walelgn Dagnaw, Abebe Merchaw Faris, Tilahun Wodaynew, Amsalu Baylie Taye, Gebrie Kassaw Yirga, Yabibal Asfaw Desro. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 7.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Multidimensional Assessment of Recovery After Total Knee Arthroplasty in Clinical Practice: Critical Narrative Review

Abderrahmane Boukabache[1], BSc (Hons), MSc; Nimalan Maruthainar[2], BSc (Hons), MBBS; Vikrant Manhas[3], MBBS, MS; Darren Player[1], BSc (Hons), PhD

[1]Division of Surgery and Interventional Science, Faculty of Medical Sciences, University College London, Bloomsbury Campus: 2nd Floor, Charles Bell House, 43-45 Foley Street, London, United Kingdom

[2]Department of Trauma and Orthopaedics, Royal Free Hospital, London, United Kingdom

[3]Department of Orthopaedics, All India Institute of Medical Sciences, New Delhi, India

Corresponding Author:
Abderrahmane Boukabache, BSc (Hons), MSc
Division of Surgery and Interventional Science, Faculty of Medical Sciences, University College London, Bloomsbury Campus: 2nd Floor, Charles Bell House, 43-45 Foley Street, London, United Kingdom

Abstract

Background: Total knee arthroplasty (TKA) is the primary treatment for advanced knee osteoarthritis. Despite its clinical success and favorable patient-reported outcome measures (PROMs), approximately 20% to 30% of patients continue to experience persistent functional limitations and muscle weakness. This highlights the need for a comprehensive evaluation of recovery parameters beyond pain and range of motion. Given the wide range of methods available for assessing TKA outcomes, clinicians often select tools based on personal preference and understanding, which may affect accuracy and consistency; for example, the Knee Injury and Osteoarthritis Outcome Score may overestimate function compared to gait analysis studies.

Objective: The aim of this study was to conduct a narrative review focusing on the use, strengths, and limitations of different outcome measures used in routine orthopedic practice to optimize post-TKA evaluation.

Methods: A literature search was conducted in February 2025 across 2 databases (PubMed and Web of Science). Eligible studies included original research articles, systematic reviews, and meta-analyses that focused on validated measures used to evaluate TKA. Case reports, conference abstracts, and studies focused exclusively on surgical techniques were excluded. Themes were identified across studies to structure the results according to types of assessments and clinical applicability.

Results: A total of 6831 studies were retrieved and screened in this review, with 4 themes emerging around muscle mass, strength, performance, and PROMs. The Oxford Knee Score is favored for its ease of use and minimal ceiling effects. Broader tools like the Knee Injury and Osteoarthritis Outcome Score and Western Ontario and McMaster Universities Osteoarthritis Index provide detailed insights but are less practical clinically. For muscle strength, the portable fixed dynamometer showed high reliability and comparability to isokinetic dynamometry. Dual-energy X-ray absorptiometry remains the gold standard for assessing muscle mass, while bioelectrical impedance analysis offers a practical alternative. The 5-Repetition Sit-to-Stand test effectively evaluates lower limb power and speed.

Conclusions: Clinicians should integrate both objective (muscle mass, strength, and performance) and subjective (PROMs) measures to improve TKA recovery assessment. This multidimensional approach has the potential to enhance the accuracy of patient evaluation and supports the development of tailored rehabilitation strategies that address individual deficits and optimize functional outcomes.

(JMIR Perioper Med 2026;9:e84011) doi:10.2196/84011

KEYWORDS

arthroplasty; osteoarthritis; function; outcome; total knee replacement

Introduction

Osteoarthritis is a leading cause of pain and disability worldwide, affecting 528 million people as of 2019 [1,2]. Among various treatment options for osteoarthritis, total knee arthroplasty (TKA) stands out for its substantial impact on alleviating chronic knee pain [3]. With advancements in prosthetic design and surgical techniques, TKA has demonstrated high survival rates and relatively low complication risks, making it one of the most successful orthopedic procedures. However, despite these clinical successes, postoperative functional outcomes can vary among patients [4], with approximately 20% to 30% experiencing persistent functional limitations and muscle weakness [5].

XSL•FO
RenderX

Several factors influence recovery following TKA, including surgeon-related aspects such as surgical volume and technique, as well as patient-related variables like preoperative physical conditioning and psychological status. Among these, preoperative quadriceps strength has been identified as a key predictor of postoperative function, directly impacting mobility and the ability to perform daily activities [6]. Nonetheless, frequently used tools to evaluate TKA in orthopedics often rely on measures such as pain assessment and range of motion (ROM) [7], which, while valuable, may not fully capture postoperative recovery or functional capacity.

Evidence suggests that these metrics correlate poorly with objective measures of physical function and do not adequately reflect muscle weakness or biomechanical impairments that persist post-TKA [8]. While patient-related outcome measures (PROMs) provide subjective insights into perceived recovery, they may overestimate functional improvements compared to more objective assessments, such as gait analysis and performance-based tests [9]. This discrepancy underscores the need for a more comprehensive evaluation framework that integrates multiple dimensions of recovery, including muscle mass, muscle strength, and physical performance.

This review aims to critically examine the use, strengths, and limitations of different outcome measures used in routine practice when assessing recovery post-TKA, emphasizing the importance of incorporating objective physical performance metrics alongside PROMs. We hypothesize that a multidimensional approach, combining assessments of muscle mass, strength, physical performance, and PROMs, will provide a more accurate and clinically meaningful evaluation of TKA recovery, ultimately guiding more effective rehabilitation strategies.

Methods

A literature search of studies focusing on methods assessing TKA outcomes was performed using PubMed and Web of Science databases in February 2025. No limits were applied to publication dates. Inclusion criteria comprised original research, systematic reviews, and meta-analyses. After the removal of duplicates and cross-referencing of articles, we screened titles and abstracts for relevance. Studies were excluded if they were case reports, conference abstracts, or focused exclusively on surgical techniques. Only studies using standardized and validated assessment tools were included to ensure clinical relevance. The methodological quality of included studies was evaluated based on study design, sample size, and clarity of outcome measures.

One author conducted the initial screening of all retrieved abstracts against the inclusion criteria. Full texts of potentially relevant studies were then independently reviewed for eligibility and data extraction. Discrepancies between reviewers were resolved through discussion until consensus was achieved. Extracted data were thematically organized (muscle mass, muscle strength, physical performance, and PROMs), allowing patterns across assessment domains to be identified and synthesized into a structured narrative aligned with the review's objectives.

Given the narrative nature of this review, our selection was not strictly systematic; instead, we documented the total number of retrieved and screened articles (n=6831).

Results

Patient Reported Outcome Measures

There are numerous PROMs available to assess the outcomes of TKA, including: Knee Society Clinical Rating System (KSS), Western Ontario and McMaster Universities Osteoarthritis (WOMAC), Knee Injury and Osteoarthritis Outcome Score (KOOS), and the Oxford Knee Score (OKS) [10-13]. In clinical practice and research, it is essential to prioritize the use of sound PROMs over frequently used ones. The ultimate goal herein is to find a balance between standardized measures and specific contextually relevant PROMs in the field of TKA.

The KSS and OKS PROMs are shorter than the KOOS and WOMAC, with 10 and 12 items, respectively. The OKS uses a 4-week recall period, while the patient-reported component of the KSS is designed to reflect current knee status without a fixed recall window, which may reduce recall bias but limit comparability across time points. The OKS is solely derived from patient input, while the KSS comprises a PROM section completed by the patient and an informational-clinical section for the surgeon. Only the PROM section of the KSS is used to produce a psychometrically valid knee score. This separation sought to allow the functional PROM section to be independently assessed from the clinical section and confounding factors such as age or comorbidities. Notably, Martimbianco et al [14] reported considerable inter- and intraexaminer variability in the clinical-reported section of the KSS, which raises concerns about its reliability and consistency in clinical use. In contrast, the KOOS and WOMAC are larger instruments, with 42 and 33 items, respectively. They assess symptoms, stiffness, pain, and activities of daily living (ADL), but also include additional domains such as sports or recreation and quality of life. Like the OKS, these instruments are totally focused on the patient and their self-reported experiences. The WOMAC uses a 48-hour recall period, which minimizes recall bias and improves measurement precision; however, this narrow timeframe may fail to reflect symptom variability and functional fluctuation over longer recovery periods following TKA. In contrast, the KOOS uses a 1-week recall period, providing a broader representation of knee symptoms and functional limitations. Nevertheless, this longer recall window may increase susceptibility to recall bias and potentially reduce sensitivity to short-term clinical change.

All 4 PROMs assess pain and activity function, but they differ in how they capture pain information. The OKS and KSS both include a single item that captures the patient's general level of pain. However, the OKS asks further pain information in the context of general activities, such as work interference and walking. This is particularly important to assess as movement is associated with changes in pain, from subtle changes in muscle coordination to complete avoidance of joint function [15]. In contrast, the KOOS and WOMAC both include a section dedicated to assessing pain experienced during specific activities, with nine identical items. This encourages the patient

XSL•FO
RenderX

to recall and report on the pain experience during knee maneuvers such as knee twisting and straightening. The KOOS is an extension of the WOMAC, which justifies the similarities in the items. It is used in younger and/or more active patients who typically perform demanding twisting movements in their ADL. By assessing knee pain during movements like twisting and straightening, changes in functional outcomes over time can be tracked. This is especially relevant post-TKA, where pain during specific activities is directly linked to the functionality of the joint and the surrounding muscles [16].

The only resemblance when assessing function across the 4 PROMs is the patient's comfort level when handling stair ascent and descent. This likely reflects the universal importance of these activities in daily life and the significant impact that knee function has on an individual's mobility and well-being. However, each PROM also has unique items that assess different aspects of knee function and quality of life.

The OKS, KOOS, and WOMAC all assess the knee concerning various ADL, such as kneeling, transportation, domestic work, and bathroom activities. The OKS and KSS share only the item related to the distance the patient can walk. The KSS and items within the symptoms section of both KOOS and WOMAC ask about the patient's ability to fully bend and extend the knee, as well as objective measures of flexion contracture and extension lag. The use of similar items and domains across these scores can lead to greater consistency in assessing knee function and symptoms. Furthermore, consistency is important for tracking changes over time and comparing outcomes across different patients and studies. The KOOS is unique in that it assesses ADL related to sports and recreation and includes items that ask about quality of life. While similarities can provide consistency, differences between PROMs may capture important nuances specific to each condition.

The KSS evaluates 4 distinct domains: clinical, functional, satisfaction, and fulfillment of expectation. A notable feature is the inclusion of both high-demand tasks and 3 patient-selected priority activities from a predefined list, aiming to tailor the assessment to individual goals and enhance the relevance of functional evaluation. This patient-led component is designed to offer a more individualized perspective on recovery and may help inform more targeted postoperative rehabilitation. However, despite these strengths, the validity of the KSS has been questioned, particularly in the context of TKA outcome measurement. Several studies have noted limitations, including Ghanem et al [17], who highlighted concerns in revision TKA scenarios, and Vogel et al [18], who found the system less responsive than broader tools like the 36-Item Short Form Survey and WOMAC. One key criticism lies in the item-selection process: patients were not involved in the development of the tool, which may result in content that does not fully reflect patient priorities or lived experiences. Bach et al [19] further noted that the relatively limited item pool may constrain the scale's ability to capture the full spectrum of functional outcomes. This lack of patient input and limited item scope could contribute to a misalignment between what the tool measures and what matters most to patients, ultimately impacting its content validity and clinical use [17,18,20].

Ghanem et al [17] highlight that 2 very different patients could receive the same clinical score on the KSS. For instance, a patient with a stiff, pain-free, well-aligned knee and a patient with mild pain, excellent knee motion, and normal alignment may both receive a similar score, even though they have very different levels of function in their daily lives. To address these concerns, a revised Knee Society Scoring System (2011 Knee Society Score) was developed and validated as a more reliable measure for assessing outcomes in TKA procedures [21,22]. Albeit, a revised outcome measure, PROMs have inherent limitations in that they rely primarily on self-reported data, which may not always align perfectly with objective measures of physical function. This is where muscle function analysis could come into play. It offers an objective assessment of a patient's physical capabilities, including muscle strength, range of motion, and muscle activation patterns. By quantifying these aspects, health care providers can validate and complement the patient's self-reported symptoms and limitations with tangible data. This objective information not only supports the patient's reported experiences but also provides a more complete picture of their health status.

Despite its validity issues, the KSS is still popular among clinical researchers [23], perhaps for including alignment and ROM measurement. Proper coronal alignment of the components in TKA has been shown in the literature to be critical for implant survival rate. Additionally, knee ROM is a crucial indicator of successful TKA and is necessary for many ADLs [24]. While ROM is commonly measured and reported, the significance of isolated muscle function has been overlooked here. Capin et al [25] found quadriceps strength in particular to be a critical factor in TKA recovery and considered a rate-limiting step. This limited quantification of isolated muscle function is an area that warrants attention and improvement in the care of patients who underwent TKA.

The OKS was developed specifically for evaluating the outcomes of TKA. The simplicity and shortness of the questionnaire make it an attractive option for clinicians, and this may in part have contributed to its broad use in cohort studies and joint replacement registers [26,27]. This intentional oversimplification of the questionnaire highlights a lack of scope and sophistication to adequately capture the complex, interrelated issues experienced by many patients, such as joint stiffness, muscle weakness, and instability. To address these limitations, the use of a combination of assessment tools and methods has been suggested [28]. This could include performance-based functional tests, isolated muscle function tests, muscle mass measurement, and imaging studies.

Another consideration is that OKS has a high response rate when compared to other PROMs [10]. Analysis of the National Health Service PROMs dataset of 72,154 OKS concluded that OKS does not exhibit a ceiling or a floor effect at 6 months [29]. Marx et al [30] reported a ceiling effect of 22% at 12 months following surgery, but this could be attributed to patients achieving an optimal outcome rather than a limitation of the OKS. Nevertheless, this could indicate that the OKS may be more appropriate for short-term outcomes and might inadequately reflect the long-term burden. Although it has been

XSL•FO
RenderX

used in randomized controlled trials to assess knee arthroplasty long-term outcomes [31].

The WOMAC, developed in 1982, has undergone multiple revisions since and has been validated for TKA clinical trials [32-34], with ceiling effects at 6 months and 12 months for patients who underwent TKA [35]. The KOOS was an upgrade to the WOMAC to effectively capture the requirements of younger and more active individuals with knee injury or osteoarthritis. Several studies have shown that the KOOS is more sensitive and responsive than the WOMAC in younger or more active patients with knee injury or knee osteoarthritis [36]. Roos and Toksvig-Larsen [35] evaluated the outcome of 105 patients (mean age of 71) after TKA and found 74% of all Sport/Recreation items were considered to be "not applicable." The floor effects of approximately 48% likely reflect the high-demand activities, such as sports and recreation, which may be more relevant for younger, more active adults undergoing TKA.

A systematic review of the literature performed by Collins et al [37] found, as it was intended, the KOOS's ADL subscale has better content validity for older patients, and Sport/Recreation has better content validity for younger patients. Increasing ceiling effects from 6 to 12 months, particularly in the pain domain, may suggest that some patients have reached a plateau in their recovery, but also raises questions about the sensitivity of the measurement tool.

According to the Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN), which define internationally accepted criteria for evaluating the validity, reliability, and responsiveness of PROMs, the OKS meets COSMIN requirements and is recommended for use as a TKA outcome measure. Pain and function subscales of WOMAC and KOOS also demonstrate adequate measurement properties when evaluated as standalone subscales. However, the KSS does not consistently fulfill all COSMIN criteria and is not classified among the instruments meeting minimum standards for psychometric validation [38]. Several studies have evaluated the performance of different PROMs in measuring outcomes following lower extremity joint replacement surgery. Harris et al [39] identified the OKS and WOMAC as the best-performing PROMs specific to the lower extremity. The study assessed the validity, reproducibility, and acceptability of the scoring systems. Similarly, Alviar et al [40] in their review found OKS and WOMAC to be the best PROMs after assessing the quality of patient-reported outcome scoring systems. Collins and Roos [41] analyzed attributes of the 11 most frequently used PROMs for total hip replacement and TKA and considered KOOS, WOMAC, and OKS to be good PROMs. These PROMs aid in monitoring progress, guiding interventions, and facilitating shared decision-making. Nevertheless, there are limitations, such as the risk of subjective bias and variability in interpretation. Patient responses may be influenced by personal perceptions, mood, cognitive state, or cultural differences. Obtaining consistent and accurate data can be challenging, especially if patients struggle to recall specific details over time. To gain a more thorough understanding of the impact of TKA, it is crucial to integrate objective measures of muscle strength and functional assessments, particularly in addressing muscle weakness.

Strength Measurements

The indirect assessment of muscle function in any of the PROMs may not provide sufficient indication of dysfunction, necessitating a comprehensive analysis of objective assessment tools in TKA (Table 1)(Multimedia Appendix 1). To this end, this section will focus on the types of muscle strength assessments that are feasible in a TKA clinical setting.

XSL•FO
RenderX

Table . Comparative analysis of outcome measures used in total knee arthroplasty (TKA), including patient-reported outcomes, muscle strength, muscle mass, and physical performance tools. Comparison criteria include measurement type, reliability, validity, ease of use, clinical relevance, limitations, and best use case.

Comparison	PROMs[a] (KSS[b], WOMAC[c], KOOS[d], and OKS)[e]	Muscle strength tests (IKD[f], PFD[g], and HHD[h])	Muscle mass tests (CT[i], MRI[j], DXA[k], and BIA[l])	Physical performance tests (TUG[m], 6MWT[n], 5R-STS[o], and SCT[p])
Measurement type	Subjective (patient-reported pain, function)	Objective (muscle force output)	Objective (muscle cross-sectional area or mass)	Objective (functional movement performance)
Reliability	High test-retest reliability but can be influenced by patient perception	High reliability but dependent on test consistency and patient effort	High for CT, MRI, and DXA; moderate for BIA (affected by hydration)	High for TUG, 6MWT, and 5R-STS
Validity	Valid for subjective function but may not correlate with actual muscle strength or mobility	IKD, highest validity. PFD, good validity with proper fixation. HHD, lower validity	Highly valid for assessing atrophy or hypertrophy	Strong correlation with mobility and functional independence
Ease of use	Simple, noninvasive, and patient-friendly	Requires equipment; portable dynamometers are easier than isokinetic ones	CT or MRI is expensive; DXA is more accessible; BIA is easiest	Quick, requires minimal equipment (eg, chair and stopwatch)
Clinical relevance	Useful for tracking patient-perceived recovery	Directly assesses quadriceps and hamstring strength, critical for TKA recovery	Helps detect muscle loss post-TKA, which can affect long-term function	Strong predictor of fall risk, mobility, and independence
Limitations	Subjective, may not reflect real functional capacity	Requires patient cooperation; costly for isokinetic dynamometers	Expensive (CT and MRI), radiation exposure (CT and DXA), less precise (BIA)	Can be influenced by patient motivation, fatigue, or comorbidities
Best use case	Tracking patient-perceived progress	Measuring post-TKA quadriceps or hamstring recovery	Evaluating long-term muscle loss and sarcopenia	Assessing mobility and real-world function

[a]PROM: patient-reported outcome measure.

[b]KSS: Knee Society Score.

[c]WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index.

[d]KOOS: Knee Injury and Osteoarthritis Outcome Score.

[e]OKS: Oxford Knee Score.

[f]IKD: isokinetic dynamometer.

[g]PFD: portable fixed dynamometer.

[h]HHD: handheld dynamometry.

[i]CT: computed tomography.

[j]MRI: magnetic resonance imaging.

[k]DXA: dual-energy X-ray absorptiometry.

[l]iBIA: bioelectrical impedance analysis.

[m]TUG: Time Up and Go.

[n]6MWT: 6-Minute Walk Test.

[o]5R-STS: 5-Repetition Sit-to-Stand.

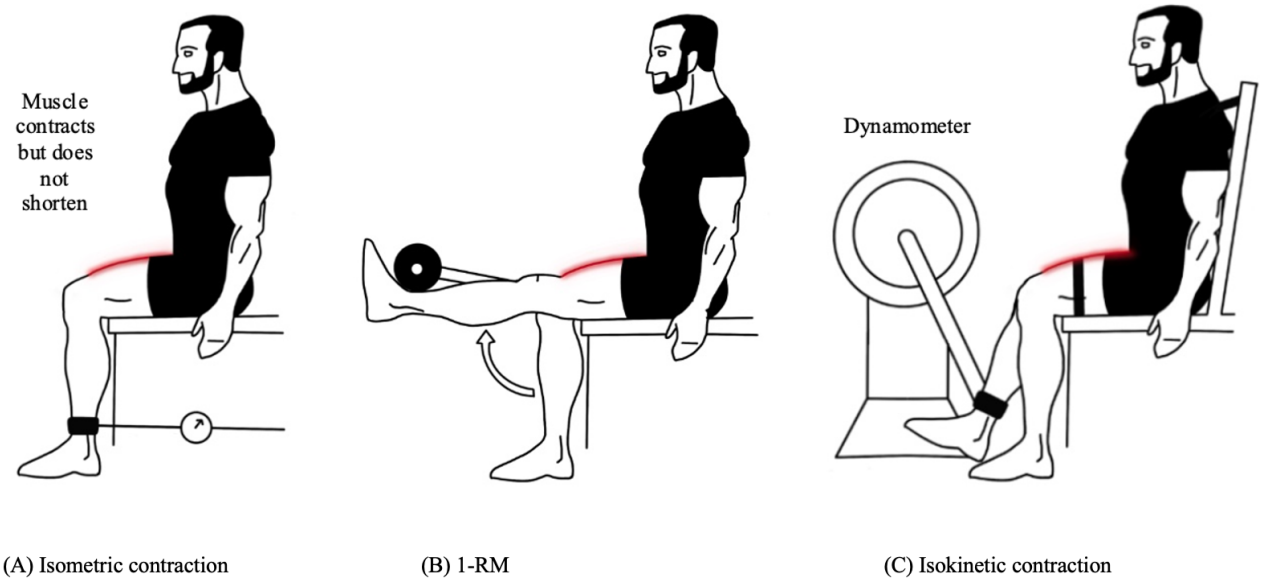
[p]SCT: Stair Climb Test.

When evaluating muscle function, the measured outcomes commonly include muscle strength and power [42]. Various methodologies assess these parameters, ranging from manual muscle testing to more complex and expensive isokinetic assessments. However, no general consensus exists regarding the preferred method of assessing muscle function in orthopedic practice, as each approach has distinct advantages and limitations [43].

Isometric strength assessment measures the maximum force generated during a contraction in which the muscle length remains constant (Figure 1A). Although isometric contractions are relatively uncommon in daily activities, their clinical relevance lies in their ability to predict functional capacity, particularly in older individuals and those with significant functional impairments [44]. Furthermore, isometric strength assessments have demonstrated high reliability across different orthopedic populations [45,46]. Isometric strength assessments can be easily performed using a dynamometer or force plate, and the individual is instructed to push or pull against an immovable object. Portable fixed dynamometer (PFD) has demonstrated high reliability (intraclass correlation coefficient [ICC] >0.90) and validity in measuring strong muscle contractions such as knee extension [47-49]. Importantly, studies have shown that knee muscle strength measurements using a PFD highly correlate with those obtained from an IKD, making PFD a viable alternative for clinical assessment [50].

XSL•FO
RenderX

Figure 1. Illustration of 3 common muscle strength assessment methods used in total knee arthroplasty (TKA) recovery: (A) isometric contraction—muscle generates force without joint movement; (B) one-repetition maximum (1-RM)—maximal weight lifted once with proper form; and (C) isokinetic contraction—muscle contracts at a constant speed throughout the range of motion using a dynamometer.



Toonstra and Mattacola [51] compared the reliability of IKD, PFD, and handheld dynamometry for isometric knee strength assessment. Their findings revealed high test-retest reliability for both IKD and PFD, while handheld dynamometry showed fair to poor reliability.

The strength of IKD lies in its ability to provide stabilization during testing and standardized protocols, making it the gold standard. However, PFD demonstrated comparable reliability while offering greater portability, ease of use, and cost-effectiveness. This makes PFD an attractive option for routine clinical evaluations, particularly in settings where IKD is not feasible.

Another key advantage of isometric assessments is their safety. Unlike dynamic strength tests, which involve movement and may place stress on healing tissues, isometric assessments minimize joint strain. This makes them particularly suitable for early postoperative assessments, where patient safety is a priority. Given these benefits, isometric strength measurements serve as a practical and reliable method for monitoring recovery following TKA.

One-repetition maximum (1-RM): Muscle strength can also be assessed through the one–repetition maximum (1-RM) test, which determines the maximum load that can be lifted in a single attempt, such as during a squat or leg extension (Figure 1B). This method evaluates muscle strength in a more functional and dynamic context, offering insight into the ability to generate force during real-world movements.

Despite its functional relevance, 1-RM testing has limitations. The requirement for specialized equipment (eg, weights and resistance machines) and the time-intensive nature of testing make it less practical for large patient cohorts. Additionally, 1-RM testing carries a risk of injury, particularly in postoperative populations where patients may not yet be able to safely perform maximal-effort movements. To mitigate this risk, researchers often use submaximal testing protocols to estimate 1-RM values [52].

Isokinetic testing measures muscle strength as the peak torque generated during a contraction performed at a constant angular velocity (Figure 1C). This method provides resistance that matches the individual's effort throughout the ROM, allowing for maximal force production at different joint angles. Isokinetic assessments are particularly valuable for evaluating muscle imbalances, monitoring rehabilitation progress, and understanding strength deficits following TKA.

Despite these advantages, isokinetic testing has notable limitations. The high cost and lack of portability make it less accessible in many clinical settings [53,54]. Additionally, isokinetic assessments require specialized equipment and trained personnel, making them impractical for routine postoperative evaluations. The potential for joint irritation and discomfort during testing further limits its suitability for early rehabilitation stages.

A key consideration when choosing between isometric and isokinetic testing is clinical feasibility. While isokinetic assessments provide detailed torque-angle relationships, they are not always necessary for functional recovery monitoring. In contrast, isometric testing provides a simple, reliable, and cost-effective means of assessing muscle strength, making it better suited for routine evaluations.

Lauermann et al [55] compared quadriceps strength assessments using isometric, isokinetic, and 1-RM methods in patients who underwent TKA. All 3 methods showed similar validity in detecting strength deficits, suggesting that isometric testing is an effective alternative to more complex assessments. Similarly, Lienhard et al [45] found comparable test-retest reliability across

XSL•FO
RenderX

these methods in patients who underwent TKA, reinforcing the viability of isometric assessments in clinical practice.

Muscle Mass Measurements

Studies have shown a significant correlation between skeletal muscle mass and cross-sectional skeletal muscle area in the extremities with muscle strength or power [56,57]. Hence, assessing these morphological features could be a good way to gauge muscle function. Furthermore, with certain imaging modalities, it may also be possible to determine the extent of sarcopenia, muscle wasting, and the level of fibrosis and fat infiltration—all factors that play a significant role in muscle function.

Computed tomography (CT) is an imaging modality that was introduced over 50 years ago, as the first clinically accepted tool for body composition measurement and served as a gold standard [58]. This method uses X-ray beams to create cross-sectional images of the body, allowing estimation of total body fat, visceral fat, and skeletal muscle mass [59,60]. Despite its early acceptance and high validity for assessing limb muscle cross-sectional area, with excellent test–retest and inter-observer reliability (ICC 0.98 - 1.00 for thigh muscle measurements), CT has general limitations such as high costs, the need for skilled technicians, and radiation exposure [61,62]. However, in current practice, CT is rarely used for measuring muscle mass, likely due to its associated limitations and the emergence of alternative, more practical techniques such as magnetic resonance imaging (MRI).

The introduction of MRI in the 1980s replaced the CT scan as the gold standard [63]. Its 3D images of skeletal muscle, fat tissue, and other organs are created by emitting radiofrequency energy from hydrogen atom nuclei in magnetic fields, with signal variations indicating different tissue types [63,64]. This development brought about high-resolution images without radiation exposure, making it suitable for tracking small changes over time, which is beneficial for intervention and observational studies. Advancements in MRI techniques have notably reduced image acquisition times, and modern scanners can accommodate obese individuals. Validated against direct anatomical measurements (r=0.97), MRI exhibits high test-retest and inter-observer reliability (2.9%, r=0.99% and 2.6%, r=0.99) in healthy populations [63,64]. However, limitations in clinical and research settings include high costs, the need for technical expertise, space requirements, infeasibility for patients with claustrophobia or those with MRI-incompatible implanted devices (eg, cardiac pacemakers), and standardization is hampered by the existence of various data acquisition protocols [61,65].

Dual-energy X-ray absorptiometry (DXA) has emerged as a strong alternative to the gold standard for assessing body composition, being relatively cheap compared with CT scan and MRI, and easy to perform. Initially designed for measuring bone mineral density, DXA is now widely used for examining overall body composition and muscle mass [63]. DXA operates by using 2 X-ray beams to differentiate between fat, bone, and lean tissue based on their X-ray absorption properties [66]. This technique has shown a high correlation with both MRI and CT in estimating skeletal muscle mass, indicating its reliability (r=0.88 and r=0.77 - 0.95, respectively) [67-69]. Additionally, DXA test-retest reliability for measures of fat-free mass demonstrated a high correlation (r=0.99), with low precision errors ranging from below 1% to 3% [70].

However, DXA does present some limitations. First, it does not directly quantify muscle mass; instead, it calculates lean soft tissue mass or fat-free mass based on the different gray tones observed in the DXA scan. Assumptions are made during this calculation, and factors like dehydration and edema, which can be common in populations with obesity, may interfere with the accuracy of the measurements. Additionally, the lack of standardization across devices, software packages, and versions can yield different results, affecting the reliability and comparability of measurements. Moreover, while DXA involves less radiation exposure compared to CT, there is still some exposure [71,72].

Bioelectrical impedance analysis (BIA) has become widely used for assessing body composition in both clinical and nonclinical settings [73]. BIA operates on the principle that tissues with higher water and electrolyte content, like skeletal muscle, offer less resistance to the passage of a low-voltage electrical current compared to lipid-rich adipose tissue, such as bone. This conductivity difference is exploited by BIA systems to quantify different body compartments. Measurements obtained along with parameters like sex, age, and body weight are then integrated into the population-specific body composition prediction equations [64,72]. The advantages of BIA stem from its noninvasive nature, cost-effectiveness, and ease of use, but limitations include sensitivity to factors like body position, physical exercise, food intake, hydration status, and the need for population-specific equations [74]. A study by Buckinx et al [75] attempted to gauge the reliability of BIA in assessing appendicular lean mass. The results showed high intraoperator reliability, indicated by an ICC of 0.89 when performed by the same operator, and interoperator reliability was also relatively high with an ICC of 0.77 when performed by different operators. However, the study revealed a notable discrepancy when comparing appendicular lean mass assessed by DXA and predicted by BIA, as evidenced by a low ICC of 0.37. Importantly, there exists a potential for a significant prediction error at the individual level with BIA, coupled with a systematic positive bias leading to an overall underestimation of lean body mass measurements [76]. Additionally, hydration status is a frequent confounder in clinical settings, as BIA relies on electrical conductivity that is highly sensitive to fluctuations in body water content, potentially compromising accuracy and limiting its reliability in certain patient populations. Despite these limitations, BIA remains a viable alternative in situations where more precise methods are not feasible.

Performance Measurements

While measurements of muscle mass and function are important for understanding physiological dysfunction and pathology, assessing physical performance is essential. In this context, functional performance-based tests use an individual's body weight as resistance, quantifying performance by the time taken or the number of repetitions completed. Several authoritative groups, including The Osteoarthritis Research Society

XSL•FO
RenderX

International, Rehabilitative Care Alliance, and the American Physical Therapy Association, recommend performance-based measures for individuals with osteoarthritis and arthroplasty [77-79]. These measures typically include assessments of gait speed tests, Stair Climb Tests (SCT), and Sit-to-Stand (STS) tests, each with distinct yet overlapping focuses.

Gait speed tests primarily assess mobility, endurance, and functional independence by measuring the time required to walk a set distance. They are simple, reliable, and highly predictive of overall recovery, fall risk, and mortality [80]. SCTs focus on functional strength, coordination, and balance, as stair negotiation requires greater knee flexion and quadriceps activation than level walking [81]. SCTs are more sensitive to persistent functional deficits than gait speed tests, as stair climbing is often more challenging postoperatively. STS tests, on the other hand, primarily evaluate lower limb strength and endurance by measuring the ability to transition from sitting to standing multiple times within a set duration or repetition count. STS tests strongly correlate with quadriceps strength and are useful for monitoring functional recovery [82].

While all 3 measures provide valuable insights into TKA recovery, they differ in their primary focus and clinical applicability. Gait speed and SCTs assess mobility in dynamic movement tasks, while STS focuses on lower limb muscle strength in a controlled, stationary task. Gait speed tests are most effective for evaluating overall mobility and endurance, whereas SCTs are ideal for assessing functional strength and dynamic balance [83]. In contrast, STS tests specifically target muscle power and endurance without assessing walking ability or balance [84]. Despite these differences, all three tests share a common goal of evaluating lower limb function and have been shown to be reliable, valid, and responsive to post-TKA functional improvements [78]. However, it is worth noting that both gait speed and SCTs may present logistical challenges in smaller clinical settings due to space constraints and increased time requirements for setup and execution, potentially limiting their routine use in busy outpatient environments.

Dobson et al [78] put forth a set of recommended performance-based measures for individuals diagnosed with hip and knee osteoarthritis or following joint replacement. This selection was based on expert surveys and systematic reviews, considering the evidence supporting the measurement properties of the tests, their feasibility, scoring methods, and expert consensus. The resulting recommended measures included the 30-Second Chair Stand Test (30CST), 40 m Fast-paced Walk Test, SCT, Time Up and Go (TUG), and 6-Minute Walk Test (6MWT). In a subsequent update by Dobson et al [85], the 10 Meter Fast-paced Walk Test and 20-second stair climb test were suggested as alternatives due to complexities in administering the 40-meter Fast-paced Walk and scoring the 11 Step Stair Climb Test.

A task force led by Westby et al [86] created the Total Joint Arthroplasty and Outcome Measures toolkit to be used before and after arthroplasty, which included: 30-second STS, gait speed, stair climb test, single leg stance test, 6MWT, TUG, and functional reach. Prior to the creation of this toolkit, Zeni et al [87] developed the Delaware Osteoarthritis Profile, a comprehensive set of tests that have been effectively used to measure functional performance pre and post knee arthroplasty, which include: TUG, SCT, and 6MWT.

While the literature contains numerous tests, the focus here will be on muscle strength performance tests. STS with its versions has been identified by Bergquist et al [88] in their systematic literature review to be the most appropriate performance-based clinical muscle strength test.

The 30CST is a component of the Senior Fitness Test and involves counting the number of sit-to-stand repetitions completed within a 30-second timeframe. Individuals who are more than halfway through a repetition at the 30-second mark are credited with completing the final repetition. This test has shown good to excellent reliability, with ICCs of approximately 0.84 to 0.92 in clinical populations, and its validity is supported by moderate to strong correlations with leg press strength measures [89].

The 5-Repetition Sit-to-Stand (5R-STS) test has gained widespread use as either a component of the short physical performance battery or as a standalone assessment tool in numerous studies. This test involves performing 5 sit-to-stand repetitions from a standard chair (44 - 48 cm height), with timing commencing upon a specific command or the initiation of the first movement and ending when the fifth stand-up is accomplished or when the patient sits down following the fifth stand-up [90]. The reliability of the 5R-STS test has been estimated across 10 studies, yielding a coefficient of 0.81, as reported by Bohannon (2011) [84]. The validity of this test as a measure of lower limb functional muscle strength is supported by its strong correlation with knee extension strength, as demonstrated by Lord et al [82]. Furthermore, the correlation between performance on the 5R-STS test and the TUG test, as well as gait speed tests, adds further support for its validity, as highlighted by Schaubert and Bohannon [91].

Although the 5R-STS and 30CST tests involve identical movements, they are not interchangeable. The 5R-STS indicates lower limb speed and power for those with lower physical function post-TKA, while the 30CST measures lower limb endurance for those with higher physical function [92].

While outcome measures are often viewed as purely psychometric, their clinical relevance is established through their role as important drivers of long-term surgical success. For instance, a recent systematic review and meta-analysis by Sumbal et al [93] demonstrated that loss of muscle mass, or sarcopenia, is a significant risk factor for prosthetic loosening, which remains a primary indication for revision surgery. Furthermore, physical performance levels and quadriceps strength have been shown to be strongly associated with patient satisfaction and restoration of functional ability after TKA [94]. More recently, Akatsuka et al [95] reinforced this clinical relevance by reporting significant correlations between quadriceps muscle mass and postoperative satisfaction, suggesting that these objective measures are predictive of how a patient perceives their surgical result.

XSL•FO
RenderX

Discussion

Principal findings

A range of assessment tools is available to evaluate recovery following TKA, each with distinct advantages and limitations. This review highlights the use, strengths, and constraints of these measures and recommends a comprehensive, multidimensional evaluation framework. Such an approach enhances the accuracy and clinical relevance of recovery assessments by triangulating data across patient-reported outcomes, strength measures, functional performance, and muscle mass.

Among muscle mass assessment techniques, DXA remains the gold standard due to its high precision and reliability. However, alternative methods such as BIA offer a noninvasive, lower-cost option suitable for clinical settings, despite reduced accuracy and sensitivity to minor changes in muscle composition.

PROMs provide valuable insights into patient perspectives. The OKS is widely employed due to its ease of use, absence of ceiling or floor effects in the short term, and extensive validation in clinical research. KOOS, while broader in scope, often demonstrates ceiling effects and a higher completion burden. WOMAC overlaps conceptually but lacks OKS's responsiveness in early recovery. KSS combines patient and clinician input but is less feasible for routine use due to scoring complexity and reliance on in-person assessment. OKS offers the optimal balance of validity, efficiency, and clinical use in TKA follow-up. Nevertheless, PROMs alone do not provide objective strength assessments, which are critical for a comprehensive evaluation.

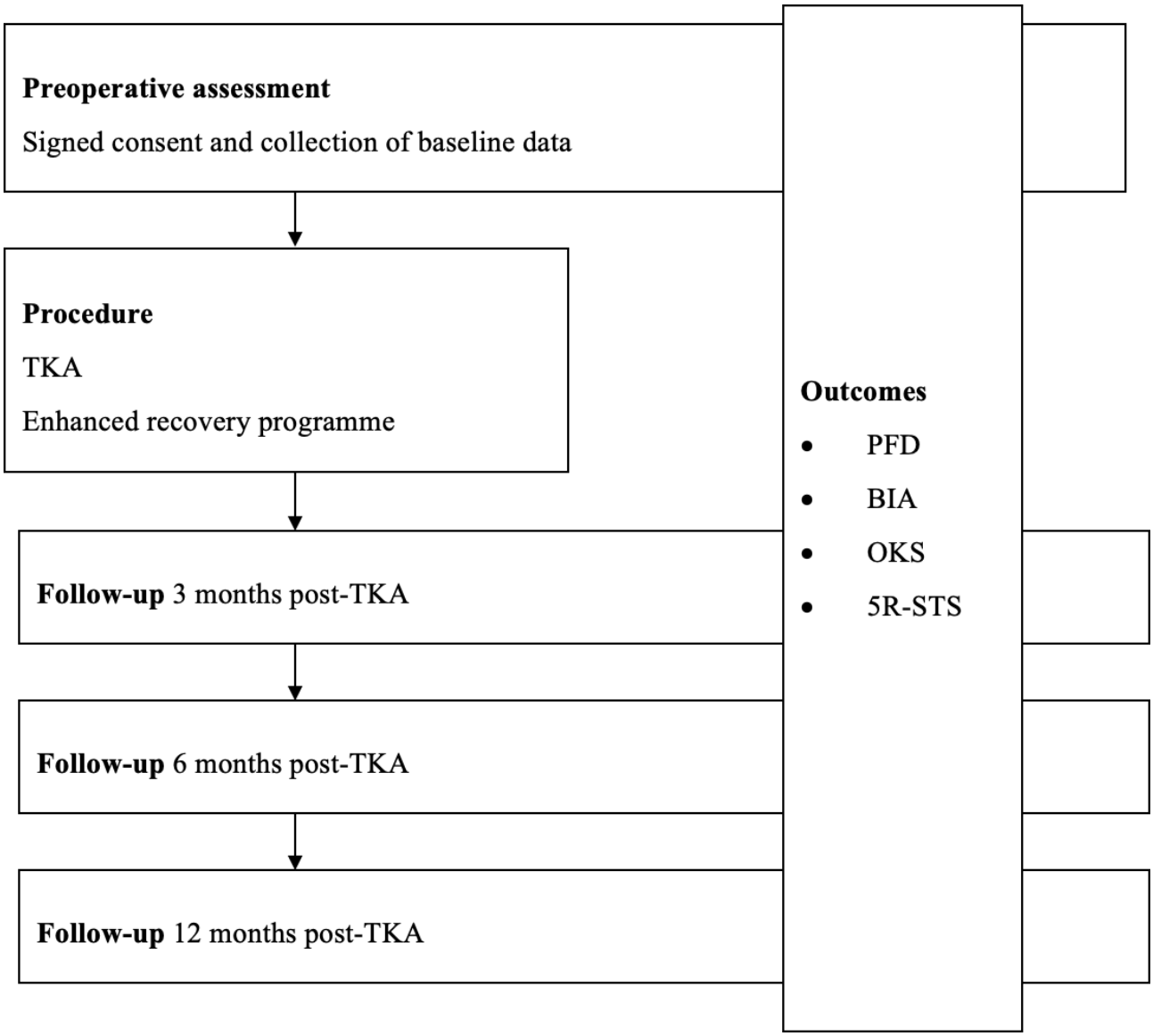
In contrast, IKD offers precise quantification of muscle strength, yet its high cost and operational complexity restrict its widespread application. As a practical alternative, PFD provides a reliable, cost-effective, and portable means of assessing muscle strength, making it highly suitable for routine clinical practice.

Performance-based measures such as the 5R-STS test offer objective, time-efficient insights into lower limb function. Compared to the TUG, 6MWT, or SCT, the 5R-STS requires less space, is less influenced by cardiovascular limitations, and is more feasible for patients with early post-op mobility impairments. It also shows strong correlation with knee extension strength and mobility metrics, enhancing its predictive use in post-TKA recovery.

This specific combination, OKS, PFD, BIA, and 5R-STS is what we have recommended in post-TKA follow-up (Figure 2) because it balances psychometric validity, clinical feasibility, and comprehensive recovery profiling. It provides a more complete view of TKA recovery than any individual tool. While alternatives like the 6MWT or SCT yield valuable data, their requirements for time, space, or greater cardiopulmonary reserve make them less practical for routine follow-up or for patients with limited mobility or comorbidities. Likewise, the TUG, though simple, primarily reflects balance and walking ability but lacks sensitivity to muscle power or patient satisfaction domains.

XSL•FO
RenderX

Figure 2. Clinical workflow algorithm for integrating recommended outcome measures throughout the total knee arthroplasty (TKA) pathway. The flowchart outlines the timing and selection of assessment tools from the preoperative assessment appointment to the final follow-up. 5R-STS: 5-Repetition Sit-to-Stand; BIA: bioelectrical impedance analysis; OKS: Oxford Knee Score; PFD: portable fixed dynamometer.



Limitations

Finally, it is important to note that the interpretability and use of these tools may vary across populations. For instance, sarcopenia, obesity, or cardiopulmonary conditions may skew performance-based results or BIA readings, requiring clinicians to contextualize outcomes based on age, BMI, and comorbid burden. By triangulating across these assessments, clinicians can reduce bias from any single measure and develop a more individualized recovery evaluation.

Conclusions

Throughout this paper, evidence suggests that incorporating muscle-based and performance-based measures alongside PROMs is essential for a comprehensive and clinically relevant assessment of TKA recovery. However, future research should ensure there are multicenter trials to validate the integration of the proposed assessments, as well as consensus agreement with clinicians (ie, through a Delphi approach) which translates into guidelines. Additionally, a greater understanding of intrinsic skeletal muscle (mal)adaptations following TKA may provide valuable insight into the mechanisms underlying persistent functional limitations.

Authors' Contributions

Conceptualization: DP, AB
Formal analysis: AB
Investigation: AB
Methodology: AB, DP

XSL•FO
RenderX

Supervision: NM, DP
Visualization: AB
Writing – original draft: AB
Writing – review & writing: AB, VM
All authors approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparative analysis of outcome measures used in total knee arthroplasty (TKA), including patient-reported outcomes, muscle strength, muscle mass, and physical performance tools. Comparison criteria include measurement type, reliability, validity, ease of use, clinical relevance, limitations, and best use case.
[DOCX File, 14 KB - periop_v9i1e84011_app1.docx]

References

1. Neogi T. The epidemiology and impact of pain in osteoarthritis. Osteoarthritis Cartilage 2013 Sep;21(9):1145-1153. [doi: 10.1016/j.joca.2013.03.018] [Medline: 23973124]
2. Vos T, Lim SS, Abbafati C. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1204-1222. [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]
3. Jüni P, Reichenbach S, Dieppe P. Osteoarthritis: Rational approach to treating the individual. Best Pract Res Clin Rheumatol 2006 Aug;20(4):721-740. [doi: 10.1016/j.berh.2006.05.002] [Medline: 16979535]
4. Rissolio L, Sabatini L, Risitano S, et al. Is it the surgeon, the patient, or the device? A comprehensive clinical and radiological evaluation of factors influencing patient satisfaction in 648 total knee arthroplasties. J Clin Med 2021 Jun 12;10(12):2599. [doi: 10.3390/jcm10122599] [Medline: 34204628]
5. Lewis GN, Rice DA, Rashid U, McNair PJ, Kluger MT, Somogyi AA. Trajectories of pain and function outcomes up to 5 to 8 years following total knee arthroplasty. J Arthroplasty 2023 Aug;38(8):1516-1521. [doi: 10.1016/j.arth.2023.02.020] [Medline: 36805116]
6. Tungtrongjit Y, Weingkum P, Saunkool P. The effect of preoperative quadriceps exercise on functional outcome after total knee arthroplasty. J Med Assoc Thai 2012 Oct;95 Suppl 10:S58-S66. [Medline: 23451440]
7. Vajapey SP, Morris J, Spitzer AI, Glassman AH, Greco NJ, Li M. Outcome reporting patterns in total knee arthroplasty: a systematic review. J Clin Orthop Trauma 2020 Jul;11(Suppl 4):S464-S471. [doi: 10.1016/j.jcot.2020.05.014] [Medline: 32774013]
8. Mizner RL, Petterson SC, Clements KE, Zeni JA, Irrgang JJ, Snyder-Mackler L. Measuring functional improvement after total knee arthroplasty requires both performance-based and patient-report assessments: a longitudinal analysis of outcomes. J Arthroplasty 2011 Aug;26(5):728-737. [doi: 10.1016/j.arth.2010.06.004] [Medline: 20851566]
9. Naili JE, Iversen MD, Esbjörnsson AC, et al. Deficits in functional performance and gait one year after total knee arthroplasty despite improved self-reported function. Knee Surg Sports Traumatol Arthrosc 2017 Nov;25(11):3378-3386. [doi: 10.1007/s00167-016-4234-7] [Medline: 27436196]
10. Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about total knee replacement. J Bone Joint Surg Br 1998 Jan;80(1):63-69. [doi: 10.1302/0301-620x.80b1.7859] [Medline: 9460955]
11. Insall JN, Dorr LD, Scott RD, Scott WN. Rationale of the Knee Society clinical rating system. Clin Orthop Relat Res 1989 Nov(248):13-14. [Medline: 2805470]
12. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol 1988 Dec;15(12):1833-1840. [Medline: 3068365]
13. Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. Knee Injury and Osteoarthritis Outcome Score (KOOS)—development of a self-administered outcome measure. J Orthop Sports Phys Ther 1998 Aug;28(2):88-96. [doi: 10.2519/jospt.1998.28.2.88] [Medline: 9699158]
14. Martimbianco ALC, Calabrese FR, Iha LAN, Petrilli M, Lira Neto O, Carneiro Filho M. Reliability of the “American Knee Society Score” (AKSS) [Article in Portuguese]. Acta Ortop Bras 2012;20(1):34-38. [doi: 10.1590/S1413-78522012000100007] [Medline: 24453578]
15. Hodges PW, Smeets RJ. Interaction between pain, movement, and physical activity: short-term benefits, long-term consequences, and targets for treatment. Clin J Pain 2015 Feb;31(2):97-107. [doi: 10.1097/AJP.0000000000000098] [Medline: 24709625]
16. Merkle SL, Sluka KA, Frey-Law LA. The interaction between pain and movement. J Hand Ther 2020;33(1):60-66. [doi: 10.1016/j.jht.2018.05.001] [Medline: 30025839]

XSL•FO
RenderX

17. Ghanem E, Pawasarat I, Lindsay A, et al. Limitations of the Knee Society Score in evaluating outcomes following revision total knee arthroplasty. J Bone Joint Surg Am 2010 Oct 20;92(14):2445-2451. [doi: 10.2106/JBJS.I.00252] [Medline: 20962195]
18. Vogel N, Kaelin R, Rychen T, Wendelspiess S, Müller-Gerbl M, Arnold MP. Comparison of responsiveness of patient-reported outcome measures after total knee arthroplasty. J Arthroplasty 2024 Jun;39(6):1487-1495.e2. [doi: 10.1016/j.arth.2023.11.029] [Medline: 38043745]
19. Bach CM, Nogler M, Steingruber IE, et al. Scoring systems in total knee arthroplasty. Clin Orthop Relat Res 2002 Jun;399(399):184-196. [doi: 10.1097/00003086-200206000-00022] [Medline: 12011708]
20. Karimijashni M, Abtahi F, Abbasalipour S, et al. Functional patient-reported outcome measures after hip or knee arthroplasty: a systematic review and content analysis using the International Classification of Functioning, Disability, and Health. Arthritis Care Res (Hoboken) 2024 Dec;76(12):1703-1722. [doi: 10.1002/acr.25413] [Medline: 39119739]
21. Noble PC, Scuderi GR, Brekke AC, et al. Development of a new Knee Society scoring system. Clin Orthop Relat Res 2012 Jan;470(1):20-32. [doi: 10.1007/s11999-011-2152-z] [Medline: 22065240]
22. Scuderi GR, Bourne RB, Noble PC, Benjamin JB, Lonner JH, Scott WN. The new Knee Society knee scoring system. Clin Orthop Relat Res 2012 Jan;470(1):3-19. [doi: 10.1007/s11999-011-2135-0] [Medline: 22045067]
23. Siljander MP, McQuivey KS, Fahs AM, Galasso LA, Serdahely KJ, Karadsheh MS. Current trends in patient-reported outcome measures in total joint arthroplasty: a study of 4 major orthopaedic journals. J Arthroplasty 2018 Nov;33(11):3416-3421. [doi: 10.1016/j.arth.2018.06.034] [Medline: 30057269]
24. Ritter MA, Davis KE, Davis P, et al. Preoperative malalignment increases risk of failure after total knee arthroplasty. J Bone Joint Surg Am 2013 Jan 16;95(2):126-131. [doi: 10.2106/JBJS.K.00607] [Medline: 23324959]
25. Capin JJ, Bade MJ, Jennings JM, Snyder-Mackler L, Stevens-Lapsley JE. Total knee arthroplasty assessments should include strength and performance-based functional tests to complement range-of-motion and patient-reported outcome measures. Phys Ther 2022 Jun 3;102(6):pzac033. [doi: 10.1093/ptj/pzac033] [Medline: 35358318]
26. Ben-Shlomo Y, Blom A, Boulton C, et al. The National Joint Registry 19th annual report 2022. : National Joint Registry; 2022 URL: https://www.njrcentre.org.uk/wp-content/uploads/NJR-Supported-Research.pdf [accessed 2025-03-18]
27. Lee MK, Naessens JM, Eton DT, O'Byrne TJ, Nyman MA. Functional outcomes and health-related quality of life before and after primary total knee replacement for patients from diverse geographic regions. J Arthroplasty 2021 Jun;36(6):1958-1965. [doi: 10.1016/j.arth.2021.01.043] [Medline: 33581972]
28. Beard DJ, Knezevic K, Al-Ali S, Dawson J, Price AJ. The use of outcome measures relating to the knee. Orthop Trauma 2010 Aug;24(4):309-316. [doi: 10.1016/j.mporth.2010.04.001]
29. Harris K, Lim CR, Dawson J, Fitzpatrick R, Beard DJ, Price AJ. The Oxford Knee Score and its subscales do not exhibit a ceiling or a floor effect in knee arthroplasty patients: an analysis of the National Health Service PROMs data set. Knee Surg Sports Traumatol Arthrosc 2017 Sep;25(9):2736-2742. [doi: 10.1007/s00167-015-3788-0] [Medline: 26519187]
30. Marx RG, Jones EC, Atwan NC, Closkey RF, Salvati EA, Sculco TP. Measuring improvement following total hip and knee arthroplasty using patient-based measures of outcome. J Bone Joint Surg Am 2005 Sep;87(9):1999-2005. [doi: 10.2106/JBJS.D.02286] [Medline: 16140815]
31. Beard DJ, Davies LJ, Cook JA, et al. The clinical and cost-effectiveness of total versus partial knee replacement in patients with medial compartment osteoarthritis (TOPKAT): 5-year outcomes of a randomised controlled trial. Lancet 2019 Aug 31;394(10200):746-756. [doi: 10.1016/S0140-6736(19)31281-4] [Medline: 31326135]
32. Escobar A, Gonzalez M, Quintana JM, et al. Patient acceptable symptom state and OMERACT-OARSI set of responder criteria in joint replacement. Identification of cut-off values. Osteoarthritis Cartilage 2012 Feb;20(2):87-92. [doi: 10.1016/j.joca.2011.11.007] [Medline: 22155074]
33. Berth A, Urbach D, Awiszus F. Improvement of voluntary quadriceps muscle activation after total knee arthroplasty. Arch Phys Med Rehabil 2002 Oct;83(10):1432-1436. [doi: 10.1053/apmr.2002.34829] [Medline: 12370881]
34. Bellamy N. WOMAC: A 20-year experiential review of a patient-centered self-reported health status questionnaire. J Rheumatol 2002 Dec;29(12):2473-2476. [Medline: 12465137]
35. Roos EM, Toksvig-Larsen S. Knee injury and Osteoarthritis Outcome Score (KOOS)—validation and comparison to the WOMAC in total knee replacement. Health Qual Life Outcomes 2003 May 25;1(1):17. [doi: 10.1186/1477-7525-1-17] [Medline: 12801417]
36. Antosh IJ, Svoboda SJ, Peck KY, Garcia EJ, Cameron KL. Change in KOOS and WOMAC scores in a young athletic population with and without anterior cruciate ligament injury. Am J Sports Med 2018 Jun;46(7):1606-1616. [doi: 10.1177/0363546518768753] [Medline: 29733680]
37. Collins NJ, Prinsen CAC, Christensen R, Bartels EM, Terwee CB, Roos EM. Knee injury and Osteoarthritis Outcome Score (KOOS): systematic review and meta-analysis of measurement properties. Osteoarthritis Cartilage 2016 Aug;24(8):1317-1329. [doi: 10.1016/j.joca.2016.03.010] [Medline: 27012756]
38. Wang Y, Yin M, Zhu S, Chen X, Zhou H, Qian W. Patient-reported outcome measures used in patients undergoing total knee arthroplasty. Bone Joint Res 2021 Mar;10(3):203-217. [doi: 10.1302/2046-3758.103.BJR-2020-0268.R1] [Medline: 33734821]

XSL•FO
RenderX

39. Harris K, Dawson J, Gibbons E, et al. Systematic review of measurement properties of patient-reported outcome measures used in patients undergoing hip and knee arthroplasty. Patient Relat Outcome Meas 2016;7:101-108. [doi: 10.2147/PROM.S97774] [Medline: 27524925]
40. Alviar MJ, Olver J, Brand C, et al. Do patient-reported outcome measures in hip and knee arthroplasty rehabilitation have robust measurement attributes? A systematic review. J Rehabil Med 2011 Jun;43(7):572-583. [doi: 10.2340/16501977-0828] [Medline: 21607295]
41. Collins NJ, Roos EM. Patient-reported outcomes for total hip and knee arthroplasty: commonly used instruments and attributes of a "good" measure. Clin Geriatr Med 2012 Aug;28(3):367-394. [doi: 10.1016/j.cger.2012.05.007] [Medline: 22840304]
42. Gerrits KH, Beltman MJ, Koppe PA, et al. Isometric muscle function of knee extensors and the relation with functional performance in patients with stroke. Arch Phys Med Rehabil 2009 Mar;90(3):480-487. [doi: 10.1016/j.apmr.2008.09.562] [Medline: 19254615]
43. Maffiuletti NA. Assessment of hip and knee muscle function in orthopaedic practice and research. J Bone Joint Surg Am 2010 Jan;92(1):220-229. [doi: 10.2106/JBJS.I.00305] [Medline: 20048117]
44. Hyun CW, Han EY, Im SH, et al. Hemiparetic knee extensor strength and balance function are predictors of ambulatory function in subacute stroke patients. Ann Rehabil Med 2015 Aug;39(4):577-585. [doi: 10.5535/arm.2015.39.4.577] [Medline: 26361594]
45. Norris R, Morrison S, Price A, et al. Inline dynamometry provides reliable measurements of quadriceps strength in healthy and ACL-reconstructed individuals and is a valid substitute for isometric electromechanical dynamometry following ACL reconstruction. Knee 2024 Jan;46:136-147. [doi: 10.1016/j.knee.2023.12.006] [Medline: 38142660]
46. Lienhard K, Lauermann SP, Schneider D, Item-Glatthorn JF, Casartelli NC, Maffiuletti NA. Validity and reliability of isometric, isokinetic and isoinertial modalities for the assessment of quadriceps muscle strength in patients with total knee arthroplasty. J Electromyogr Kinesiol 2013 Dec;23(6):1283-1288. [doi: 10.1016/j.jelekin.2013.09.004] [Medline: 24113423]
47. Ieiri A, Tushima E, Ishida K, Inoue M, Kanno T, Masuda T. Reliability of measurements of hip abduction strength obtained with a hand-held dynamometer. Physiother Theory Pract 2015 Feb;31(2):146-152. [doi: 10.3109/09593985.2014.960539] [Medline: 25264015]
48. Bohannon RW, Kindig J, Sabo G, Duni AE, Cram P. Isometric knee extension force measured using a handheld dynamometer with and without belt-stabilization. Physiother Theory Pract 2012 Oct;28(7):562-568. [doi: 10.3109/09593985.2011.640385] [Medline: 22191509]
49. Bohannon RW, Pritchard RO, Glenney SS. Portable belt–stabilized hand-held dynamometry set-up for measuring knee extension force. Isokinet Exerc Sci 2013 Nov;21(4):325-329. [doi: 10.3233/IES-130517]
50. Martins J, da Silva JR, da Silva MRB, Bevilaqua-Grossi D. Reliability and validity of the belt-stabilized handheld dynamometer in hip- and knee-strength Tests. J Athl Train 2017 Sep;52(9):809-819. [doi: 10.4085/1062-6050-52.6.04] [Medline: 28787180]
51. Toonstra J, Mattacola CG. Test-retest reliability and validity of isometric knee-flexion and -extension measurement using 3 methods of assessing muscle strength. J Sport Rehabil 2013 Feb 1;22(1):10.1123/jsr.2013.TR7. [doi: 10.1123/jsr.2013.TR7] [Medline: 22951307]
52. Abdalla PP, Carvalho ADS, Dos Santos AP, et al. One-repetition submaximal protocol to measure knee extensor muscle strength among older adults with and without sarcopenia: a validation study. BMC Sports Sci Med Rehabil 2020;12(1):29. [doi: 10.1186/s13102-020-00178-9] [Medline: 32391159]
53. Horstmann H, Medico P, Lasch F, Krutsch W, Weber-Spickschen TS. Simplified measurement of maximum strength after knee surgery: application-based knee-training device compared to isokinetic testing. Open Access J Sports Med 2020;11:43-49. [doi: 10.2147/OAJSM.S214598] [Medline: 32158284]
54. Whiteley R, Jacobsen P, Prior S, Skazalski C, Otten R, Johnson A. Correlation of isokinetic and novel hand-held dynamometry measures of knee flexion and extension strength testing. J Sci Med Sport 2012 Sep;15(5):444-450. [doi: 10.1016/j.jsams.2012.01.003] [Medline: 22424705]
55. Lauermann SP, Lienhard K, Item-Glatthorn JF, Casartelli NC, Maffiuletti NA. Assessment of quadriceps muscle weakness in patients after total knee arthroplasty and total hip arthroplasty: methodological issues. J Electromyogr Kinesiol 2014 Apr;24(2):285-291. [doi: 10.1016/j.jelekin.2013.10.018] [Medline: 24290027]
56. Tsukasaki K, Matsui Y, Arai H, et al. Association of muscle strength and gait speed with cross-sectional muscle area determined by mid-thigh computed tomography—a comparison with skeletal muscle mass measured by dual-energy X-ray absorptiometry. J Frailty Aging 2020;9(2):82-89. [doi: 10.14283/jfa.2020.16] [Medline: 32259181]
57. Riviati N, Indra B. Relationship between muscle mass and muscle strength with physical performance in older adults: a systematic review. SAGE Open Med 2023;11:20503121231214650. [doi: 10.1177/20503121231214650] [Medline: 38033420]
58. Heymsfield SB, Adamek M, Gonzalez MC, Jia G, Thomas DM. Assessing skeletal muscle mass: historical overview and state of the art. J Cachexia Sarcopenia Muscle 2014 Mar;5(1):9-18. [doi: 10.1007/s13539-014-0130-5] [Medline: 24532493]
59. Binay Safer V, Safer U. Usefulness and limitations of single-slice computed tomography analysis at the third lumbar region in the assessment of sarcopenia. Crit Care 2013 Nov 20;17(6):466. [doi: 10.1186/cc13123] [Medline: 24252497]

XSL•FO
RenderX

60. Gibson DJ, Burden ST, Strauss BJ, Todd C, Lal S. The role of computed tomography in evaluating body composition and the influence of reduced muscle mass on clinical outcome in abdominal malignancy: a systematic review. Eur J Clin Nutr 2015 Oct;69(10):1079-1086. [doi: 10.1038/ejcn.2015.32] [Medline: 25782424]
61. Lemos T, Gallagher D. Current body composition measurement techniques. Curr Opin Endocrinol Diabetes Obes 2017 Oct;24(5):310-314. [doi: 10.1097/MED.0000000000000360] [Medline: 28696961]
62. Strandberg S, Wretling ML, Wredmark T, Shalabi A. Reliability of computed tomography measurements in assessment of thigh muscle cross-sectional area and attenuation. BMC Med Imaging 2010 Aug 11;10(1):18. [doi: 10.1186/1471-2342-10-18] [Medline: 20701775]
63. Erlandson MC, Lorbergs AL, Mathur S, Cheung AM. Muscle analysis using pQCT, DXA and MRI. Eur J Radiol 2016 Aug;85(8):1505-1511. [doi: 10.1016/j.ejrad.2016.03.001] [Medline: 27005009]
64. Heymsfield SB, Gonzalez MC, Lu J, Jia G, Zheng J. Skeletal muscle mass and quality: evolution of modern measurement concepts in the context of sarcopenia. Proc Nutr Soc 2015 Nov;74(4):355-366. [doi: 10.1017/S0029665115000129] [Medline: 25851205]
65. Prado CMM, Heymsfield SB. Lean tissue imaging: a new era for nutritional assessment and intervention. JPEN J Parenter Enteral Nutr 2014 Nov;38(8):940-953. [doi: 10.1177/0148607114550189] [Medline: 25239112]
66. Lustgarten MS, Fielding RA. Assessment of analytical methods used to measure changes in body composition in the elderly and recommendations for their use in phase II clinical trials. J Nutr Health Aging 2011 May;15(5):368-375. [doi: 10.1007/s12603-011-0049-x] [Medline: 21528163]
67. Maden-Wilkinson TM, Degens H, Jones DA, McPhee JS. Comparison of MRI and DXA to measure muscle size and age-related atrophy in thigh muscles. J Musculoskelet Neuronal Interact 2013 Sep;13(3):320-328. [Medline: 23989253]
68. Kim J, Wang Z, Heymsfield SB, Baumgartner RN, Gallagher D. Total-body skeletal muscle mass: estimation by a new dual-energy X-ray absorptiometry method. Am J Clin Nutr 2002 Aug;76(2):378-383. [doi: 10.1093/ajcn/76.2.378] [Medline: 12145010]
69. Taylor AE, Kuper H, Varma RD, et al. Validation of dual energy X-ray absorptiometry measures of abdominal fat by comparison with magnetic resonance imaging in an Indian population. PLoS One 2012;7(12):e51042. [doi: 10.1371/journal.pone.0051042] [Medline: 23272086]
70. Hangartner TN, Warner S, Braillon P, Jankowski L, Shepherd J. The official positions of the International Society for Clinical Densitometry: acquisition of dual-energy X-ray absorptiometry body composition and considerations regarding analysis and repeatability of measures. J Clin Densitom 2013;16(4):520-536. [doi: 10.1016/j.jocd.2013.08.007] [Medline: 24183641]
71. Toombs RJ, Ducher G, Shepherd JA, De Souza MJ. The impact of recent technological advances on the trueness and precision of DXA to assess body composition. Obesity (Silver Spring) 2012 Jan;20(1):30-39. [doi: 10.1038/oby.2011.211] [Medline: 21760631]
72. Johnson Stoklossa CA, Forhan M, Padwal RS, Gonzalez MC, Prado CM. Practical considerations for body composition assessment of adults with class ii/iii obesity using bioelectrical impedance analysis or dual-energy x-ray absorptiometry. Curr Obes Rep 2016 Dec;5(4):389-396. [doi: 10.1007/s13679-016-0228-5] [Medline: 27591783]
73. Khalil SF, Mohktar MS, Ibrahim F. The theory and fundamentals of bioimpedance analysis in clinical status monitoring and diagnosis of diseases. Sensors (Basel) 2014 Jun 19;14(6):10895-10928. [doi: 10.3390/s140610895] [Medline: 24949644]
74. Marra M, Sammarco R, De Lorenzo A, et al. Assessment of body composition in health and disease using bioelectrical impedance analysis (BIA) and dual energy x-ray absorptiometry (DXA): a critical overview. Contrast Media Mol Imaging 2019;2019:3548284. [doi: 10.1155/2019/3548284] [Medline: 31275083]
75. Buckinx F, Reginster JY, Dardenne N, et al. Concordance between muscle mass assessed by bioelectrical impedance analysis and by dual energy x-ray absorptiometry: a cross-sectional study. BMC Musculoskelet Disord 2015 Mar 18;16(1):60. [doi: 10.1186/s12891-015-0510-9] [Medline: 25887598]
76. Ling CHY, de Craen AJM, Slagboom PE, et al. Accuracy of direct segmental multi-frequency bioimpedance analysis in the assessment of total body and segmental body composition in middle-aged adult population. Clin Nutr 2011 Oct;30(5):610-615. [doi: 10.1016/j.clnu.2011.04.001] [Medline: 21555168]
77. Jette DU, Hunter SJ, Burkett L, et al. Physical therapist management of total knee arthroplasty. Phys Ther 2020 Aug 31;100(9):1603-1631. [doi: 10.1093/ptj/pzaa099] [Medline: 32542403]
78. Dobson F, Hinman RS, Roos EM, et al. OARSI recommended performance-based tests to assess physical function in people diagnosed with hip or knee osteoarthritis. Osteoarthritis Cartilage 2013 Aug;21(8):1042-1052. [doi: 10.1016/j.joca.2013.05.002] [Medline: 23680877]
79. Rehabilitative care best practices for patients with primary hip & knee replacements. : Rehabilitative Care Alliance; 2017 URL: https://rehabcarealliance.ca/wp-content/uploads/2025/03/TJR_Framework.pdf [accessed 2026-02-13]
80. Studenski S, Perera S, Patel K, et al. Gait speed and survival in older adults. JAMA 2011 Jan 5;305(1):50-58. [doi: 10.1001/jama.2010.1923] [Medline: 21205966]
81. Fenner V, Behrend H, Kuster MS. Whole body gait function during stair ascending and level walking in patients following total knee arthroplasty. Int J Phys Med Rehabil 2014;02(5). [doi: 10.4172/2329-9096.1000231]

XSL•FO
RenderX

82. Lord SR, Murray SM, Chapman K, Munro B, Tiedemann A. Sit-to-stand performance depends on sensation, speed, balance, and psychological status in addition to strength in older people. J Gerontol A Biol Sci Med Sci 2002 Aug;57(8):M539-M543. [doi: 10.1093/gerona/57.8.m539] [Medline: 12145369]
83. Mizner RL, Petterson SC, Snyder-Mackler L. Quadriceps strength and the time course of functional recovery after total knee arthroplasty. J Orthop Sports Phys Ther 2005 Jul;35(7):424-436. [doi: 10.2519/jospt.2005.35.7.424] [Medline: 16108583]
84. Bohannon RW. Test-retest reliability of the five-repetition sit-to-stand test: a systematic review of the literature involving adults. J Strength Cond Res 2011 Nov;25(11):3205-3207. [doi: 10.1519/JSC.0b013e318234e59f] [Medline: 21904240]
85. Dobson F, Hinman RS, Hall M, et al. Reliability and measurement error of the Osteoarthritis Research Society International (OARSI) recommended performance-based tests of physical function in people with hip and knee osteoarthritis. Osteoarthritis Cartilage 2017 Nov;25(11):1792-1796. [doi: 10.1016/j.joca.2017.06.006] [Medline: 28647467]
86. Hoens A, Westby M, Longstaff S, Duggan M. Total Joint Arthroplasty and Outcome Measures (TJAOM) toolkit. The University of British Columbia. 2016. URL: https://physicaltherapy.med.ubc.ca/physical-therapy-knowledge-broker/tjaom-toolkit/ [accessed 2026-02-13]
87. Zeni JA, Axe MJ, Snyder-Mackler L. Clinical predictors of elective total joint replacement in persons with end-stage knee osteoarthritis. BMC Musculoskelet Disord 2010 May 6;11(1):86. [doi: 10.1186/1471-2474-11-86] [Medline: 20459622]
88. Bergquist R, Weber M, Schwenk M, et al. Performance-based clinical tests of balance and muscle strength used in young seniors: a systematic literature review. BMC Geriatr 2019 Jan 9;19(1):9. [doi: 10.1186/s12877-018-1011-0] [Medline: 30626340]
89. Gill S, Hely R, Page RS, Hely A, Harrison B, Landers S. Thirty second chair stand test: test-retest reliability, agreement and minimum detectable change in people with early-stage knee osteoarthritis. Physiother Res Int 2022 Jul;27(3):e1957. [doi: 10.1002/pri.1957] [Medline: 35592902]
90. Klukowska AM, Staartjes VE, Vandertop WP, Schröder ML. Five-repetition sit-to-stand test performance in healthy individuals: reference values and predictors from 2 prospective cohorts. Neurospine 2021 Dec;18(4):760-769. [doi: 10.14245/ns.2142750.375] [Medline: 35000330]
91. Schaubert KL, Bohannon RW. Reliability and validity of three strength measures obtained from community-dwelling elderly persons. J Strength Cond Res 2005 Aug;19(3):717-720. [doi: 10.1519/R-15954.1] [Medline: 16095431]
92. McCarthy EK, Horvat MA, Holtsberg PA, Wisenbaker JM. Repeated chair stands as a measure of lower limb strength in sexagenarian women. J Gerontol A Biol Sci Med Sci 2004 Nov;59(11):1207-1212. [doi: 10.1093/gerona/59.11.1207] [Medline: 15602077]
93. Sumbal R, Abbas M, Sheikh SM, Sumbal A. Prevalence and clinical impact of sarcopenia in patients undergoing total joint arthroplasty: a systematic review and a meta-analysis. J Arthroplasty 2024 Dec;39(12):3128-3135. [doi: 10.1016/j.arth.2024.06.021] [Medline: 38901711]
94. Furu M, Ito H, Nishikawa T, et al. Quadriceps strength affects patient satisfaction after total knee arthroplasty. J Orthop Sci 2016 Jan;21(1):38-43. [doi: 10.1016/j.jos.2015.10.002] [Medline: 26755384]
95. Akatsuka Y, Teramoto A, Takashima H, Okada Y, Watanabe K, Yamashita T. Relationships of cross-sectional area of the thigh muscles before or after total knee arthroplasty with postoperative pain or patient satisfaction: a retrospective, exploratory study. Asia Pac J Sports Med Arthrosc Rehabil Technol 2023 Jul;33:20-24. [doi: 10.1016/j.asmart.2023.08.006] [Medline: 37680195]

Abbreviations

1-RM: 1–repetition maximum
30CST: 30-Second Chair Stand Test
5R-STS: 5-Repetition Sit-to-Stand
6MWT: 6-Minute Walk Test
ADL: activities of daily living
BIA: bioelectrical impedance analysis
COSMIN: Consensus-Based Standards for the Selection of Health Measurement Instruments
CT: computed tomography
DXA: dual-energy X-ray absorptiometry
ICC: intraclass correlation coefficient
KOOS: Knee Injury and Osteoarthritis Outcome Score
KSS: Knee Society Clinical Rating System
MRI: magnetic resonance imaging
OKS: Oxford Knee Score
PFD: portable fixed dynamometer
PROM: patient-related outcome measure
ROM: range of motion

XSL•FO
RenderX

SCT: Stair Climb Test
STS: Sit-to-Stand
TKA: total knee arthroplasty
TUG: Time Up and Go
WOMAC: Western Ontario and McMaster Universities Osteoarthritis

Edited by N Rohatgi; submitted 12.Sep.2025; peer-reviewed by J Most; revised version received 30.Jan.2026; accepted 04.Feb.2026; published 25.Feb.2026.

Please cite as:
Boukabache A, Maruthainar N, Manhas V, Player D
Multidimensional Assessment of Recovery After Total Knee Arthroplasty in Clinical Practice: Critical Narrative Review
JMIR Perioper Med 2026;9:e84011
URL: https://periop.jmir.org/2026/1/e84011
doi:10.2196/84011

© Abderrahmane Boukabache, Nimalan Maruthainar, Vikrant Manhas, Darren Player. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Forced-Air Warming Temperature Settings for Treating Postoperative Hypothermia in the Postanesthesia Care Unit: Randomized Controlled Trial

Koravee Pasutharnchat, MD; Rattaphol Seangrung, MD; Sirikarn Sirisophaphong, MD; Wilailuck Wongkum, BNS

Department of Anesthesiology, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, 270 Rama 6 Road, Ratchathewi, Bangkok, Thailand

Corresponding Author:
Koravee Pasutharnchat, MD
Department of Anesthesiology, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, 270 Rama 6 Road, Ratchathewi, Bangkok, Thailand

Related Article:

Correction of: https://periop.jmir.org/2026/1/e85045

(JMIR Perioper Med 2026;9:e110801) doi:10.2196/110801

In "Forced-Air Warming Temperature Settings for Treating Postoperative Hypothermia in the Postanesthesia Care Unit: Randomized Controlled Trial" [1], the authors noted 9 errors. The following corrections have been made:

First, in the Methods section of the abstract, the following sentence was revised:

> *A total of 132 patients undergoing elective surgery at Ramathibodi Hospital between April 2023 and May 2024 were randomized into 3 groups (n=44 per group): group C (warming set to 38 °C), group F1 (warming set to 42 °C), and group F2 (warming set to 42 °C, reduced to 38 °C after achieving 36 °C).*

The corrected sentence now reads:

> *A total of 132 patients undergoing elective surgery at Ramathibodi Hospital between April 2023 and May 2024 were randomized into 3 groups (n=44 per group): group C (warming set to 38 °C), group F1 (warming set to 43 °C), and group F2 (warming set to 43 °C, reduced to 38 °C after achieving 36 °C).*

Second, in the Conclusions section of the abstract, the following sentence was revised:

> *While rewarming times were similar across groups, the protocol using an initial setting of 42 °C followed by a reduction to 38 °C (group F2) effectively minimized temperature drops after normothermia, suggesting its superiority for managing postoperative hypothermia in the postanesthesia care unit.*

The corrected sentence now reads:

> *While rewarming times were similar across groups, the protocol using an initial setting of 43 °C followed by a reduction to 38 °C (group F2) effectively minimized temperature drops after normothermia, suggesting its superiority for managing postoperative hypothermia in the postanesthesia care unit.*

Third, in the final paragraph of the Introduction, the following sentences were revised:

> *Current practice in our PACUs uses FAW with adjustable temperature settings ranging from 38 °C to 42 °C; however, no standardized protocol exists to guide optimal temperature selection. This variability reflects broader uncertainty regarding the most effective strategy for postoperative rewarming and underscores the need for evidence-based guidance. To our knowledge, no prior randomized trial has evaluated a step-down temperature protocol (42 °C to 38 °C) in a mixed adult surgical population.*

The corrected sentences now read:

> *Current practice in our PACUs uses FAW with adjustable temperature settings ranging from 38 °C to 43 °C; however, no standardized protocol exists to guide optimal temperature selection. This variability reflects broader uncertainty regarding the most effective strategy for postoperative rewarming and underscores the need for evidence-based guidance. To our knowledge, no prior randomized trial has evaluated a step-down temperature protocol (43 °C to 38 °C) in a mixed adult surgical population.*

Fourth, in the Rewarming section of the Methods, the following sentence was revised:

> *Only those with a core temperature lower than 36 °C were enrolled and randomly allocated into 3 groups (n=44 per group): group C (forced-air warming set to 38 °C), group F1 (forced-air warming set to 42 °C), and group F2 (forced-air warming initially set to 42 °C, then reduced to 38 °C once the core temperature reached 36 °C).*

The corrected sentence now reads:

> *Only those with a core temperature lower than 36 °C were enrolled and randomly allocated into 3 groups (n=44 per group): group C (forced-air warming set*

XSL•FO
RenderX

to 38 °C), group F1 (forced-air warming set to 43 °C), and group F2 (forced-air warming initially set to 43 °C, then reduced to 38 °C once the core temperature reached 36 °C).

Fifth, in the Baseline Characteristics section of the Results, the following sentence was revised:

Finally, 132 patients were included and equally divided into 3 groups (Figure 1): group C (rewarming set to 38 °C, n=44), group F1 (rewarming set to 42 °C, n=44), and group F2 (rewarming set to 42 °C until reaching a core temperature of 36 °C, then reduced to 38 °C, n=44).

The corrected sentence now reads:

Finally, 132 patients were included and equally divided into 3 groups (Figure 1): group C (rewarming set to 38 °C, n=44), group F1 (rewarming set to 43 °C, n=44), and group F2 (rewarming set to 43 °C until reaching a core temperature of 36 °C, then reduced to 38 °C, n=44).

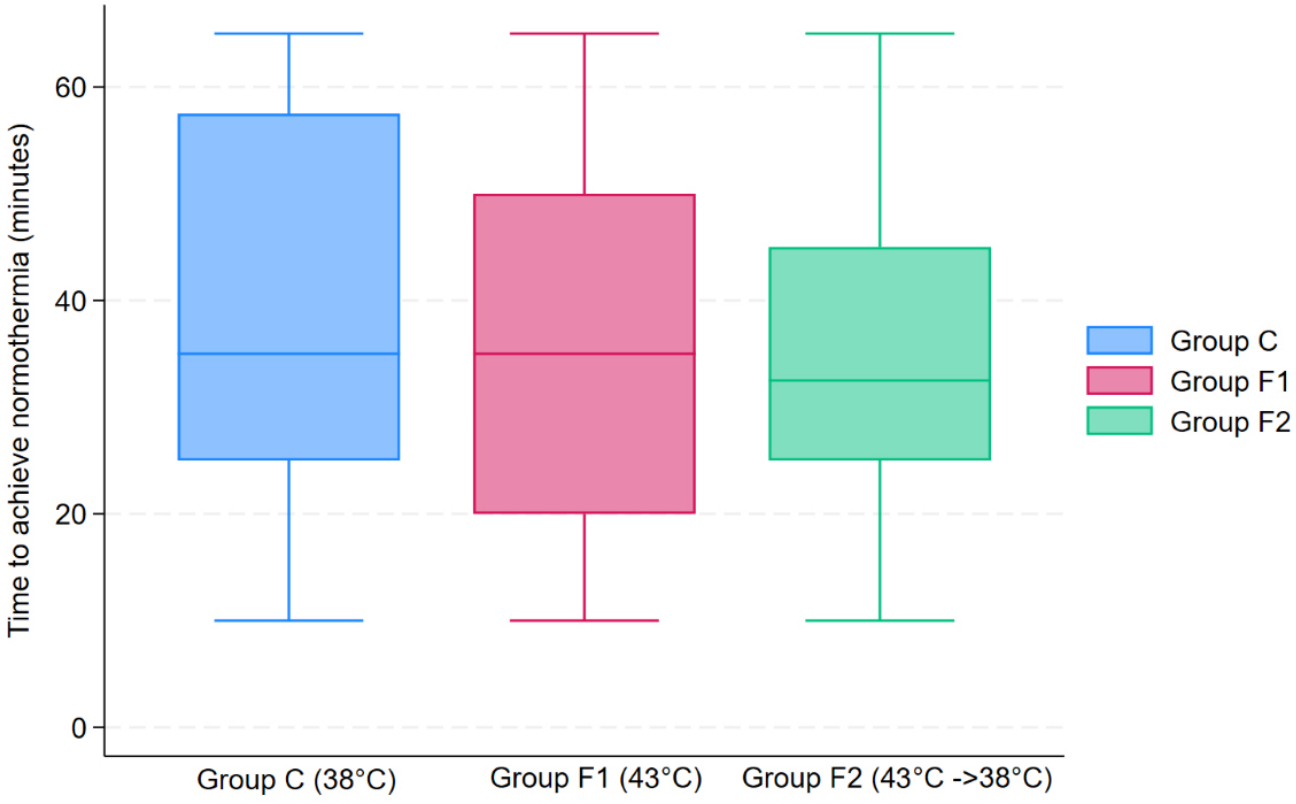
Sixth, in the originally published Figure 2, all labels within the figure image that referred to the high-temperature setting displayed 42 °C. The figure caption read:

Control group (group C): forced-air warming at 38 °C; group F1: forced-air warming at 42 °C; group F2: forced-air warming initially set at 42 °C, then reduced to 38 °C upon reaching 36 °C.

In the corrected Figure 2 (included here as Figure 1), all labels within the figure image that refer to the high-temperature setting display 43 °C. The corrected version of Figure 2 has replaced the originally published figure. The figure caption now reads:

Control group (group C): forced-air warming at 38 °C; group F1: forced-air warming at 43 °C; group F2: forced-air warming initially set at 43 °C, then reduced to 38 °C upon reaching 36 °C.

Figure 1. Revised version of Figure 2: Comparison of rewarming outcomes among the 3 study groups. Box plot showing the distribution of time to achieve normothermia. The horizontal line within each box represents the median rewarming time. The top and bottom boundaries of the boxes indicate the IQR, and the whiskers extend to the minimum and maximum values. No statistically significant differences were observed (*P*=.460). Control group (group C): forced-air warming at 38 °C; group F1: forced-air warming at 43 °C; group F2: forced-air warming initially set at 43 °C, then reduced to 38 °C upon reaching 36 °C.



Seventh, in the originally published Figure 3, all labels within the figure image that referred to the high-temperature setting displayed 42 °C. The figure caption read:

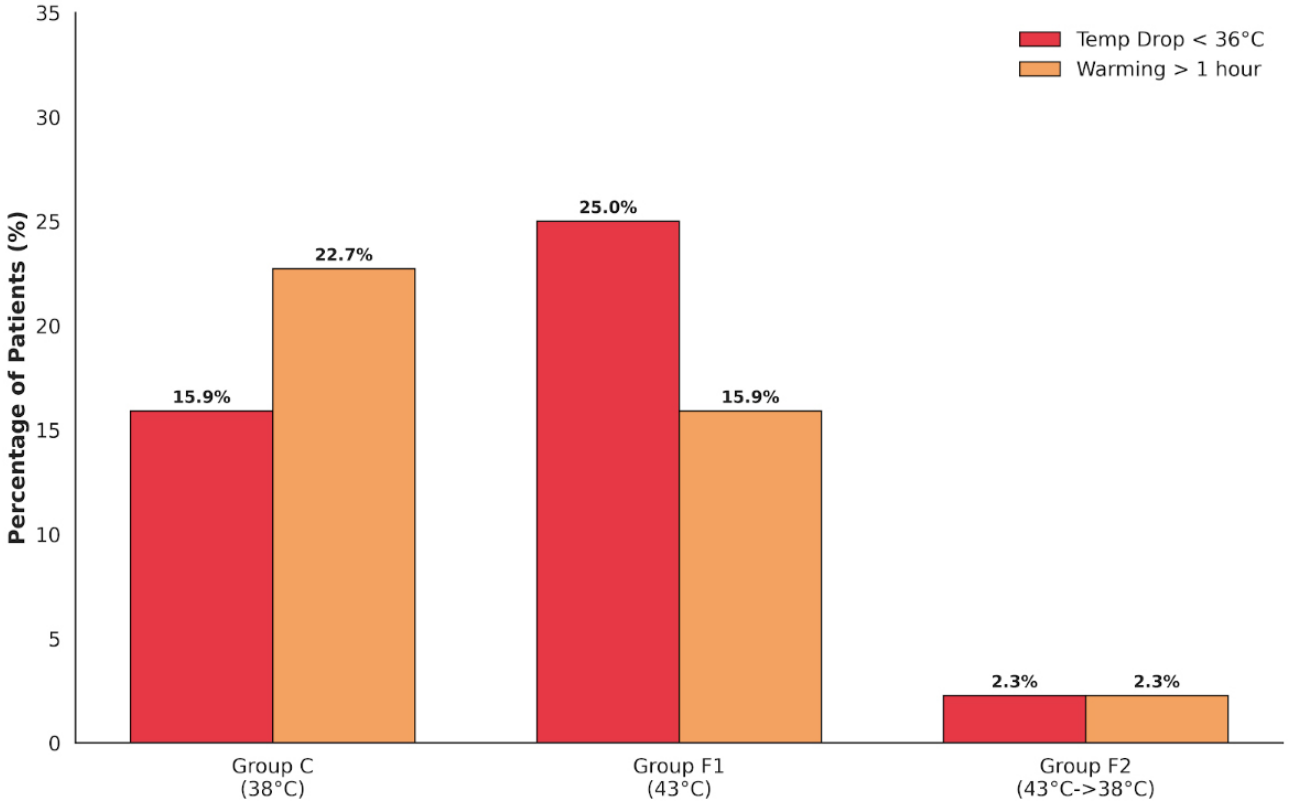
Control group (group C): forced-air warming at 38 °C; group F1: forced-air warming at 42 °C; group F2: forced-air warming initially set at 42 °C, then reduced to 38 °C upon reaching 36 °C.

In the corrected Figure 3 (included here as Figure 2), all labels within the figure image that refer to the high-temperature setting display 43 °C. The corrected version of Figure 3 has replaced the originally published figure. The figure caption reads:

Control group (group C): forced-air warming at 38 °C; group F1: forced-air warming at 43 °C; group F2: forced-air warming initially set at 43 °C, then reduced to 38 °C upon reaching 36 °C.

XSL•FO
RenderX

Figure 2. Revised version of Figure 3: Comparison of rewarming outcomes among the 3 study groups. Clustered bar chart illustrating the incidence of recurrent hypothermia (core temperature dropping <36 °C after achieving normothermia) and the proportion of patients requiring active warming for more than 1 h. Group F2 demonstrated significantly lower rates for both outcomes compared to groups C and F1 (*P*=.009 and *P*=.017, respectively). Control group (group C): forced-air warming at 38 °C; group F1: forced-air warming at 43 °C; group F2: forced-air warming initially set at 43 °C, then reduced to 38 °C upon reaching 36 °C.



Eighth, in the Principal Findings section of the Discussion, the following sentence was revised:

> *The principal finding of this study is that increasing the FAW setting from 38 to 42 °C did not yield a statistically significant reduction in the overall rewarming time to normothermia.*

The corrected sentence now reads:

> *The principal finding of this study is that increasing the FAW setting from 38 °C to 43 °C did not yield a statistically significant reduction in the overall rewarming time to normothermia.*

Finally, in the Conclusions section of the Discussion, the following sentence was revised:

> *In summary, although varying temperature settings of forced-air warming systems produced comparable rewarming times, the protocol involving an initial setting of 42 °C followed by a reduction to 38 °C (group F2) was associated with superior maintenance of normothermia and a significantly lower incidence of postoperative hypothermia recurrence.*

The corrected sentence now reads:

> *In summary, although varying temperature settings of forced-air warming systems produced comparable rewarming times, the protocol involving an initial setting of 43 °C followed by a reduction to 38 °C (group F2) was associated with superior maintenance of normothermia and a significantly lower incidence of postoperative hypothermia recurrence.*

The correction will appear in the online version of the paper on the JMIR Publications website together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Pasutharnchat K, Seangrung R, Sirisophaphong S, Wongkum W. Forced-air warming temperature settings for treating postoperative hypothermia in the postanesthesia care unit: randomized controlled trial. JMIR Perioper Med 2026 Jan 28;9:e85045. [doi: 10.2196/85045] [Medline: 41604636]

XSL•FO
RenderX

Submitted 31.Aug.2026; this is a non–peer-reviewed article;accepted 09.Sep.2026; published 27.Oct.2026.

Please cite as:
Pasutharnchat K, Seangrung R, Sirisophaphong S, Wongkum W
Correction: Forced-Air Warming Temperature Settings for Treating Postoperative Hypothermia in the Postanesthesia Care Unit: Randomized Controlled Trial
JMIR Perioper Med 2026;9:e110801
URL: https://periop.jmir.org/2026/1/e110801
doi:10.2196/110801

© Koravee Pasutharnchat, Rattaphol Seangrung, Sirikarn Sirisophaphong, Wilailuck Wongkum. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 27.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Radiology-to-Surgical Consultation Timing for Suspected Appendicitis in the Emergency Department: Retrospective Observational Study

Phillip Jenkins[1,2], MD, MS; Obert Xu[3], MBBS, MBA, BSN; Shelby Willis, MD; Ethan Kimball[3], DO; Ruchi Thanawala[1,2,4], MD, MS; Beech Burns[3], MD, MCR; Julie Doberne[1,2,4], MD, PhD

[1]Surgical Data and Decision Sciences Lab, Department of Surgery, Oregon Health & Science University, 3181 Sam Jackson Park Road, Portland, OR, United States

[2]Division of Informatics and Clinical Epidemiology, Department of Medicine, Oregon Health & Science University, Portland, OR, United States

[3]Department of Emergency Medicine, Oregon Health & Science University, Portland, OR, United States

[4]Division of Cardiothoracic Surgery, Department of Surgery, Oregon Health & Science University, Portland, United States

Corresponding Author:
Julie Doberne, MD, PhD
Surgical Data and Decision Sciences Lab, Department of Surgery, Oregon Health & Science University, 3181 Sam Jackson Park Road, Portland, OR, United States

Abstract

Background: Acute appendicitis is a common condition that typically requires surgery, yet the workflow linking diagnostic imaging to surgical consultation varies substantially across emergency departments (EDs). Delays between imaging completion and consultation ordering may prolong care and contribute to avoidable clinical and operational inefficiencies.

Objective: This study aimed to quantify real-world consultation delays among patients with radiology-confirmed appendicitis, identify factors associated with prolonged delays, and establish a manually adjudicated dataset linking radiology impression language to consultation behavior as an empirical foundation for future workflow automation. This study did not develop or evaluate a natural language processing (NLP) system; rather, the NLP and automation applications discussed represent potential areas for future work.

Methods: We performed a retrospective observational study of ED encounters at a single academic medical center from January 1, 2020, to December 31, 2025, during which abdominal imaging was obtained to evaluate possible appendicitis (N=1422). All radiology impressions were manually adjudicated and classified as positive, indeterminate, or negative. The primary timing measure was the interval from imaging completion to the placement of a surgical consultation order, which was analyzed within a predefined plausible window of −60 to +360 minutes. Mann-Whitney *U* tests were used to compare delays across imaging modalities and age groups. Logistic regression was used to assess predictors of prolonged delay (>30 minutes).

Results: Among 1422 encounters, radiology findings were classified as positive for appendicitis in 566 (39.8%), indeterminate in 675 (47.5%), and negative in 181 (12.7%). Surgical teams evaluated 99.8% (565/566) of patients with radiology-positive findings. Among 524 radiology-positive encounters with complete time stamps within the plausible window, the median time from imaging completion to consultation order was 30.8 (IQR 17.8-48.5) minutes, and 268 (51.1%) encounters exceeded 30 minutes. Delays were longer for computed tomography (CT) than for ultrasound (median 34.9 [IQR 21.7-53.0] vs 21.2 [IQR 10.9-36.4] minutes; *P*<.001). In adjusted analysis, CT was associated with prolonged delay (odds ratio [OR] 2.29, 95% CI 1.08-4.86; *P*=.03), whereas adult age group was not (OR 1.09, 95% CI 0.52-2.28; *P*=.81). Across a typical year, cumulative imaging-to-consultation waiting time totaled approximately 58 patient-hours.

Conclusions: Radiology-confirmed appendicitis reliably triggered surgical evaluation, yet meaningful delays remained. This foundational timing analysis and the accompanying adjudicated dataset quantify the opportunity for future workflow improvements and can inform the development and evaluation of NLP-driven tools to automate consultation activation following clear radiologic diagnoses. Prospective studies are needed to develop and validate such systems.

(JMIR Perioper Med 2026;9:e91385) doi:10.2196/91385

KEYWORDS

appendicitis; emergency department; surgical consultation; workflow; radiology reporting

XSL•FO
RenderX

Introduction

Background

Acute appendicitis remains one of the most common conditions requiring urgent surgical intervention in both children and adults, and timely evaluation is essential to reducing the risk of perforation and postoperative complications [1,2]. Contemporary practice relies heavily on imaging, with computed tomography (CT) and ultrasound now central to diagnostic pathways [3]. Although imaging improves diagnostic confidence, it also introduces delays in the emergency department (ED) [1-4]. Recent workflow studies confirm that imaging remains a major contributor to ED length of stay: CT-related actions account for approximately 30% of total ED time for abdominal pain evaluations, and radiology report turnaround represents approximately one-third of the CT-related time interval [5]. These findings highlight how imaging-driven diagnostic workflows can prolong the time to definitive surgical consultation and management.

Variability in Consultation Workflows

Clinical pathways differ widely in the degree and timing of surgical team involvement, and imaging modality selection is itself shaped by patient factors such as age and concerns about radiation exposure [6,7]. In many pediatric and adult pathways, imaging is obtained before surgical evaluation, even when the pretest probability of appendicitis is high. Consequently, surgical consultation depends on ED clinicians initiating communication after reviewing radiology results, with 1 study demonstrating a median interval of more than 2 hours between ED physician evaluation and surgical consultation, even under structured protocols [8]. Imaging modality choice can further prolong evaluation; for example, abdominal magnetic resonance imaging (MRI) in children or pregnant patients increases the time to surgical involvement because of the need for specialized radiology interpretation [6]. In contrast, several studies show that earlier surgical involvement can expedite definitive care. An electronic health record (EHR)–integrated decision support tool in 2 pediatric EDs prompted surgical consultation before CT in high-risk cases and decreased CT use by more than half without increasing the rates of missed appendicitis [9]. The FASTPASS direct-to-surgery protocol produced similar gains and halved ED-to-operation times for appendicitis by simplifying communication steps [10]. Together, these studies demonstrate that communication workflows among the ED, radiology, and surgery are variable, subject to delay, and potentially modifiable.

Radiology Reporting and Informatics Approaches

Communication challenges may also arise from variability in radiology reporting. Radiology impressions are largely unstructured, and free-text wording differs among radiologists [11,12]. Because ED clinicians rely on impression text to determine the urgency of surgical evaluation, inconsistent or ambiguous language may create avoidable delays [13].

Informatics approaches have been proposed to address this variability. Natural language processing (NLP) systems have demonstrated high accuracy in identifying appendicitis in radiology impressions, with reported accuracies of 0.97 to 0.99 and F_1-scores near 0.90 [14], and large language models (LLMs) can now enhance these information extraction capabilities [15]. Computerized consult systems and automated paging tools have shortened ED length of stay by improving the reliability of interservice communication [16], while automated closed-loop notification systems have improved the timeliness and overall documentation of critical imaging result communication [17,18]. Similarly, decision support interventions that modify consult timing have been shown to change workflow behavior and improve outcomes [19].

Study Objectives

Despite these promising developments, no published work has established a clinically adjudicated dataset that links radiology impression language to actual surgical consultation initiation for suspected appendicitis. A validated dataset and an understanding of existing variation in consultation timing are prerequisites for designing, implementing, or evaluating automated triggers or NLP-driven tools. The aims of this study were therefore to (1) manually classify radiology impressions obtained for suspected appendicitis, (2) quantify the real-world interval between imaging completion and surgical consultation, and (3) identify factors associated with prolonged delay. We emphasize that this study is foundational: it does not develop, validate, or benchmark an NLP model, and the automation applications discussed are prospective.

Methods

Study Design and Setting

We conducted a retrospective observational study of ED encounters at a single academic medical center in which abdominal imaging was performed to evaluate suspected appendicitis. All eligible encounters from January 1, 2020, to December 31, 2025, were extracted from the institution's clinical data warehouse.

Ethical Considerations

This study was reviewed by the Oregon Health & Science University Institutional Review Board (STUDY00029543) and was determined not to constitute human subjects research; therefore, informed consent was not required. Data were handled in accordance with institutional privacy and security policies.

Data Source and Cohort Construction

The dataset included demographic characteristics, encounter time stamps, imaging orders, radiology impressions, and surgical consultation orders. All primary imaging modalities capable of evaluating appendicitis were included, including CT of the abdomen and pelvis, ultrasound of the appendix, limited abdominal ultrasound, and pediatric ultrasound of the appendix. Encounters missing an imaging report, consultation order, or specific time stamps were retained in the dataset but were excluded from analyses that required those elements. Chart review was performed for patients with radiology impressions positive for appendicitis but no consultation order.

XSL•FO
RenderX

Manual Classification of Radiology Impressions

Each radiology impression was reviewed and assigned to 1 of 3 categories based on explicit language regarding appendicitis. Impressions were labeled positive when they described acute appendicitis or complications such as perforation; negative when they described a normal appendix, the absence of secondary signs, or a clear alternative diagnosis; and indeterminate when the appendix was not visualized, visualization was incomplete, or the impression stated that appendicitis could not be confirmed or excluded. A single reviewer, a surgical resident with 4 years of experience and graduate training in clinical informatics (PJ), adjudicated all impressions using these predefined rules; ambiguous cases were resolved by reviewing the full radiology report and encounter documentation. This adjudication established the reference dataset used to characterize real-world consultation behavior.

Identification of Surgical Consultation Events

Surgical consultation events were identified by reviewing all order names associated with general surgery and pediatric surgery. Because consultation naming conventions varied, a normalization process was used to map all relevant orders to a single binary variable. Only consultations placed during the ED encounter were included.

Timing Measures

Several time stamps were available to measure the interval between radiologic diagnosis and consultation, including the time the examination ended, the time the preliminary interpretation was entered, the time the final interpretation was signed, and the time the consultation order was placed for the surgical service.

Because clinical decision-making often occurs before the final radiologist signature, the primary timing measure was the interval between imaging completion (examination end) and the placement of the consultation order. We selected this interval, rather than total ED length of stay or time to operation, because it isolates the discrete interservice communication step that automated consult activation would directly target. These downstream measures may be influenced by many additional factors (eg, operating room availability, boarding, and disposition logistics) that would not be addressed by automation of consult activation.

For the primary analysis, intervals were restricted to a predefined plausible window of −60 to +360 minutes. We retained negative values as low as −60 minutes because consultation orders are sometimes placed shortly before the recorded end of imaging acquisition, for example, when clinicians respond to preliminary communication from radiology or act on clinical concerns while imaging is still in progress. The upper bound of 360 minutes was chosen because consultation orders placed more than 6 hours after imaging completion may reflect a change in the patient's clinical status or an indication unrelated to the index imaging study, rather than a response to its results. This upper bound also approximates the upper limit of typical total ED visit durations in our region [20]. We also performed a sensitivity analysis restricted to nonnegative intervals (0 to 360 minutes).

Timing Analysis Using Radiology Documentation Milestones

To quantify how consultation delays vary depending on the radiology milestone used, we calculated the time from each of 3 radiology documentation milestones to the surgical consultation order: end of examination (the moment the scan itself was completed), preliminary interpretation (the time the first documented radiologist interpretation was entered), and final interpretation (the time the final interpretation was signed). The analysis included only radiology-positive encounters with a documented surgical consultation. For each milestone, we included encounters only if the consultation order occurred at or after the corresponding radiology time stamp, because a consultation placed before that milestone would not reflect the normal ED workflow for appendicitis; we applied no upper time limit in this descriptive analysis. Because these inclusion rules differ from the plausibility window used in the primary analysis, the number of encounters included differs between the 2 analyses (n=513 for the examination-end milestone vs n=524 in the primary analysis).

Statistical Analysis

Descriptive statistics summarized classification frequencies, consultation rates, and time intervals. Continuous variables were nonnormally distributed and were compared using Mann-Whitney *U* tests. Comparison groups included CT vs ultrasound and adult vs pediatric patients.

To evaluate factors associated with delayed consultation, a logistic regression model was constructed with a binary dependent variable indicating whether the interval from examination end to consultation order exceeded 30 minutes. This threshold was prespecified because it approximates the cohort median, yielding a balanced outcome distribution for regression, and because 30 minutes is an established benchmark for timely surgical response in American College of Surgeons verification standards: emergency general surgery standards require a surgeon or surgical team member to be physically available within 30 minutes of request and reserve a 30-minute highest-level consultation response for the most acute patients (Standards PSR.3 and PC.2) [21], whereas trauma center standards require trauma surgeon bedside response to the highest-level activation within 15 minutes (level I or II centers) or 30 minutes (level III centers; Standard 5.4) [22]. Predictor variables included age group (adult or pediatric) and imaging modality (CT or ultrasound). In an exploratory sensitivity model, indicator variables for overnight imaging completion (7 PM to 7 AM) and weekend imaging completion were added to assess confounding by the time of imaging. Odds ratios (ORs) with 95% CIs were calculated from model coefficients.

To estimate cumulative operational burden, we summed imaging-to-consultation intervals within the plausible window and divided the total by the 6-year study period to yield average patient-hours per year.

Software and Reproducibility

We conducted all data preprocessing, merging of manual classifications, and statistical analyses in Python (version 3.12; Python Software Foundation) using the *pandas*, *NumPy*, and

XSL•FO
RenderX

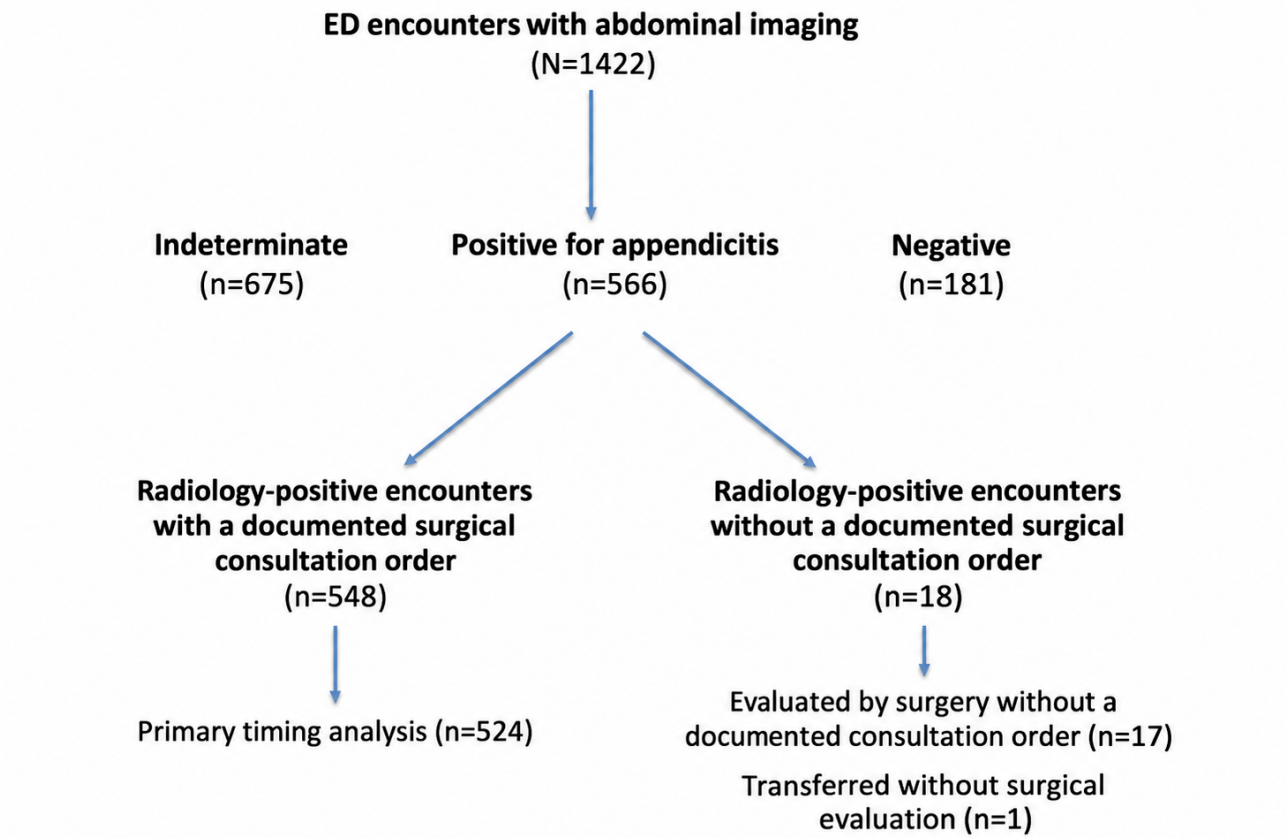
Matplotlib packages, with *SciPy* for univariate testing and *statsmodels* for multivariable regression.

Results

Cohort and Classification of Radiology Impressions

A total of 1422 ED encounters involving abdominal imaging for suspected appendicitis were included. Imaging modalities consisted of CT and ultrasound examinations intended to assess the appendix. After manual review of all radiology impressions, 566 (39.8%) studies were classified as positive for appendicitis, 675 (47.5%) as indeterminate, and 181 (12.7%) as negative. Figure 1 summarizes cohort construction and final classification counts.

Figure 1. Cohort construction and manual classification of radiology impressions among emergency department (ED) encounters involving abdominal imaging for suspected appendicitis (N=1422).



Surgical Consultation After Positive Imaging

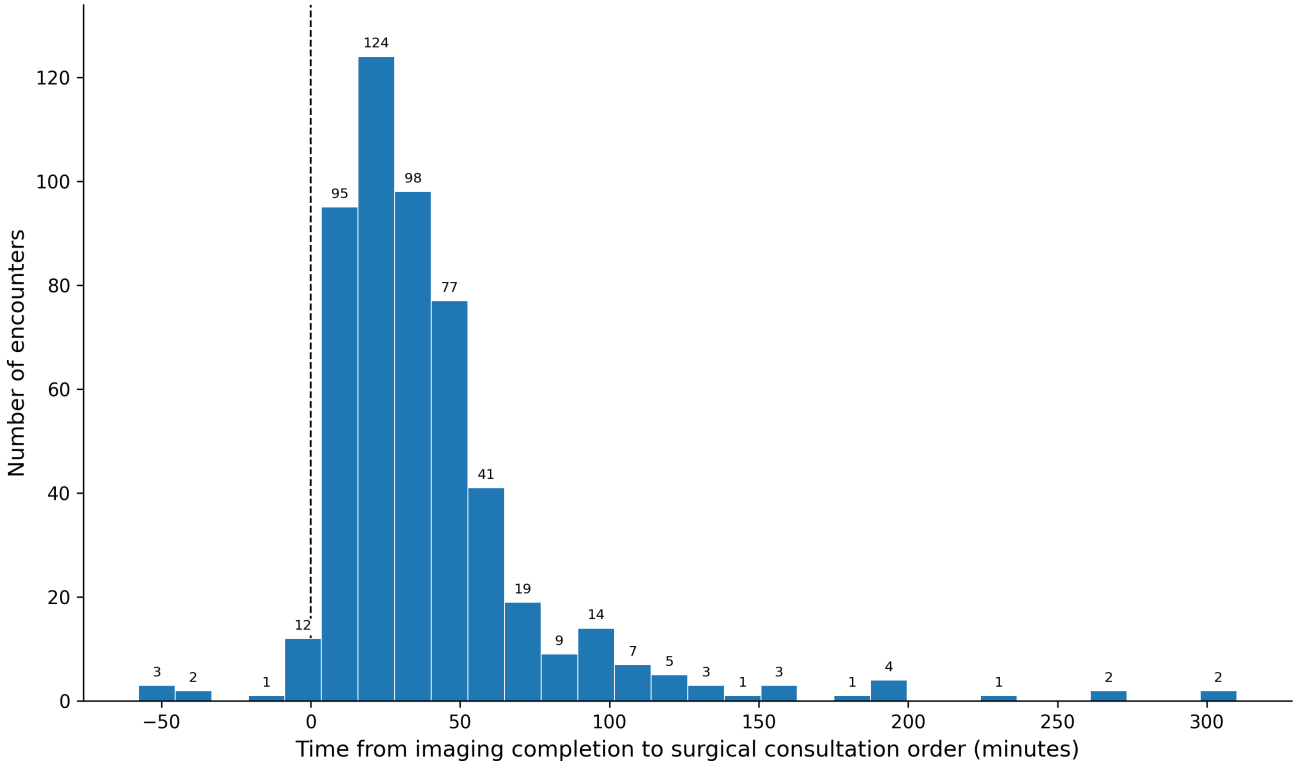
Among the 566 radiology-positive encounters, 548 (96.8%) had a formal surgical consultation order documented in the EHR. Of the 18 encounters without a recorded consultation order, chart review revealed that 17 patients were nevertheless evaluated by the surgical service without a discrete consultation order, most commonly through direct communication leading to admission, including 1 patient who was assessed by the surgical service in the ED before discharge without hospitalization. One patient was transferred to another facility without being seen by surgery at our institution. In total, 565 (99.8%) patients with radiology-confirmed appendicitis received surgical evaluation.

Time From Imaging Completion to Consultation

Of the 548 radiology-positive encounters with complete imaging and consultation time stamps, 24 fell outside the predefined plausible window (n=21 with intervals below −60 minutes and n=3 with intervals above +360 minutes), leaving 524 encounters for the primary analysis. The median interval from imaging completion to consultation order was 30.8 (IQR 17.8-48.5) minutes. A total of 14 (3%) encounters had negative intervals, representing situations in which consultation was initiated shortly before the end of imaging acquisition, typically when preliminary communication or another clinical concern prompted the consultation. In the sensitivity analysis restricted to nonnegative intervals (n=510), the median interval was 31.2 (IQR 18.5-49.4) minutes, which was nearly identical to that in the primary analysis. The overall distribution of consultation delays is shown in Figure 2.

XSL•FO
RenderX

Figure 2. Distribution of time from imaging completion to surgical consultation order among radiology-positive encounters within the plausible window (−60 to +360 minutes; N=524). The dashed line marks 0 minutes.



Consultation Delays by Radiology Documentation Milestone

The number of encounters available for analysis varied across the 3 radiology milestones because of differences in the consistency with which each time stamp was documented. The end of imaging acquisition was available for almost all radiology-positive encounters, whereas the preliminary interpretation time was available less frequently, and the final signature time was the least consistently documented. When a time stamp was unavailable, the corresponding interval could not be calculated, and the encounter was omitted from that comparison. Because examination completion, preliminary interpretation, and final interpretation are generally sequential, consultation delays differed depending on which time stamp was used. Delays calculated from the end of imaging acquisition were the most complete and operationally meaningful; delays calculated from the preliminary interpretation reflected earlier clinical review; and delays calculated from the final signature were heavily skewed by limited after-hours availability of attending radiologists. A full summary of delay distributions for each milestone is provided in Table 1.

Table . Consultation delays measured from 3 radiology documentation milestones among radiology-positive encounters (N=566).

Metrics	Examination end to consultation	Preliminary interpretation to consultation	Final interpretation to consultation
Encounters included[a], n (%)	513 (90.1)	220 (38.9)	148 (26.1)
Delay (minutes), median (IQR)	31.4 (18.5-49.5)	10.2 (2.1-31.1)	10.2 (4.8-25.1)
Delay (minutes), mean (SD)	45.8 (79.2)	29.1 (61.1)	33.3 (119.0)

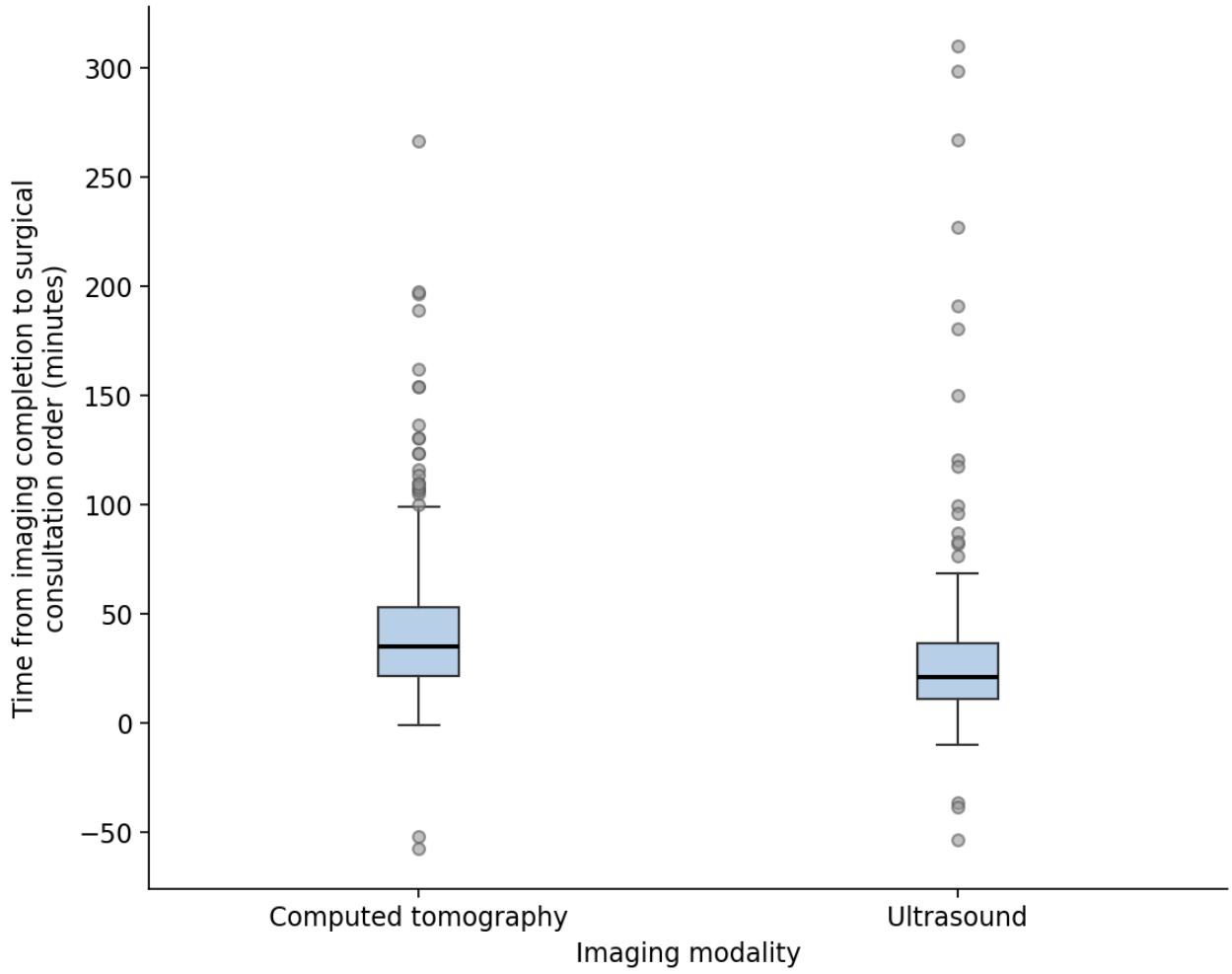
[a]Encounters were included in each column only if the consultation order occurred at or after the corresponding radiology time stamp; no upper time limit was applied in this descriptive analysis. Preliminary and final interpretation time stamps were documented less consistently than examination end times.

Comparisons by Imaging Modality and Age Group

Consultation timing differed by imaging modality. For CT examinations, the median time from imaging completion to consultation order was 34.9 (IQR 21.7-53.0) minutes (n=348), compared with 21.2 (IQR 10.9-36.4) minutes for ultrasound (n=176). A Mann-Whitney *U* test demonstrated a statistically significant difference between modalities (U=41,204; P<.001). These modality-specific patterns are shown in Figure 3.

XSL•FO
RenderX

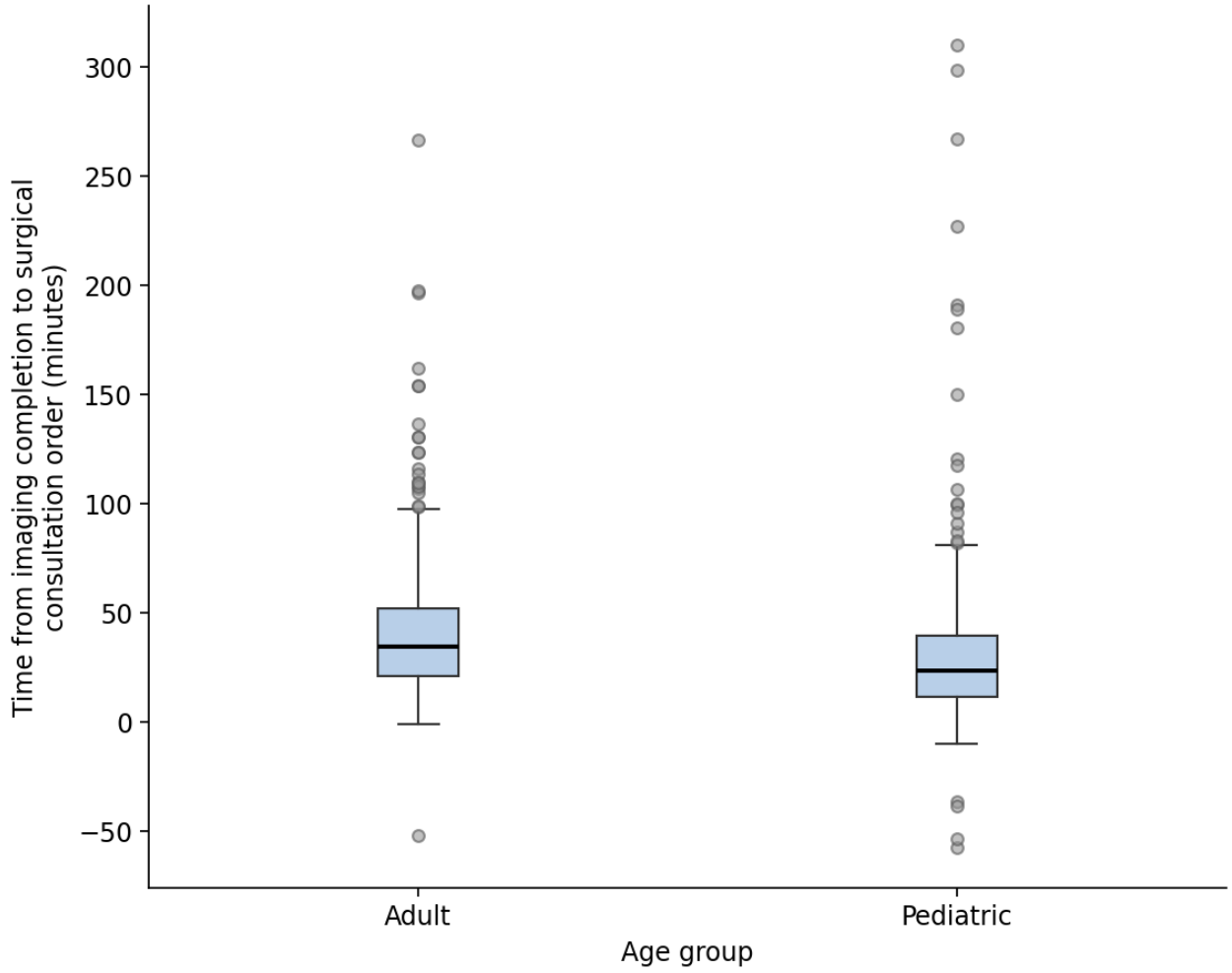
Figure 3. Distribution of time from imaging completion to surgical consultation order, stratified by imaging modality (computed tomography: n=348; ultrasound: n=176).



Timing also differed by age group in the unadjusted comparison. Adults had a median delay of 34.8 (IQR 21.3-51.9) minutes (n=335), while pediatric patients had a median delay of 23.6 (IQR 11.6-39.6) minutes (n=189). This difference was statistically significant (U=40,477; P<.001), as shown in Figure 4.

XSL•FO
RenderX

Figure 4. Distribution of time from imaging completion to surgical consultation order, stratified by age group (adult: n=335; pediatric: n=189).



Predictors of Prolonged Delay

Of the 524 encounters, 268 (51.1%) exceeded the 30-minute threshold. In the adjusted logistic regression analysis, adult age was not significantly associated with prolonged delay (OR 1.09, 95% CI 0.52-2.28; *P*=.81), whereas CT was associated with significantly higher odds of delay than ultrasound (OR 2.29, 95% CI 1.08-4.86; *P*=.03). The overall model was statistically significant (likelihood ratio *P*<.001), indicating that imaging modality contributed meaningfully to variation in consultation timing. In the exploratory model additionally adjusted for overnight imaging completion and weekend, the association with CT persisted (OR 2.32, 95% CI 1.09-4.94; *P*=.03), while neither overnight completion (OR 1.11, 95% CI 0.78-1.57; *P*=.58) nor weekend imaging (OR 0.84, 95% CI 0.57-1.25; *P*=.40) was associated with prolonged delay.

Cumulative Operational Burden

Across the 524 analyzable encounters, imaging-to-consultation waiting time totaled approximately 346 patient-hours over the 6-year study period, corresponding to approximately 58 patient-hours of additional ED occupancy per year attributable to this single diagnosis.

Discussion

Principal Findings

In this single-center retrospective study of 1422 ED encounters involving imaging for suspected appendicitis, 3 principal findings emerged. First, radiology-confirmed appendicitis was almost uniformly followed by surgical evaluation (565/566, 99.8%), indicating that the decision to involve surgery after a positive impression is effectively deterministic. Second, this reliable step was nonetheless delayed: the median interval from imaging completion to consultation order was 30.8 (IQR 17.8-48.5) minutes, 51.1% (268/524) of encounters exceeded 30 minutes, and delays were longer after CT than after ultrasound. Finally, these delays accumulated to approximately 58 patient-hours of additional ED occupancy annually. As health care systems move toward more intelligent and automated workflows, identifying processes that already occur reflexively but are performed slowly is a critical first step toward safely optimizing clinical care.

Comparison With Prior Work

Our findings extend prior work on imaging-driven ED workflow. Structured pediatric appendicitis pathways still exhibited median physician-to-consultation intervals exceeding 2 hours [8], and our data show that even the shorter imaging-to-consultation

XSL•FO
RenderX

segment accounted for a median of half an hour. Interventions that restructure this communication step have consistently improved timeliness: a computerized consultation management system reduced ED length of stay [16], EHR-integrated decision support altered consultation and imaging behavior in pediatric appendicitis [9], the FASTPASS direct-to-surgery pathway halved ED-to-operation times [10], and automated closed-loop notification systems for critical imaging results improved the speed and completeness of documentation of radiologist-to-clinician communication [17,18]. Our study complements this literature by quantifying, at the encounter level, the specific interval that such systems would target for radiographically confirmed appendicitis and by contributing a manually adjudicated dataset that links impression language to consultation behavior.

Automation vs Standardized Protocols

A reasonable question is whether these delays would be better addressed by a standardized clinical protocol than by automation. We view the 2 approaches as complementary rather than competing. Protocolized pathways still depend on a clinician noticing the imaging result and manually initiating consultation; prior work shows that substantial delays persist even under structured protocols [8], and our own data demonstrate that a step that is already effectively obligatory (99.8% of positive studies led to surgical evaluation) still incurred a median delay of half an hour. Automation directly removes the dependence on human vigilance for a task with essentially no decision content, while protocols remain essential for defining who responds and how. Because more than half of encounters exceeded the 30-minute threshold, even modest reductions in activation latency would affect a large share of patients.

Toward Automated Consult Activation: Prospective Agenda

The broader significance of this work lies in its implications for future clinical systems, and we emphasize that these implications are prospective: the present study establishes the empirical foundation for automation but does not develop, validate, or benchmark an NLP model. Radiology impressions already function as triggers for action, even when documented as free text, and NLP systems, including modern LLMs, can extract actionable findings from such text with high accuracy [14,15]. Appendicitis is a favorable prototype condition because imaging has a high signal-to-noise ratio, treatment pathways are well defined, and downstream clinical actions are consistent; in such settings, automation does not replace clinical judgment but rather formalizes and accelerates decisions that are already being made reliably. Future work should develop and prospectively evaluate tools that use radiology-confirmed appendicitis findings to automatically initiate surgical consultation, trigger standardized order sets, and notify relevant care teams, with clinician oversight retained for radiology-negative or indeterminate cases that require more nuanced decision-making. The adjudicated dataset created here is designed to support the training and validation of such systems. Similar imaging-driven activation points exist across emergency and surgical care, each offering an opportunity to better align operational workflows with predictable clinical decisions.

Operational and Economic Implications

Although this analysis was not designed as a formal economic evaluation, the operational impact of these delays is meaningful. At a single institution, the cumulative imaging-to-consultation delay for appendicitis accounted for approximately 58 patient-hours of additional ED occupancy annually. When considered alongside conservative estimates of ED cost per patient-hour, this translates to several thousand dollars per year attributable to a single common diagnosis [20,23]. More importantly, these delays consume limited ED capacity, contribute to crowding, and increase the cognitive burden on clinicians who must manually track and act on imaging results. When these delays are multiplied across institutions and other imaging-driven conditions, the system-level impact is likely substantial.

Limitations

Several limitations warrant consideration. First, this was a single-center study, and consultation workflows, staffing patterns, and radiology coverage differ across institutions; in particular, our findings partly reflect local factors such as after-hours attending radiologist availability and surgical service structure, which may limit their generalizability. Second, all impressions were classified by a single reviewer, a surgical resident with clinical informatics training; although predefined classification rules were applied and ambiguous cases were resolved by reviewing the full report, interrater reliability could not be assessed, and some subjectivity in classification is possible. Third, EHR time stamps are imperfect proxies for clinical events: they do not capture informal or verbal communication, documentation practices varied across milestones (with preliminary and final interpretation times recorded inconsistently), and consultation orders may be entered before or after the actual conversation with the surgical service, all of which may bias measured intervals in either direction. Fourth, although an exploratory model found no association between prolonged delay and overnight or weekend imaging completion, other potentially relevant variables, including ED workload and crowding, radiology turnaround time, individual clinician staffing, and patient complexity, were not reliably available in the dataset and could have confounded the observed associations. Fifth, the retrospective design limits our ability to assess downstream clinical outcomes such as time to operation, perforation, or complication rates. Finally, the cost estimate is a coarse extrapolation intended only to illustrate the order of magnitude.

Conclusions

This study demonstrates that radiographic identification of appendicitis is almost uniformly followed by surgical consultation but that the process incurs a consistent, measurable delay. By quantifying existing delays and establishing a manually adjudicated reference dataset, this work reframes workflow automation as an opportunity to better align clinical systems with decisions clinicians are already making. Future prospective studies should focus on developing and evaluating

XSL•FO
RenderX

automated activation tools that leverage radiology impressions to improve timeliness, reduce operational inefficiency, and serve as scalable models for system-wide improvement in surgical care.

Generative AI Use

The authors declare that no generative AI tool was used in the study design, data collection, statistical analysis, interpretation, or preparation of the manuscript.

Funding

The authors declare that no financial support was received for this work.

Data Availability

The dataset generated and analyzed during this study contains protected health information and is not publicly available. Deidentified data may be made available from the corresponding author upon reasonable request and subject to institutional approval.

Authors' Contributions

Conceptualization: PJ, OX, SW, RT, BB, JD
Data curation: PJ, SW, BB, JD
Formal analysis: PJ, SW, EW, RT, JD
Investigation: PJ
Methodology: PJ
Project administration: OX
Resources: OX, EW, JD
Software: PJ
Supervision: RT, BB, JD
Validation: PJ, JD
Visualization: PJ
Writing—original draft: PJ, JD
Writing—review and editing: PJ, OX, SW, EW, RT, BB, JD

Conflicts of Interest

None declared.

References

1. Lotfollahzadeh S, Lopez RA, Deppen JG. Appendicitis. In: StatPearls: StatPearls Publishing; 2024. URL: https://www.ncbi.nlm.nih.gov/books/NBK493193/ [accessed 2026-08-17]
2. Tang G, Zhang L, Xia L, Zhang J, Chen R, Zhou R. Preoperative in-hospital delay increases postoperative morbidity and mortality in patients with acute appendicitis: a meta-analysis. Int J Surg 2025 Jan 1;111(1):1275-1284. [doi: 10.1097/JS9.0000000000001938] [Medline: 38990284]
3. Di Saverio S, Podda M, De Simone B, et al. Diagnosis and treatment of acute appendicitis: 2020 update of the WSES Jerusalem guidelines. World J Emerg Surg 2020 Apr 15;15(1):27. [doi: 10.1186/s13017-020-00306-3] [Medline: 32295644]
4. Dhanik A, Stenson BA, Levenson RB, Antkowiak PS, Sanchez LD, Chiu DT. Root cause analysis of delayed emergency department computed tomography scans. West J Emerg Med 2024 Mar;25(2):226-229. [doi: 10.5811/westjem.17831] [Medline: 38596923]
5. Wang DC, Parry CR, Feldman M, Tomlinson G, Sarrazin J, Glanc P. Acute abdomen in the emergency department: is CT a time-limiting factor? AJR Am J Roentgenol 2015 Dec;205(6):1222-1229. [doi: 10.2214/AJR.14.14057] [Medline: 26587929]
6. Marin JR, Lyons TW, Claudius I, et al. Optimizing advanced imaging of the pediatric patient in the emergency department: technical report. Pediatrics 2024 Jul 1;154(1):e2024066855. [doi: 10.1542/peds.2024-066855] [Medline: 38932719]
7. Durupt G, Ranzenberger LR, Tafti D, Carter KR. Appendix imaging. In: StatPearls: StatPearls Publishing; 2026. URL: https://www.ncbi.nlm.nih.gov/books/NBK549903/ [accessed 2026-08-17]
8. Saucier A, Huang EY, Emeremni CA, Pershad J. Prospective evaluation of a clinical pathway for suspected appendicitis. Pediatrics 2014 Jan;133(1):e88-e95. [doi: 10.1542/peds.2013-2208] [Medline: 24379237]
9. Kharbanda AB, Madhok M, Krause E, et al. Implementation of electronic clinical decision support for pediatric appendicitis. Pediatrics 2016 May;137(5):e20151745. [doi: 10.1542/peds.2015-1745] [Medline: 27244781]
10. Kulvatunyou N, Zimmerman SA, Adhikhari S, et al. The impact of FASTPASS: a collaboration with emergency department to improve management of patients with gallbladder disease and acute appendicitis. J Surg Res 2021 Apr;260:293-299. [doi: 10.1016/j.jss.2020.11.018] [Medline: 33360754]

XSL•FO
RenderX

11. Pesapane F, Tantrige P, De Marco P, et al. Advancements in standardizing radiological reports: a comprehensive review. Medicina (Kaunas) 2023 Sep 17;59(9):1679. [doi: 10.3390/medicina59091679] [Medline: 37763797]
12. Schwartz LH, Panicek DM, Berk AR, Li Y, Hricak H. Improving communication of diagnostic radiology findings through structured reporting. Radiology 2011 Jul;260(1):174-181. [doi: 10.1148/radiol.11101913] [Medline: 21518775]
13. Choi K, Choi JY, Kim HJ, Kim HJ, Jang SK. Added value of structured reporting for US of the pediatric appendix: additional CT examinations and negative appendectomy. J Korean Soc Radiol 2023 May;84(3):653-662. [doi: 10.3348/jksr.2022.0027] [Medline: 37324982]
14. Li MD, Wood PA, Alkasab TK, Lev MH, Kalpathy-Cramer J, Succi MD. Automated tracking of emergency department abdominal CT findings during the COVID-19 pandemic using natural language processing. Am J Emerg Med 2021 Nov;49:52-57. [doi: 10.1016/j.ajem.2021.05.057] [Medline: 34062318]
15. Reichenpfader D, Müller H, Denecke K. A scoping review of large language model based approaches for information extraction from radiology reports. NPJ Digit Med 2024 Aug 24;7(1):222. [doi: 10.1038/s41746-024-01219-0] [Medline: 39182008]
16. Cho SJ, Jeong J, Han S, et al. Decreased emergency department length of stay by application of a computerized consultation management system. Acad Emerg Med 2011 Apr;18(4):398-402. [doi: 10.1111/j.1553-2712.2011.01039.x] [Medline: 21496143]
17. Lacson R, O'Connor SD, Sahni VA, et al. Impact of an electronic alert notification system embedded in radiologists' workflow on closed-loop communication of critical results: a time series analysis. BMJ Qual Saf 2016 Jul;25(7):518-524. [doi: 10.1136/bmjqs-2015-004276] [Medline: 26374896]
18. Morris RW, Williams NM, Gordji RR, et al. Increasing utilization and improving documentation in a radiology critical alert system. Curr Probl Diagn Radiol 2023;52(4):230-232. [doi: 10.1067/j.cpradiol.2023.03.013] [Medline: 37032290]
19. Rosen JE, Monsell SE, DePaoli SC, et al. The use and impact of a decision support tool for appendicitis treatment. Ann Surg 2024 Oct 1;280(4):616-622. [doi: 10.1097/SLA.0000000000006412] [Medline: 38916104]
20. Average Oregon emergency department visit is 5.2 hours, dashboard shows. Oregon Health Authority. URL: https://www.oregon.gov/oha/erd/pages/average-oregon-emergency-department-visit-is-5.2-hours-dashboard-shows.aspx [accessed 2026-08-17]
21. Optimal resources for emergency general surgery: 2022 standards. : American College of Surgeons; 2022 URL: https://www.facs.org/media/evyh55dj/optimal-resources-for-emergency-general-surgery.pdf [accessed 2026-08-17]
22. Resources for optimal care of the injured patient: 2022 standards. : American College of Surgeons; 2025 URL: https://www.facs.org/quality-programs/acs-quality-program-standards-manuals/ [accessed 2026-08-19]
23. Moore BJ, Stocks C, Owens PL. Trends in emergency department visits, 2006-2014. : Agency for Healthcare Research and Quality; 2017 URL: https://hcup-us.ahrq.gov/reports/statbriefs/sb227-Emergency-Department-Visit-Trends.pdf [accessed 2026-08-17]

Abbreviations

CT: computed tomography
ED: emergency department
EHR: electronic health record
LLM: large language model
MRI: magnetic resonance imaging
NLP: natural language processing
OR: odds ratio

Edited by I Steenstra, K Cummings; submitted 13.Jan.2026; peer-reviewed by M Ordu, P Shingru; revised version received 23.Oct.2026; accepted 26.Oct.2026; published 16.Nov.2026.

Please cite as:
Jenkins P, Xu O, Willis S, Kimball E, Thanawala R, Burns B, Doberne J
Radiology-to-Surgical Consultation Timing for Suspected Appendicitis in the Emergency Department: Retrospective Observational Study
JMIR Perioper Med 2026;9:e91385
URL: https://periop.jmir.org/2026/1/e91385
doi:10.2196/91385

© Phillip Jenkins, Obert Xu, Shelby Willis, Ethan Kimball, Ruchi Thanawala, Beech Burns, Julie Doberne. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 16.Nov.2026. This is an open-access article distributed under the terms

XSL•FO
RenderX

of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Forced-Air Warming Temperature Settings for Treating Postoperative Hypothermia in the Postanesthesia Care Unit: Randomized Controlled Trial

Koravee Pasutharnchat, MD; Rattaphol Seangrung, MD; Sirikarn Sirisophaphong, MD; Wilailuck Wongkum, BNS

Department of Anesthesiology, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, 270 Rama 6 Road, Ratchathewi, Bangkok, Thailand

Corresponding Author:
Koravee Pasutharnchat, MD
Department of Anesthesiology, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, 270 Rama 6 Road, Ratchathewi, Bangkok, Thailand

Related Article:

This is a corrected version. See correction statement: https://periop.jmir.org/2026/1/e110801

Abstract

Background: Hypothermia, defined as a core body temperature below 36 °C, is a common postoperative complication associated with adverse outcomes, including delayed wound healing, infections, and increased bleeding.

Objective: This randomized controlled trial evaluated the efficacy of different forced-air warming system temperature settings in treating postoperative hypothermia in the postanesthesia care unit.

Methods: A total of 132 patients undergoing elective surgery at Ramathibodi Hospital between April 2023 and May 2024 were randomized into 3 groups (n=44 per group): group C (warming set to 38 °C), group F1 (warming set to 43 °C), and group F2 (warming set to 43 °C, reduced to 38 °C after achieving 36 °C). Tympanic temperature was recorded at 5-minute intervals during rewarming and every 10 minutes after normothermia (≥36 °C) was achieved. The primary outcome was rewarming time. Secondary outcomes included the incidence of temperature drops, hemodynamic parameters, adverse events, and patient comfort scores.

Results: Baseline characteristics and clinical variables, including vital signs, were comparable among groups (P>.05). Group F2 achieved the shortest mean rewarming time of 33.3 (SD 13.81) min; however, differences between groups were not statistically significant (P=.460). Group F2 had the lowest incidence of temperature drops below 36 °C after normothermia (1/44, 2.27%; P=.009). Group C had the highest incidence of rewarming exceeding 1 hour (10/44, 22.73%; P=.017).

Conclusions: While rewarming times were similar across groups, the protocol using an initial setting of 43 °C followed by a reduction to 38 °C (group F2) effectively minimized temperature drops after normothermia, suggesting its superiority for managing postoperative hypothermia in the postanesthesia care unit.

Trial Registration: Thaiclinicaltrials.org TCTR20231012004; https://www.thaiclinicaltrials.org/show/TCTR20231012004

(JMIR Perioper Med 2026;9:e85045) doi:10.2196/85045

KEYWORDS

postoperative hypothermia; forced-air warming; effectiveness; postanesthesia care unit; temperature settings

Introduction

Hypothermia, defined as a core body temperature below 36 °C, is a frequent complication in patients undergoing elective surgery [1]. Intraoperative hypothermia, if uncorrected, often leads to postoperative hypothermia, particularly in the recovery room, where insufficient warming measures can exacerbate the condition. The prevalence of postoperative hypothermia in the postanesthesia care unit (PACU) has been reported to range from 20% to 28% at arrival and from 18.5% to 26% within 30 minutes after arrival [2]. Postoperative hypothermia is clinically significant, as it has been associated with impaired wound healing, increased risk of surgical site infection, greater blood loss, cardiac arrhythmias, and prolonged hospitalization [3,4]. These adverse consequences highlight the importance of effective temperature management strategies throughout the perioperative period. Recent guidelines and reviews, including Enhanced Recovery After Surgery pathways and the clinical recommendations from the Royal College of Anesthesiologists of Thailand, emphasize the critical role of maintaining normothermia as a core component to reduce surgical site infections and hospital stay [5-8].

XSL•FO
RenderX

Active warming techniques, particularly forced-air warming (FAW), are widely implemented to reduce the incidence of perioperative hypothermia. FAW devices deliver warmed air (32 °C - 47 °C) through a specialized blanket, with built-in safety mechanisms to prevent overheating [9,10]. Systematic reviews have demonstrated that FAW is superior to conventional blankets, reducing the time to restore normothermia by more than an hour [11]. While these findings confirm its effectiveness in facilitating rewarming, the literature remains inconclusive regarding the optimal temperature setting for postoperative use. Most previous studies have focused on preoperative or intraoperative warming [12-18], whereas evidence for postoperative FAW application remains limited. Xu et al [19] reported that FAW at 42 °C was more effective than at 38 °C or conventional blankets in elderly patients undergoing joint replacement. However, the generalizability of that study was restricted by the narrow patient population, short operative times, and limited assessment of adverse events.

At our institution, the prevalence of postoperative hypothermia has remained notable despite the routine availability of FAW systems. Pisitsak et al [20] documented hypothermia in 20% of patients under regional anesthesia and in 16% under general anesthesia. More recent institutional data from 2019 to 2022 indicate an incidence of 23% among surgical patients recovering in the PACU. Furthermore, between 2022 and 2024, the prevalence of hypothermia ranged from 10.8% to 13.8% despite widespread FAW use across multiple surgical specialties, including general surgery, orthopedics, otolaryngology, obstetrics and gynecology, and cardiac surgery (Department of Anesthesiology, Faculty of Medicine Ramathibodi Hospital, Mahidol University. Internal statistical data analyzed via Power BI dashboard, unpublished data, January 2025). These findings suggest that, in addition to patient- and procedure-related factors, variability in FAW temperature settings contributes to inconsistent outcomes.

Current practice in our PACUs uses FAW with adjustable temperature settings ranging from 38 °C to 43 °C; however, no standardized protocol exists to guide optimal temperature selection. This variability reflects broader uncertainty regarding the most effective strategy for postoperative rewarming and underscores the need for evidence-based guidance. To our knowledge, no prior randomized trial has evaluated a step-down temperature protocol (43 °C to 38 °C) in a mixed adult surgical population. By addressing this gap, the present study examines the effectiveness of different FAW temperature settings to inform a pragmatic and standardized PACU warming approach, with the goal of improving consistency in clinical practice and enhancing patient safety.

Methods

Study Design

This study was designed as a prospective randomized controlled trial.

Patients

A total of 132 patients scheduled for elective surgery across various specialties, including general surgery, orthopedics, urology, otolaryngology, obstetrics and gynecology, and cardiac surgery, were enrolled between April 2023 and May 2024. The inclusion criteria consisted of patients aged 18 to 80 years, American Society of Anesthesiologists (ASA) physical status I to III, who were undergoing elective procedures under either general or regional anesthesia, with an expected operating time of at least 2 hours.

Exclusion criteria included patients with a core temperature exceeding 37.5 °C, evidence of infection (eg, sepsis), conditions precluding the use of forced-air warming (eg, burns, agitation, or delirium), those unable to communicate or complete the trial questionnaire, and patients who declined participation.

Sample Size Calculation

A priori sample size calculation was conducted to ensure adequate statistical power for the study's primary outcome: the duration of forced-air warming required for a patient's core temperature to reach ≥36 °C. Based on a previous randomized controlled trial by Xu et al [19], utilizing a 2-sided significance level (α=.05), adjusted for multiple comparisons among the 3 groups (α/3=.017), corresponding to a z score of 2.41, a statistical power of 80% (z for β=0.84), an estimated SD of 6.45 minutes, and a clinically meaningful difference in rewarming time of 5 minutes, the calculation determined that 36 participants were required per group. To accommodate an anticipated 20% participant dropout rate, the sample size was prudently inflated to 44 participants for each of the 3 intervention groups. This resulted in a total sample size of 132 participants, ensuring robust statistical inference for our findings.

Randomization

Randomization was performed using stratified block randomization with proportional allocation based on the type of anesthesia (general vs regional) to ensure balanced distribution of thermoregulatory impairment mechanisms across groups. A research assistant not involved in patient recruitment generated the computer-based random sequence using permuted blocks of variable size. Allocation was concealed using sequentially numbered, sealed opaque envelopes, which were opened only after participant enrollment. The study personnel responsible for enrollment were different from those assigning participants to groups to ensure the integrity of allocation concealment.

Rewarming

Intraoperative management followed our institution's standard of care, which included the routine use of fluid warmers, application of forced-air warming blankets, and continuous core temperature monitoring for all patients. Upon arrival at the PACUs, patients who met the preliminary criteria were assessed. Only those with a core temperature lower than 36 °C were enrolled and randomly allocated into 3 groups (n=44 per group): group C (forced-air warming set to 38 °C), group F1 (forced-air warming set to 43 °C), and group F2 (forced-air warming initially set to 43 °C, then reduced to 38 °C once the core temperature reached 36 °C).

All participants received identical warming systems and core temperature monitoring devices at the PACU. Rewarming was

XSL•FO
RenderX

carried out using a forced-air warming system (Bair Hugger) with a blanket and a core temperature measurement device (Braun ThermoScan ear thermometer).

The rewarming process was monitored and recorded every 5 minutes during the active warming phase. In groups C and F1, the forced-air warmer was discontinued once the core temperature reached ≥36 °C, at which point patients were covered with a regular blanket and monitored every 10 minutes. In group F2, the setting was reduced to 38 °C upon reaching a core temperature of ≥36 °C, and patients were similarly monitored every 10 minutes until discharge from the PACU. Rewarming time was calculated as the time taken for the core temperature to rise from baseline to ≥36 °C, measured in minutes.

Outcome Measures

The primary outcome was the rewarming time, defined as the duration from the initiation of rewarming to the recovery of normothermia (core temperature ≥36 °C). Additionally, the incidence of a decrease in core temperature after achieving normothermia was recorded in each group.

Secondary outcomes included the incidence of adverse events—such as hypotension, hypertension, arrhythmias, nausea or vomiting, pain, and shivering—and patient satisfaction. Patient satisfaction was evaluated using 2 validated instruments: the 5-point Patient Comfort Scale, which measures overall comfort and satisfaction, and the 7-point Thermal Comfort Scale, which assesses subjective thermal sensation ranging from –3 (cold) to +3 (hot), with 0 representing thermal neutrality.

Data Collection

Preoperative and intraoperative data were collected, including patient demographics, surgical procedure, operative time, anesthetic technique, blood loss, and fluid and blood product administration. Upon PACU admission, core temperature was recorded every 5 minutes during the rewarming phase by trained PACU nurses. To ensure consistency, the same nurse performed all assessments for a given patient using the same device and the same ipsilateral ear. Blinding of these nurses was not feasible because the FAW device displayed temperature settings during operation; consequently, the nurses were aware of group allocation, although patients remained blinded. Once the core temperature reached ≥36 °C, measurements continued every 10 minutes until discharge based on the Modified Aldrete scoring system. Throughout the PACU stay, adverse events were monitored continuously, and patient comfort and thermal comfort scores were assessed by the nurses at the time of discharge.

Statistical Analysis

Data were analyzed using SPSS software version 27 (IBM Corp.). Continuous variables were expressed as mean (SD) or median (IQR), depending on the distribution, which was assessed using the Shapiro-Wilk test. Categorical variables were presented as counts and percentages.

For comparisons between groups, one-way ANOVA was used for normally distributed continuous variables, while the Kruskal-Wallis test was applied to non-normally distributed data. Post hoc analyses were performed using Tukey honest significant difference test for ANOVA and Dunn test for the Kruskal-Wallis test, as appropriate. Categorical variables were compared using the chi-square test or Fisher exact test, as required.

Monte Carlo simulation was utilized for the Fisher exact test extension in contingency tables larger than 2×2 where cell counts were sparse (expected count <5), ensuring robust *P* value estimation without violating asymptotic assumptions. Relative risks with 95% CI were reported for significant categorical outcomes. For continuous variables, effect sizes were expressed as Cohen *d* to ensure consistency and enhance clinical interpretability. This was an intention-to-treat analysis, and all randomized patients were analyzed in their assigned groups. A *P* value of <.05 was considered statistically significant.

Ethical Considerations

This study was approved by the Human Research Ethics Committees of Ramathibodi Hospital, Mahidol University (approval number MURA2023/202) and registered at Thaiclinicaltrials.org on March 14, 2023 (approval number TCTR20231012004). Written informed consent was obtained from all participants before enrollment. Participants' privacy and confidentiality were strictly protected, and all data were deidentified before analysis. No financial compensation was provided to participants for their participation in the study.

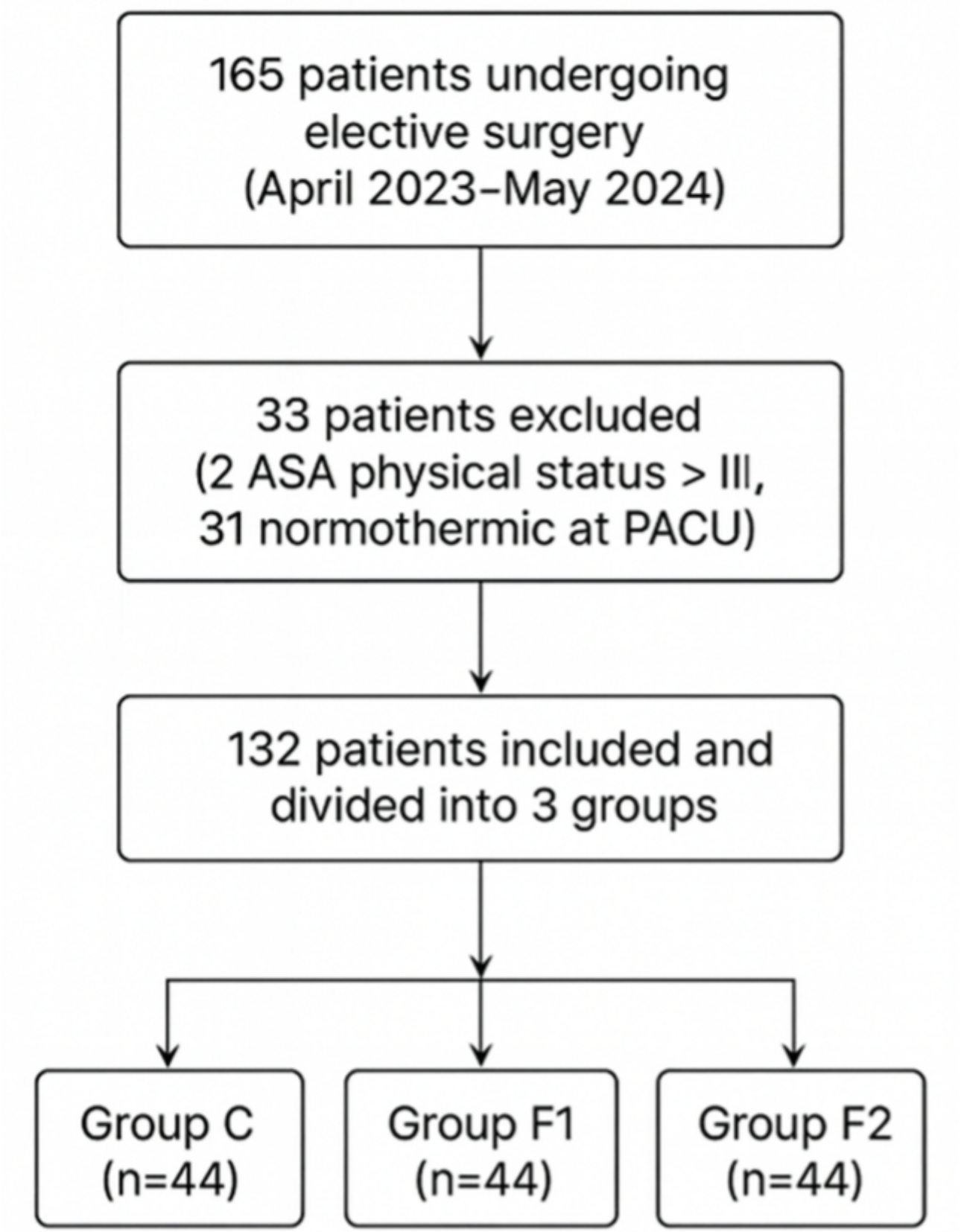
Results

Baseline Characteristics

A total of 165 patients were assessed for eligibility between April 2023 and May 2024. Thirty-three patients were excluded (2 with ASA physical status >III and 31 normothermic at PACU admission). Finally, 132 patients were included and equally divided into 3 groups (Figure 1): group C (rewarming set to 38 °C, n=44), group F1 (rewarming set to 43 °C, n=44), and group F2 (rewarming set to 43 °C until reaching a core temperature of 36 °C, then reduced to 38 °C, n=44). The baseline characteristics, including age, gender, ASA physical status, BMI, and underlying diseases, were comparable across all groups, as detailed in Table 1.

XSL•FO
RenderX

Figure 1. Flow diagram of patient enrollment. This diagram shows the screening and allocation process for all patients. ASA: American Society of Anesthesiologists; PACU: postanesthesia care unit.



XSL•FO
RenderX

Table . Baseline characteristics of patients in the 3 groups.

Characteristic	Group C (n=44)	Group F1 (n=44)	Group F2 (n=44)	*P* value
Age, n (%)				.265
Elderly (age ≥65 y)	10 (22.73)	13 (29.55)	17 (38.64)	
Nonelderly (age <65 y)	34 (77.27)	31 (70.45)	27 (61.36)	
Gender (male/female)				.083
Male	17	18	9	
Female	27	26	35	
ASA[a] Physical status, n (%)				.965
>2	13 (29.55)	13 (29.55)	14 (31.82)	
≤2	31 (70.45)	31 (70.45)	30 (68.18)	
BMI (kg/m^2), median (IQR)	23.85 (22 - 26.95)	24 (22.43 - 26.3)	23.67 (21.05 - 27)	.738
Underlying diseases, n (%)				
Diabetes mellitus	3 (6.82)	8 (18.18)	7 (15.91)	.259
Hypertension	14 (31.82)	18 (40.91)	19 (43.18)	.511
Obesity	5 (11.36)	6 (13.64)	4 (9.08)	.798
Extreme age	10 (22.73)	12 (27.27)	15 (34.09)	.490
Heart disease	3 (6.82)	4 (9.09)	5 (11.36)	.927
Cerebrovascular disease	3 (6.82)	5 (11.36)	1 (2.27)	.295
Chronic kidney disease	4 (9.09)	5 (11.36)	2 (4.55)	.622
Cancer	5 (11.36)	8 (18.18)	11 (25)	.253
Respiratory disease	2 (4.55)	3 (6.82)	2 (4.55)	>.999
Others	11 (25)	14 (31.82)	7 (15.91)	.217

[a]ASA: American Society of Anesthesiologists.

Intraoperative Data

No significant differences were observed among the 3 groups regarding operative time, type of operation, anesthetic technique, estimated blood loss, total fluid administration, total blood components used, or recorded temperatures, as summarized in Table 2.

XSL•FO
RenderX

Table . Intraoperative parameters of patients in the 3 groups.

Parameter	Group C (n=44)	Group F1 (n=44)	Group F2 (n=44)	*P* value
Operative time (min), median (IQR)	165 (120 - 221.25)	180 (123.75 - 223.75)	195 (137.5 - 233.75)	.153
Anesthetic technique, n (%)				.926
Regional	15 (34.09)	14 (31.82)	14 (31.82)	
General	22 (50)	21 (47.73)	24 (54.55)	
Combined	7 (15.91)	9 (20.45)	6 (13.64)	
Operation, n (%)				.731
Open orthopedic surgery	18 (40.91)	18 (40.91)	17 (38.64)	
Arthroscopic/laparoscopic orthopedic surgery	8 (18.18)	7 (15.91)	11 (25)	
Open gynecologic surgery	3 (6.82)	4 (9.09)	4 (9.09)	
Laparoscopic gynecologic surgery	6 (13.64)	3 (6.82)	8 (18.18)	
Breast surgery	3 (6.82)	0 (0)	2 (4.55)	
Open general surgery	2 (4.55)	2 (4.55)	1 (2.27)	
Urological surgery	1 (2.27)	1 (2.27)	0 (0)	
Thoracic surgery	1 (2.27)	1 (2.27)	0 (0)	
Plastic surgery	18 (40.91)	1 (2.27)	0 (0)	
Vascular surgery	0 (0)	2 (4.55)	0 (0)	
Laparoscopic general surgery	2 (4.55)	4 (9.09)	1 (2.27)	
Otolaryngologic surgery	0 (0)	1 (2.27)	0 (0)	
Intraoperative period				
Large estimated blood loss (≥500 ml), n (%)	1 (2.27)	3 (6.82)	1 (2.27)	.435
Total fluid administered (ml), median (IQR)	975 (762.5 - 1425)	900 (612.5 - 1387.5)	1000 (712.5 - 1487.5)	.808
Total blood component (ml), median (IQR)	0 (0)	0 (0)	0 (0)	.173
Temperature recorded (°C), mean (SD)	36.07 (0.43)	35.93 (0.37)	36.1 (0.32)	.139

Postoperative Data

Postoperative outcomes recorded in the PACU are summarized in Table 3. No significant differences were observed among the 3 groups regarding tympanic temperature upon arrival or hemodynamic parameters, including blood pressure, heart rate, respiratory rate, and SpO_2 (P>.05). However, a statistically significant difference was observed in the duration of PACU stay (P=.015). Post hoc comparisons revealed significant differences between group C versus group F2 and group F1 versus group F2, indicating a more favorable distribution of discharge times in group F2. Regarding electrocardiogram findings, adverse events were rare, with only 1 patient exhibiting bradycardia.

XSL•FO
RenderX

Table . Postoperative outcomes of patients in the 3 groups.

Parameter	Group C (n=44)	Group F1 (n=44)	Group F2 (n=44)	*P* value
Duration in PACU[a] (min),[c] median (IQR)	60 (60 - 65)	60 (60 - 60)	60 (60 - 60)	.015
Tympanic temperature <36 °C upon arrival in the PACU, n (%)				.165
<35 °C	0 (0)	1 (2.27)	3 (6.82)	
≥35 °C	44 (100)	43 (97.73)	41 (93.18)	
Systolic blood pressure (mmHg), mean (SD)	134.07 (22.57)	134.53 (18.21)	135.95 (21.86)	.910
Diastolic blood pressure (mmHg), mean (SD)	76.32 (31.85)	79.09 (12.5)	94.45 (106.34)	.341
Respiratory rate (per min), mean (SD)	17.07 (3.25)	18.09 (3.2)	17.95 (3.43)	.290
Heart rate (per min), mean (SD)	68.61 (12.12)	72.55 (13.15)	74.75 (13.31)	.081
SpO_2[b] (%), median (IQR)	100 (99 - 100)	100 (99 - 100)	100 (99 - 100)	.688

[a]PACU: postanesthesia care unit.

[b]SpO_2: peripheral capillary oxygen saturation.

[c]A statistically significant difference was observed only in the duration of stay in PACU (*P*=.015; post hoc comparisons revealed significant differences for group C vs group F2 [Cohen *d*=6.29] and group F1 vs group F2 [Cohen *d*=8.55]).

Effect of Different Rewarming Methods

Rewarming outcomes across all dimensions are summarized in Table 4. The time to achieve normothermia is illustrated in Figure 2. Figure 3 shows the number of patients who experienced a drop in core temperature below 36 °C after achieving normothermia, and those who required rewarming for more than 1 hour.

No significant differences were observed among the 3 groups in hemodynamic parameters recorded in the PACU, including systolic and diastolic blood pressure, respiratory rate, heart rate, peripheral capillary oxygen saturation, and electrocardiogram findings (*P*>.05).

While the rewarming time did not differ significantly among the groups, the incidence of patients experiencing a drop in core temperature below 36 °C after achieving normothermia was significantly lower in group F2 compared to groups C and F1 (*P*=.009, Table 4). Patients exhibiting temperature decline required extended thermal support to restore or maintain normothermia. Consequently, a significantly higher proportion of patients in groups C and F1 required active warming for more than 1 hour compared to group F2 (*P*=.017). However, this prolonged warming requirement did not lead to a clinically relevant delay in discharge, as the median duration of PACU stay remained 60 minutes across all groups (Table 3).

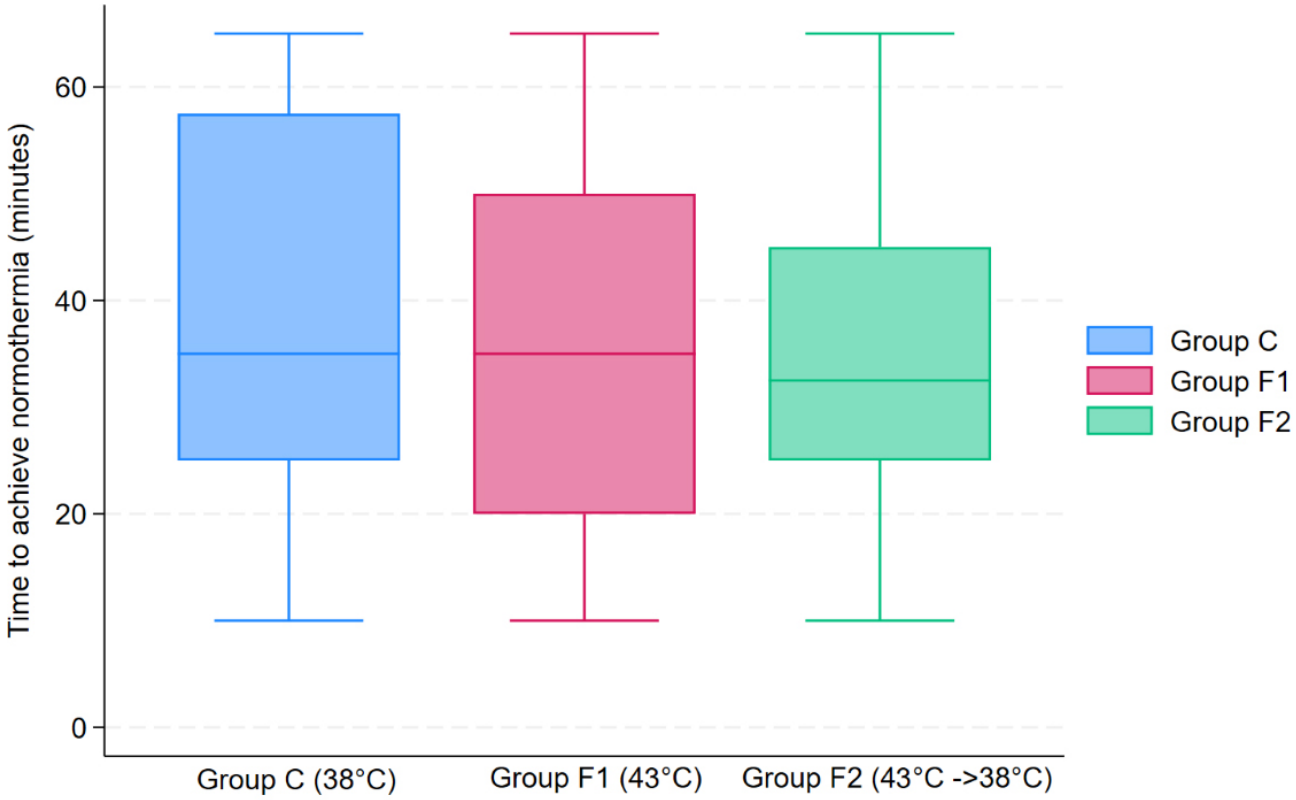
Table . Rewarming outcomes of patients in the 3 groups.

Parameter	Group C (n=44)	Group F1 (n=44)	Group F2 (n=44)	*P* value
Mean rewarming time[a] (min), mean (SD)	37.39 (16.58)	35.11 (15.64)	33.30 (13.81)	.460
Decrease temperature below 36 °C after achieving normothermia, n (%)	7 (15.91)	11 (25)	1 (2.27)	.009
Warming more than 1 h, n (%)	10 (22.73)	7 (15.91)	1 (2.27)	.017

[a]For the mean rewarming time, no significant differences were observed among groups; the mean differences (95% CI) compared to group C were −2.28 (−9.03 to 4.47) for group F1 and −4.09 (−10.46 to 2.28) for group F2. However, a statistically significant difference was observed in the proportion of patients with temperature decrease below 36 °C (*P*=.009; significant pairwise differences were observed for group F1 vs group F2 [relative risk (RR)=11.00; 95% CI 1.48 - 81.61]) and those requiring warming for more than 1 h (*P*=.017; significant pairwise differences were observed for group C vs group F2 [RR=10.00; 95% CI 1.34 - 74.84]).

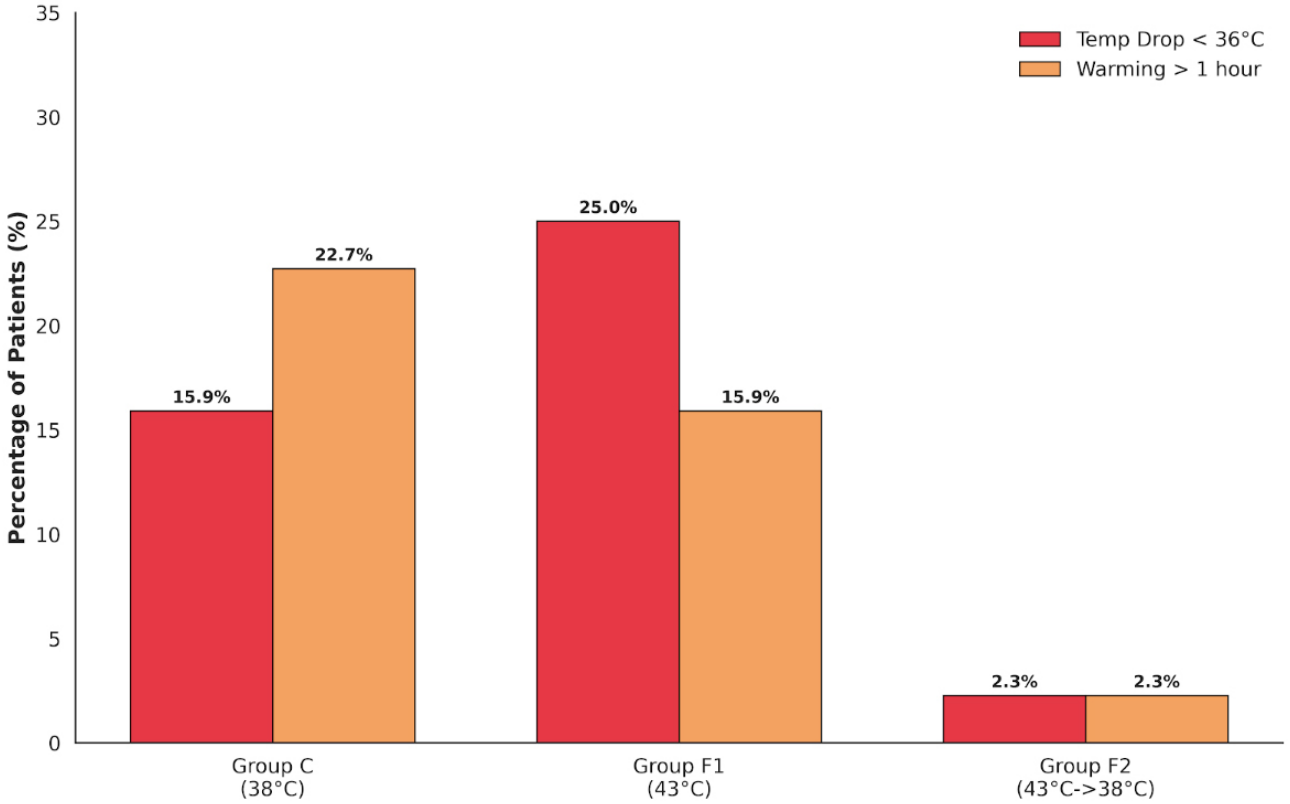
XSL•FO
RenderX

Figure 2. Comparison of rewarming outcomes among the 3 study groups. Box plot showing the distribution of time to achieve normothermia. The horizontal line within each box represents the median rewarming time. The top and bottom boundaries of the boxes indicate the IQR, and the whiskers extend to the minimum and maximum values. No statistically significant differences were observed (*P*=.460). Control group (group C): forced-air warming at 38 °C; group F1: forced-air warming at 43 °C; group F2: forced-air warming initially set at 43 °C, then reduced to 38 °C upon reaching 36 °C.



XSL•FO
RenderX

Figure 3. Comparison of rewarming outcomes among the 3 study groups. Clustered bar chart illustrating the incidence of recurrent hypothermia (core temperature dropping <36 °C after achieving normothermia) and the proportion of patients requiring active warming for more than 1 h. Group F2 demonstrated significantly lower rates for both outcomes compared to groups C and F1 (*P*=.009 and *P*=.017, respectively). Control group (group C): forced-air warming at 38 °C; group F1: forced-air warming at 43 °C; group F2: forced-air warming initially set at 43 °C, then reduced to 38 °C upon reaching 36 °C.



Adverse Events

In terms of postoperative adverse events, no significant differences were observed among the 3 groups. Pain was the most frequently reported complication, affecting 22.7% (10/44) of patients in group C and 18.2% (8/44) in both group F1 and group F2 (*P*=.826). Nausea and vomiting occurred infrequently, with an incidence ranging from 2.3%(1/44) to 4.6%(2/44) across the groups (*P*>.999). Shivering was reported in 4.6% (2/44) of patients in group C, 6.8% (3/44) in group F1, and 2.3% (1/44) in group F2 (*P*=.871). Hemodynamic events were rare, comprising 1 case of hypertension in group C, no events in group F1, and 1 case each of hypotension and hypertension in group F2 (all *P*>.999). No patients experienced arrhythmia or other adverse effects. Overall, the incidence of postoperative complications was low and comparable among the groups, supporting the safety of forced-air warming across different temperature settings.

Patient comfort, as measured by satisfaction levels, showed a significant difference among the groups (*P*=.049), along with the average comfort scores (*P*=.039). The proportion of patients reporting being “very much satisfied” was 27.27% (12/44) in group C, 43.18% (19/44) in group F1, and 52.27% (23/44) in group F2 (Table 5). However, there was no significant difference in the thermal comfort scale among the groups (*P*=.131).

XSL•FO
RenderX

Table . Patient satisfaction in the 3 groups.

Parameter	Group C (n=44)	Group F1 (n=44)	Group F2 (n=44)	*P* value
Patient's comfort, n (%)				.049
Very much satisfied	12 (27.27)	19 (43.18)	23 (52.27)	
Somewhat satisfied	29 (65.91)	21 (47.73)	21 (47.73)	
Undecided	2 (4.55)	4 (9.09)	0 (0)	
Not really satisfied	0 (0)	0 (0)	0 (0)	
Not at all satisfied	1 (2.27)	0 (0)	0 (0)	
Patient's comfort (average score), median (IQR)	4 (4-5)	4 (4-5)	5 (5-5)	.039
Thermal comfort scale, n (%)				.131
Hot	1 (2.27)	4 (9.09)	2 (4.55)	
Warm	34 (77.27)	32 (72.73)	31 (70.45)	
Slightly warm	6 (13.64)	2 (4.55)	9 (20.45)	
Neutral	2 (4.55)	6 (13.64)	2 (4.55)	
Slightly cold	1 (2.27)	0 (0)	0 (0)	
Cool	0 (0)	0 (0)	0 (0)	
Cold	0 (0)	0 (0)	0 (0)	
Thermal comfort scale (average score), median (IQR)	2.00(2-2)	2.00(2-2)	2.00 (1.25 - 2)	.676

Discussion

Principal Findings

Postoperative hypothermia is a frequent complication of both general and regional anesthesia, primarily resulting from thermoregulatory impairment and internal heat redistribution [21]. While previous research identified FAW at 42 °C as effective for elderly patients [19], evidence regarding the optimal temperature setting for the general surgical population remains limited. Consequently, this trial aimed to evaluate the most effective and efficient rewarming protocol for patients undergoing various surgical procedures.

In this study, hypothermia was defined as a core temperature <36 °C upon PACU admission. Tympanic thermometry was selected over invasive nasopharyngeal or rectal probes, as used in previous studies [19,22], to prioritize patient comfort during the awake recovery phase.

The principal finding of this study is that increasing the FAW setting from 38 °C to 43 °C did not yield a statistically significant reduction in the overall rewarming time to normothermia. While group F2 achieved the target temperature approximately 3 to 4 minutes faster than the control group, the precision estimates provided by the 95% CI suggest that this difference is negligible. Given that all groups achieved normothermia within a comparable timeframe, the variation in rewarming speed appears to lack clinical relevance for PACU throughput.

Several physiological factors may explain why higher settings did not produce faster rewarming, a finding that contrasts with some previous studies [19]. Peripheral vasoconstriction can limit the rate of convective heat transfer from the skin to the core, creating a "plateau effect" regardless of the external heat gradient provided by higher FAW settings. Moreover, as the core temperature approaches the normal thermoregulatory threshold, the body initiates vasodilation to redistribute heat, preventing a linear increase in core temperature [21,23]. Device-specific factors, such as automatic safety regulation at higher settings or variability in blanket positioning, may have further minimized the actual difference in heat delivery.

Although rewarming rates were comparable, the group F2 protocol demonstrated superior thermal stability. Unlike group C, which exhibited a significantly higher incidence of prolonged rewarming, the step-down protocol (group F2) effectively minimized the number of "outliers"—patients requiring extended care due to thermal instability. The prolonged rewarming observed in group C is likely due to several physiological and thermodynamic factors. At a lower temperature (38 °C), the gradient between the patient's core temperature and the surrounding warming environment is reduced, leading to a slower rate of heat transfer [24]. Additionally, peripheral vasoconstriction limits blood flow to the skin and extremities, impeding the transport of externally applied heat to the core [25]. Furthermore, the reduced metabolic rate associated with hypothermia decreases endogenous heat generation, collectively contributing to the extended recovery time [21]. Consequently, the requirement for prolonged active warming in groups C and F1 likely contributed to the statistical difference observed in the total duration of PACU stay (*P*=.015). Although the median stay was consistent at 60 minutes across all groups, the distribution of discharge times suggests that while

XSL•FO
RenderX

group F2 may not shorten the mandatory minimum recovery time, it optimizes unit throughput by reducing the incidence of prolonged stays.

Clinical Implications

Regarding safety and comfort, the incidence of adverse events—including pain, nausea, vomiting, hemodynamic changes, and shivering—did not differ significantly among the groups, and no severe adverse events were observed. These results align with the safety profile reported by Xu et al [19] However, regarding patient experience, group F2 reported higher satisfaction scores related to comfort during rewarming compared to groups C and F1. This suggests that an initial high-temperature setting effectively enhances thermal comfort, while the subsequent reduction prevents the discomfort associated with overheating.

Intraoperative factors—including ambient cooling, fluid administration, and anesthesia-induced thermoregulatory impairment—are known to significantly impact rewarming. In this study, potential confounding was minimized through a standardized intraoperative care protocol that included routine fluid and forced-air warming. Furthermore, randomization successfully balanced these physiological stressors across study arms; as shown in Table 2, there were no significant differences in operative duration, total fluid volume, or anesthetic technique. Consequently, the observed differences in PACU outcomes can be primarily attributed to the specific postoperative warming protocols rather than intraoperative disparities.

Strengths and Limitations

This study has notable strengths and limitations. A key strength is the rigorous randomization and standardized intraoperative care, which successfully balanced potential confounders such as operative duration and fluid volume across study arms. However, several limitations exist. First, strict environmental control of the PACU was challenging due to the open-plan nature of the unit. The ambient temperature fluctuated between 22 and 24 °C, which serves as a potential environmental confounder influencing convective heat loss. Nevertheless, this variation reflects real-world clinical conditions, potentially enhancing the ecological validity of our results. Second, regarding measurement, reliance on tympanic thermometry introduces inherent variability compared to the gold standard of invasive core monitoring. We acknowledge that readings can be affected by factors such as probe positioning, cerumen obstruction, and post-anesthetic peripheral vasoconstriction. To mitigate these inaccuracies, we strictly standardized the technique by using the same device and assessing the ipsilateral ear throughout the study, aiming to balance measurement precision with patient comfort in the awake state. Third, data collection involved intermittent recordings at 5-minute intervals rather than continuous electronic monitoring. While this frequency is clinically practical, it may lack the temporal resolution to capture rapid, transient temperature fluctuations during the active rewarming phase, potentially masking the true extent of thermal variability. Finally, we did not perform a formal cost-effectiveness analysis. Although the step-down protocol (group F2) showed potential for optimizing PACU throughput, future studies including economic evaluations are needed to confirm the financial implications of these warming strategies.

Conclusions

In summary, although varying temperature settings of forced-air warming systems produced comparable rewarming times, the protocol involving an initial setting of 43 °C followed by a reduction to 38 °C (group F2) was associated with superior maintenance of normothermia and a significantly lower incidence of postoperative hypothermia recurrence. These findings underscore the potential benefits of implementing optimized warming protocols to enhance patient outcomes in the PACU.

Acknowledgments

The authors would like to thank all patients who willingly participated in this study, without whom the research would not have been possible. We gratefully acknowledge the contribution of Ms. Rojnarin Komonhirun, M.Ed., for her valuable assistance with data analysis. We also thank the PACU staff of Ramathibodi Hospital for their support in facilitating the data collection process.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

Due to privacy concerns, the datasets generated or analyzed during this study are not publicly available. However, deidentified data may be provided by the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: KP (lead), SS (equal), RS (supporting), WW (supporting)
Data curation: KP (lead), SS (equal)
Formal analysis: KP (lead), SS (equal)
Investigation: KP (lead), SS (equal)
Methodology: KP (lead), SS (equal)

XSL•FO
RenderX

Resources: KP (lead), SS (equal), WW (supporting)
Supervision: KP (lead), RS (supporting)
Validation: KP (lead), SS (equal)
Visualization: KP (lead), SS (equal), RS (supporting), WW (supporting)
Writing – original draft: KP (lead), SS (equal)
Writing – review & editing: KP (lead), SS (equal), RS (supporting), WW (supporting)

Conflicts of Interest

None declared.

References

1. Shenoy L, Krishna HM, Kalyan N, Prasad KH. A prospective comparative study between prewarming and cowarming to prevent intraoperative hypothermia. J Anaesthesiol Clin Pharmacol 2019;35(2):231-235. [doi: 10.4103/joacp.JOACP_353_17] [Medline: 31303714]
2. Ashoobi MT, Shakiba M, Keshavarzmotamed A, Ashraf A. Prevalence of postoperative hypothermia in the post-anesthesia care unit. Anesth Pain Med 2023 Oct;13(5):e136730. [doi: 10.5812/aapm-136730] [Medline: 38028114]
3. Torossian A, Bräuer A, Höcker J, Bein B, Wulf H, Horn EP. Preventing inadvertent perioperative hypothermia. Dtsch Arztebl Int 2015 Mar 6;112(10):166-172. [doi: 10.3238/arztebl.2015.0166] [Medline: 25837741]
4. Shaw CA, Steelman VM, DeBerg J, Schweizer ML. Effectiveness of active and passive warming for the prevention of inadvertent hypothermia in patients receiving neuraxial anesthesia: a systematic review and meta-analysis of randomized controlled trials. J Clin Anesth 2017 May;38:93-104. [doi: 10.1016/j.jclinane.2017.01.005] [Medline: 28372696]
5. Rauch S, Miller C, Bräuer A, Wallner B, Bock M, Paal P. Perioperative hypothermia - a narrative review. Int J Environ Res Public Health 2021 Aug 19;18(16):8749. [doi: 10.3390/ijerph18168749] [Medline: 34444504]
6. Munday J, Duff J, Wood FM, Sturgess D, Ralph N, Ramis MA. Perioperative hypothermia prevention: development of simple principles and practice recommendations using a multidisciplinary consensus-based approach. BMJ Open 2023 Nov 14;13(11):e077472. [doi: 10.1136/bmjopen-2023-077472] [Medline: 37963694]
7. Grant MC, Engelman DT. Enhanced recovery after surgery: overarching themes of the ERAS® Society Guidelines & Consensus Statements for Adult Specialty Surgery. Perioper Med (Lond) 2025;14(1):120. [doi: 10.1186/s13741-025-00590-0] [Medline: 41168801]
8. Wongsripuemtet P, Leelanokrom R, Tangwiwat S, Vorapaluk P, Karnjanarachata C, Karnjanawanichkul O, et al. Clinical guidelines for prevention and management of perioperative hypothermia by the Royal College of Anesthesiologists of Thailand. Thai J Anesthesiol 2024;50:287-300 [FREE Full text]
9. Lynch S, Dixon J, Leary D. Reducing the risk of unplanned perioperative hypothermia. AORN J 2010 Nov;92(5):553-562. [doi: 10.1016/j.aorn.2010.06.015] [Medline: 21040819]
10. Bräuer A, Quintel M. Forced-air warming: technology, physical background and practical aspects. Curr Opin Anaesthesiol 2009 Dec;22(6):769-774. [doi: 10.1097/ACO.0b013e328331d134] [Medline: 19734783]
11. Warttig S, Alderson P, Campbell G, Smith AF. Interventions for treating inadvertent postoperative hypothermia. Cochrane Database Syst Rev 2014 Nov 20;2014(11):CD009892. [doi: 10.1002/14651858.CD009892.pub2] [Medline: 25411963]
12. Pruitthitada S, Sinchai S. Effective of preoperative force air warming on perioperative hypothermia in patients undergoing caesarean section under spinal anesthesia. Thai J Anesthesiol 2022;48(2):87-98 [FREE Full text]
13. Becerra Á, Valencia L, Villar J, Rodríguez-Pérez A. Short-periods of pre-warming in laparoscopic surgery. A non-randomized clinical trial evaluating current clinical practice. J Clin Med 2021 Mar 3;10(5):1047. [doi: 10.3390/jcm10051047] [Medline: 33802512]
14. Jun JH, Chung MH, Kim EM, et al. Effect of pre-warming on perioperative hypothermia during holmium laser enucleation of the prostate under spinal anesthesia: a prospective randomized controlled trial. BMC Anesthesiol 2018 Dec 22;18(1):201. [doi: 10.1186/s12871-018-0668-4] [Medline: 30579334]
15. Yoo JH, Ok SY, Kim SH, et al. Effects of 10-min of pre-warming on inadvertent perioperative hypothermia in intraoperative warming patients: a randomized controlled trial. Anesth Pain Med (Seoul) 2020 Jul 31;15(3):356-364. [doi: 10.17085/apm.20027] [Medline: 33329836]
16. Parameeanon M, Poomepraman W, Sarachai W. Effect of pre-warming by forced air warmer during total hip and knee arthroplasties for perioperative and postoperative hypothermia. Thai J Anesthesiol 2020;46(2):88-95 [FREE Full text]
17. Zhang R, Chen X, Xiao Y. The effects of a forced-air warming system plus electric blanket for elderly patients undergoing transurethral resection of the prostate: a randomized controlled trial. Medicine (Baltimore) 2018 Nov;97(45):e13119. [doi: 10.1097/MD.0000000000013119] [Medline: 30407328]
18. Tyvold SS. Preventing hypothermia in outpatient plastic surgery by self-warming or forced-air-warming blanket: a randomised controlled trial. Eur J Anaesthesiol 2019 Nov;36(11):843-850. [doi: 10.1097/EJA.0000000000001087] [Medline: 31567576]
19. Xu H, Xu G, Ren C, Liu L, Wei L. Effect of forced-air warming system in prevention of postoperative hypothermia in elderly patients: a prospective controlled trial. Medicine (Baltimore) 2019;98(22):e15895. [doi: 10.1097/MD.0000000000015895]

XSL•FO
RenderX

20. Pisitsak C, Virankabutra T, Deewong K, Pornprasertsuk M. Postoperative hypothermia: incidence and associated factors. Thai J Anesthesiol 2011;37:93-103.
21. Sessler DI. Temperature monitoring and perioperative thermoregulation. Anesthesiology 2008 Aug;109(2):318-338. [doi: 10.1097/ALN.0b013e31817f6d76] [Medline: 18648241]
22. Liu X, Shi Y, Ren C, Li X, Zhang Z. Effect of an electric blanket plus a forced-air warming system for children with postoperative hypothermia: a randomized controlled trial. Medicine (Baltimore) 2017;96(26):e7389. [doi: 10.1097/MD.0000000000007389]
23. Sessler DI. Complications and treatment of mild hypothermia. Anesthesiology 2001 Aug;95(2):531-543. [doi: 10.1097/00000542-200108000-00040] [Medline: 11506130]
24. Taylor EE, Carroll JP, Lovitt MA, et al. Active intravascular rewarming for hypothermia associated with traumatic injury: early experience with a new technique. Proc (Bayl Univ Med Cent) 2008 Apr;21(2):120-126. [doi: 10.1080/08998280.2008.11928375] [Medline: 18382749]
25. Ereth MH, Lennon RL, Sessler DI. Limited heat transfer between thermal compartments during rewarming in vasoconstricted patients. Aviat Space Environ Med 1992 Dec;63(12):1065-1069. [Medline: 1456917]

Abbreviations

ASA: American Society of Anesthesiologists
FAW: forced-air warming
PACU: postanesthesia care unit

Edited by J Shiffermiller; submitted 10.Oct.2025; peer-reviewed by P Sawaddiruk; revised version received 10.Dec.2025; accepted 25.Dec.2025; published 28.Jan.2026.

Please cite as:
Pasutharnchat K, Seangrung R, Sirisophaphong S, Wongkum W
Forced-Air Warming Temperature Settings for Treating Postoperative Hypothermia in the Postanesthesia Care Unit: Randomized Controlled Trial
JMIR Perioper Med 2026;9:e85045
URL: https://periop.jmir.org/2026/1/e85045
doi:10.2196/85045

© Koravee Pasutharnchat, Rattaphol Seangrung, Sirikarn Sirisophaphong, Wilailuck Wongkum. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Virtual Reality for the Management of Postoperative Pain and Anxiety in Children and Adolescents Undergoing Nuss Repair of Pectus Excavatum: Randomized Controlled Trial

Charlotte M Walter[1], MD; Dillon Froass[2], BS; Nora Bell[3], MD; Lauren Haack[4], MD; Chloe Boehmer[1], MA; Claudia Bruguera Torres[1], MD; Rachel Spivak[1], BA; Max Chou[5], MD; Kristie Geisler[1], BS; Keith O'Conor[6], MD; Sara E Williams[7], PhD; Lili Ding[1], PhD; Christopher D King[1], PhD; Vanessa A Olbrecht[8,9], MBA, MD

[1]Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave, Cincinnati, OH, United States

[2]The Ohio State University Wexner Medical Center, Columbus, OH, United States

[3]Department of Pediatrics, Vanderbilt University, Nashville, TN, United States

[4]Department of Surgery, Wayne State University, Detroit, MI, United States

[5]Department of Anesthesiology, University of Cincinnati Medical Center, Cincinnati, OH, United States

[6]Department of Anesthesiology, Penn State Milton S. Hershey Medical Center, Hershey, PA, United States

[7]Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University, Palo Alto, CA, United States

[8]Department of Anesthesiology, Nemours Children’s Hospital, Wilmington, DE, United States

[9]Sidney Kimmel Medical College, Thomas Jefferson University Hospital, Philadelphia, PA, United States

Corresponding Author:
Charlotte M Walter, MD
Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave, Cincinnati, OH, United States

Abstract

Background: Virtual reality (VR) is a novel technology with implications for pain and sensory processing. VR may serve as a novel, scalable method to deliver clinically validated therapy for pain management as an alternative or adjunct to opioids for acute pain. Given that psychological factors and pain perception are both components of postoperative pain, it may also be beneficial to incorporate modalities that decrease anxiety, such as active relaxation and guided meditation with VR. Unfortunately, these therapies are not widely available due to multiple barriers. VR has the potential to deliver pain-reducing, psychologically based therapy to children, thereby enhancing multimodal analgesia and potentially decreasing opioid use. This study investigates the role of VR in reducing pain and anxiety after surgery. Given the substantial risks associated with opioid use, particularly in younger populations, alternative pain management strategies are crucial.

Objective: The primary aim of this study was to evaluate the efficacy of VR as a nonpharmacological intervention for managing postoperative pain intensity, pain unpleasantness, anxiety, and opioid use in children and adolescents undergoing Nuss repair of pectus excavatum.

Methods: A single-center, prospective, randomized, controlled trial was conducted at a tertiary care children’s hospital and research center. Ninety children and adolescents (8-18 y) undergoing the Nuss procedure were randomized to guided relaxation or mindfulness VR (n=30) and distraction-based gaming VR (n=30), combined to form the VR group (n=60), and a control group using a passive 360° video (n=30). Patients received a 10-minute session on postoperative days 1 and 2. Pain intensity, pain unpleasantness, and anxiety were evaluated before and 0-, 15-, and 30-minute post-session. In-hospital pain scores, anxiety scores, and opioid use were collected.

Results: Children and adolescents who participated in VR reported a significantly greater decrease in pain intensity from baseline (0.41, SE 0.23) compared with those in the 360° video group at 30 minutes (P=.04) before multiplicity adjustment but not after multiplicity adjustment. There were no significant differences in pain scores or opioid use between the VR and control groups on postoperative day 1 or 2, nor were there changes in pain unpleasantness or anxiety at any time after the intervention.

Conclusions: Daily, 10-minute VR sessions provided some trends toward transient analgesic and anxiolytic effects, albeit none that were statistically significant. VR did not significantly decrease overall pain scores or opioid usage, possibly due to the limited intervention duration and high standardized opioid use. Future studies should investigate extended and more frequent VR sessions and the integration of VR with other therapeutic modalities.

Trial Registration: ClinicalTrials.gov NCT04351776; https://clinicaltrials.gov/study/NCT04351776

(JMIR Perioper Med 2026;9:e80902) doi:10.2196/80902

XSL•FO
RenderX

KEYWORDS

analgesia; anxiety; distraction-based virtual reality; pediatric anesthesia; prospective studies; child; human; virtual reality; pediatric pain medicine; acute postoperative pain

Introduction

Background

Multimodal pain management techniques for acute postoperative pain are commonly studied and utilized [1]; opioids continue to be the cornerstone of postoperative pain management. Opioid misuse continues to be a major public health issue in the United States, with children and adolescents particularly vulnerable, as many are initially exposed to opioids prescribed for pain management [2-4]. Furthermore, the risk of future opioid overdose significantly increases with the quantity of pills prescribed; adolescents receiving 30 or more pills have a 35% higher rate of overdose than those prescribed 18 or fewer pills [5]. The prescription of opioid analgesics is a well-documented pathway to misuse, opioid use disorder, and overdose [6].

The Nuss procedure, performed to repair pectus excavatum, is associated with severe postoperative pain [7]. Effective management of postoperative pain after this surgery is crucial, as alleviating pain can enhance patient satisfaction and reduce complication rates [8]. Effective pain management techniques and regimens vary across pediatric institutions and have begun including intercostal nerve cryoablation [8,9]. Opioid use during recovery from the Nuss procedure is common, with one study finding that patients used opioids for a median of 8 days with an IQR of 6 - 10 days [10]. Given the absence of standardized postoperative pain management protocols and the high usage of opioids following the Nuss procedure, it is essential to explore nonpharmacologic pain control adjuncts for these patients.

Virtual reality (VR) technology provides an immersive, multisensory, and 3D environment that enables individuals to experience a modified reality, creating a sense of “presence” for each individual [11]. There is a clear need for alternative pain management methods, including nonpharmacologic techniques. VR has been shown to be effective in reducing perioperative and postoperative anxiety in pediatric patients. Studies show significant reductions in anxiety in pediatric patients immediately after distraction-based gaming virtual reality (VR-D) sessions, with some effects lasting for at least 15 minutes post-intervention [12,13]. Two approaches—VR-D and guided relaxation–based virtual reality (VR-GR)—are being researched for their effectiveness in reducing pain and anxiety following surgery.

VR-D immerses patients in engaging experiences that help divert attention from pain or anxiety, providing effective short-term relief. Integration of these techniques is challenging in the perioperative space, with limited providers and resources and high costs limiting its feasibility. VR can be used anywhere, anytime, with access to a headset.

Gate control theory suggests that distraction can be a valuable tool for pain management, as attentional load is fixed, and distraction toward a pleasant experience means less attention to pain [14,15]. It has been associated with immediate and short-term reductions in postoperative pain intensity and unpleasantness. VR-D techniques have been shown to decrease acute pain in children and adults [16-18]. Single sessions of VR-D have been shown to reduce postoperative pain for up to 30 minutes in some cases, regardless of baseline pain catastrophizing levels, suggesting broad applicability across pediatric populations experiencing postoperative pain [19]. The use of VR-D has also demonstrated pain reduction comparable to opioid use in burn injury patients during wound cleaning [20]. While VR-D is particularly effective for short-term pain management, additional strategies like guided relaxation may be needed for longer-lasting pain relief [21].

VR-GR seeks to provide more sustained pain relief by combining distraction with mind-body techniques, such as guided relaxation or mindfulness within the VR environment. Psychological factors—including calmness, fear, anxiety, and depression—affect the subjective experience of pain [22]. Resilience has been negatively associated with pain unpleasantness, potentially serving as a protective factor in patients with higher baseline anxiety [22,23]. Incorporating active relaxation and guided meditation techniques may significantly contribute to pain reduction. This combination of settling the mind to increase resilience and distraction from acute pain may play a significant role in acute pain reduction [21]. VR-GR may further improve anxiety reduction, especially in children with higher anxiety sensitivity [12,24]. Although VR-GR may offer additional benefits for sustained pain relief compared to distraction alone, its effects were also primarily transient [21]. Using VR to perform guided relaxation could expand the benefits of these nonpharmacological pain management techniques to more children, including those having surgery.

Overall, VR is a promising nonpharmacologic tool for managing postoperative pain and anxiety in children and adolescents. It can potentially enhance the perioperative experience, reduce reliance on pharmacological interventions, and increase patient and family satisfaction. However, randomized controlled trials are needed to establish standardized protocols and explore VR integration with other therapies, such as biofeedback, for more durable outcomes [13,21,25].

Aim

In this prospective, randomized, controlled clinical trial, we compare the short-term efficacy of immersive VR in decreasing acute postoperative pain (primary outcome), anxiety, and opioid consumption following pectus excavatum repair. We hypothesize that the use of VR will be more effective at reducing pain, anxiety, and opioid use as compared to the control group in this patient population.

Methods

The original protocol for this study has been published [26].

XSL•FO
RenderX

Study Design and Setting

This single-center, randomized, unblinded clinical trial was conducted at Cincinnati Children's Hospital Medical Center (CCHMC), a 670-bed tertiary care academic children's hospital. The recruitment began on July 10, 2020, and the study was completed on July 30, 2023. The COVID-19 pandemic delayed study completion. We recruited children and adolescents undergoing Nuss repair of pectus excavatum to investigate the role of VR in the management of postoperative pain and anxiety.

Ethical Considerations

This study complies with the Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) statement [27] and the Consolidated Standard of Reporting Trials (CONSORT) statement [28]. The CCHMC Institutional Review Board approved this study (IRB 2019 - 1090) on November 26, 2019, and it was conducted per the rules and regulations for ethical research. This study was registered at ClinicalTrials.gov on April 3, 2020 (NCT04351776). Written informed parental consent and patient assent (children >11 years) were obtained from all participants before enrollment into this study. Patients received a small stipend for participation. All identifying patient information was kept private and confidential.

Patients and Recruitment

Patients

This study recruited 90 patients (30 patients per group), ages 8 to 18 years, undergoing Nuss repair of pectus excavatum surgery. Informed parental consent and patient assent were obtained before enrollment into this study.

The inclusion criteria were as follows: patients were (1) between the ages of 8 and 18 years, (2) able to read, understand, and speak English, (3) presenting for Nuss repair of pectus excavatum, and (4) followed by the acute pain service following surgery.

The exclusion criteria were as follows: patients with (1) a history of developmental delay, uncontrolled psychiatric conditions, or neurological conditions, (2) a history of seizures, epilepsy, vertigo, or significant motion sickness/nausea/vomiting, or (3) any condition that would preclude the application of the VR headset, such as craniofacial abnormalities.

Recruitment

Approximately 150 Nuss repair surgeries are performed at CCHMC each year. Therefore, our recruitment target of 90 patients was well within achievable limits. During the study, patients who underwent Nuss repair were recruited continuously until we met the targeted enrollment. The operating room schedule and surgical patient list were reviewed for potentially eligible patients, who were approached for recruitment before surgery. If patients wished to participate, consent (and assent for patients >11 years of age) was obtained, and eligibility criteria were verified. We recruited about 2 patients per week. Recruiting stopped during the COVID-19 pandemic, when elective surgeries were not performed, delaying study completion.

Randomization

Potential patients were identified using the operating room schedule and the pectus surgery list provided by the surgery team. Eligible participants were randomized (1:1:1) into three groups: active distraction-based guided relaxation virtual reality (VR-DGR, n=30) and active VR-D (n=30)—collectively the VR group (n=60)—and a control group—passive 360° video (360-V) without instructions, sound, guided relaxation, or active patient involvement (n=30).

VR Technology

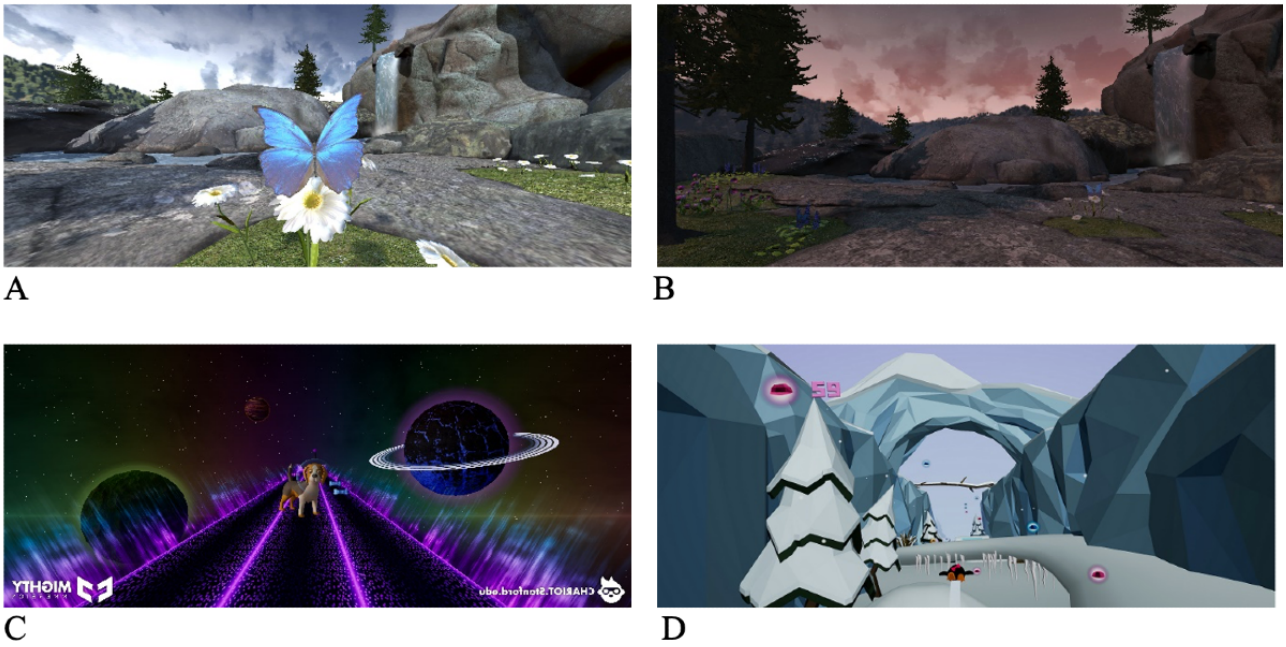
All participants used a Starlight Xperience VR all-in-one device and software developed specifically for hospital settings. This technology is a customized version of the Lenovo Mirage Solo with a Daydream VR headset. It is easy to disinfect to comply with hospital infection safety protocols. Importantly, an integrated headphone device provides audio content, and the patients use head movements and a handheld controller for interaction and navigation. It is commercially available (not Food and Drug Administration–regulated) and was supplied by the Starlight Children's Foundation.

VR-DGR and 360-V participants used the Mindful Aurora application, developed by the Stanford University Childhood Anxiety Reduction through Innovation and Technology program, to deliver relaxation/mindfulness content, which presents a relaxing nature scene with prompts instructing patients to actively slow and pace their breathing in conjunction with the movement of objects in the VR environment. 360°-V participants experienced the same relaxing nature scene without guided relaxation prompts; the 360°-V group also did not receive any audio and thus did not experience an immersive environment.

VR-D participants had the option to choose and play one of the three games: Space Pups, in which the participant controls a puppy in space and collects treats to music; Pebbles the Penguin, in which the participant controls a penguin sliding down a mountain to collect pebbles; or Wonderglade, in which the participant can play five different mini-carnival games (Figure 1).

XSL•FO
RenderX

Figure 1. Scenes from the Mindful Aurora application used in distraction-based guided relaxation virtual reality (VR-DGR) and 360° video (360-V) (A and B), and scenes from Space Pups (C) and Pebbles the Penguin (D) used in distraction-based gaming virtual reality (VR-D).



Procedure

Consent and assent were obtained before the visit. Patient characteristics, demographics, weight, and pain scores were collected preoperatively. All patients enrolled in this study received standard postoperative care via the CCHMC Pectus Surgery Pain Management Protocol, which standardizes all medications received by all pectus patients. This includes non-opioid pain medication such as pregabalin, acetaminophen, ketorolac, methocarbamol, and diazepam. All participants received the same non-opioid medications. Before the first session, patients completed the Childhood Anxiety Sensitivity Index to establish baseline anxiety levels and the Pain Catastrophizing Scale (PCS) for children. Patients were visited daily for one 10-minute session. Every effort was made to ensure the consistent timing of the visits for all patients. Sessions were completed beginning on postoperative day (POD) 1, then daily until the day of discharge, or until POD 3.

Patients were trained to use the technology before the first VR session. All participants received a device tutorial that taught them how to use the device and introduced them to the VR software. Patients received a script about VR-GR, VR-D, or 360-V, depending on the group to which they are assigned. During each session, patients completed a 10-minute session of either VR-GR, VR-D, or 360-V, per assigned group. Patients were asked to rate their pain intensity, pain unpleasantness, and anxiety via Numerical Rating Scale (NRS), before, immediately after, and 15 and 30 minutes after each session. Pain and anxiety scores and opioid use/day were recorded in REDCap (Research Electronic Data Capture).

Data Collection

The primary outcome measure was pain intensity (NRS), measured before, immediately after, and 15 and 30 minutes following each session on POD1 and POD2. Secondary outcomes included opioid use on POD0, POD1, and POD2 and pain area under the curve (AUC) on POD1 and POD2, pain unpleasantness, and anxiety scores before and 0, 15, and 30 minutes after each session on POD1 and POD2 to establish change from baseline.

For each eligible participant, data were collected from their patient history/interview and the electronic medical record in a standardized case report form in the REDCap system. Inpatient opioid use was identified from the patient's electronic medical record based on documentation in the medication administration record and transferred to REDCap. All opioid quantities were translated to morphine milligram equivalents (MME) and summed to determine total morphine equivalents per 24-hour period (midnight-to-midnight) during the patients' inpatient stay. Measures used in the study are summarized in Table 1.

XSL•FO
RenderX

Table . Scales and questionnaires used in the study.

Scales and questionnaires	Definition
Pain intensity and pain unpleasantness	
Numerical Rating Scale (NRS) [29]	The NRS is the most common validated self-report measure of pain intensity and pain unpleasantness. It involves verbally asking for an estimate of pain intensity using numbers from 0 (no pain) to 10 (maximal pain). Pain was described as being like listening to music; pain intensity is the volume, and pain unpleasantness is how much the music is disliked [30]. It requires no equipment to administer or score.
Pain intensity across all postoperative days	Area under the time-pain score (NRS) curve using the trapezoidal rule (pain AUC[a]) measured pain intensity across all postoperative days 1 and 2.
Anxiety	
Pain Catastrophizing Scale for Children (PCS-C) [31]	A validated 13-item questionnaire (each rated on a 5-point scale, 0-4) designed to measure pain catastrophizing in children of age 8 - 17 years. It is adapted from the adult version and assesses three key aspects of pain-related negative thinking: rumination, magnification, and helplessness.
Child Anxiety Sensitivity Index (CASI) [32]	A validated 18-item survey that measures perceived anxiety symptoms. Participants respond to each item on a 3-point scale (eg, “none,” “some,” and “a lot”). The total score is calculated by summing the responses, with higher scores indicating greater anxiety sensitivity. The total scores range from 18 to 54. CASI has been used in VR[b] studies in adolescents of age 10 - 21 years [33].
Numerical Rating Scale-Anxiety (NRS-A) [34]	A validated self-report numeric 0 - 10 anxiety scale that is easy to administer to children. The NRS-A is easy to administer and can be used quickly to assess anxiety levels.
Opioid use	
NIH morphine milligram equivalents (MME) per day [35,36]	Standardizes a metric for quantifying and comparing doses of different opioids. However, MMEs serve as a common metric for comparing different opioids.

[a]AUC: area under the curve.

[b]VR: virtual reality.

Statistical Analysis

Sample Size Calculation

Sample size was based on the feasibility of conducting this clinical study and unpublished preliminary data that assessed the impact of a single VR-D session on pain intensity in children and adolescents after surgery, with a goal of 80% power to detect differences in mean changes of 1 between VR and 360-V (given pilot data which showed average change in pain intensity of −1 [SD 1.2] and correlation of 0.88). Assuming similar results in the passive control group, a sample size of 30 per group will have 80% power to detect differences in mean changes of 1 between VR-GR and the two control groups. Significance (α) is .025 to control for 2 comparisons.

The VR-DGR and VR-D groups were combined into a single VR group for data analysis because both groups utilized active, distraction-based, immersive VR experiences. VR-DGR did not provide participants with feedback on their respiratory or heart rates. Consequently, it functioned as a distraction-based technique and did not significantly differ from the VR-D experience. Therefore, we combined the two groups, as both were fundamentally distraction-based.

A sample size of 60 for the treatment group and 30 for the control group will have 80% power to detect differences in mean changes of 1 between VR and control.

Data Analysis

All statistical analyses were performed using SAS 9.4 (SAS Institute). Patient demographics were described using mean (SD) or median (IQR) for continuous variables, depending on data distribution, and frequency (percentage) for categorical variables and compared between groups using *t* tests, Wilcoxon rank-sum tests, chi-square tests, or Fisher exact tests, as appropriate. Pain AUCs on POD1 and POD2 were calculated as the area under the time-pain score curve using the trapezoidal rule. MMEs, a metric for quantifying and comparing doses of different opioids, were derived for POD0, POD1, and POD2. Change from baseline on pain intensity, pain unpleasantness, and anxiety immediately after and 15 and 30 minutes following each session was calculated as the postinterval value minus the baseline (before session) value on POD 1 and POD2. Mixed effects models for repeated measures were used for pain AUC, MME, and change from baseline on pain intensity, pain unpleasantness, and anxiety outcomes. All mixed effects models included the intervention group and POD as fixed effects and the participant as a random effect. Models for the change from baseline outcomes also included baseline value, time (0-, 15-,

XSL•FO
RenderX

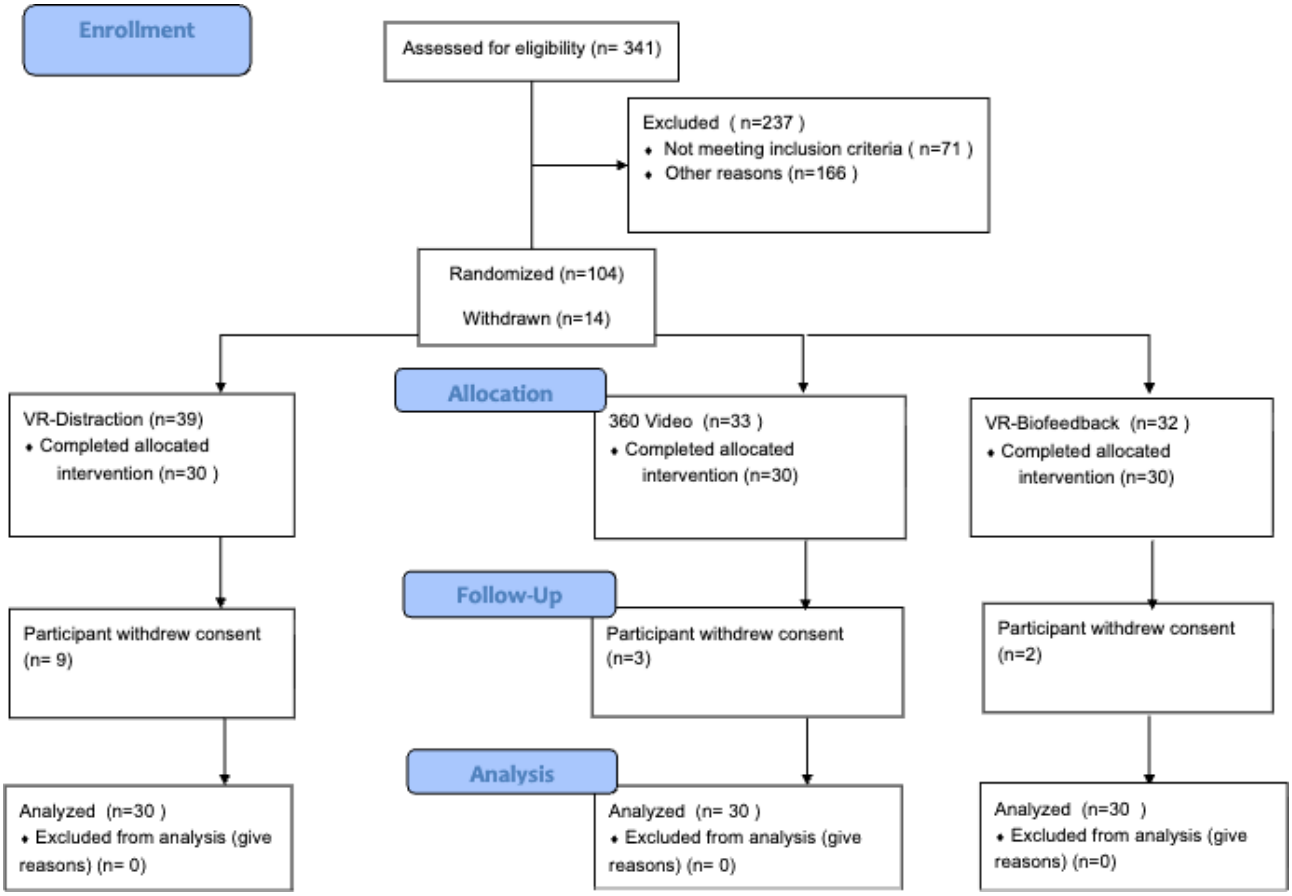
and 30-min post-intervention), and group and time interaction as fixed effects. Missing data in the outcomes were examined for pattern and assumed missing at random and handled using full information maximum likelihood (FIML) for mixed effects models. Sidak adjustment for multiplicity was used for change from baseline in pain intensity between intervention groups at 3 time points (immediately after and 15 and 30 minutes following each session).

Results

Participants

Ninety patients were enrolled in the study (60 in VR and 30 in 360-V; Figure 2). The participants comprised 73 male and 17 female patients and had an American Society of Anesthesiologists (ASA) Physical Status Classification System score of 1 - 3, with a mean age of 15.5 (SD 1.4) years. The 2 groups had no differences demographically except for a difference in PCS scores (VR: median 18, IQR 15 - 23; VR-360: median 24, IQR 16 - 28; *P*=.04). Patients were primarily male, adolescent, and Caucasian. This is consistent with the demographics of patients with pectus excavatum [37], and these are the patients most likely to undergo the Nuss procedure [38] (Table 2). All patients had at least 1 observation on all repeated-measured outcomes (pain AUC, MME, pain intensity, pain unpleasantness, and anxiety), and all available data were included in the mixed effects models for the outcomes.

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) 2010 flow diagram. VR: virtual reality.



XSL•FO
RenderX

Table . Patient characteristics.

Characteristic	VR[a]	360° video	Overall	*P* value (test)
Age (y), mean (SD)	15.6 (1.4)	15.1 (1.5)	15.5 (1.4)	.10
ASA[b] physical status, n (%)				.34
I	3 (5)	4 (13.3)	7 (7.8)	
II	44 (73.3)	21 (70.0)	65 (72.2)	
III	13 (21.7)	5 (16.7)	18 (20.0)	
Race, n (%)				.55
Caucasian	57 (95)	30 (100)	87 (96.7)	
African American	0 (0)	0 (0)	0 (0)	
Asian	0 (0)	0 (0)	0 (0)	
Other	3 (5)	0 (0)	3 (3.3)	
Ethnicity, n (%)				.25
Hispanic	3 (5)	0 (0)	3 (3.3)	
Non-Hispanic	56 (93.3)	28 (93.3)	94 (93.3)	
Unknown	1 (1.7)	2 (6.7)	3 (3.3)	
Sex, n (%)				.70
Male	48 (80)	25 (83.3)	73 (81.1)	
Female	12 (20)	5 (16.7)	17 (18.9)	
Weight (kg), mean (SD)	58.9 (9.8)	56.3 (10)	58.1 (9.9)	.24
CASI[c] score, mean (SD)	29.3 (5)	30.9 (4.8)	29.8 (5)	.14
Pain Catastrophizing Scale (PCS), median (IQR)	18 (15-23)	24 (16-28)	19 (15-26)	.04

[a]VR: virtual reality.

[b]ASA: American Society of Anesthesiologists.

[c]CASI: Child Anxiety Sensitivity Index.

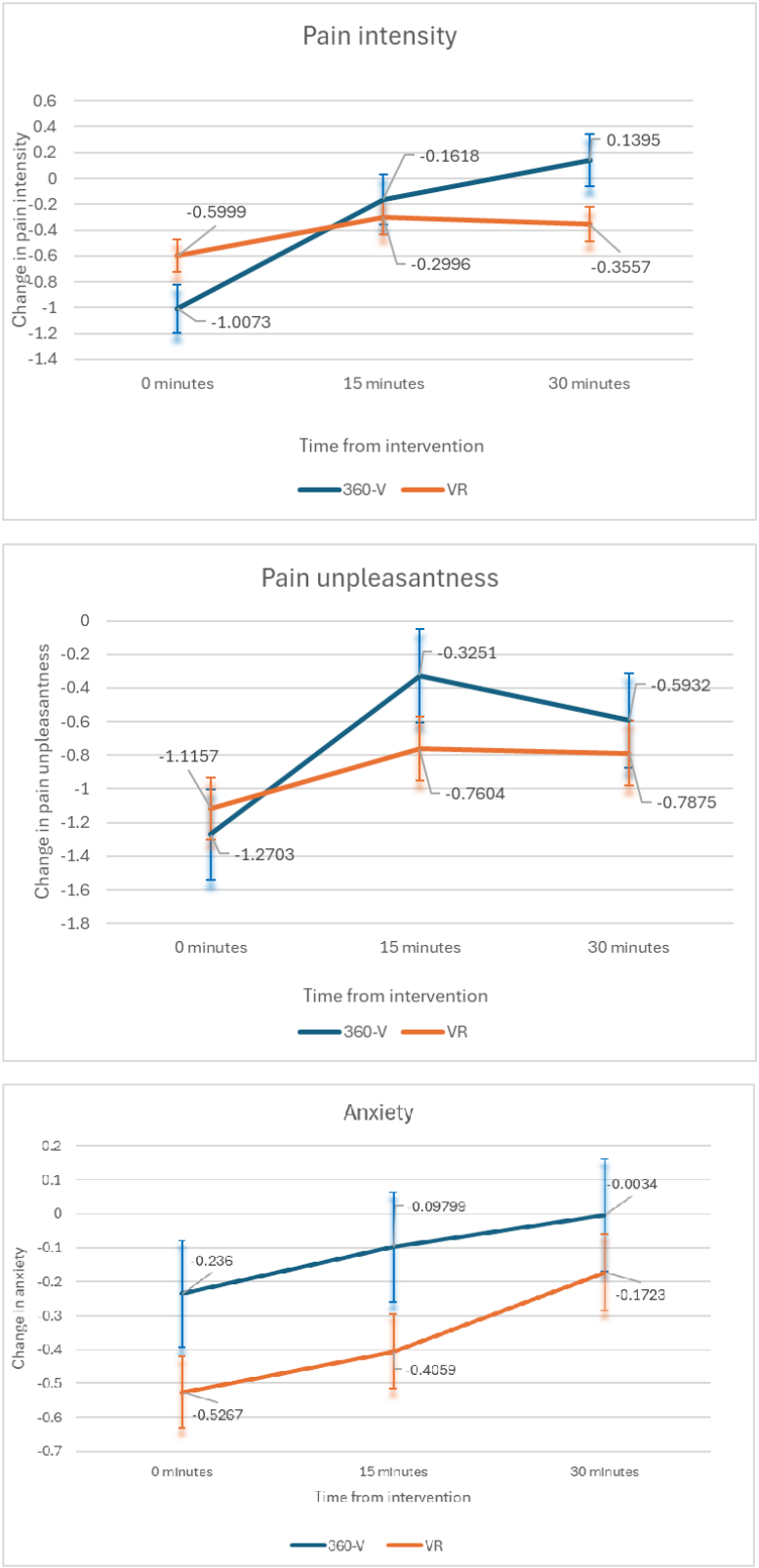
Changes From Baseline (VR vs Control)

Pain Intensity

Patients who participated in VR reported significantly decreased pain intensity from baseline (0.41 more decrease in pain intensity with SE 0.23) compared with those in the 360-V group at 30 minutes (*P*=.04) before multiplicity adjustment but not after multiplicity adjustment. There was no significant difference from baseline in reported pain intensity between VR vs 360-V immediately following the session (*P*=.08) or after 15 minutes (*P*=.56; Figure 3).

XSL•FO
RenderX

Figure 3. Changes in baseline in pain intensity, pain unpleasantness, and anxiety in time points following 360° video (360-V) and virtual reality (VR) in a mixed effect model with standard error bars.



Pain Unpleasantness

There was no significant difference in the reported pain unpleasantness between patients who participated in VR versus 360-V immediately following the VR session (P=.64), after 15 minutes (P=.20), or after 30 minutes (P=.57; Figure 3).

XSL•FO
RenderX

Anxiety

There were no significant differences in reported anxiety from baseline between patients who participated in the VR versus 360-V immediately following the session (*P*=.13), after 15 minutes (*P*=.12), or after 30 minutes (*P*=.40; Figure 3).

Inpatient Pain and Opioid Use

There were no significant differences in mean AUC pain scores between VR and 360-V (*P*=.60). There were also no significant differences in inpatient opioid use (MME/kg/day) between VR and 360-V (*P*=.26; Table 3).

Table . Inpatient pain and opioid use.

Characteristic	VR[a], mean (SD)	360-V[b], mean (SD)	*P* value[c]
Inpatient pain (AUC[d])			.60
POD[e]1	110.0 (34.3)	107.9 (31.8)	
POD2	95.9 (32.5)	90.6 (36.0)	
Inpatient opioid use (MME[f]/kg/day)			.26
POD0	0.18 (0.29)	0.18 (0.24)	
POD1	0.58 (0.22)	0.64 (0.28)	
POD2	0.50 (0.17)	0.51 (0.20)	

[a]VR: virtual reality.

[b]360-V: 360° video.

[c]*P* value from mixed effects models.

[d]AUC: area under the curve.

[e]POD: postoperative day.

[f]MME: morphine milligram equivalents.

Discussion

Principal Results

In our study, we found that active, immersive VR experiences had some trends to transient effects on both acute pain and anxiety compared to a nonimmersive 360-V control; however, these effects did not meet statistical significance. Patients who participated in VR reported a significantly greater decrease in pain intensity from baseline (0.41 with SE 0.23) compared with those in the 360-V group at 30 minutes (*P*=.04) before multiplicity adjustment but not after multiplicity adjustment. The trends in reduction in pain and anxiety were small; these trends did not achieve clinical significance either. Current literature indicates that a reduction of at least 2 points on the NRS for pain intensity or a 30% reduction in pain is considered clinically significant [39]. We did not see effects on overall AUC pain scores or opioid use. In this research, the VR-GR experience was likely distraction-based, as we could not document or assess feedback on patients' ability to perform the guided relaxation techniques correctly. In spite of not reaching statistical significance, these trends are not an absence of evidence of the effectiveness of VR to reduce postoperative pain and anxiety. We had relatively similar treatment conditions in small samples with attrition. The trends point us in the direction of future work.

Demographically, our two groups showed no significant differences except for a difference in PCS scores (VR: median 18, IQR 15 - 23; VR-360: median 24, IQR 16 - 28; *P*=.04). However, this result, while statistically significant, may not be clinically significant, as only PCS scores above 30 are clinically relevant, and neither group's median score exceeded 30 [40].

Although we noted some trends toward reduction in acute pain and anxiety from immersive VR following Nuss repair of pectus excavatum, these effects did not result in a significant change in AUC pain scores or inpatient opioid usage between the two groups. Several factors may account for the lack of significance. At our institution, the standard postoperative pain management protocol for following pectus surgery involves the scheduled administration of opioids, meaning all patients receive a standardized, weight-based dosage of opioids during their hospital stay, regardless of their actual pain level, with variations only in as-needed doses [10]. Consequently, opioid consumption may not accurately represent the patients' pain levels and opioid requirements. Home medication use might better reflect patients' pain and opioid needs. This study limited VR experiences to hospitalized patients. Extending its use past hospital discharge may have yielded different results. Future studies should investigate the integration of VR therapy into postoperative pain management both during and after hospitalization.

Limitations

Although this study was a prospective, randomized clinical trial, which can provide the best clinical evidence and support for VR, it has several limitations due to both study design and factors outside the control of the research team. Our patient population was somewhat homogeneous, as most individuals undergoing pectus excavatum repair are adolescent white males [38]. Therefore, our findings may not be generalizable to a broader population. Additionally, our control group may have been too similar to our intervention group. While the 360-V group did not receive audio instructions for guided relaxation, they still used a VR headset and experienced some level of distraction and immersion. Hence, the difference between the

XSL•FO
RenderX

two groups may not have been substantial enough to detect a meaningful difference. This lack of significant differences in the treatment groups may account for the lack of statistical differences between the two groups when comparing pain and anxiety. Future research will use nonimmersive, non-headset control and retrospectively compare historical data to better assess the impact of VR on these outcomes.

Reporting bias may have also played a role, as study participants may have felt inclined to report decreased pain and/or anxiety after treatment due to their perception of receiving an intervention, regardless of actual changes. The self-reported 1 to 10 rating scale is limited and has been shown to have a moderate correlation with clinical indicators of pain; thus, we cannot rely exclusively on this measure for pain evaluation [41]. Future studies should consider additional outcome measures.

The limited use of the intervention (10 min per d) makes it unlikely to produce meaningful changes in the severe postoperative pain often experienced by these patients. This suggests a need for further investigation of VR for pain relief, potentially incorporating VR interventions more systematically throughout the perioperative period. This could include preoperative VR exposure, repeated interventions multiple times per day, for additional consecutive days following surgery. We accounted for our attrition in our statistical analysis. However, the attrition with small groups to start likely was one reason for our inability to find statistically significant differences in results between the two groups.

Comparison With Prior Work

Controlling pain after surgery is important; uncontrolled postoperative pain can lead to increased morbidity, decreased function, prolonged recovery, and higher costs [42]. Severe acute postoperative pain can also lead to chronic postoperative pain, with rates of chronic postsurgical pain reported to be about 20% in pediatric populations [43]. Furthermore, opioid use after surgery also has risks, including persistent opioid use postoperatively; one study found a rate of 4.8% of persistent opioid use in postoperative adolescents as compared to a 0.1% rate of persistent opioid use in their nonsurgical matched cohorts [44].

Our study's results align with prior research suggesting that VR-D may play a role in transiently reducing acute pain and that relaxation via guided imagery can promote reductions in both pain and anxiety [20,45]. Research consistently shows that VR is effective in lowering procedural pain, anxiety, and fear in pediatric patients, particularly during needle-related and other painful interventions [46-55]. A few studies have also demonstrated the feasibility of using VR to alleviate acute postoperative pain [21,56]. In adults, a meta-analysis found that patients receiving perioperative VR had lower pain scores than those receiving usual care (mean deviation −0.64, 95% CI −1.05 to −0.22; $P<.02$). Additionally, patients receiving VR postoperatively experienced a significant reduction in pain scores (mean deviation −0.50, 95% CI −0.76 to −0.24; $P=.002$) [57]. One pediatric study indicated that a single preoperative VR experience reduced the need for rescue analgesics in the recovery unit for painful procedures [58].

Relaxation-guided imagery has been shown to reduce both pain and anxiety in children undergoing minor surgery [59]. VR-GR has also demonstrated effectiveness in reducing pain and anxiety in children during medical procedures. These effects were immediate but transient, with some studies reporting reductions lasting up to 30 minutes after a session [21,60,61]. Additionally, a small (n=51) single-center, prospective study evaluated a single VR-GR session for acute postoperative pain and anxiety in children and adolescents. This study showed similar transient reductions in pain intensity and anxiety [21].

Using guided relaxation, our study aimed to harness the benefits of using mind-body techniques to manage postoperative pain. However, there are several potential reasons for the lack of clinically significant lasting effects on acute pain and anxiety in this population or opioid consumption. First, our study did not include a non-headset control group. Previous studies assessing the effects of VR on opioid use compared VR experiences to standard care without VR [55,62,63]. In our study, every patient utilized a VR headset and had some level of an immersive experience, even with 360-V (a nonimmersive option). As a result, all patients, including the control group, likely experienced some distraction.

Moreover, the effectiveness of the VR sessions may have been limited by the short duration of each session and the low number of sessions. Ten minutes per day may be an inadequate time to achieve lasting pain relief of severe acute pain, especially as compared to other therapeutic and/or pharmacologic interventions that are administered more frequently. Future studies should investigate the optimal timing and length of intervention for maximal benefit of VR.

While our results indicated that active, immersive VR experiences had some trends toward effects on both acute pain and anxiety compared to the nonimmersive 360-V control, these effects were not clinically significant. The study may not have produced clear positive results due to experimental factors such as the similarity of the VR and control treatments, the short duration of treatments, and the protocol-driven high use of opioids that were not adequately considered in the study design. Nevertheless, the findings provide a valuable framework for designing future VR studies. We can learn from our null results to design future VR studies with a control treatment that does not use a VR headset and use a population that only receives opioids on an as-needed basis. Longer treatment duration and less subject attrition could also lead to more significant results.

Conclusions

This study found that daily, 10-minute VR sessions had trends toward transiently reducing pain and anxiety compared to a 360-V experience in participants following Nuss repair of pectus excavatum. These results were not clinically significant. Due to the limited duration of the intervention and the standardized, scheduled, high utilization of opioids in this population, VR was not sufficient in significantly decreasing opioid use and overall AUC pain scores. Despite these conclusions, exploring guided relaxation VR as an adjunct to, rather than a replacement for, postoperative pharmacologic analgesics may prove valuable. Increasing the length and frequency of VR experiences per day, along with a policy of not automatically administering opioids

XSL•FO
RenderX

unless requested, may help decrease opioid usage and AUC pain scores. A systematic integration of VR into perioperative care is likely necessary to impact the pain trajectory and opioid usage in postoperative patients. Furthermore, improving the VR experience to incorporate true guided relaxation would likely enhance effectiveness compared to a purely distraction-based approach. Future studies are needed to further explore the use of this therapy in postoperative pain management.

Acknowledgments

The authors gratefully acknowledge Maria E Ashton, MS, RPh, MBA, medical writer, Department of Anesthesiology, Cincinnati Children's Hospital Medical Center, for providing writing assistance, editing, and proofreading.

The authors gratefully acknowledge Luke Wilson and the Stanford CHARIOT program for the development and use of the Mindful Aurora application.

Funding

Support was provided by the Department of Anesthesiology at Cincinnati Children's Hospital Medical Center and a Research Innovation/Pilot Award from Cincinnati Children's Hospital Medical Center.

Conflicts of Interest

None declared.

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 1122 KB - periop_v9i1e80902_app1.pdf]

References

1. Yaster M, McNaull PP, Davis PJ. The opioid epidemic in pediatrics: a 2020 update. Curr Opin Anaesthesiol 2020 Jun;33(3):327-334. [doi: 10.1097/ACO.0000000000000865] [Medline: 32371640]
2. Drug overdose deaths: facts and figures. National Institute on Drug Abuse (NIDA). URL: https://www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates [accessed 2026-02-16]
3. Hudgins JD, Porter JJ, Monuteaux MC, Bourgeois FT. Trends in opioid prescribing for adolescents and young adults in ambulatory care settings. Pediatrics 2019 Jun;143(6):e20181578. [doi: 10.1542/peds.2018-1578] [Medline: 31138669]
4. Johnston LD, O'Malley PM, et al. Monitoring the future" national survey results on drug use, 1975–2017: 2017 overview, key findings on adolescent drug use. : Institute for Social Research, University of Michigan; 2017 URL: https://monitoringthefuture.org/wp-content/uploads/2022/08/mtf-overview2017.pdf [accessed 2026-02-16]
5. Groenewald CB, Zhou C, Palermo TM, Van Cleve WC. Associations between opioid prescribing patterns and overdose among privately insured adolescents. Pediatrics 2019 Nov;144(5):e20184070. [doi: 10.1542/peds.2018-4070] [Medline: 31575622]
6. Battaglia M, Groenewald CB, Campbell F, et al. We need to talk: the urgent conversation on chronic pain, mental health, prescribing patterns and the opioid crisis. J Psychopharmacol 2023 May;37(5):437-448. [doi: 10.1177/02698811221144635] [Medline: 37171242]
7. Breglio AM, Fitzgerald TN, Moore CB, Einhorn LM. Evaluation of analgesic practice changes following the Nuss procedure in pediatric patients. J Surg Res 2023 Nov;291:289-295. [doi: 10.1016/j.jss.2023.06.018] [Medline: 37481964]
8. Densmore JC, Peterson DB, Stahovic LL, et al. Initial surgical and pain management outcomes after Nuss procedure. J Pediatr Surg 2010 Sep;45(9):1767-1771. [doi: 10.1016/j.jpedsurg.2010.01.028] [Medline: 20850618]
9. Lai K, Notrica DM, McMahon LE, et al. Cryoablation in 350 Nuss procedures: evolution of hospital length of stay and opioid use. J Pediatr Surg 2023 Aug;58(8):1435-1439. [doi: 10.1016/j.jpedsurg.2022.10.051] [Medline: 36494205]
10. Walter CM, Lee CS, Moore DL, et al. Retrospective study comparing outcomes of multimodal epidural and erector spinae catheter pain protocols after pectus surgery. J Pediatr Surg 2023 Mar;58(3):397-404. [doi: 10.1016/j.jpedsurg.2022.06.017] [Medline: 35907711]
11. Tashjian VC, Mosadeghi S, Howard AR, et al. Virtual reality for management of pain in hospitalized patients: results of a controlled trial. JMIR Ment Health 2017 Mar 29;4(1):e9. [doi: 10.2196/mental.7387] [Medline: 28356241]
12. Simonetti V, Tomietto M, Comparcini D, Vankova N, Marcelli S, Cicolini G. Effectiveness of virtual reality in the management of paediatric anxiety during the peri-operative period: a systematic review and meta-analysis. Int J Nurs Stud 2022 Jan;125:104115. [doi: 10.1016/j.ijnurstu.2021.104115] [Medline: 34781118]
13. Olbrecht VA, O'Conor KT, Williams SE, et al. Transient reductions in postoperative pain and anxiety with the use of virtual reality in children. Pain Med 2021 Nov 26;22(11):2426-2435. [doi: 10.1093/pm/pnab209] [Medline: 34175959]

XSL•FO
RenderX

14. Melzack R, Wall PD. Pain mechanisms: a new theory. Science 1965 Nov 19;150(3699):971-979. [doi: 10.1126/science.150.3699.971] [Medline: 5320816]
15. Smith KL, Wang Y, Colloca L. Impact of virtual reality technology on pain and anxiety in pediatric burn patients: a systematic review and meta-analysis. Front Virtual Real 2022 Jan;2:2022. [doi: 10.3389/frvir.2021.751735] [Medline: 36247202]
16. Indovina P, Barone D, Gallo L, Chirico A, De Pietro G, Giordano A. Virtual reality as a distraction intervention to relieve pain and distress during medical procedures: a comprehensive literature review. Clin J Pain 2018 Sep;34(9):858-877. [doi: 10.1097/AJP.0000000000000599] [Medline: 29485536]
17. Ahmadpour N, Randall H, Choksi H, Gao A, Vaughan C, Poronnik P. Virtual reality interventions for acute and chronic pain management. Int J Biochem Cell Biol 2019 Sep;114:105568. [doi: 10.1016/j.biocel.2019.105568] [Medline: 31306747]
18. Hadjiat Y, Marchand S. Virtual reality and the mediation of acute and chronic pain in adult and pediatric populations: research developments. Front Pain Res (Lausanne) 2022;3:840921. [doi: 10.3389/fpain.2022.840921] [Medline: 35599969]
19. Sánchez-Caballero E, Ortega-Donaire L, Sanz-Martos S. Immersive virtual reality for pain and anxiety management associated with medical procedures in children and adolescents: a systematic review. Children (Basel) 2024 Aug 13;11(8):975. [doi: 10.3390/children11080975] [Medline: 39201910]
20. Hoffman HG, Rodriguez RA, Gonzalez M, et al. Immersive virtual reality as an adjunctive non-opioid analgesic for pre-dominantly Latin American children with large severe burn wounds during burn wound cleaning in the intensive care unit: a pilot study. Front Hum Neurosci 2019;13(262):262. [doi: 10.3389/fnhum.2019.00262] [Medline: 31440148]
21. Olbrecht VA, O'Conor KT, Williams SE, et al. Guided relaxation–based virtual reality for acute postoperative pain and anxiety in a pediatric population: pilot observational study. J Med Internet Res 2021 Jul 12;23(7):e26328. [doi: 10.2196/26328] [Medline: 34048358]
22. Hemington KS, Cheng JC, Bosma RL, Rogachov A, Kim JA, Davis KD. Beyond negative pain-related psychological factors: resilience is related to lower pain affect in healthy adults. J Pain 2017 Sep;18(9):1117-1128. [doi: 10.1016/j.jpain.2017.04.009] [Medline: 28506777]
23. Agoston AM, Sieberg CB. Nonpharmacologic treatment of pain. Semin Pediatr Neurol 2016 Aug;23(3):220-223. [doi: 10.1016/j.spen.2016.10.005] [Medline: 27989329]
24. Turgut A, Özcan İlçe A, Öztürk H. The effect of immersive virtual reality application on anxiety, pain, and parental satisfaction in the perioperative process of children: a randomized controlled trial. Pain Manag Nurs 2024 Dec;25(6):584-590. [doi: 10.1016/j.pmn.2024.06.002] [Medline: 38955550]
25. Orgil Z, Johnson L, Karthic A, et al. Feasibility and acceptability of perioperative application of biofeedback-based virtual reality versus active control for pain and anxiety in children and adolescents undergoing surgery: protocol for a pilot randomised controlled trial. BMJ Open 2023 Jan 25;13(1):e071274. [doi: 10.1136/bmjopen-2022-071274] [Medline: 36697053]
26. Olbrecht VA, Williams SE, O'Conor KT, et al. Guided relaxation-based virtual reality versus distraction-based virtual reality or passive control for postoperative pain management in children and adolescents undergoing Nuss repair of pectus excavatum: protocol for a prospective, randomised, controlled trial (FOREVR Peds trial). BMJ Open 2020 Dec 30;10(12):e040295. [doi: 10.1136/bmjopen-2020-040295] [Medline: 33380482]
27. Chan AW, Tetzlaff JM, Gøtzsche PC, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ 2013 Jan 8;346:e7586. [doi: 10.1136/bmj.e7586] [Medline: 23303884]
28. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomized trials. JAMA 2025 Jun 10;333(22):1998-2005. [doi: 10.1001/jama.2025.4347] [Medline: 40228499]
29. Tsze DS, von Baeyer CL, Pahalyants V, Dayan PS. Validity and reliability of the verbal numerical rating scale for children aged 4 to 17 years with acute pain. Ann Emerg Med 2018 Jun;71(6):691-702. [doi: 10.1016/j.annemergmed.2017.09.009] [Medline: 29107409]
30. Price TJ, Ray PR. Recent advances toward understanding the mysteries of the acute to chronic pain transition. Curr Opin Physiol 2019 Oct;11:42-50. [doi: 10.1016/j.cophys.2019.05.015] [Medline: 32322780]
31. Pielech M, Ryan M, Logan D, Kaczynski K, White MT, Simons LE. Pain catastrophizing in children with chronic pain and their parents: proposed clinical reference points and reexamination of the Pain Catastrophizing Scale measure. Pain 2014 Nov;155(11):2360-2367. [doi: 10.1016/j.pain.2014.08.035] [Medline: 25180013]
32. Silverman WK, Fleisig W, Rabian B, Peterson RA. Childhood anxiety sensitivity index. J Clin Child Psychol 1991 Jun;20(2):162-168. [doi: 10.1207/s15374424jccp2002_7]
33. Gold JI, Mahrer NE. Is virtual reality ready for prime time in the medical space? A randomized control trial of pediatric virtual reality for acute procedural pain management. J Pediatr Psychol 2018 Apr 1;43(3):266-275. [doi: 10.1093/jpepsy/jsx129] [Medline: 29053848]
34. Proczkowska M, Ericsson E. Validity of the modified-distraction-short-scale and verbal-numeric-anxiety-fear-rating-scale for children in a preoperative setting. Paediatr Anaesth 2024 Feb;34(2):121-129. [doi: 10.1111/pan.14765] [Medline: 37728169]

XSL•FO
RenderX

35. Adams MCB, Sward KA, Perkins ML, Hurley RW. Standardizing research methods for opioid dose comparison: the NIH HEAL morphine milligram equivalent calculator. Pain 2025 Feb 3;166(8):1729-1737. [doi: 10.1097/j.pain.0000000000003529] [Medline: 39907478]
36. Morphine milligram equivalent (MME) calculator: frequently asked questions (FAQs) for HEAL researchers. : National Institutes of Health; 2025 URL: https://www.nih.gov/sites/default/files/2026-01/MME%20Calculator%20FAQ_508.pdf [accessed 2026-02-16]
37. Pitt JB, Zeineddin S, Carter M, et al. Demographics of anterior chest wall deformity patients: a tertiary children's hospital experience. J Surg Res 2024 Jan;293:451-457. [doi: 10.1016/j.jss.2023.09.040] [Medline: 37827024]
38. Mack SJ, Till BM, Huang C, et al. National trends in pectus excavatum repair: patient age, facility volume, and outcomes. J Thorac Dis 2022 Apr;14(4):952-961. [doi: 10.21037/jtd-21-1671] [Medline: 35572899]
39. Rowbotham MC. What is a "clinically meaningful" reduction in pain? Pain 2001 Nov;94(2):131-132. [doi: 10.1016/S0304-3959(01)00371-2] [Medline: 11690725]
40. Sullivan MJ. The pain catastrophizing scale: user manual. : McGill University; 2009 URL: https://drsandler.in/PCSmanual.pdf [accessed 2026-02-16]
41. Nugent SM, Lovejoy TI, Shull S, Dobscha SK, Morasco BJ. Associations of pain numeric rating scale scores collected during usual care with research administered patient reported pain outcomes. Pain Med 2021 Oct 8;22(10):2235-2241. [doi: 10.1093/pm/pnab110] [Medline: 33749760]
42. Gan TJ. Poorly controlled postoperative pain: prevalence, consequences, and prevention. J Pain Res 2017;10:2287-2298. [doi: 10.2147/JPR.S144066] [Medline: 29026331]
43. Rabbitts JA, Fisher E, Rosenbloom BN, Palermo TM. Prevalence and predictors of chronic postsurgical pain in children: a systematic review and meta-analysis. J Pain 2017 Jun;18(6):605-614. [doi: 10.1016/j.jpain.2017.03.007] [Medline: 28363861]
44. Harbaugh CM, Lee JS, Hu HM, et al. Persistent opioid use among pediatric patients after surgery. Pediatrics 2018 Jan;141(1):e20172439. [doi: 10.1542/peds.2017-2439] [Medline: 29203521]
45. Chen SF, Wang HH, Yang HY, Chung UL. Effect of relaxation with guided imagery on the physical and psychological symptoms of breast cancer patients undergoing chemotherapy. Iran Red Crescent Med J 2015 Nov;17(11):e31277. [doi: 10.5812/ircmj.31277] [Medline: 26734485]
46. Dumoulin S, Bouchard S, Ellis J, et al. A randomized controlled trial on the use of virtual reality for needle-related procedures in children and adolescents in the emergency department. Games Health J 2019 Aug;8(4):285-293. [doi: 10.1089/g4h.2018.0111] [Medline: 31135178]
47. Iannicelli AM, Vito D, Dodaro CA, et al. Does virtual reality reduce pain in pediatric patients? A systematic review. Ital J Pediatr 2019 Dec 30;45(1):171. [doi: 10.1186/s13052-019-0757-0] [Medline: 31888710]
48. Eijlers R, Utens EMWJ, Staals LM, et al. Systematic review and meta-analysis of virtual reality in pediatrics: effects on pain and anxiety. Anesth Analg 2019 Nov;129(5):1344-1353. [doi: 10.1213/ANE.0000000000004165] [Medline: 31136330]
49. Czech O, Wrzeciono A, Rutkowska A, Guzik A, Kiper P, Rutkowski S. Virtual reality interventions for needle-related procedural pain, fear and anxiety-a systematic review and meta-analysis. J Clin Med 2021 Jul 23;10(15):3248. [doi: 10.3390/jcm10153248] [Medline: 34362032]
50. Walther-Larsen S, Petersen T, Friis SM, Aagaard G, Drivenes B, Opstrup P. Immersive virtual reality for pediatric procedural pain: a randomized clinical trial. Hosp Pediatr 2019 Jul;9(7):501-507. [doi: 10.1542/hpeds.2018-0249] [Medline: 31160472]
51. Jenabi E, Bashirian S, Salehi AM, Rafiee M, Bashirian M. Virtual reality for pain reduction during intravenous injection in pediatrics: a systematic review and meta-analysis of controlled clinical trials. Clin Exp Pediatr 2023 Dec;66(12):533-537. [doi: 10.3345/cep.2022.01193] [Medline: 37321586]
52. Addab S, Hamdy R, Thorstad K, Le May S, Tsimicalis A. Use of virtual reality in managing paediatric procedural pain and anxiety: an integrative literature review. J Clin Nurs 2022 Nov;31(21-22):3032-3059. [doi: 10.1111/jocn.16217] [Medline: 35068011]
53. Alshatrat SM, Sabarini JM, Hammouri HM, Al-Bakri IA, Al-Omari WM. Effect of immersive virtual reality on pain in different dental procedures in children: a pilot study. Int J Paediatr Dent 2022 Mar;32(2):264-272. [doi: 10.1111/ipd.12851] [Medline: 34133809]
54. Gao Y, Xu Y, Liu N, Fan L. Effectiveness of virtual reality intervention on reducing the pain, anxiety and fear of needle-related procedures in paediatric patients: a systematic review and meta-analysis. J Adv Nurs 2023 Jan;79(1):15-30. [doi: 10.1111/jan.15473] [Medline: 36330583]
55. Camacho-Cruz J, Palacios-Ariza MA, Orrego-Celestino L, et al. Effectiveness of non-immersive virtual reality in the management of procedure-related pain in preschool children: a randomized clinical trial. Eur J Pediatr 2023 Sep;182(9):4103-4112. [doi: 10.1007/s00431-023-05070-5] [Medline: 37407711]
56. Mosso JL, Rizzo S, Wiederhold B, et al. Cybertherapy--new applications for discomfort reductions. Surgical care unit of heart, neonatology care unit, transplant kidney care unit, delivery room-cesarean surgery and ambulatory surgery, 27 case reports. Stud Health Technol Inform 2007;125:334-336. [Medline: 17377297]
57. Ding L, Hua H, Zhu H, et al. Effects of virtual reality on relieving postoperative pain in surgical patients: a systematic review and meta-analysis. Int J Surg 2020 Oct;82:87-94. [doi: 10.1016/j.ijsu.2020.08.033] [Medline: 32882400]

XSL•FO
RenderX

58. Eijlers R, Dierckx B, Staals LM, et al. Virtual reality exposure before elective day care surgery to reduce anxiety and pain in children: a randomised controlled trial. Eur J Anaesthesiol 2019 Oct;36(10):728-737. [doi: 10.1097/EJA.0000000000001059] [Medline: 31356373]
59. Vagnoli L, Bettini A, Amore E, De Masi S, Messeri A. Relaxation-guided imagery reduces perioperative anxiety and pain in children: a randomized study. Eur J Pediatr 2019 Jun;178(6):913-921. [doi: 10.1007/s00431-019-03376-x] [Medline: 30944985]
60. Bernaerts S, Bonroy B, Daems J, et al. Virtual reality for distraction and relaxation in a pediatric hospital setting: an interventional study with a mixed-methods design. Front Digit Health 2022;4:866119. [doi: 10.3389/fdgth.2022.866119] [Medline: 35712230]
61. Hoag JA, Karst JS, Bingen KM, Palou-Torres A, Yan K. Distracting through procedural pain and distress using virtual reality and guided imagery in pediatric, adolescent, and young adult patients: randomized controlled trial. J Med Internet Res 2022 Apr 18;24(4):e30260. [doi: 10.2196/30260] [Medline: 35436209]
62. Xiang H, Shen J, Wheeler KK, et al. Efficacy of smartphone active and passive virtual reality distraction vs standard care on burn pain among pediatric patients: a randomized clinical trial. JAMA Netw Open 2021 Jun 1;4(6):e2112082. [doi: 10.1001/jamanetworkopen.2021.12082] [Medline: 34152420]
63. Gold JI, SooHoo M, Laikin AM, Lane AS, Klein MJ. Effect of an immersive virtual reality intervention on pain and anxiety associated with peripheral intravenous catheter placement in the pediatric setting: a randomized clinical trial. JAMA Netw Open 2021 Aug 2;4(8):e2122569. [doi: 10.1001/jamanetworkopen.2021.22569] [Medline: 34432011]

Abbreviations

360-V: 360° video
AUC: area under the curve
CCHMC : Cincinnati Children's Hospital Medical Center
CONSORT: Consolidated Standard of Reporting Trials
MME: morphine milligram equivalents
NRS: numerical rating scale
PCS : Pain Catastrophizing Scale
POD : postoperative day
REDCap: Research Electronic Data Capture
SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials
VR: virtual reality
VR-D : distraction-based gaming virtual reality
VR-DGR: distraction-based guided relaxation virtual reality
VR-GR: guided relaxation–based virtual reality

Edited by N Rohatgi; submitted 18.Jul.2025; peer-reviewed by I Cornelisz, JW Gander; revised version received 20.Jan.2026; accepted 06.Feb.2026; published 10.Mar.2026.

Please cite as:
Walter CM, Froass D, Bell N, Haack L, Boehmer C, Bruguera Torres C, Spivak R, Chou M, Geisler K, O'Conor K, Williams SE, Ding L, King CD, Olbrecht VA
Virtual Reality for the Management of Postoperative Pain and Anxiety in Children and Adolescents Undergoing Nuss Repair of Pectus Excavatum: Randomized Controlled Trial
JMIR Perioper Med 2026;9:e80902
URL: https://periop.jmir.org/2026/1/e80902
doi:10.2196/80902

© Charlotte M Walter, Dillon Froass, Nora Bell, Lauren Haack, Chloe Boehmer, Claudia Bruguera Torres, Rachel Spivak, Max Chou, Kristie Geisler, Keith O'Conor, Sara E Williams, Lili Ding, Christopher D King, Vanessa A Olbrecht. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating the Impact of Virtual Reality on Orthopedic Trauma Skills Acquisition Among Surgical Residents: Randomized Crossover Study

Sandipan Chatterjee[1], PhD; Khairul Faizi Mohammad[2], MD; Monica Ghidinelli[1], PhD

[1]AO Education Institute, AO Foundation, Clavadelerstrasse 8, Davos, Switzerland

[2]Department of Orthopedics, Pantai Hospital Cheras, Kuala Lumpur, Malaysia

Corresponding Author:
Sandipan Chatterjee, PhD
AO Education Institute, AO Foundation, Clavadelerstrasse 8, Davos, Switzerland

Abstract

Background: Orthopedic trauma skills training is time-consuming and expensive. Current training modalities rely heavily on synthetic bone models, anatomical laboratory simulations, or assistance in surgeries (the apprenticeship model). Virtual reality (VR) appears to present a promising complement to current training modalities.

Objective: This study evaluated the effectiveness of VR training on surgical performance and gauged learning preferences among orthopedic trauma residents in Malaysia.

Methods: In total, 123 orthopedic residents were randomly assigned to 2 groups. One group practiced for about 30 minutes using VR glasses, followed by conventional nailing exercises on synthetic bones, while the other group first performed the nailing exercise, followed by VR practice. Performance was measured by time to completion of the exercise, and participants completed a postexercise survey.

Results: Participants who completed VR training before the synthetic bone nailing exercise were significantly faster, completing the task between 4 (P=.05) and 7 (P=.002) minutes more quickly than without VR training. In addition, VR training improved self-assessed performance during the exercise. Survey data revealed that while 43% (50/117) of participants preferred conventional methods of learning (lectures, discussions, and hands-on simulations), 89% (104/117) of participants supported VR use as an adjunct to conventional methods of learning. Less than 2% (2/117, 1.7%) of participants indicated that conventional methods of learning were outdated.

Conclusions: A single session of VR training significantly reduced completion times and improved self-assessment of competence in orthopedic trauma simulation exercises. Although learners continue to value conventional training modalities, there is a strong desire to include VR as a supplementary tool. Its integration into surgical curricula may accelerate skill acquisition, especially in low-resource settings with limited access to high-fidelity simulation labs. In addition, the availability of VR training modules in hospitals could help residents and junior consultants prepare for surgery.

(JMIR Perioper Med 2026;9:e79343) doi:10.2196/79343

KEYWORDS

virtual reality; orthopedic procedures; graduate medical education; simulation training; fracture fixation

Introduction

Background

The growing complexity of orthopedic trauma surgery has heightened the demand for advanced and effective training modalities. Conventional approaches such as synthetic bone models, anatomical laboratory simulations, and the apprenticeship model of learning through assisting in surgeries are time-consuming, expensive, and limited in availability [1,2]. Anatomical laboratory simulations provide high anatomical fidelity and tissue realism but are logistically demanding (eg, cadaver procurement, facility access, higher per-learner cost, and lower frequency). These constraints, particularly pronounced in low-resource settings, limit sufficient exposure to trauma case variations and present barriers to skill acquisition.

Prior Work

Alternatives are being explored to complement traditional training modalities. Virtual reality (VR) has emerged as a promising tool that offers immersive, repeatable, and cost-effective simulations in a controlled environment [3-5]. In contrast to anatomical laboratories, immersive VR has standardized scenarios, providing immediate feedback and repetition at scale. VR also offers a lower marginal cost and greater scheduling flexibility, albeit with reduced tactile realism

XSL•FO
RenderX

and haptic resistance compared with cadaveric tissue. Recent studies have highlighted the benefits of VR for orthopedic surgery, where procedural repetition and real-time feedback, which are critical for skill acquisition, can help reduce the learning curve [6]. VR in orthopedic training appears particularly effective in simulating procedures such as intramedullary nailing, arthroscopy, and arthroplasty [7-10]. Beyond technical skill acquisition, VR can also support the development of nontechnical skills, including decision-making and teamwork—critical elements in high-stakes surgical environments [11,12].

One of VR's primary advantages is its scalability, particularly in low-resource settings [13]. The increasing availability of VR platforms such as Oculus Rift and HTC Vive has the potential to increase training opportunities in such regions [5]. The ability to simulate a broad range of clinical scenarios, including rare or complex cases, provides a comprehensive training experience that would otherwise be difficult to achieve through traditional methods.

Although prior studies have established VR's general utility in surgical training, our study is the first to observe immediate skill transfer after a single VR session for both tibial and femoral nailing procedures, in a Malaysian residency context. Additionally, we examined the effect of VR training on self-assessed performance and gauged learner preferences for different training modalities.

Training Challenges in Malaysia

In Malaysia, orthopedic residency is a 4-year training program under the Ministry of Health and the Ministry of Education. Upon graduation, residents serve in hospitals run by either the Ministry of Health or the Ministry of Education. Selection is based on open entrance examinations and interviews [14].

Surgical training in Malaysia is generally considered "in work training." Hence, the level of training gained depends on the level provided by the trainer at the hospital in which the candidate is serving. The training level also varies depending on the opportunity to perform surgery and the types of cases available within a particular hospital. Other constraints particular to the region are the lack of sufficient hospital- or university-based courses. Anatomical laboratories are available but are expensive and infrequent [15].

According to a report published by the University of Malaya, there were an estimated 900 orthopedic surgeons in Malaysia in 2018, giving a ratio of 1 orthopedic surgeon per 40,000 population [14]. To meet the suggested World Health Organization ratio of 1:30,000 population, orthopedics in Malaysia faces the challenge of increasing surgeon numbers accordingly. Although approximately 70 residents are recruited annually and placed at 30 accredited training centers, they must also navigate economic realities to provide the best treatment within the limitations of public funding and patient affordability [14].

Goal of This Study

This study aimed to evaluate the impact of VR-based training on surgical performance and to explore learning preferences of orthopedic trauma residents in Malaysia.

Methods

Educational Intervention

The educational intervention was integrated into a 3-day course that combined lectures, small group discussions, and hands-on practical exercises. Participants practiced a common tibia nailing procedure and a femoral nailing procedure on both VR headsets and synthetic bones. VR training was conducted using Oculus Rift S headsets (Meta Platforms Inc), running Johnson & Johnson MedTech simulation software. The headset recorded procedure completion time, procedural steps, and errors, where applicable.

Each practical exercise workstation had 3 participants per synthetic bone. For VR training, participants practiced the nailing procedure once. In total, 30 VR headsets were available for this study, divided across 6 rooms, each with 5 participants, 1 faculty member, 1 laptop, and 1 technical staff member. The rooms were divided into individual grids of 1.5 m^2 (Figure 1). The hotel Wi-Fi network was used to connect the headsets.

XSL•FO
RenderX

Figure 1. Setup of the virtual reality training session. Consent for using this image was obtained via course registration.



Tibia Nailing Exercise

The tibia nail used for the synthetic bone exercise was the Expert Tibia Nail (Johnson & Johnson). For the VR module, the Tibia Nail Advanced (Johnson & Johnson) module was used. As the principal and major steps of both nailing systems are similar, they were deemed equivalent for comparison in this study. The tibia nailing procedure was performed during an AO (Arbeitsgemeinschaft für Osteosynthesefragen) Trauma Basic Principles course (cohort 1).

Femoral Nailing Exercise

The femoral nail used for the synthetic bone exercise was the Trochanteric Fixation Nail Advanced Proximal Femoral Nailing System (Johnson & Johnson). The same nailing procedure was used for the VR module. The femoral nailing procedure was performed during the AO Trauma Basic Principles course (cohort 1) and the AO Trauma Advanced Principles course (cohort 2).

Study Design

We used a 2-period crossover design. Participants were randomized to one of two sequences: (1) VR practice followed by the synthetic bone exercise or (2) the synthetic bone exercise followed by VR practice. After completing the first period, participants crossed over to the other modality. The primary end point was time (in minutes) to complete the synthetic bone exercise, with or without prior VR practice. These comparisons are reported separately for tibial and femoral procedures. Completion times for both modalities (VR and synthetic bone) are reported.

A qualitative postexercise survey assessed self-perceived performance and learning preferences. The questionnaire was developed, revised, and pilot-tested with 5 surgeons and administered via Microsoft Forms (version 16.0; Microsoft Corp).

Population

Participants in this study were those who attended the AO Trauma Basic and Advanced Principles courses on the Principles of Fracture Management in Malaysia in March 2023. Beyond course attendance and consent, no inclusion or exclusion criteria were applied. All participants were able to complete VR training, as no cases of nausea or dizziness were observed while using the VR headsets.

Data Collection

Participant demographic data were collected at baseline. For the synthetic bone exercises, each participant captured and reported the synthetic bone exercise completion time using a stopwatch. The completion times of the VR module were automatically captured via the VR headset software. Data were collected after each exercise and analyzed using Microsoft Excel (version 16.0; Microsoft Corp). Survey responses were collected after the exercise through Microsoft Forms.

XSL•FO
RenderX

Analysis

A descriptive analysis was performed to determine the frequencies and percentages of the response selections. The tables and figures display frequency counts and/or percentages for each response option. A paired 2-tailed *t* test was used to test the effect of VR training on time to completion of the synthetic bone practical exercise. A Bonferroni post hoc test was conducted to correct for multiple testing.

Ethical Considerations

According to the ethics committee of the Canton of Zurich, this study did not require ethical committee authorization (Req-2025-00605). Informed consent was obtained from participants prior to the start of the course and no compensation was provided to study participants for taking part in this study. In the survey, we included the following statement of purpose, which disclosed our intended use of the data: "The information you provide will be anonymized and made available to the study group in aggregate form. The data will be used for research purposes."

Results

Participant Demographics

This study included 2 cohorts with 63 and 60 participants from the AO Trauma Basic and Advanced Principles courses, respectively (Table 1). The average age between these two cohorts was significantly different (*P*<.001), as was the average years of orthopedic trauma training (*P*<.001) (Table 1). There was no significant difference in the number of femoral nailing cases performed by the participants of each cohort.

Table . Participant demographics and experience level.[a]

Participant demographics	Cohort 1[b]	Cohort 2[c]	*P* value
Sex, n (%)			—[d]
Female	15 (24)	5 (8)	
Male	48 (76)	55 (92)	
Age (years), mean (SD)	33.0 (3.6)	37.36 (2.8)	*<.001*
Career stage, n (%)			—[d]
Specialist	14 (22)	57 (95)	
Master's trainee	9 (14)	2 (3)	
Service medical officer (resident)	40 (64)	1 (2)	
Experience, mean (SD)			*<.001*
Duration of orthopedic training (years)	4.5 (3.2)	9.6 (3.3)	
Number of tibia nailing cases performed as a primary surgeon	10 (12)	—	
Number of tibia nailing cases performed as an assistant surgeon	15 (19)	—	
Number of femoral nailing cases performed as a primary surgeon	10 (13)	15 (20)	
Number of femoral nailing cases performed as an assistant surgeon	13 (16)	16 (24)	

[a]Averages include SDs, and *P* values when significant are italicized.

[b]Cohort 1: n=63.

[c]Cohort 2: n=60.

[d]Not applicable.

Completion Time for Tibia Nailing Exercise

For each procedure, we compared the synthetic bone completion time among participants who performed VR practice before the synthetic bone task versus those who performed it after. For cohort 1, the survey response rates (proportion of participants who completed the postexercise survey) varied from 80% (50/63) to 92% (58/63). For cohort 2, the response rate was approximately 93% (55/60).

For the tibia nailing exercise, participants who trained on the VR module prior to performing the synthetic bone practical exercise were, on average, 5 minutes faster in completing the nailing exercise (P=.007; Table 2). A significant decrease in time needed to finish the VR module was also noted when participants first completed the synthetic bone practical exercise (*P*=.02; Table 2).

XSL•FO
RenderX

Table . Completion times for the tibia nailing exercise.

Cohort 1	VR[a] module then synthetic bone exercise, mean (95% CI)	Synthetic bone exercise then VR module, mean (95% CI)	*P* [b] *value*
Time (min) to complete synthetic bone exercise	18.48 (16.19-20.77)	23.43 (20.8-26.01)	*.007*
Time (min) to complete VR module	34.97 (31.51-38.44)	29.82 (27.43-32.2)	*.02*

[a]VR: virtual reality.

[b]Italicized *P* values are significant.

Completion Time for Femoral Nailing Exercise

Participants from cohort 1 who trained on the VR module prior to performing the synthetic bone practical exercise were, on average, approximately 7 minutes faster in completing the femoral nailing exercise (*P*<.01; Table 3). No significant difference in the VR module completion time was noted after participants performed the synthetic bone exercise (*P*=.09; Table 3). Similarly, in cohort 2, participants were on average approximately 4 minutes faster in completing the synthetic bone practical exercise after training on the VR module (*P*=.05; Table 3), and no significant difference in the VR module completion time was noted (*P*=.54; Table 3).

Table . Completion times for the femoral nailing exercise.

Cohort and outcome	VR[a] module then synthetic bone exercise, mean (95% CI)	Synthetic bone exercise then VR module, mean (95% CI)	*P*[b] value
Cohort 1			
Time (min) to complete synthetic bone exercise	20.71 (18.11-23.32)	27.55 (24.25-30.86)	*<.01*
Time (min) to complete VR module	18.65 (16.82-20.48)	16.69 (15.36-18.01)	.09
Cohort 2			
Time (min) to complete synthetic bone exercise	16.51 (15.03-17.98)	20.73 (16.89-24.58)	*.05*
Time (min) to complete VR module	21.57 (19.65-23.49)	22.45 (20.42-24.48)	.54

[a]VR: virtual reality.

[b]Italicized *P* values are significant.

Impact of VR Training on Self-Assessed Performance

Analysis of the qualitative questionnaire showed that after VR training, a higher number of participants rated the ease of understanding the nailing practical exercise in the “extremely easy” and/or “very easy” categories. This trend (results not statistically significant) was also noticed when participants were asked to evaluate their performance during the synthetic bone exercise (with the exception of cohort 1 for femoral nailing). This observation was seen for both the tibial and femoral nailing exercises (Table 4 and Table 5).

XSL•FO
RenderX

Table . Impact of virtual reality (VR) training on understanding and performance during the tibia synthetic bone nailing exercise.

Tibia nailing outcomes	Synthetic bone exercise without VR training, n (%)	Synthetic bone exercise with VR training, n (%)
How would you rate the ease of understanding the practical?		
Extremely easy	0 (0)[a]	4 (13)[b]
Very easy	9 (30)[c]	10 (33)[d]
Easy	17 (57)[d]	10 (33)[d]
Somewhat easy	4 (13)[b]	6 (20)[c]
Not at all easy	0 (0)[a]	0[a] (0)
How would you rate your performance in this practical?		
Excellent	2 (7)[b]	4 (13)[b]
Very good	9 (30)[c]	13 (43)[d]
Good	19 (63)[d]	13 (43)[d]
Poor	0[a] (0)	0[a] (0)
Very poor	0[a] (0)	0[a] (0)

[a]Lowest frequency of responses.

[b]Moderate frequency of responses.

[c]High frequency of responses.

[d]Highest frequency of responses.

Table . Impact of prior virtual reality (VR) training on understanding and performance during the femoral synthetic bone nailing exercise.

Femoral nailing outcomes	Cohort 1		Cohort 2	
	Synthetic bone without VR training, n (%)	Synthetic bone with VR training, n (%)	Synthetic bone without VR training, n (%)	Synthetic bone with VR training, n (%)
How would you rate the ease of understanding the practical?				
Extremely easy	0 (0)[a]	4 (13)[b]	1 (3)[a]	4 (15)[b]
Very easy	3 (10)[c]	5 (16)[b]	8 (29)[c]	9 (33)[c]
Easy	21 (73)[d]	13 (42)[d]	13 (46)[d]	13 (48)[d]
Somewhat easy	3 (10)[c]	8 (26)[c]	5 (18)[b]	1 (4)[a]
Not at all easy	2[b] (7)	1 (3)[a]	1 (3)[a]	0 (0)[a]
How would you rate your performance in this practical?				
Excellent	0[a] (0)	2 (7)[b]	5 (18)[c]	4 (15)[b]
Very good	10 (35)[c]	7 (23)[c]	3 (11)[b]	10 (37)[c]
Good	16 (55)[d]	19 (61)[d]	18 (64)[d]	12 (44)[d]
Poor	3 (10)[b]	2 (7)[b]	2 (7)[b]	1 (4)[b]
Very poor	0[a] (0)	1 (3)[a]	0 (0)[a]	0 (0)[a]

[a]Lowest frequency of responses.

[b]Moderate frequency of responses.

[c]High frequency of responses.

[d]Highest frequency of responses.

XSL•FO
RenderX

Learning Preferences

When asked to compare VR with conventional methods of learning (lectures and synthetic bone practical exercises), participants were strongly in favor of VR as a supplement to conventional methods of learning. Only 3.4% (4/117) chose VR modules alone as their preferred method of learning, 89% (104/117) preferred VR as an adjunct to conventional methods, and only 8% (9/117) believed VR added no value (Table 6).

When asked about the reverse, evaluating conventional methods of learning versus VR, 43% (50/117) preferred conventional methods of learning, 56% (65/117) preferred conventional methods as an adjunct to VR, and less than 2% (2/117, 1.7%) considered conventional methods to be outdated (Table 6).

Table . Participant learning preferences.

Learning preference	Participants, n (%)
How do you rate the VR[a] module vs conventional methods of learning (lectures and hands-on training with synthetic bones)?	
Prefer the VR module	4 (3.42)
Prefer VR as an adjunct to conventional methods of learning	104 (88.89)
VR does not add to conventional methods of learning	9 (7.69)
How do you rate conventional methods of learning (lectures and hands-on training with synthetic bones) vs the VR module?	
Prefer conventional methods of learning	50 (42.74)
Prefer conventional methods of learning as an adjunct to VR teaching	65 (55.56)
Conventional methods of learning are outdated	2 (1.71)

[a]VR: virtual reality.

Association of Age and Experience Level With Completion Times of the Femoral Nailing Exercise and VR Module

As the average age of the 2 cohorts were significantly different (33.0, SD 3.6 years for cohort 1 and 37.6, SD 2.8 years for cohort 2; *P*<.001; Table 1) and the average number of years of orthopedic training was also significantly different (4.5, SD 3.2 years vs 9.6, SD 3.3 years; *P*<.001; Table 1), we conducted an unplanned subanalysis examining the influence of age and experience on completion times of the femoral nailing exercise. Participants in cohort 2 (older, with more years of training) were significantly faster in completing the synthetic bone nailing exercise than those in cohort 1 (20.7 vs 27.5 minutes; *P*=.01; Table 7). However, when comparing the VR module completion times, cohort 1 (younger, with fewer years of training) was significantly faster than cohort 2 (18.6 vs 21.8 minutes; *P*=.03; Table 7).

Table . Effect of age and experience on completion time of nailing exercises.

Outcomes	Cohort 1 (time in minutes)	Cohort 2 (time in minutes)	*P*[a] value
Time to complete the femoral nailing *synthetic bone* exercise (without virtual reality training)	27.5	20.73	*.01*
Time to complete the femoral nailing *virtual reality module* (without prior synthetic bone exercise)	18.65	21.57	*.03*

[a]Italicized *P* values are significant.

Discussion

Principal Results

This study demonstrated immediate skill transfer from a single VR session to synthetic bone performance across 2 orthopedic procedures in a Malaysian residency context.

Our primary finding suggests that VR supports rapid skill acquisition and transfer to synthetic bone models. Additionally, we explored learning preferences, which could inform future curriculum design in low-resource countries. Although traditional hands-on training remains highly valued, most participants viewed VR positively as a tool complementary to conventional training methods. Our qualitative data further support VR's effectiveness, with participants reporting a better understanding of the practical exercises and higher self-rated performance after VR training.

An interesting secondary finding emerged from our unplanned subanalysis comparing cohorts of different experience levels. The more experienced participants performed the conventional synthetic bone femoral nailing exercise significantly faster than the less-experienced group. This result confirms that conventional surgical skills improve with years of practice.

In contrast, the younger and less-experienced cohort completed the VR module significantly faster. This finding suggests that younger trainees may adapt more quickly to digital learning environments, possibly due to greater familiarity with technology. However, given baseline differences and the

XSL•FO
RenderX

pragmatic study design, the analysis of age and experience is exploratory and should not be interpreted as causal.

Several practical implications emerge from our findings. First, in the context of Malaysia's orthopedic training landscape, where there is a need to train more surgeons to meet the World Health Organization–recommended ratio of 1:30,000 population, VR offers a scalable supplement that could help standardize training while optimizing the use of limited resources. With limited access to cadaveric specimens and economic constraints, VR platforms could provide residents with additional repetitions of procedures before they enter the operating room.

Second, the demonstrated reduction in time to complete practical exercises after VR training shows the transferability of skills from VR to real-life settings. Incorporating VR into preprocedural preparation could improve operating room efficiency. This is particularly valuable in resource-constrained settings where operating room time is limited and maximizing throughput is essential.

Third, the improved self-assessment of performance after VR training indicates enhanced procedural confidence, which could translate to reduced stress and potentially better outcomes in actual surgical situations. The combination of faster performance and increased understanding suggests that VR training may help flatten the learning curve for complex surgical procedures and offers a solution that complements rather than replaces traditional training methods [16,17].

Comparison With Prior Work

The growing body of evidence on the use of VR simulations for orthopedic surgery skills training and education has resulted in VR being included as part of the third technological wave in orthopedics [18]. Although there are still some questions concerning cost-effectiveness and skill retention using VR simulations, there is no doubt about VR's effectiveness in surgical training, particularly in orthopedics, where procedural repetition and spatial understanding are critical [19,20]. Our work aligns well with this literature, as we demonstrate immediate skill transfer after a single VR training intervention for not 1 but 2 orthopedic procedures.

A strong preference for a mixed-learning approach (conventional and VR) among participants reflects the need to rethink the role of VR in surgical skills training, especially in the context of low-resource settings [21,22], and to offer a scalable solution that complements rather than replaces traditional training methods. Acquiring VR headsets would be a 1-time investment for the training facility or could be sourced from companies willing to sponsor the event. On the other hand, skills training on synthetic bones requires the purchase of bone models for every participant and training event.

Strengths

This study benefits from several methodological strengths. The crossover design allowed each participant to experience both training modalities, reducing individual variation as a confounding factor. The relatively large sample size (N=123) and the inclusion of 2 distinct cohorts with different experience levels enhance the generalizability of our findings. Additionally, testing 2 different nailing procedures (tibia and femoral) demonstrates that the observed effects are not procedure specific.

Limitations

Several limitations warrant consideration when interpreting our results. First, completion times for the synthetic bone exercises were self-reported, whereas VR module completion times were captured by the software. This could introduce a reporting bias between the 2 modalities. Although we provided participants with standardized instructions, reporting errors cannot be excluded. Future studies could benefit from objective timing methods or video recording of procedures for more accurate measurement. Second, we evaluated effectiveness primarily through completion time, which, while important, represents only 1 dimension of surgical competence. Additional outcome measures such as procedural accuracy, error rates, or expert assessment of technique would provide a more comprehensive evaluation of VR's impact on skill acquisition. Third, our study assessed immediate skill transfer but did not include follow-up assessments to determine the durability of learned skills. Longitudinal studies would be valuable to assess whether VR-acquired skills persist over time and translate to improved performance in actual surgical cases. Finally, this was a pragmatic, course-embedded evaluation. Beyond course attendance and consent, we did not apply formal inclusion or exclusion criteria, which may contribute to baseline heterogeneity. Using immersive VR headsets has been shown to cause dizziness or nausea (cybersickness) [23]. Although we did not experience any such cases in this study, it needs to be considered while planning similar interventions.

Conclusions

This study provides new evidence that even a single, brief VR training session can significantly decrease completion times of simulated nailing procedures. This observation, coupled with improved self-reported procedural understanding and performance, suggests that VR represents a valuable educational tool for orthopedic skills training.

Although conventional hands-on training remains an integral part of residency training, there is a strong preference for a mixed approach incorporating VR technologies. In the context of Malaysia's need to train more orthopedic surgeons, VR training could offer a scalable complement to traditional methods that could help standardize training while optimizing resource use. This study also highlights the need for tailored implementation strategies that account for varying levels of technological familiarity. Finally, as VR technology continues to evolve and become more accessible, its integration into orthopedic training curricula represents a promising approach to enhance surgical education.

XSL•FO
RenderX

Acknowledgments

The authors thank the AO (Arbeitsgemeinschaft für Osteosynthesefragen) Foundation for organizing and running the courses where this study took place and the participants for agreeing to take part in this study. The authors thank Johnson & Johnson MedTech Malaysia for providing the virtual reality headsets and for in-kind support to run the course. The authors also thank the course chairperson, Dr Nazari Ahmad Tarmuzi, for making faculty available to support the study and Iris Lam for logistic support.

Funding

Financial and logistic support for conducting the educational event was provided by the AO Foundation, through an education grant from Johnson & Johnson MedTech to the AO Foundation.

Data Availability

All data generated and analyzed during this study are included in this published article.

Authors' Contributions

SC and KFM designed the study and collected the data. All authors contributed to data analysis, manuscript drafting, and revision of the manuscript.

Conflicts of Interest

None declared.

References

1. Babineau TJ, Becker J, Gibbons G, et al. The “cost” of operative training for surgical residents. Arch Surg 2004 Apr;139(4):366-369. [doi: 10.1001/archsurg.139.4.366] [Medline: 15078701]
2. Marchand LS, Sciadini MF. Simulation training in fracture surgery. J Am Acad Orthop Surg 2020 Nov 1;28(21):e939-e947. [doi: 10.5435/JAAOS-D-20-00076] [Medline: 32796368]
3. Hasan LK, Haratian A, Kim M, Bolia IK, Weber AE, Petrigliano FA. Virtual reality in orthopedic surgery training. Adv Med Educ Pract 2021;12:1295-1301. [doi: 10.2147/AMEP.S321885] [Medline: 34785971]
4. Li T, Yan J, Gao X, et al. Using virtual reality to enhance surgical skills and engagement in orthopedic education: systematic review and meta-analysis. J Med Internet Res 2025 May 30;27:e70266. [doi: 10.2196/70266] [Medline: 40446337]
5. Ntakakis G, Plomariti C, Frantzidis C, Antoniou PE, Bamidis PD, Tsoulfas G. Exploring the use of virtual reality in surgical education. World J Transplant 2023 Feb 18;13(2):36-43. [doi: 10.5500/wjt.v13.i2.36] [Medline: 36908307]
6. Makransky G, Mayer RE. Benefits of taking a virtual field trip in immersive virtual reality: evidence for the immersion principle in multimedia learning. Educ Psychol Rev 2022;34(3):1771-1798. [doi: 10.1007/s10648-022-09675-4] [Medline: 35475019]
7. Blumstein G, Zukotynski B, Cevallos N, et al. Randomized trial of a virtual reality tool to teach surgical technique for tibial shaft fracture intramedullary nailing. J Surg Educ 2020;77(4):969-977. [doi: 10.1016/j.jsurg.2020.01.002] [Medline: 32035854]
8. Walbron P, Common H, Thomazeau H, et al. Virtual reality simulator improves the acquisition of basic arthroscopy skills in first-year orthopedic surgery residents. Orthop Traumatol Surg Res 2020 Jun;106(4):717-724. [doi: 10.1016/j.otsr.2020.03.009] [Medline: 32284277]
9. Unal YC, Adanas C, Kaya S, Ozkan S, Ismailov U. Virtual reality headset versus traditional training in orthopedic surgery: a comparative study on real surgical performance in total knee arthroplasty. Medicine (Baltimore) 2024 Nov 22;103(47):e40615. [doi: 10.1097/MD.0000000000040615] [Medline: 39809210]
10. Crockatt WK, Confino JE, Kopydlowski NJ, Jobin CM, Levine WN. Comparing skill acquisition and validity of immersive virtual reality with cadaver laboratory sessions in training for reverse total shoulder arthroplasty. JB JS Open Access 2023;8(3):e22.00141. [doi: 10.2106/JBJS.OA.22.00141] [Medline: 37415724]
11. Wan T, Liu K, Li B, Wang X. Validity of an immersive virtual reality training system for orthognathic surgical education. Front Pediatr 2023;11:1133456. [doi: 10.3389/fped.2023.1133456] [Medline: 37033170]
12. Hiemstra LA, Williams H, Sasyniuk TM, Kerslake S. Orthopaedic sport medicine surgeons and fellows value immersive virtual reality for improving surgical training, procedural planning, and distance learning. J ISAKOS 2024 Aug;9(4):769-773. [doi: 10.1016/j.jisako.2024.05.002] [Medline: 38734310]
13. Please H, Narang K, Bolton W, et al. Virtual reality technology for surgical learning: qualitative outcomes of the first virtual reality training course for emergency and essential surgery delivered by a UK-Uganda partnership. BMJ Open Qual 2024 Jan 29;13(1):e002477. [doi: 10.1136/bmjoq-2023-002477] [Medline: 38286564]
14. Orthopaedic surgery postgraduate training in Malaysia. Majlis Dekan Fakulti Perubatan Universiti Awam Malaysia. 2020. URL: https://medicine.um.edu.my/pdf/postgraduate/NPMC/NPMC_OrthopaedicSurgeryPGTraining.pdf [accessed 2026-04-18]

XSL•FO
RenderX

15. Ibrahim S. The evolution of postgraduate orthopaedic training in Malaysia. International Society of Orthopaedic Surgery and Traumatology (SICOT). 2015. URL: https://www.sicot.org/enewsletter-75-editorial [accessed 2026-04-18]
16. Seymour NE, Gallagher AG, Roman SA, et al. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Ann Surg 2002 Oct;236(4):458-463. [doi: 10.1097/00000658-200210000-00008] [Medline: 12368674]
17. Vaishya R, Misra S, Singh LP, et al. Surgical training and learning curves in orthopedic surgery – a review. J Orthop Rep 2025 Aug;4(2):100584. [doi: 10.1016/j.jorep.2025.100584]
18. Zhang J, Ye Z. Digital orthopedics: the third technological wave of orthopedics. Intell Med 2025 May;5(2):91-94. [doi: 10.1016/j.imed.2024.09.003]
19. Lohre R, Bois AJ, Pollock JW, et al. Effectiveness of immersive virtual reality on orthopedic surgical skills and knowledge acquisition among senior surgical residents: a randomized clinical trial. JAMA Netw Open 2020 Dec 1;3(12):e2031217. [doi: 10.1001/jamanetworkopen.2020.31217] [Medline: 33369660]
20. Cevallos N, Zukotynski B, Greig D, Silva M, Thompson RM. The utility of virtual reality in orthopedic surgical training. J Surg Educ 2022;79(6):1516-1525. [doi: 10.1016/j.jsurg.2022.06.007] [Medline: 35821110]
21. Cate G, Barnes J, Cherney S, et al. Current status of virtual reality simulation education for orthopedic residents: the need for a change in focus. Global Surg Educ 2023;2(1):46. [doi: 10.1007/s44186-023-00120-w] [Medline: 38013875]
22. Heo K, Cheng S, Joos E, Joharifard S. Use of innovative technology in surgical training in resource-limited settings: a scoping review. J Surg Educ 2024 Feb;81(2):243-256. [doi: 10.1016/j.jsurg.2023.11.004] [Medline: 38161100]
23. Cossio S, Chiappinotto S, Dentice S, et al. Cybersickness and discomfort from head-mounted displays delivering fully immersive virtual reality: a systematic review. Nurse Educ Pract 2025 May;85:104376. [doi: 10.1016/j.nepr.2025.104376] [Medline: 40267853]

Abbreviations

AO: Arbeitsgemeinschaft für Osteosynthesefragen
VR: virtual reality

Edited by N Rohatgi; submitted 19.Jun.2025; peer-reviewed by G Rodriguez-Fuentes, OR Babalola; revised version received 11.Mar.2026; accepted 19.Mar.2026; published 06.May.2026.

Please cite as:
Chatterjee S, Mohammad KF, Ghidinelli M
Evaluating the Impact of Virtual Reality on Orthopedic Trauma Skills Acquisition Among Surgical Residents: Randomized Crossover Study
JMIR Perioper Med 2026;9:e79343
URL: https://periop.jmir.org/2026/1/e79343
doi:10.2196/79343

© Sandipan Chatterjee, Khairul Faizi Mohammad, Monica Ghidinelli. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 6.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Novel Customizable Datamart and Tableau Dashboard to Monitor Multiple Enhanced Recovery After Surgery Programs: Development and Validation Study

Sunitha Margaret Singh[1], MD; Susannah Oster[2], BA; Efrat Bolze[3], BA; Aaron Sasson[4], MD; James Nicholson[5], MD; Elliott Bennett-Guerrero[2], MD

[1]Department of Perioperative Surgical Services, Stony Brook University Medical Center, 101 Nichols Road, Stony Brook, NY, United States

[2]Department of Anesthesiology, Stony Brook University Medical Center, Stony Brook, NY, United States

[3]Enterprise Analytics, Stony Brook Medicine Information Technology (SBMIT), St. James, NY, United States

[4]Department of Surgery, Stony Brook University Medical Center, Stony Brook, NY, United States

[5]Department of Orthopedics, Stony Brook University Medical Center, Stony Brook, NY, United States

Corresponding Author:
Sunitha Margaret Singh, MD
Department of Perioperative Surgical Services, Stony Brook University Medical Center, 101 Nichols Road, Stony Brook, NY, United States

Abstract

Background: Enhanced recovery after surgery (ERAS) programs bundle evidence-based interventions to standardize care, expedite recovery, and improve outcomes. As ERAS programs have expanded, it has become clear that a major challenge is monitoring the compliance of bundle elements and outcomes to feedback performance to stakeholders and guide changes. Manual data abstraction is onerous and not feasible. Reliance on receiving new reports from busy health system IT groups is challenging. Therefore, we sought to address this unmet need at our hospital by developing a novel ERAS Datamart system.

Objective: Our objectives were to develop a novel Datamart and Tableau dashboard to (1) enable continuous analysis of data, harvested directly from the electronic medical record (EMR), measure compliance and outcomes, and (2) enable end users (e.g., an ERAS coordinator) to create reports customized based on surgical procedure types, requested data variables, and custom date ranges.

Methods: After "buy-in" from hospital leadership and other stakeholders, data metrics were identified and categorized according to phase of care, that is, preoperative, intraoperative, and postoperative. A multidisciplinary team reviewed *International Classification of Diseases, Tenth Revision* procedure codes to capture EMR data for patients undergoing ERAS procedures. IT was given a master list with metric names, definitions, and screenshots of the discrete field in the EMR to assist with building the metrics. Validations of the novel Datamart were done against known ERAS patient populations maintained by the surgery clinic.

Results: The Datamart and Tableau dashboard has been built, is functional, and contains over 17,000 patients across 5 ERAS service lines: colorectal (n=1742), joint replacement (n=4235), surgical oncology (n=941), bariatric (n=1130), and cesarean section (n=9390). Currently, 56 metrics spanning the perioperative period have been validated across these populations. Reports can be tailored according to patients, time frames, and metrics. If desired, patient-level raw data can be exported for statistical analyses. Two use cases (total joint replacement and surgical oncology ERAS programs) are presented showing how the Datamart can be used.

Conclusions: Discrete fields within an EMR can be successfully captured into a novel Datamart and visualized using a custom Tableau dashboard for providing stakeholder feedback, facilitating quality improvement analyses, and auditing pathways.

(JMIR Perioper Med 2026;9:e82472) doi:10.2196/82472

KEYWORDS

data monitoring; enhanced recovery; enhanced recovery after surgery; ERAS; perioperative outcomes; quality improvement; web platform

XSL•FO
RenderX

Introduction

Enhanced recovery after surgery (ERAS) programs have transformed perioperative care by implementing evidence-based interventions that aim to standardize patient care and management, decrease resource utilization, expedite recovery, and improve patient outcomes [1-3]. The success and efficacy of ERAS programs are most likely achieved through the implementation of a comprehensive approach that bundles care for patients undergoing elective surgery, encompassing approximately 20 care elements [1,4].

It is increasingly recognized that as ERAS programs increase in size, it is very challenging to monitor and track bundle elements to feedback performance and guide outcomes. Traditional methods usually rely on manual data abstraction, frequent reports generated by hospital IT systems, or the use of third-party data warehouses. As ERAS programs grow in size and complexity, manual data abstraction becomes impractical due to time demands, error risk, and challenges with real-time analysis. Basic data points, such as length of stay, are easier to track, but capturing complex metrics, such as total opioid use (e.g., oral morphine equivalents), is often not feasible. Reliance on an IT report strategy is typically limited by very long delays in obtaining data reports from hospital IT workers who are usually burdened with many requests. The use of third-party data warehouses, for example, ERAS Interactive Audit System (EIAS), offers an alternative but raises concerns about data security, control, costs, system downtimes, and limited flexibility [5]. To address the above limitations, our institution created a novel dynamic Datamart dashboard.

Methods

Overview

Stony Brook University Hospital is a tertiary care academic medical center on Long Island, New York. Its first ERAS program (lumbar spine fusion) began in 2016, and an additional 9 ERAS programs were subsequently added. As the program grew, this revealed the unmet need for how to efficiently capture and monitor compliance and outcomes across a large number of patients.

As described in more detail below, the process for creating this system included: (1) leadership support and data governance, (2) validated identification of relevant patients to be included in the Datamart, (3) metric identification and validation, and (4) Tableau visualization as the user interface.

Leadership Support and Data Governance

Under an institutional quality assurance program, a guideline was developed to map the creation of a novel Datamart and Tableau dashboard and govern the data extracted. To prioritize this effort, a value statement was presented to institutional leadership. This statement provided background information on the institution's ERAS programs, highlighted their prior success, and outlined the intended purpose of the Datamart and Tableau dashboard, such as monitoring and improving compliance, reducing errors associated with manual data abstraction, and limiting frequent report requests made to IT.

After approval, a Global ERAS Data Governance plan was conceived with policies and procedures for protecting and using ERAS data. These included how the data would be stored and protected, who would have access to data, and how data would be managed (e.g., requests for aggregate and patient-level reports, quality assurance (QA) analyses, and institutional review board–approved research projects). The Global ERAS Data Governance plan was subsequently signed by applicable departmental chairpersons to ensure data analysis was conducted in accordance with institutional standards and to protect against breaches of protected health information.

Identification of Relevant Patients

Choice of International Classification of Diseases, Tenth Revision Codes Methods

We considered several possible strategies for identifying relevant patients for a given ERAS pathway. The hospital's operating room schedule, that is, planned surgical procedure, provides information on the "planned" procedure; however, it does not accurately reflect the "actual" surgery performed. Current procedural terminology professional billing codes were not used since the hospital's IT department did not have direct access to them. Therefore, we decided to use the *International Classification of Diseases, Tenth Revision* (*ICD-10*) Procedure Coding System (PCS) since this was feasible and these codes are believed to be accurate as they are used for hospital billing purposes.

To ensure the accurate identification of the ERAS patient population, the team collaborated closely with surgical leads from each ERAS pathway. Surgical procedures were sent to the coding department to identify the *ICD-10* PCS associated with specific procedures. The coding department supplied the leading 4 digits of all *ICD-10* PCS for the specified procedures. These digits encompass the section, body part, root operation, and where relevant, the body part of the given procedure. This preliminary list was forwarded to the IT department, which then appended the remaining digits of the *ICD-10* PCS, corresponding to the approach, device, and qualifier, for the procedures performed. Subsequently, a multidisciplinary team reviewed the complete *ICD-10* PCS to ensure the accurate capture of electronic medical record (EMR) data for patients undergoing ERAS procedures. This task required careful attention due to the complexities and overlaps within *ICD-10* PCS across multiple procedure types.

To ensure a comprehensive approach, a strategy was developed that included both *ICD-10* procedure codes and additional criteria, such as surgical case procedure name and other associated details (eg, associated *ICD-10* PCS, ambulatory surgical center vs main operating room, surgeon). This established a robust, multistep process for accurately pinpointing the desired patient population. This method ensured that only patients who received care associated with an ERAS pathway were included, enhancing the precision and reliability of the data captured for monitoring and analysis.

Population Validation Methods

Next, we validated the accuracy of using these *ICD-10* procedure codes to identify the desired patient population. This

XSL•FO
RenderX

validation aimed to identify instances where incorrect patients (non-ERAS patients) were erroneously included or, at the other extreme, ERAS patients were missing (i.e., not included). Admit type was then utilized to refine the patient population based on the urgency of admission, categorized as urgent, emergent, or elective. Since ERAS patients almost always fall into the "elective" category, this refinement allowed for an additional method for validating patient populations. Exception reports were generated and scheduled for automated reporting of patients who met the *ICD-10* PCS criteria but were admitted urgently or emergently. These records were then cross-referenced with the EMR to verify their eligibility for inclusion in the Datamart, ensuring only accurate ERAS populations were maintained. Several validations of the Datamart data were then performed by comparing the dataset against known ERAS patient populations maintained by the surgery clinic or stored in a REDCap (Research Electronic Data Capture) database.

Metric Identification and Validation Methods

ERAS programs usually include many (e.g., ≥20) best practice elements. For example, the use of nonopioid analgesics, such as acetaminophen and nonsteroidal medications, is prioritized to minimize opioid use [6]. Since there are no national benchmarks for these programs to identify metrics of interest, key stakeholders (e.g.surgeons, anesthesiologists, hospital quality department personnel) were engaged. Metrics were selected, defined, and aligned with institutional interest, key performance, and patient safety indicators, and the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) by these stakeholders. We also sought to identify metrics that would be broadly applicable to most elective surgical procedures since the proposed Datamart system would be used for all ERAS programs.

Given the variations in documentation that can occur across an institution, it was crucial to investigate how and when an EMR field was completed. For example, nursing staff across different surgical specialties might record specifics related to urinary catheterization in separate discrete fields within the EMR. Additionally, there could be discrepancies in other charting practices as well, such as documenting ambulation as the "number of steps" taken versus "number of laps" taken. Metrics were validated with each ERAS population to identify discrepancies and ensure that the discrete field identified could be applied to the majority of surgical populations. This validation of metrics was carried out in multiple phases (e.g., 6 - 10 metrics at a time with each population) to alleviate the burden of mass validation. Metric validations were performed against known ERAS populations using manual chart review to ensure accurate data extraction into the Datamart from discrete EMR fields.

Once finalized, the metrics were categorized according to phase of care: preoperative, intraoperative, postoperative, and discharge. IT was provided with metric names, definitions, and screenshots of the discrete fields in the EMR to assist with building the metrics. To compartmentalize the data in the Datamart, these phases were bucketed into 7 categories, encompassing various areas of care. These categories included patient characteristics, preoperative care, operating room, postoperative fluid, postoperative multimodal analgesia, postoperative opioid, and postoperative patient experience.

User Interface (Tableau Visualization) Methods

Tableau was chosen as the software for enabling a web-based dashboard user interface due to its on-demand filtering options and Health Insurance Portability and Accountability Act–compliant capabilities. An open query was established between our institution's EMR relational database and the Tableau dashboard. This setup allowed automated data extracts from the EMR to the Tableau dashboard according to the desired export rate. Multiple viewpoints and designs were trialed through various builds created by IT. Essential filtering options, such as specified time frames, surgical specialty, surgical procedures, and other specific metrics, were established and selected based on their relevance and informativeness for patients receiving care related to ERAS.

Ethical Considerations

This project was conducted as part of an institutionally approved quality improvement/quality assurance initiative aimed at optimizing perioperative care processes and monitoring ERAS program performance. In accordance with our institution's policies on human participants protections, quality improvement/quality assurance activities that are designed solely for internal program evaluation are not considered human participants research.

Results

Results are presented in the same sequence of events (1-4) as described in the Methods section.

Leadership Support and Data Governance

An ERAS data governance guidance plan was established. The document outlined the appropriate use of the ERAS data warehouse, including storage, protection from data breach and leak of protected health information, access to data, and ensuring data analysis is in accordance with institutional standards. Data access is managed by a data access group. Requests for data must be submitted in writing to the data access group for review and approval. Data can be used for quality and research scholarly activities.

Validation of Patient Selection Using *ICD-10* Procedures Codes

Initially, 50 metrics were trialed with *ICD-10* PCS for ERAS colorectal procedures. The patient population yielded from these codes returned many patients, more than 2-fold, compared to the known ERAS colorectal population. Preliminary validations revealed the need to (1) exclude surgery types such as emergent or urgent, (2) exclude overly broad or ambiguous *ICD 10* procedure codes, and (3) screen for potential inclusion of the planned surgical case procedure. After validating the colorectal population, the initial metrics were tested in the surgical oncology and joint replacement ERAS populations. We then backfilled data to 2015 to include pre-ERAS populations.

XSL•FO
RenderX

After final validations, the Datamart (last 6.5 y) contains more than 17,000 patients, consisting of colorectal (n=1742), surgical oncology (n=941), joint replacement (n=4235), bariatric (n=1130), and cesarian section (n=9390). In addition, the Datamart contains over 3000 archived cases from 2015 to 2018.

Metric Identification and Validation

Over 100 metrics of potential interest, spanning the perioperative period and including patient characteristics, were identified. Several of these metrics represent the same metric measured at different time points over several days, such as peak pain on postoperative day (POD) 0, peak pain on POD 1, and peak pain on POD 2. The metrics were categorized according to the phase of care: preoperative, intraoperative, and postoperative. Currently, there are 56 metrics (demographics, surgical and anesthesia care, and postoperative endpoints) in the Datamart Textbox 1.

XSL•FO
RenderX

Textbox 1. Current metrics active in the Datamart.

1. Patient characteristics
   - Patient age, mean or median
   - Body mass index, mean or median
   - Diabetes diagnosis, #/%
   - Current smoker, #/%
   - Chronic opioid use (regular opioid use as listed on home medications), #/%
   - Associated diagnosis code, cancer, #/%
   - Associated diagnosis code, inflammatory bowel disease, #/%
   - Actual procedure completed, #/%
2. Preoperative care
   - Preprocedural bowel prep, #/%
   - Preprocedural oral antibiotics, #/%
   - Total functional status score (sum of activities of daily living score on the day of surgery in the preoperative area), mean or median
   - Hemoglobin A_{1c} within 90 days (closest value prior to surgery start time), mean or median
   - Hemoglobin within 30 days (closest value prior to surgery start time), mean or median
3. Operating room
   - Laparoscopic procedure, #/%
   - Rectal procedure, #/%
   - Length of surgery, minutes, mean or median
   - Total intravenous (IV) fentanyl, mcg, mean or median
   - Received epidural or spinal, #/%
   - Intraoperative crystalloid IV fluids, mL (sum of normal saline, lactated ringers, and D5W), mean or median
4. Postoperative opioid
   - Day of surgery (DOS): oral morphine equivalent, mg, mean or median
   - Postoperative day (POD) 1: oral morphine equivalent, mg, mean or median
   - Postoperative day (POD) 2: oral morphine equivalent, mg, mean or median
5. Postoperative multimodal analgesia
   - Day of surgery
     - Oral acetaminophen, #/% of patients receiving any amount
     - IV acetaminophen, #/% of patients receiving any amount
     - IV or oral nonsteroidal anti-inflammatory drug (NSAID), #/% of patients receiving any amount
   - Postoperative day 1:
     - Oral acetaminophen, #/% of patients receiving any amount
     - IV acetaminophen, #/% of patients receiving any amount
     - IV or oral NSAID, #/% of patients receiving any amount
     - Multimodal agents, # of agents received, mean or median
   - Postoperative day 2:
     - Oral acetaminophen, #/% of patients receiving any amount
     - IV acetaminophen, #/% of patients receiving any amount
     - IV or oral NSAID, #/% of patients receiving any amount

XSL•FO
RenderX

- Multimodal agents, # of types of agents received, mean/median

- Postoperative day 3:
  - Oral acetaminophen, #/% of patients receiving any amount
  - IV acetaminophen, #/% of patients receiving any amount
- Total IV acetaminophen (DOS through POD 3), mg, mean or median
- Total oral acetaminophen (DOS through POD 3), mg, mean or median

6. Postoperative fluid
   - Day of surgery
     - Net input and output, mL, mean or median
     - Total input, mL, mean or median
   - · Postoperative day 1:
     - Net input and output, mL, mean or median
     - Total input, mL, mean or median

7. Postoperative patient experience
   - Postoperative patient-controlled analgesia use, hours, mean or median
   - Peak pain score on day of surgery, mean or median
   - Peak pain score on postoperative day 1, mean or median
   - Peak pain score on postoperative day 2, mean or median
   - Postoperative urinary straight catheterization on day of surgery, #/%
   - Postoperative urinary straight catheterization on postoperative day 1, #/%·
   - Postoperative insertion of urinary catheter (insertion occurs more than 2 hours after surgery stop or removal in operating room), #/%
   - Postoperative duration of urinary catheter, hours, mean or median
   - Time to first flatus or bowel movement, hours, mean or median
   - Postoperative duration of nasogastric tube, hours, mean or median
   - Postoperative insertion of nasogastric tube (insertion occurs more than 2 hours after surgery stop or removal in operating room), #/%
   - Delta creatinine (peak value within 72 hours postoperatively minus preoperative value), mean or median
   - Postoperative length of stay, days, mean or median
   - Discharge opioids (oral morphine equivalents) prescribed, mg, mean or median

Unless specified otherwise, units of measurement for continuous variables are mg and categorical end points are yes/no.

Tableau Visualization

Overview

The Datamart dashboard was designed to incorporate a user-friendly interface, allowing users to easily generate customizable reports. It offers tools to filter data by patient population, time frames, and specific metrics, enabling a focused analysis of ERAS outcomes (e.g., compare before and after implementation outcomes, track changes in protocols over specific time frames).

The user is first presented with a screening interface where patient population, time frames, and age parameters can be defined (Figure 1). By hovering over the colored square, the user can also review certain patient and surgery-specific information, such as associated *ICD-10* PCS, surgeon, and date of surgery (Figure 2). Patients can be deselected from the population if, for example, erroneous classification of an emergency trauma patient as an elective case.

XSL•FO
RenderX

Figure 1. User interface (screening). User interface depicting screening options. The user can select time frames, patient age range, and surgical population. The user is presented with the patient name, arrival and discharge date and time, patient age on day of surgery, qualifier for inclusion in the Datamart (*International Classification of Diseases, Tenth Revision* [*ICD-10*] *Procedure Coding System* [PCS] and/or surgical case procedure), and qualifier description. Patients can be selected or deselected on the right. Hovering over the colored square yields additional screening data (see Figure 2).

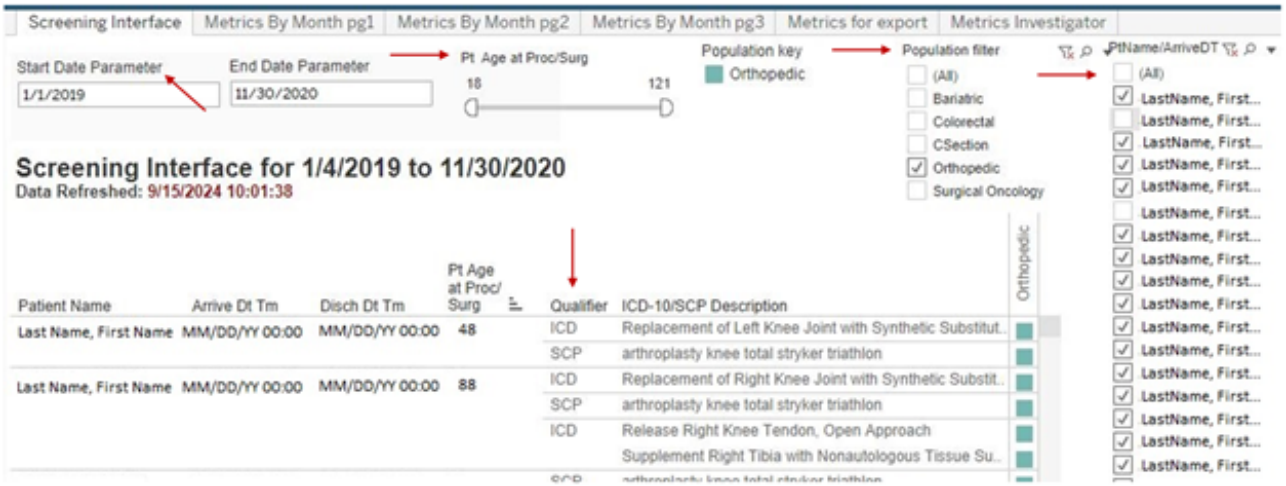
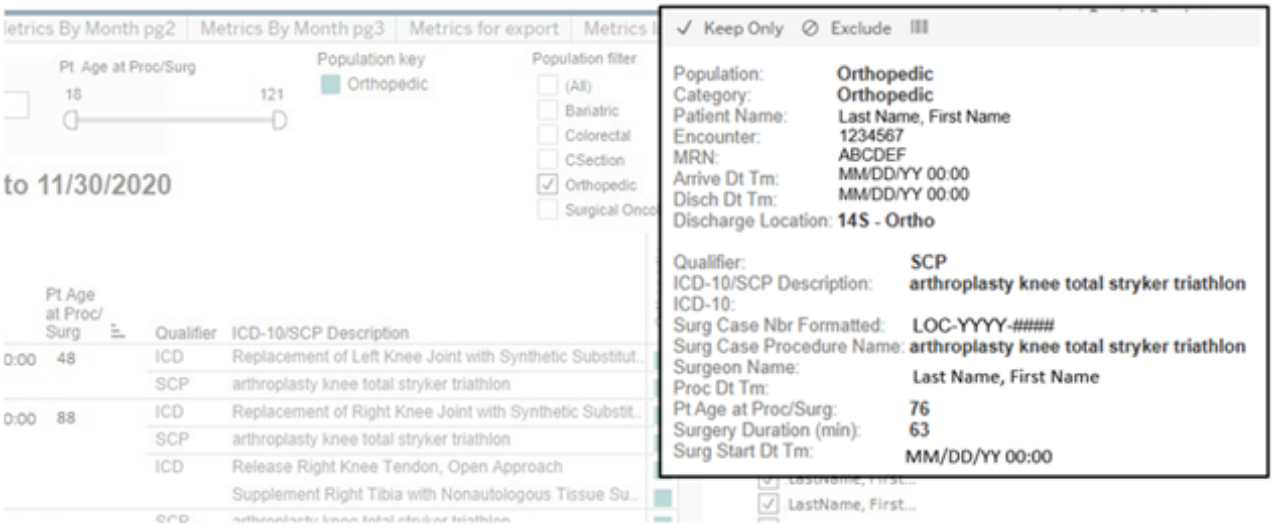


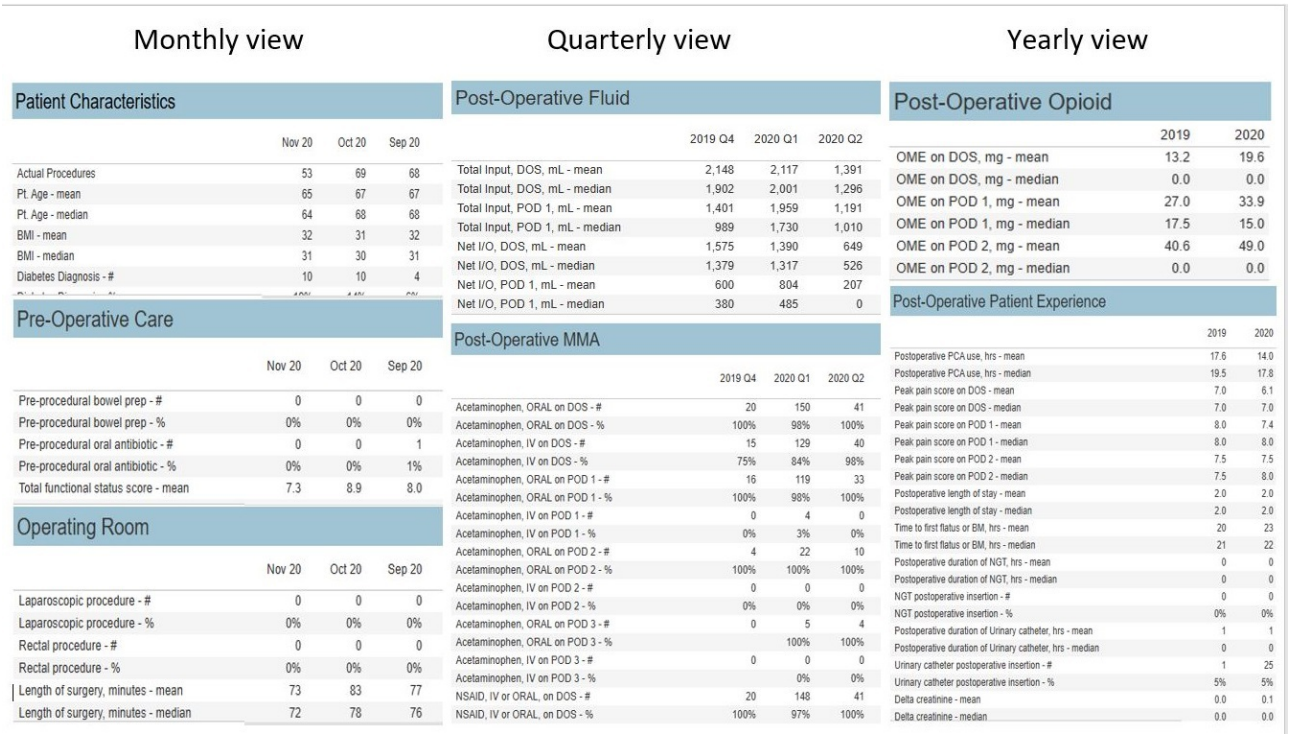
Figure 2. User interface (related screening). Additional screening information is available to the user. Hovering over the colored square next to the *International Classification of Diseases, Tenth Revision* (*ICD 10*)/surgical case procedure (SCP) description "pops out" information related to the selected case. This includes discharge location, surgical case number, surgeon name, and length of surgery.



On subsequent screens, the user can determine the metrics for analysis and how to view the data (Figure 3). Filters can be customized to focus on specific metrics or patient outcomes. For example, to focus on total acetaminophen use, one can view more details, such as "total IV acetaminophen on post-op day 1 (mean or median values)." Individual elements, such as opioid administration or fluid management details, can also be isolated for deeper analysis. Additionally, data can be displayed in weekly, monthly, quarterly, or yearly summaries and saved and exported as tables.

XSL•FO
RenderX

Figure 3. User interface (metrics). After procedures have been selected for review, the user can determine the metrics for analysis. Metrics can be viewed weekly, monthly, quarterly, or annually. Individual elements, such as opioid administration or fluid management details, can also be isolated for deeper analysis. DOS: day of surgery; IV: intravenous;nasogastric tube; MMA: multimodal analgesia; NGT: NSAID: nonsteroidal anti-inflammatory drug; OME: oral morphine equivalent; PCA: patient-controlled analgesia; POD: postoperative day.



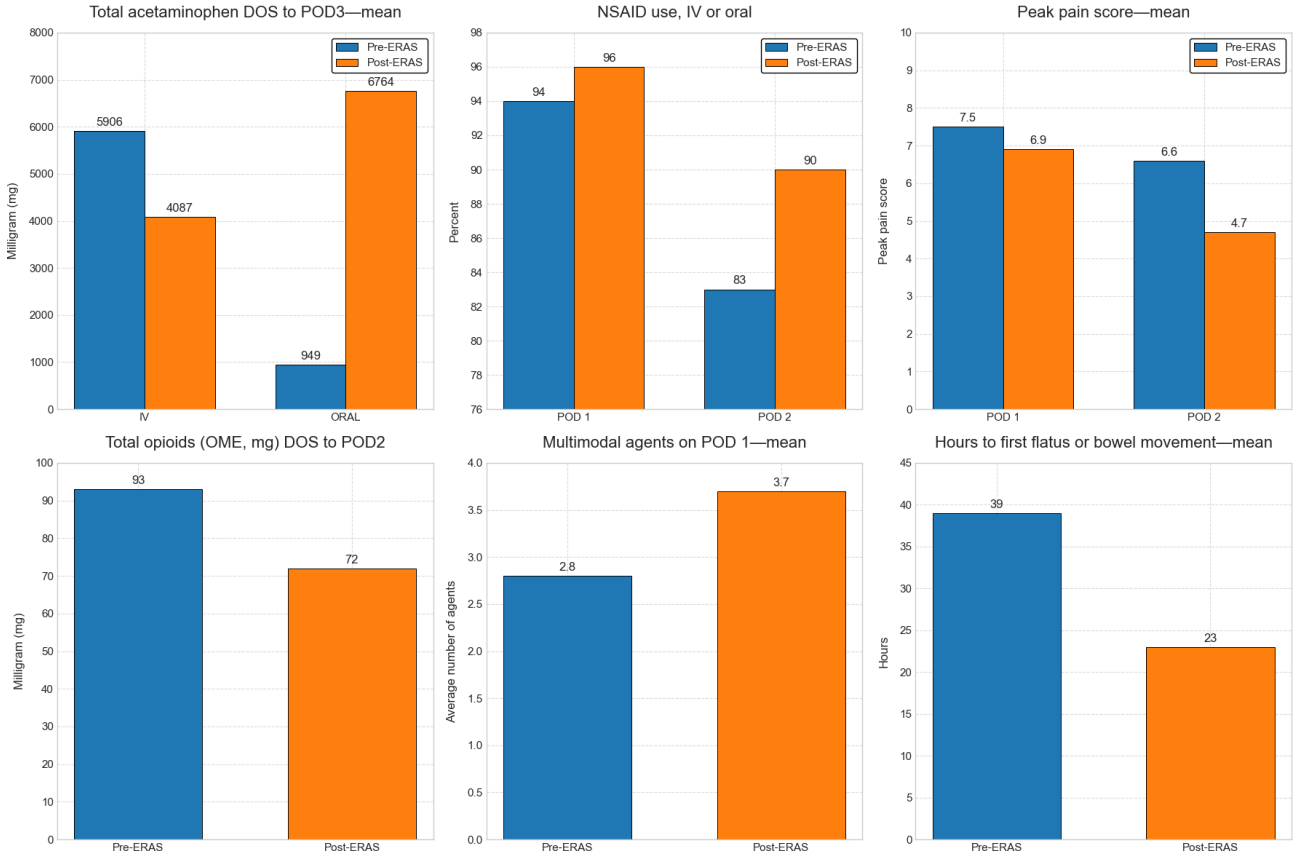
For further in-depth analyses, for example, inferential statistical analysis, users can export patient-level raw data in either comma-separated value or Excel format. These data can be exported in a deidentified manner, safeguarding against the unintentional disclosure of protected health information when exiting the secure Datamart dashboard. By using these features, the Datamart permits users to analyze ERAS metrics with precision, adapt to different populations, and investigate trends and outcomes.

Use Case Examples: ERAS Total Joint Replacement and Surgical Oncology

Using total joint replacement as a use case example, we sought to compare patient care prior to the initiation of ERAS (pre-ERAS n=693 cases) and 12 months post-ERAS implementation (n=563 cases). Figures 4 and 5 show the impact of an ERAS program on pain using the data harvested from the Datamart, notably an apparent improvement in several pain-related metrics. Of note, the Tableau dashboard does not calculate or show error bars for continuous variables; the calculation of these requires export of the raw data for statistical analysis. Therefore, we opted to present mean and median, since if they are similar, it suggests that the data are normally distributed and the mean value is not inflated due to a few large outliers. Therefore, for most day-to-day QA purposes, the presentation of mean and median is sufficient, but for more rigorous quantitative analysis, for example, inferential statistics for hypothesis testing, the export of patient-level raw data allows that capability.

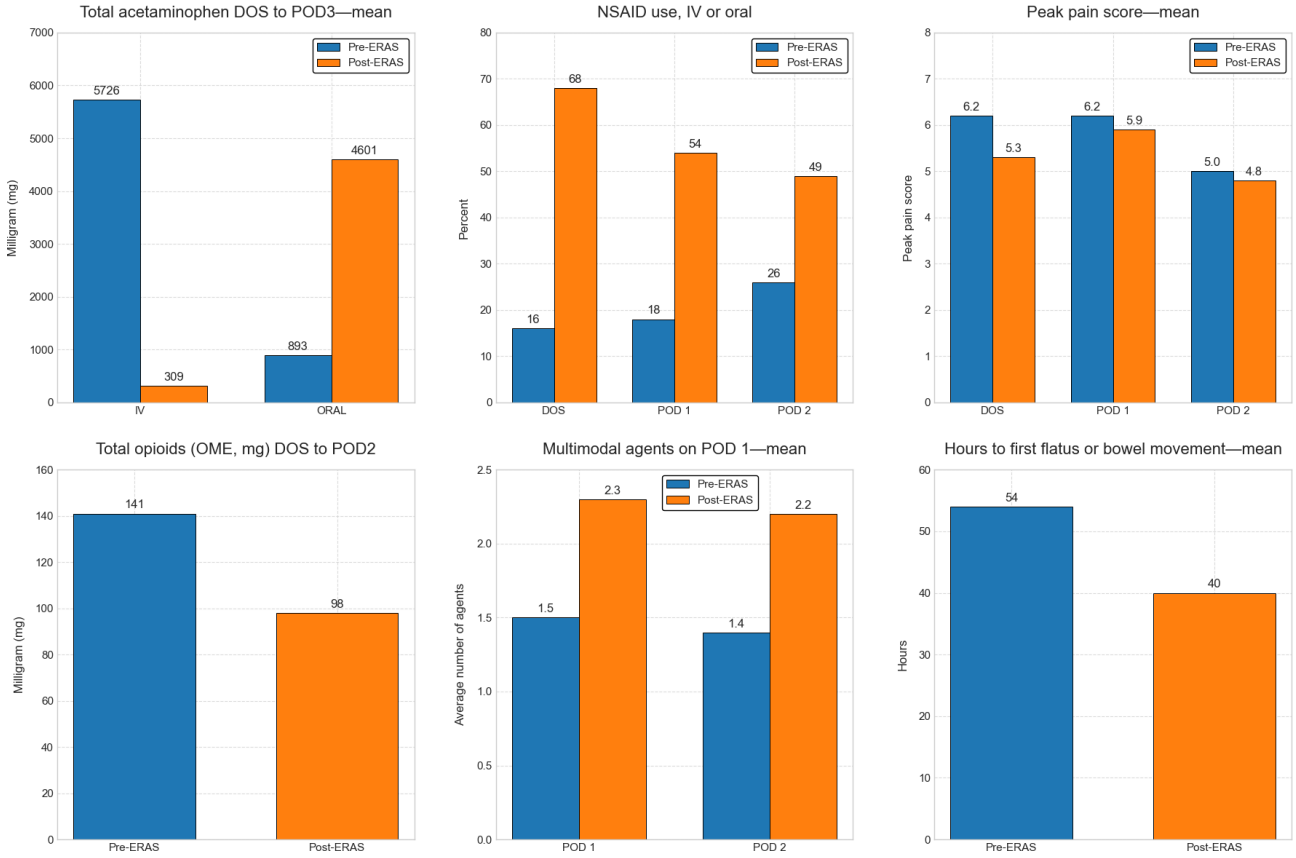
XSL•FO
RenderX

Figure 4. Enhanced recovery after surgery (ERAS) total joint replacement program: impact on pain. The Tableau dashboard does not display error bars for continuous variables; these require the export of patient-level raw data for statistical analysis, which is a capability of the Datamart. As ERAS pathways are primarily a quality assurance (QA) initiative, we report mean and median values, as their similarity suggests a roughly normal distribution without major outliers. DOS: day of surgery; IV: intravenous; NSAID: nonsteroidal anti-inflammatory drug; OME: oral morphine equivalent; POD: postoperative day.



XSL•FO
RenderX

Figure 5. Enhanced recovery after surgery (ERAS) surgical oncology program: impact on pain. The Tableau dashboard does not display error bars for continuous variables; these require export of patient-level raw data for statistical analysis, which is a capability of the Datamart. As ERAS pathways are primarily a quality assurance (QA) initiative, we report mean and median values, as their similarity suggests a roughly normal distribution without major outliers. DOS: day of surgery; IV: intravenous; NSAID: nonsteroidal anti-inflammatory drug; OME: oral morphine equivalent; POD: postoperative day.



We noticed similar findings for our Surgical Oncology ERAS program (Figure 5).

Discussion

Principal Results

The novel Datamart dashboard has been a transformative tool at our hospital by enabling continuous analysis of ERAS programs and patient outcomes. By automating data extraction from the EMR, the Datamart eliminates the need for manual data abstraction, likely reducing errors and improving data accuracy. This centralized system currently serves as the primary data resource for our institutional ERAS programs, streamlining the monitoring of bundle elements and outcomes at both the individual and aggregate levels. Successful adoption required early stakeholder engagement, quarterly dashboard reviews, and review at quarterly stakeholder meetings. Barriers included initial unfamiliarity with Tableau capabilities and competing clinical priorities. These were mitigated through targeted training and dashboard use in routine quality meetings.

Key features of the Datamart include categorized metrics across perioperative phases, tailored reporting capabilities, and the ability to export data. The level of granularity of the Datamart provides actionable insights into ERAS pathway efficacy, supporting evidence-based decision-making.

Limitations

Despite its many advantages, the Datamart has some limitations that warrant mention. As with any EMR-based form of data capture, it relies on accurate charting by clinicians, which is not infallible. For example, postoperative ambulation was not included in this iteration, as documentation practices are inconsistent. Moreover, patient-reported outcomes, such as satisfaction and functional recovery, are not currently documented in the EMR, which prevents their inclusion in the Datamart.

Accurately identifying patients on an ERAS pathway is another challenge. The Datamart relies on *ICD-10* PCS for surgical procedure identification, which can lead to the inadvertent inclusion of non-ERAS cases. For example, procedures such as hemorrhoidectomy (non-ERAS) and hemicolectomy (ERAS) share the same *ICD-10* PCS, requiring manual filtering to exclude nonrelevant cases. While procedures, such as total knee replacement or cesarean section, are easier to identify due to constrained coding options, consistent oversight is critical for accurate data classification.

An intentional 2-month lag in data import further limits real-time analysis. However, this delay ensures the accuracy of *ICD-10* coding and surgical procedure inclusion, contributing to the integrity of the dataset.

XSL•FO
RenderX

Comparison With Prior Work

Health care auditing has significantly evolved with the advent of digital dashboards, data warehouses, and interactive audit systems. Before widespread dashboard integration, many institutions relied on manual data abstraction and semiautomated systems. Manual abstraction is cumbersome and relies on human resources to extract data from the EMR to input into a repository for analysis (e.g., REDCap, Microsoft Excel) [7-9]. Semiautomated audits (e.g. IT-generated reports), which at our institution can take 9 months or longer, add to the resource burden. Some institutions combine EMR data extraction with administrative databases (e.g., NSQIP) to build a centralized reporting structure [2,10]. However, reliance on these reports poses challenges such as delayed access to real-time data, lack of customization, and risk of system downtimes.

For some institutions with ERAS pathways, third-party data warehouses (e.g., EIAS) may be an option [5]. While they offer additional storage and analysis options, third-party warehouses present issues of data security, higher costs, and limited control over information (see Table 1). In contrast, our in-house customizable Datamart can support diverse ERAS populations and continuous improvements through iterative refinements based on user feedback and technological advancements.

Table . Comparison between the Datamart and third-party options.

Feature	Datamart+Tableau	EIAS[a]	ACS[b] NSQIP[c]
Data control	Full institutional control over data	Vendor-controlled; limited flexibility	External benchmarking database
Customization	Highly customizable by procedure type, metrics, and time frame	Some customization possible, but limited	Standardized measures; less customizable
Real-time access	Near real-time (2-month lag for data integrity)	Typically delayed, relies on upload	Reports released quarterly or semiannually
Cost	Internal development; no licensing fees	Requires subscription or license fees	Expensive participation and data access
Data security	Remains within institutional firewall; HIPAA[d]-compliant	Data housed externally; possible security concerns	Data deidentified but externally stored
Metric flexibility	Fully institution-defined (≥56 metrics currently used)	Limited to ERAS[e]-recommended fields	Fixed set of standard metrics
Scalability	Easily scalable across services and use cases	Limited by system design and vendor	Only covers specific surgeries (e.g., colectomy)

[a]EIAS: ERAS Interactive Audit System.

[b]ACS: American College of Surgeons.

[c]NSQIP: National Surgical Quality Improvement Program.

[d]HIPAA: Health Insurance Portability and Accountability Act.

[e]ERAS: enhanced recovery after surgery.

While we were unable to find reports in the literature of the use of a datamart with dashboard visualization tools to support ERAS programs, these tools have been described in other types of quality, clinical, and research programs. Institutionally developed data infrastructures have been used to aggregate disparate data sources into unified repositories to support research and clinical audits [11-13]. Interactive dashboards (e.g., Tableau and Qlik) have been implemented to visualize clinical performance and facilitate quality improvement efforts [14-16]. Thus, enabling data-driven feedback to stakeholders. The Datamart combines the filtering and query capabilities of a unified data repository with the visualization tools of interactive dashboards to improve patient outcomes, enhance adherence to protocols, improve interdisciplinary communication, and support decision-making.

Conclusions

The Datamart represents a significant advancement in ERAS program management. Unlike third-party systems, such as the EIAS or ACS NSQIP, the Datamart provides full institutional control over data, customizable metrics aligned with local priorities, and flexible reporting capabilities. Although third-party systems can be limited by fixed datasets, external data hosting, and reporting delays, the Datamart enables near real-time internal data access, tailored QA tracking, and the ability to refine metrics, as clinical needs evolve. This offers a robust, centralized, and automated approach to data monitoring and analysis without recurring licensing costs or dependence on vendor timelines. The ability to reduce manual abstraction, improve data accuracy, and evaluate intervention makes the Datamart a vital tool for enhancing perioperative care.

There are several promising directions for the enhancement and expansion of this system. Integrating predictive analytics, automated alerts, natural language processing, and machine learning into the Datamart could enable users to better anticipate complications and tailor interventions [11,12]. For example, predictive models could identify patients at high risk for ERAS noncompliance or adverse outcomes, allowing proactive adjustments to care plans. Automated alerts embedded within the dashboard could notify clinicians in real time when critical metrics fall outside expected ranges, supporting timely interventions and reducing preventable complications.

XSL•FO
RenderX

Expanding the dashboard to encompass additional surgical specialties and incorporating patient-reported outcomes, through structured EMR fields or digital surveys, into the Datamart could further enhance applicability and patient-centered care. These innovations would transform the Datamart from a retrospective monitoring tool into a dynamic, decision-support platform that drives continuous improvement in perioperative care.

As ERAS programs grow in scope and complexity, the need for scalable, adaptable solutions to implementing and monitoring evidence-based care and patient outcomes is increasingly evident. This novel Datamart dashboard addresses many of these challenges while providing a foundation for ongoing innovation. Although developed within Cerner, the Datamart framework is adaptable to other EMRs, provided discrete data fields are available. Implementation requires collaboration with nursing, anesthesiology and surgical services, institutional IT, along with Tableau or similar visualization tools. Institutions seeking to improve ERAS monitoring may consider adapting the framework described here, tailoring metric section and dashboard design to their local EMR environment and clinical priorities.

Data Availability

The datasets used are not available.

Authors' Contributions

Conceptualization: EB-G (lead), SMS (equal)
Data curation: EB
Formal analysis: AS (supporting), EB (lead), JN (supporting), SMS (supporting)
Funding acquisition: EB-G
Investigation: EB (equal), SMS (equal)
Methodology: EB (equal), EB-G (lead), SMS (equal)
Project administration: EB (supporting), EB-G (lead), SMS (equal)
Resources: EB
Software: EB
Supervision: EB-G
Validation: EB (equal), SMS (equal)
Visualization: EB-G (lead), SMS (supporting)
Writing – original draft: EB-G (equal), SO (equal), SMS (equal)
Writing – review & editing: AS (supporting), EB (supporting), EB-G (lead), JN (supporting), SMS (equal), SO (equal)

Conflicts of Interest

None declared.

References

1. Gramlich LM, Sheppard CE, Wasylak T, et al. Implementation of enhanced recovery after surgery: a strategy to transform surgical care across a health system. Implement Sci 2017 May 19;12(1):67. [doi: 10.1186/s13012-017-0597-5] [Medline: 28526041]
2. Thiele RH, Rea KM, Turrentine FE, et al. Standardization of care: impact of an enhanced recovery protocol on length of stay, complications, and direct costs after colorectal surgery. J Am Coll Surg 2015 Apr;220(4):430-443. [doi: 10.1016/j.jamcollsurg.2014.12.042] [Medline: 25797725]
3. Singh SM, Liverpool A, Romeiser JL, Thacker J, Gan TJ, Bennett-Guerrero E. Types of surgical patients enrolled in enhanced recovery after surgery (ERAS) programs in the USA. Perioper Med (Lond) 2021 Apr 27;10(1):12. [doi: 10.1186/s13741-021-00185-5] [Medline: 33902705]
4. Carey S, Hogan S. Failure in systems and culture: barriers that prevent implementation of evidence-based fasting times for patients in the acute care setting. JPEN J Parenter Enteral Nutr 2021 Jul;45(5):933-940. [doi: 10.1002/jpen.1961] [Medline: 32654214]
5. ERAS interactive audit system - EIAS by Encare. Encare Enhancing Eras. URL: https://encare.net/products-and-services/eras-interactive-audit-system-eias/ [accessed 2024-12-12]
6. Langnas EM, Matthay ZA, Lin A, et al. Enhanced recovery after surgery protocol and postoperative opioid prescribing for cesarean delivery: an interrupted time series analysis. Perioper Med (Lond) 2021 Nov 15;10(1):38. [doi: 10.1186/s13741-021-00209-0] [Medline: 34775985]
7. Carter J. Fast-track surgery in gynaecology and gynaecologic oncology: a review of a rolling clinical audit. ISRN Surg 2012;2012:368014. [doi: 10.5402/2012/368014] [Medline: 23320193]
8. Archibald LH, Ott MJ, Gale CM, Zhang J, Peters MS, Stroud GK. Enhanced recovery after colon surgery in a community hospital system. Dis Colon Rectum 2011 Jul;54(7):840-845. [doi: 10.1007/DCR.0b013e31821645bd] [Medline: 21654251]

XSL•FO
RenderX

9. Sinha R, Verma N, Bana R, Kalidindi N, Sampurna S, Mohanty GS. Intra- and post-operative outcomes in benign gynaecologic surgeries before and after the implementation of enhanced recovery after surgery protocols: a comparison. J Minim Access Surg 2023;19(1):112-119. [doi: 10.4103/jmas.jmas_42_22] [Medline: 36722536]
10. Ripollés-Melchor J, Abad-Motos A, Díez-Remesal Y, et al. Association between use of enhanced recovery after surgery protocol and postoperative complications in total hip and knee arthroplasty in the Postoperative Outcomes Within Enhanced Recovery After Surgery Protocol in Elective Total Hip and Knee Arthroplasty Study (POWER2). JAMA Surg 2020 Apr 1;155(4):e196024. [doi: 10.1001/jamasurg.2019.6024] [Medline: 32049352]
11. Marazzi F, Tagliaferri L, Masiello V, et al. GENERATOR Breast DataMart—the novel breast cancer data discovery system for research and monitoring: preliminary results and future perspectives. J Pers Med 2021 Jan 22;11(2):65. [doi: 10.3390/jpm11020065] [Medline: 33498985]
12. D'Amario D, Laborante R, Delvinioti A, et al. GENERATOR HEART FAILURE DataMart: an integrated framework for heart failure research. Front Cardiovasc Med 2023;10:1104699. [doi: 10.3389/fcvm.2023.1104699] [Medline: 37034335]
13. Hurst JH, Liu Y, Maxson PJ, Permar SR, Boulware LE, Goldstein BA. Development of an electronic health records datamart to support clinical and population health research. J Clin Transl Sci 2020 Jun 23;5(1):e13. [doi: 10.1017/cts.2020.499] [Medline: 33948239]
14. Lechner C, Rumpler M, Dorley MC, Li Y, Ingram A, Fryman H. Developing an online dashboard to visualize performance data—Tennessee newborn screening experience. Int J Neonatal Screen 2022 Sep 2;8(3):49. [doi: 10.3390/ijns8030049] [Medline: 36135349]
15. Dang QCL, Román E, Donner K, et al. Implementation and impact of a surgical dashboard on pediatric tonsillectomy outcomes: a quality improvement study. Laryngoscope Investig Otolaryngol 2024 Oct;9(5):e1315. [doi: 10.1002/lio2.1315] [Medline: 39281202]
16. Alhmoud B, Melley D, Khan N, Bonicci T, Patel R, Banerjee A. Evaluating a novel, integrative dashboard for health professionals' performance in managing deteriorating patients: a quality improvement project. BMJ Open Qual 2022 Dec;11(4):e002033. [doi: 10.1136/bmjoq-2022-002033] [Medline: 36588306]

Abbreviations

ACS: American College of Surgeons
EIAS: ERAS Interactive Audit System
EMR: electronic medical record
ERAS: Enhanced Recovery after Surgery
***ICD-10*:** *International Classification of Diseases, Tenth Revision*
NSQIP: National Surgical Quality Improvement Program
PCS: Procedure Coding System
POD: postoperative day
QA: quality assurance
REDCap: Research Electronic Data Capture

Edited by N Rohatgi; submitted 15.Aug.2025; peer-reviewed by AR Cochran, C McLane; revised version received 15.Dec.2025; accepted 18.Dec.2025; published 11.Feb.2026.

Please cite as:
Singh SM, Oster S, Bolze E, Sasson A, Nicholson J, Bennett-Guerrero E
A Novel Customizable Datamart and Tableau Dashboard to Monitor Multiple Enhanced Recovery After Surgery Programs: Development and Validation Study
JMIR Perioper Med 2026;9:e82472
URL: https://periop.jmir.org/2026/1/e82472
doi:10.2196/82472

© Sunitha Margaret Singh, Susannah Oster, Efrat Bolze, Aaron Sasson, James Nicholson, Elliott Bennett-Guerrero. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 11.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comparison Between Ultrasound and Magnetic Resonance Imaging Measurements of the Optic Nerve Sheath Diameter in Patients Undergoing Intracranial Surgery: Prospective Observational Single-Center Study

Mauricio Giraldo[1*], MD; Luz Maria Lopera[2*], MD; Aly Balbaa[3*], MD; Nelson Gonzalez[1*], MD; Raffael Pereira Cezar Zamper[1*], MD; Michael Mayich[4*], MD; Mel Boulton[5*], MD

[1]Department of Anesthesia and Perioperative Medicine, Schulich School of Medicine & Dentistry, Western University, London Health Sciences Centre, 339 Windermere Road, London, ON, Canada

[2]Department of Anesthesia and Perioperative Medicine, Faculty of Medicine, Dalhousie University, Queen Elizabeth II Health Sciences Centre, Halifax, NS, Canada

[3]Department of Family Medicine, Faculty of Medicine, McMaster University, Hamilton Health Sciences, Hamilton, ON, Canada

[4]Department of Neuroradiology, Schulich School of Medicine & Dentistry, Western University, London Health Sciences Centre, London, ON, Canada

[5]Department of Neurosurgery, Schulich School of Medicine & Dentistry, Western University, London Health Sciences Centre, London, ON, Canada

[*]all authors contributed equally

Corresponding Author:
Mauricio Giraldo, MD
Department of Anesthesia and Perioperative Medicine, Schulich School of Medicine & Dentistry, Western University, London Health Sciences Centre, 339 Windermere Road, London, ON, Canada

Abstract

Background: Measuring the optic nerve sheath diameter (ONSD) with ultrasound is a promising, noninvasive way to estimate intracranial pressure (ICP). While magnetic resonance imaging (MRI) provides high-resolution imaging, it is less accessible in urgent or perioperative settings. Comparing ONSD measurements between ultrasound and MRI may help confirm the use of ultrasound in neurosurgical patients.

Objective: The aim of this study is to evaluate how closely ultrasound and MRI measurements of ONSD align in patients undergoing surgery for supratentorial brain tumors.

Methods: This prospective, single-center observational study included 50 adult patients scheduled for elective supratentorial tumor resection. ONSD was measured preoperatively using both transorbital ultrasound and MRI. Measurements were compared using Pearson and Spearman correlation coefficients, the intraclass correlation coefficient, and Bland-Altman analysis.

Results: The average ONSD measured by ultrasound was 5.94 (0.99) mm, compared to 5.75 (SD 1.08) mm via MRI. The two methods showed a strong correlation (Pearson r=0.88, P<.001) and good agreement (intraclass correlation coefficient=0.86). Bland-Altman analysis showed a mean bias of 0.19 mm (95% limits of agreement: –0.62 to 1.00 mm).

Conclusions: Ultrasound-based ONSD measurements closely matched those obtained by MRI in this patient group. These findings support the use of ultrasound as a practical tool for noninvasive ICP assessment in the perioperative care of patients with intracranial tumors.

(JMIR Perioper Med 2026;9:e67480) doi:10.2196/67480

KEYWORDS

optic nerve; optic nerve diameter; intracranial pressure; ocular ultrasound; brain tumor

Introduction

Accurate assessment of intracranial pressure (ICP) is essential in the management of patients with a range of neurological conditions [1,2]. Several tools are available for this purpose, including invasive monitoring devices, magnetic resonance imaging (MRI), computed tomography (CT), and ultrasonography (US) [3,4]. While invasive methods remain the gold standard, they carry risks and are not always feasible. MRI and CT offer practical alternatives but can be limited by cost, access, and the need for patient transport, particularly in acute or perioperative settings.

In recent years, ultrasonographic measurement of the optic nerve sheath diameter (ONSD) has gained attention as a practical,

XSL•FO
RenderX

noninvasive method for estimating ICP. Ultrasound is widely available at the bedside, relatively inexpensive, and quick to perform. A growing body of evidence supports its use, with studies showing good correlation between ONSD measured by ultrasound and both invasive ICP measurements and imaging-based assessments [5]. This technique has been validated in adult and pediatric populations, including neonates, and across a range of pathologies such as trauma, hydrocephalus, ischemic stroke, and brain tumors. Systematic reviews have also contributed to the evidence base [6,7]. For example, a 2011 meta-analysis by Moretti et al [8] demonstrated the diagnostic accuracy of ONSD US in detecting raised ICP. A 2018 review further emphasized its value, particularly when invasive methods are unavailable or contraindicated [9].

Despite its clinical promise, some uncertainties remain. One ongoing discussion involves the optimal cutoff for diagnosing elevated ICP. While there is no universally accepted threshold, ONSD values of 0.48 - 0.50 cm (4.8 - 5.0 mm) are commonly used to indicate ICP above 20 cm H_2O or 20 mm Hg [10-15], with several studies reporting high sensitivity and specificity with these values. For instance, Kimberly et al (2008) [16] reported that an ONSD exceeding 5.0 mm was a strong predictor of raised ICP.

Technique is another area where standardization is still evolving. The typical approach involves measuring the sheath in the transverse plane, 5 mm posterior to the globe [17]. However, anatomical differences between individuals, particularly in the transverse diameter of the eyeball (ETD), may influence interpretation. Some researchers have proposed using the ONSD/ETD ratio to adjust for these variations, with values above 0.19 potentially indicating elevated ICP. While ETD ranges from 21 to 27 mm in healthy individuals, its use in routine clinical decision-making remains under investigation [17]. Recent publications also highlight the growing role of ocular ultrasound in ICP assessment [18,19]. One open-access study noted that bedside ONSD measurement can provide quick, clinically meaningful information during the perioperative period, especially when formal imaging is not immediately available. Their work supports the idea that ultrasound is becoming an increasingly practical addition to routine neurologic monitoring, helping clinicians recognize changes in intracranial dynamics without interrupting patient flow [19].

This study aims to explore the use of ultrasound-based ONSD measurement in a specific neurosurgical population: patients with supratentorial brain tumors. These patients are at risk for increased ICP due to mass effect, edema, or obstruction of cerebrospinal fluid pathways. While previous research has shown a strong correlation between ONSD measured by ultrasound and other modalities, our goal is to assess the agreement between ultrasound- and MRI-derived ONSD values obtained on the same day in a controlled preoperative setting. We also examine the feasibility and accuracy of ultrasound when performed by trained anesthesiologists, a group often involved in perioperative decision-making. Our findings may help clarify the role of ONSD US in situations where access to MRI or CT is limited or delayed.

We conducted a prospective observational study of 50 adult patients undergoing elective supratentorial brain tumor resections at London Health Sciences Center (University Hospital) between July 2021 and May 2023.

Methods

Patient Selection

Inclusion criteria were adults (≥18 y) scheduled for elective intracranial surgery involving supratentorial tumors. Patients with previous intraocular lens implantation were included in the study. Patients were excluded if they experienced hemodynamic instability, required emergent or repeat intracranial surgery, or had a history of ocular pathology, including ocular infection, ocular trauma, or prior ocular surgery. All participants provided written informed consent prior to enrollment.

Ultrasound Examination

On the day of surgery, patients underwent bedside ocular ultrasound prior to anesthetic induction. Ultrasound was performed by a cardiac anesthesiologist formally trained in transthoracic and transesophageal echocardiography and point-of-care ultrasound (POCUS), who had also received a dedicated 2-hour session in standardized ocular ultrasound technique. The session followed current consensus recommendations and included both didactic and hands-on instruction.

Patients were positioned supine with the neck in a neutral position. A high-frequency linear probe (4 - 12 MHz, L12-4; Koninklijke Philips NV) was used with sterile water-soluble gel (Aquasonic 100, Parker Laboratories Inc) applied over the closed eyelid, following standard radiologic convention. The ultrasound marker was oriented toward the patient's right. The optic nerve was visualized in the transverse plane (Figure S1 in Multimedia Appendix 1), and measurements of the ONSD were taken 5 mm posterior to the retina (Figure S2 in Multimedia Appendix 2), perpendicular to the optic nerve axis. Each eye was scanned separately.

Three ONSD measurements were obtained per eye in the transverse plane, and the average of these was used for analysis (Figure S3 in Multimedia Appendix 3). If image quality was suboptimal, additional attempts were made until three acceptable measurements were recorded. All images were stored using a Philips Sparq ultrasound system and Q-Path 5.2.434 software (Telexy Corporation) and anonymized using numeric identifiers per institutional privacy protocols. The ultrasound results were not disclosed to the neuroradiologist performing the MRI measurements.

MRI Assessment

All patients underwent preoperative brain MRI as part of their standard neurosurgical evaluation. MRI images were obtained using a Siemens Magnetom Vida 3T scanner (Siemens Healthineers). ONSD measurements were performed by a neuroradiologist using standard institutional technique in the axial plane, approximately 5 mm posterior to the globe (Figure S4 in Multimedia Appendix 4). The measurement included the

XSL•FO
RenderX

whole optic nerve sheath complex and was obtained at the widest diameter perpendicular to the nerve axis. The MRI results were not disclosed to the anesthesiologist.

Data Collection

Demographic and clinical variables collected included age, sex, height, weight, and BMI, along with ONSD measurements from both ultrasound and MRI. These covariates were included based on prior literature suggesting that factors such as sex, age, and anthropometric measurements may influence baseline ONSD values or their interpretation in the context of raised ICP. Additionally, clinical characteristics such as tumor size, tumor location, and duration of symptoms were recorded, as these may influence ICP and, by extension, ONSD.

Statistical Analysis

Data were analyzed using descriptive and inferential statistics. Continuous variables were summarized as means and standard deviations for normally distributed data or medians with interquartile ranges for nonnormal data, while categorical variables were summarized using counts and percentages. Normality of numerical variables was assessed using the Shapiro-Wilk test.

The relationship between ONSD measurements obtained by ultrasound and MRI was evaluated using the Pearson correlation coefficient for normally distributed variables and the Spearman rank correlation when normality assumptions were not met. Differences between ultrasound and MRI ONSD measurements were compared using paired *t* tests for normally distributed data or Wilcoxon signed-rank tests otherwise. Agreement between the two measurement methods was further assessed using the intraclass correlation coefficient (ICC) with a 1-way random-effects model, with significance tested using the *F* test.

A *P* value <.01 was considered statistically significant to account for multiple comparisons and reduce the risk of type I error in this small, exploratory sample. This threshold was defined a priori.

Ethical Considerations

This study was reviewed and approved by the Research Ethics Board of Western University (protocol number 115045). The study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki and applicable institutional and national research guidelines. Written informed consent was obtained from all participants prior to enrollment. No compensation was provided to participants.

Results

Patient Demographics

A total of 50 patients undergoing brain tumor resection were included in the study, comprising 25 males and 25 females. Participant ages ranged from 21 to 80 years, with a mean age of 55 years (SD 16) and a median of 58 (IQR 21 - 80) years. Anthropometric measurements showed a mean weight of 81.7 kg (SD 21.4) and a mean height of 1.71 m (SD 0.11), resulting in a mean BMI of 27.7 (SD 6.0) and a median BMI of 27.0 (IQR 17.3 - 42.9). According to World Health Organization BMI classifications, two patients (4%) were underweight, 17 (34%) had normal weight, 19 (38%) were overweight, and 12 (24%) were classified as obese; 5 had type I obesity (BMI 30 - 34.9), 5 had type II obesity (BMI 35 - 39.9), and 2 had type III obesity (BMI ≥40) (Table S1 in Multimedia Appendix 5).

ONSD Measurements

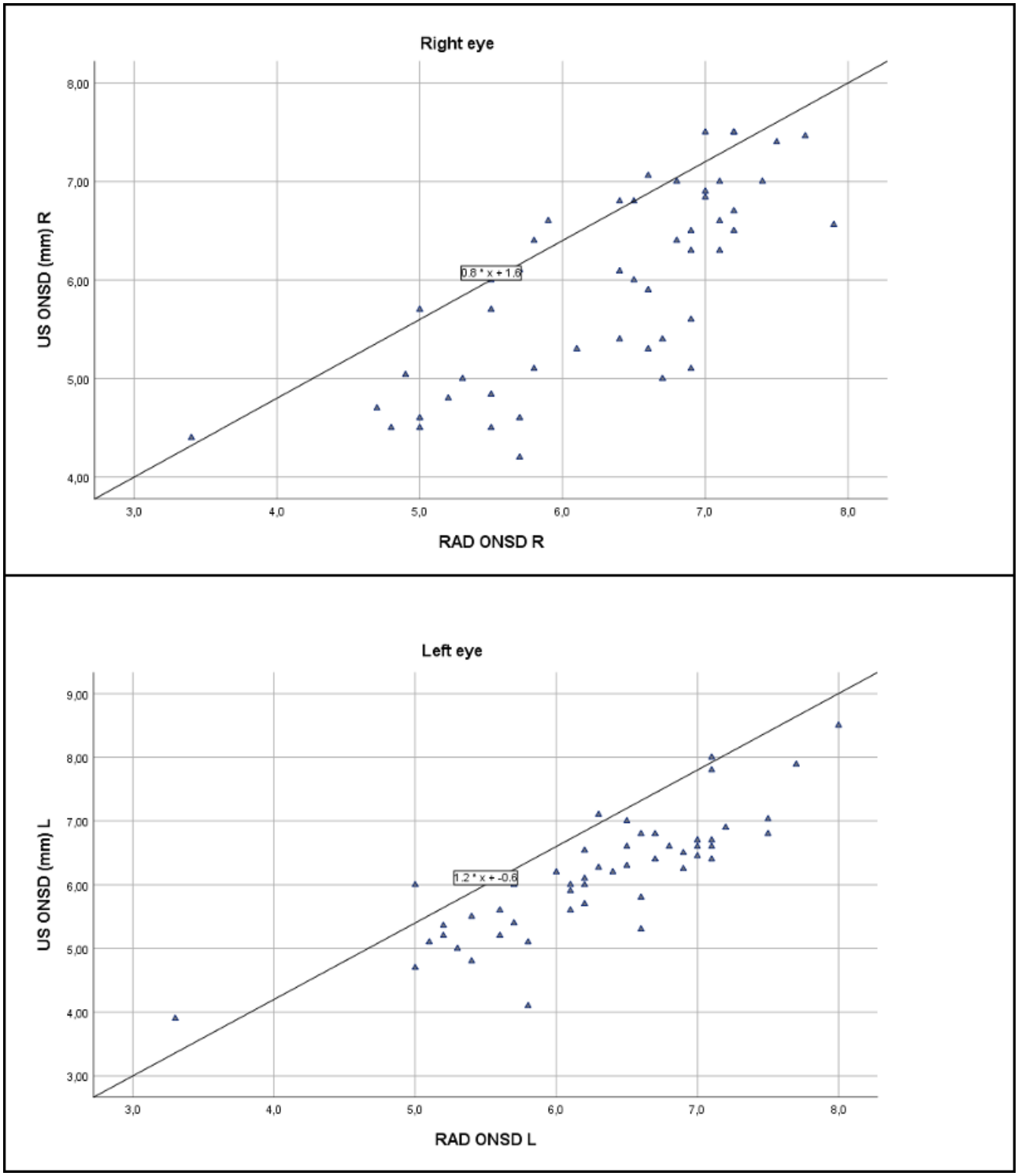
The average ONSD measured by ultrasound (US) was 5.94 mm (SD 0.99) in the right eye and 6.15 mm (SD 0.94) in the left eye. In comparison, radiological assessment (RAD) showed slightly higher values of 6.28 mm (SD 0.94) and 6.30 mm (SD 0.86) for the right and left eyes, respectively. Differences between the two methods ranged from –1 mm to 1.8 mm in the right eye and –1 mm to 1.7 mm in the left eye. Statistical analysis suggested a strong positive correlation between US and RAD measurements for both eyes, with correlation coefficients of 0.78 (right eye, Spearman) and 0.84 (left eye, Pearson), both highly significant (*P*<.001). These findings show a good agreement between ultrasound and radiological methods for assessing ONSD in patients undergoing brain tumor resection (Table S2 in Multimedia Appendix 6).

Correlation Between Imaging Modalities

There was a clear, statistically significant correlation between ultrasound-measured ONSD values and radiological imaging values. In the right eye, the correlation coefficient was 0.779 (Spearman), and in the left eye, it was 0.836 (Pearson), with both showing *P* values below .001. These findings suggest that the two methods consistently align in their assessment of ONSD, with strong agreement in both eyes (Figure 1).

XSL•FO
RenderX

Figure 1. Optic nerve sheath diameter (ONSD) correlation with ultrasound (US) and magnetic resonance imaging (RAD) in patients scheduled for brain tumor resection (University Hospital, July 2021-May 2023, right and left eyes).



Intraclass Correlation Coefficient

We used ICCs to assess the agreement between ultrasound and radiology measurements of ONSD. For the right eye, the ICC was 0.707 for individual measurements and 0.829 for averaged measurements. The agreement was even stronger in the left eye, with ICCs of 0.823 and 0.903 for single- and average-measures, respectively. All results were statistically significant (*P* values=.001), indicating strong agreement between the two methods. Full details of the ICC values and confidence intervals are available in Table S3 in Multimedia Appendix 7.

Discussion

This study evaluated the relationship between ONSD measurements obtained via ultrasound and MRI in adult patients undergoing brain tumor resection. We found a strong and statistically significant correlation between the two methods, with minor average differences and high ICCs. These findings

XSL•FO
RenderX

suggest that ultrasound may be a practical alternative to MRI for assessing ONSD in the perioperative setting, where rapid, noninvasive evaluation of ICP is often required [20].

In addition to the primary findings, we observed that ultrasound-derived ONSD measurements were generally consistent with MRI-derived measurements, with most values falling within clinically acceptable limits of agreement. Notably, the measurements showed high reproducibility in both eyes, reinforcing the potential utility of ultrasound in routine practice when MRI is unavailable or impractical in patients with supratentorial masses.

Previous research has demonstrated strong associations between increased ONSD and raised ICP [21,22]. Our findings align with those reported by Bäuerle et al [23], who found a correlation between ultrasound- and MRI-based ONSD in a small cohort of healthy volunteers (r=0.72). However, their population was younger and nonsurgical. Kerscher et al [24-26] reported even stronger agreement in pediatric patients (r=0.976), but their findings are limited to that population. In contrast, our study focuses on adults in a surgical context and provides new evidence supporting the feasibility of ultrasound-assessed ONSD in preoperative neurosurgical care.

A key strength of our approach was the use of POCUS performed by anesthesiologists familiar with the technique. While the clinicians involved had prior ultrasound experience, ONSD-specific training was brief, which reflects the method's accessibility. Notably, even minimal training has been shown to produce reliable measurements, as seen in studies involving nonphysician trainees [27]. This highlights the practicality of implementing ONSD ultrasound in various clinical settings, especially where MRI is unavailable or delayed.

It is also worth noting that while ONSD ultrasound is most often discussed in the context of trauma or critical care [28], our findings suggest its potential value in elective neurosurgery. In patients with supratentorial tumors, where intraoperative ICP changes can have serious consequences, the ability to assess ONSD in real time could support timely decision-making. Given the growing interest in using POCUS for perioperative neuromonitoring, these findings may help guide future protocols, especially in resource-limited environments.

However, several limitations must be acknowledged. First, the study did not assess interobserver variation, which limits conclusions about the method's broader applicability to other clinicians or settings. Second, although trained anesthesiologists performed the ultrasound scans, access to formal ultrasound education may not be universal, especially in remote or low-resource centers. Third, our sample included only patients with tumoral disease; thus, the findings may not extend to other intracranial pathologies, such as hemorrhage or infection. Finally, the study's sample size, while larger than some previous comparisons, remains relatively small and limits the strength of conclusions that can be drawn.

Despite these limitations, this study provides valuable insight into the use of ultrasound for optic nerve evaluation in neurosurgical patients. By demonstrating a strong correlation with MRI, our findings support further exploration of ONSD ultrasound as a screening or monitoring tool for raised ICP, particularly in settings where MRI is unavailable or impractical. Recent open access articles have also pointed out the need for more consistent training and standardized ONSD protocols [32, which aligns with our recommendation for larger, coordinated studies to refine measurement techniques.

Future studies should focus on standardizing training protocols, assessing consistency between different examiners, and evaluating ONSD thresholds across a broader range of intracranial conditions. Larger, multicenter trials could help define normal and pathological ONSD values more precisely and assess how ultrasound-guided ICP screening might impact clinical outcomes. As interest in perioperative neuromonitoring continues to grow, incorporating tools like ONSD ultrasound may help streamline care, improve responsiveness to neurological changes, and broaden access to noninvasive ICP monitoring.

Funding

The authors declared no financial support was received for this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Optic nerve sheath.
[DOCX File, 289 KB - periop_v9i1e67480_app1.docx]

Multimedia Appendix 2
Optic nerve measurement 5 mm below retinal layer as indicated in the picture. This is the level to measure the transverse diameter of the optic nerve.
[DOCX File, 292 KB - periop_v9i1e67480_app2.docx]

Multimedia Appendix 3
Optic nerve diameter measurement.

XSL•FO
RenderX

[DOCX File, 299 KB - periop_v9i1e67480_app3.docx]

Multimedia Appendix 4
Optic nerve magnetic resonance imaging measurement.
[DOCX File, 483 KB - periop_v9i1e67480_app4.docx]

Multimedia Appendix 5
Optic nerve diameter correlation with ultrasound and magnetic resonance imaging. University Hospital, July 2021-May 2023.
[PNG File, 88 KB - periop_v9i1e67480_app5.png]

Multimedia Appendix 6
Optic nerve diameter correlation with ultrasound (US) and magnetic resonance imaging (RAD) in patients scheduled for brain tumor resection. University Hospital, July 2021-May 2023.
[DOCX File, 15 KB - periop_v9i1e67480_app6.docx]

Multimedia Appendix 7
Interclass correlation coefficient.
[PNG File, 77 KB - periop_v9i1e67480_app7.png]

References

1. Byrne SF, Green RL. Ultrasound of the Eye and Orbit, 2nd edition: Mosby; 2002.
2. Bedi DG, Gombos DS, Ng CS, Singh S. Sonography of the eye. AJR Am J Roentgenol 2006 Oct;187(4):1061-1072. [doi: 10.2214/AJR.04.1842] [Medline: 16985158]
3. Shirodkar CG, Munta K, Rao SM, Mahesh MU. Correlation of measurement of optic nerve sheath diameter using ultrasound with magnetic resonance imaging. Indian J Crit Care Med 2015 Aug;19(8):466-470. [doi: 10.4103/0972-5229.162465] [Medline: 26321806]
4. Ohle R, McIsaac SM, Woo MY, Perry JJ. Sonography of the optic nerve sheath diameter for detection of raised intracranial pressure compared to computed tomography: a systematic review and meta-analysis. J Ultrasound Med 2015 Jul;34(7):1285-1294. [doi: 10.7863/ultra.34.7.1285] [Medline: 26112632]
5. Liu D, Kahn M. Measurement and relationship of subarachnoid pressure of the optic nerve to intracranial pressures in fresh cadavers. Am J Ophthalmol 1993 Nov 15;116(5):548-556. [doi: 10.1016/s0002-9394(14)73195-2] [Medline: 8238213]
6. Hayreh SS. Pathogenesis of optic disc edema in raised intracranial pressure. Prog Retin Eye Res 2016 Jan;50:108-144. [doi: 10.1016/j.preteyeres.2015.10.001] [Medline: 26453995]
7. Newman WD, Hollman AS, Dutton GN, Carachi R. Measurement of optic nerve sheath diameter by ultrasound: a means of detecting acute raised intracranial pressure in hydrocephalus. Br J Ophthalmol 2002 Oct;86(10):1109-1113. [doi: 10.1136/bjo.86.10.1109] [Medline: 12234888]
8. Moretti R, Pizzi B, Cassini F, Vivaldi N. Reliability of optic nerve ultrasound for the evaluation of patients with spontaneous intracranial hemorrhage. Neurocrit Care 2009 Dec;11(3):406-410. [doi: 10.1007/s12028-009-9250-8] [Medline: 19636971]
9. Robba C, Santori G, Czosnyka M, et al. Optic nerve sheath diameter measured sonographically as non-invasive estimator of intracranial pressure: a systematic review and meta-analysis. Intensive Care Med 2018 Aug;44(8):1284-1294. [doi: 10.1007/s00134-018-5305-7] [Medline: 30019201]
10. Rajajee V, Vanaman M, Fletcher JJ, Jacobs TL. Optic nerve ultrasound for the detection of raised intracranial pressure. Neurocrit Care 2011 Dec;15(3):506-515. [doi: 10.1007/s12028-011-9606-8] [Medline: 21769456]
11. Robba C, Donnelly J, Cardim D, et al. Optic nerve sheath diameter ultrasonography at admission as a predictor of intracranial hypertension in traumatic brain injured patients: a prospective observational study. J Neurosurg 2020 Apr 1;132(4):1279-1285. [doi: 10.3171/2018.11.JNS182077] [Medline: 30849751]
12. Bekerman I, Gottlieb P, Vaiman M. Variations in eyeball diameters of the healthy adults. J Ophthalmol 2014;2014:503645. [doi: 10.1155/2014/503645] [Medline: 25431659]
13. Chandrapatham K, Cardim D, Czosnyka M, et al. Variability of the optic nerve sheath diameter on the basis of sex and age in a cohort of healthy volunteers. Acta Neurochir Suppl 2021;131:121-124. [doi: 10.1007/978-3-030-59436-7_25] [Medline: 33839831]
14. Abbinante G, Vitiello L, Coppola A, Salerno G, Gagliardi V, Pellegrino A. Optic nerve ultrasound evaluation in children: a review. Diagnostics (Basel) 2023 Feb 1;13(3):535. [doi: 10.3390/diagnostics13030535] [Medline: 36766639]
15. Betcher J, Becker TK, Stoyanoff P, Cranford J, Theyyunni N. Military trainees can accurately measure optic nerve sheath diameter after a brief training session. Mil Med Res 2018 Dec 20;5(1):42. [doi: 10.1186/s40779-018-0189-y] [Medline: 30572931]
16. Kimberly HH, Shah S, Marill K, Noble V. Correlation of optic nerve sheath diameter with direct measurement of intracranial pressure. Acad Emerg Med 2008 Feb;15(2):201-204. [doi: 10.1111/j.1553-2712.2007.00031.x] [Medline: 18275454]

XSL•FO
RenderX

17. Shafiq F, Ali MA, Khan FA. Pre-anaesthetic assessment of intracranial pressure using optic nerve sheath diameter in patients scheduled for elective tumour craniotomy. J Ayub Med Coll Abbottabad 2018;30(2):151-154. [Medline: 29938408]
18. Stevens RRF, Gommer ED, Aries MJH, et al. Optic nerve sheath diameter assessment by neurosonology: a review of methodologic discrepancies. J Neuroimaging 2021 Sep;31(5):814-825. [doi: 10.1111/jon.12906] [Medline: 34270144]
19. Trocha G, Bonilla A, Romero C, Palacios J, Molano-Gonzalez N. Ultrasound measurement of optic nerve sheath diameter in a healthy adult Colombian population. BMC Neurol 2023 Jan 13;23(1):16. [doi: 10.1186/s12883-023-03062-4] [Medline: 36639617]
20. Jiang J, Xie X. Ultrasound measurement of optic nerve sheath diameter for intracranial pressure monitoring. Interdiscip Neurosurg 2025 Sep;41:102091. [doi: 10.1016/j.inat.2025.102091]
21. Desai VR, Sadrameli SS, Hoppe S, et al. Contemporary management of increased intraoperative intracranial pressure: evidence-based anesthetic and surgical review. World Neurosurg 2019 Sep;129:120-129. [doi: 10.1016/j.wneu.2019.05.224] [Medline: 31158533]
22. Padayachy LC, Padayachy V, Galal U, Pollock T, Fieggen AG. The relationship between transorbital ultrasound measurement of the optic nerve sheath diameter (ONSD) and invasively measured ICP in children.: Part II: age-related ONSD cut-off values and patency of the anterior fontanelle. Childs Nerv Syst 2016 Oct;32(10):1779-1785. [doi: 10.1007/s00381-016-3068-4] [Medline: 27659820]
23. Bäuerle J, Schuchardt F, Schroeder L, Egger K, Weigel M, Harloff A. Reproducibility and accuracy of optic nerve sheath diameter assessment using ultrasound compared to magnetic resonance imaging. BMC Neurol 2013 Dec 1;13:187. [doi: 10.1186/1471-2377-13-187] [Medline: 24289136]
24. Kerscher SR, Schöni D, Hurth H, et al. The relation of optic nerve sheath diameter (ONSD) and intracranial pressure (ICP) in pediatric neurosurgery practice - part I: correlations, age-dependency and cut-off values. Childs Nerv Syst 2020 Jan;36(1):99-106. [doi: 10.1007/s00381-019-04266-1] [Medline: 31256241]
25. Kerscher SR, Schöni D, Neunhoeffer F, et al. The relation of optic nerve sheath diameter (ONSD) and intracranial pressure (ICP) in pediatric neurosurgery practice - Part II: influence of wakefulness, method of ICP measurement, intra-individual ONSD-ICP correlation and changes after therapy. Childs Nerv Syst 2020 Jan;36(1):107-115. [doi: 10.1007/s00381-019-04336-4] [Medline: 31392457]
26. Kerscher SR, Zipfel J, Groeschel S, Bevot A, Haas-Lude K, Schuhmann MU. Comparison of B-scan ultrasound and mri-based optic nerve sheath diameter (ONSD) measurements in children. Pediatr Neurol 2021 Nov;124:15-20. [doi: 10.1016/j.pediatrneurol.2021.08.002] [Medline: 34508997]
27. Tayal VS, Neulander M, Norton HJ, Foster T, Saunders T, Blaivas M. Emergency department sonographic measurement of optic nerve sheath diameter to detect findings of increased intracranial pressure in adult head injury patients. Ann Emerg Med 2007 Apr;49(4):508-514. [doi: 10.1016/j.annemergmed.2006.06.040] [Medline: 16997419]
28. Herrero MP, Fernández FDP, Moreno MM, Martín AM, Fernández-Caballero S. Indirect measurement of perioperative intracranial pressure-eye and orbitary ultrasound utility. Advances in Medicine and Engineering Interdisciplinary Research 2024;2(3):78-85. [doi: 10.32629/ameir.v2i3.2615]

Abbreviations

CT: computed tomography
ETD: transverse diameter of the eyeball
ICC: intraclass correlation coefficient
ICP: intracranial pressure
MRI: magnetic resonance imaging
ONSD: optic nerve sheath diameter
POCUS: point-of-care ultrasound
RAD: radiological assessment
US: ultrasonography

Edited by N Rohatgi; submitted 17.Nov.2024; peer-reviewed by D Janitschke, MI Hirzallah, T Aslanidis; revised version received 16.Jan.2026; accepted 19.Jan.2026; published 17.Apr.2026.

Please cite as:
Giraldo M, Lopera LM, Balbaa A, Gonzalez N, Cezar Zamper RP, Mayich M, Boulton M
Comparison Between Ultrasound and Magnetic Resonance Imaging Measurements of the Optic Nerve Sheath Diameter in Patients Undergoing Intracranial Surgery: Prospective Observational Single-Center Study
JMIR Perioper Med 2026;9:e67480
URL: https://periop.jmir.org/2026/1/e67480
doi:10.2196/67480

XSL•FO
RenderX

© Mauricio Giraldo, Luz Maria Lopera, Aly Balbaa, Nelson Gonzalez, Raffael Pereira Cezar Zamper, Michael Mayich, Mel Boulton. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 17.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Survival Prediction in Patients With Bladder Cancer Undergoing Radical Cystectomy Using a Machine Learning Algorithm: Retrospective Single-Center Study

Francesco Andrea Causio[1,2], MD; Vittorio De Vita[1,2], MD; Andrea Nappi[1,3], BSc; Melissa Sawaya[4], PhD; Bernardo Rocco[5,6], MD, PhD; Nazario Foschi[5,6], MD; Giuseppe Maioriello[5,6], MD; Pierluigi Russo[1,5,6,7], MD

[1]Italian Society for Artificial Intelligence in Medicine (SIIAM - Società Italiana Intelligenza Artificiale in Medicina), Rome, Italy

[2]University Department of Life Sciences and Public Health, Section of Hygiene, Università Cattolica del Sacro Cuore, Largo F Vito 1, Rome, Italy

[3]Computer Science Department, University of Twente, Twente, The Netherlands

[4]Université Paris-Saclay, UVSQ, Inserm, Gustave Roussy, CESP, Villejuif, France

[5]University Department of Urology, Università Cattolica del Sacro Cuore, Rome, Italy

[6]University Department Medicine and Translational Surgery, Università Cattolica del Sacro Cuore, Rome, Italy

[7]Department of Life Science, Health, and Health Professions, Università Degli Studi Link, Rome, Italy

Corresponding Author:
Vittorio De Vita, MD
Italian Society for Artificial Intelligence in Medicine (SIIAM - Società Italiana Intelligenza Artificiale in Medicina), Rome, Italy

Abstract

Background: Traditional statistical models often fail to capture the complex dynamics influencing survival outcomes in patients with bladder cancer after radical cystectomy, a procedure where approximately 50% of patients develop metastases within 2 years. The integration of artificial intelligence (AI) offers a promising avenue for enhancing prognostic accuracy and personalizing treatment strategies.

Objective: This study aimed to develop and evaluate a machine learning algorithm for predicting disease-free survival (DFS), overall survival (OS), and the cause of death in patients with bladder cancer undergoing cystectomy, using a comprehensive dataset of clinical and pathological variables.

Methods: Retrospective data of 370 patients with bladder cancer who underwent radical cystectomy at Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy, were collected. The dataset comprised 20 input variables, encompassing demographics, tumor characteristics, treatment variables, and inflammatory markers. For specific analyses and models, we used patient subcohorts. The CatBoost algorithm was used for regression tasks (DFS in 346 patients, OS in 347 patients) and a binary classification task (tumor-related death in 312 patients). Model performance was assessed using mean absolute error (MAE) for regression and F_1-score for classification, prioritizing a minimum recall of 75% for tumor-related deaths. Five-fold cross-validation and Shapley additive explanations (SHAP) values were used to ensure robustness and interpretability.

Results: For DFS prediction, the CatBoost model achieved an MAE of 18.68 months, with clinical tumor stage and pathological tumor classification identified as the most influential predictors. OS prediction yielded an MAE of 17.2 months, which improved to 14.6 months after feature filtering, where tumor classification and the systemic immune-inflammation index (SII) were most impactful. For tumor-related death classification, the model achieved a recall of 78.6% and an F_1-score of 0.44 for the positive class (tumor-related deaths), correctly identifying 11 of 14 cases. Bladder tumor position was the most influential feature for cause-of-death prediction.

Conclusions: The developed machine learning algorithm demonstrates promising accuracy in predicting survival and the cause of death in patients with bladder cancer after cystectomy. The key predictors include clinical and pathological tumor staging, systemic inflammation (SII), and bladder tumor position. These findings highlight the potential of AI in providing clinicians with an objective, data-driven tool to improve personalized prognostic assessment and guide clinical decision-making.

(JMIR Perioper Med 2026;9:e86666) doi:10.2196/86666

KEYWORDS

cystectomy; disease-free survival; artificial intelligence; neoplasm staging; retrospective studies; urinary bladder neoplasms; clinical decision-making; machine learning; statistical models

XSL•FO
RenderX

Introduction

In the evolving landscape of health care, the integration of artificial intelligence (AI) into clinical decision-making has gained significant momentum, particularly in the realm of oncology [1,2]. With advancements in machine learning techniques, health care professionals are increasingly harnessing the power of AI to enhance diagnosis, prognosis, and treatment planning. The exponential growth of digital health care data, including electronic health records, medical imaging, genomic data, and real-time patient monitoring, has fueled the development of predictive algorithms [1,3].

The field of urology is complex: cancerous conditions benefit from the leverage of additional data sources and decision-making algorithms that allow physicians to plan treatment while considering several complex factors. Urological cancers, including prostate, bladder, and renal cancers, place a considerable burden on health care systems worldwide [4]. These malignancies often require complex management involving early diagnosis, accurate staging, and personalized treatment strategies to optimize patient outcomes. Traditional methods of assessing prognosis rely heavily on statistical models that may not capture the multifaceted nature of cancer behavior and patient responses to treatment. Conventional regression statistics often fail to provide the depth of analysis required to address the complexities of cancer management. In contrast, AI techniques, such as artificial neural networks, Bayesian networks, and neuro-fuzzy modeling systems, offer innovative approaches to constructing data-driven models that can adapt to the heterogeneous nature of cancer [5].

The potential of AI in predicting patient outcomes is particularly evident in its ability to analyze large datasets without the constraints of predetermined statistical distributions. By leveraging retrospective data, we can develop algorithms that not only identify patterns and correlations but also provide insights into individual patient behavior. This capability is crucial for clinicians who face the challenge of tailoring treatment plans to the unique characteristics of each patient. In the context of mortality and postoperative survival, the application of AI can provide critical insights that enhance our understanding of patient outcomes following surgical interventions. The ability to predict which patients are at higher risk of complications or recurrence can lead to more informed clinical decisions, ultimately improving the quality of care [6]. For instance, machine learning algorithms can analyze a multitude of variables, including clinical, pathological, and demographic factors, to generate individualized risk profiles that guide treatment strategies and follow-up plans [7]. Recent urological research has shown that combining hematological inflammation indexes with machine learning algorithms can improve the prediction of surgical outcomes, as demonstrated in patients who underwent urethroplasty [8].

In this study, we focus specifically on the training of an AI algorithm using retrospective data collected from patients diagnosed with bladder cancer who underwent radical cystectomy. Patients with localized muscle-invasive or recurrent non–muscle-invasive bladder cancer benefit most from radical cystectomy, which may be preceded by neoadjuvant chemotherapy in selected cases, in terms of local disease control. Even with sufficient local control achieved through cystectomy, approximately 50% of patients develop metastases within 2 years and may ultimately die from the disease. This is likely due to the existence of regional or distant microscopic metastatic disease at the time of surgery [9]. The proposed methodology will involve the comprehensive examination of variables associated with patient demographics, tumor characteristics, treatment modalities, and postoperative outcomes. Using machine learning techniques, we aim to identify key predictors of mortality and postoperative survival, ultimately constructing a predictive model with potential relevance for clinical decision-making.

Methods

Study Design

We collected retrospective data on patients with high-risk and very high-risk non–muscle-invasive bladder cancer and muscle-invasive bladder cancer who underwent radical cystectomy at Fondazione Policlinico Universitario Agostino Gemelli IRCCS in Rome, Italy. The dataset included data on various clinical and pathological variables from 370 patients.

Ethical Considerations

Ethical approval was obtained from the institutional ethical review board (protocol number 676 - 02). As primary consent for data collection covered secondary analyses, additional consent was not required for this study. The data used in this study were anonymized, and no compensation was provided to patients.

Data Collection and Preprocessing

Clinical and pathological data were extracted from medical records, including demographic, lifestyle, tumor, treatment, and laboratory variables. The dataset was split into three outcome-specific subsets to maximize the usable sample for each task:

1. DFS dataset: predicting disease-free survival (DFS) in months (346 patients in total).
2. OS dataset: predicting overall survival (OS) in months (347 patients in total).
3. Death cause dataset: for classification purposes, the classes “death from other causes” and “alive” were merged into a single negative class to create a binary variable. Therefore, the cause is defined as either “no” or “yes,” depending on whether it was tumor related (312 patients in total).

The variables included in the dataset are detailed in Table 1. All categorical variables were cast as strings to allow native handling by CatBoost (version 1.2.8; Yandex).

A total of 20 input variables were selected for model development.

XSL•FO
RenderX

Table . Variables included in the study

Variable (English)	Description	Data type	Input/output
Patient demographics and lifestyle			
AGE	Age (years)	Numerical	Input
BMI	Body mass index (kg/m^2)	Numerical	Input
SEX	Biological sex (0: man, 1: woman)	Categorical	Input
SMOKE	Smoker (0: no, 1: yes)	Categorical	Input
Patient medical history			
DM	Patient has diabetes mellitus (0: no, 1: yes)	Categorical	Input
PRIOR SURGERY	Patient had previously undergone surgery in the abdominal area (0: no, 1: yes)	Categorical	Input
PRIOR RADIOTHERAPY	Patient had previously received radiotherapy in the abdominal area (0: no, 1: yes)	Categorical	Input
PRIOR SYSTEMIC CHEMOTHERAPY	Patient had previously received systemic chemotherapy (0: no, 1: yes)	Categorical	Input
Tumor characteristic			
BLADDER TUMOR POSITION	Identifier of tumor position (0: intertrigonal zone, 1: right periosteal, 2: left periosteal, 3: dome, 4: posterior wall, 5: right lateral wall, 6: left lateral wall, 7: prostatic urethra, 8: anterior wall, 9: entire bladder, 10: bladder base)	Categorical	Input
TUMOR DIMENSION	Tumor dimension (cm)	Numerical	Input
PRE-HYDRONEPHROSIS	Hydronephrosis (0: no, 1: right hydronephrosis, 2: left hydronephrosis, 3: bilateral hydronephrosis)	Categorical	Input
H.E. TURV	Histological examination for transurethral resection of the bladder (0: localized to mucosa +/– submucosa multirecurrent, 1: muscle-invasive, 2: squamous)	Categorical	Input
LVI	Lymphovascular invasion (0: absent, 1: present)	Categorical	Input
CTS	Clinical tumor stage (0: cTa, 1: cTis, 2: cT1, 3: cT2, 4: cT3, 5: cT4)	Categorical	Input
TC	Tumor classification (1: T0, 2: Ta, 3: Tis, 4: T1, 5: T2a, 6: T2b, 7: T3a, 8: T3b, 9: T4a, 10: T4b)	Categorical	Input
Inflammatory and immune marker			
SII	Systemic immune-inflammation index (decimals)	Numerical	Input
Treatment and outcome			
UD	Urinary diversion type (0: Bricker ileal conduit, 1: ureterocutaneostomy, 2: vesicoileal pouch)	Categorical	Input
RECURRENCE	Tumor recurrence (0: no, 1: yes)	Categorical	Input
DFS	Disease-free survival after treatment (in months)	Numerical	Output

XSL•FO
RenderX

Variable (English)	Description	Data type	Input/output
OS	Overall survival: time from diagnosis/treatment start to death from any cause (in months)	Numerical	Output
DEATH CAUSE	Cause of death (X: alive, 1: other, 2: cancer); later merged (0: alive + other, 1: cancer)	Categorical	Output

Machine Learning Models

To predict clinical outcomes, we used the CatBoost algorithm for both regression (DFS and OS) and classification (cause of death) tasks, as it is effective for small and structured datasets.

For DFS and OS, we applied CatBoostRegressor models. For predicting the cause of death, we used the CatBoostClassifier, with the binary outcome of death being tumor related or not.

DFS and OS Model Evaluation

For the regression tasks (DFS and OS), we evaluated model performance using mean absolute error (MAE) to quantify the average prediction error in months.

Cause-of-Death Classification Model

For the classification task (cause of death), we evaluated performance using the F_1-score. The F_1-score is a single metric that balances precision and recall, particularly useful in cases of imbalanced classes where the positive class is of primary interest. For class 1 (tumor-related deaths), it was calculated as the harmonic mean of precision and recall:

(1)F1=2∗(Precision∗Recall)/(Precision+Recall)

Precision is the proportion of correctly identified positive predictions among all positive predictions, and recall is the proportion of correctly identified positive predictions among all actual positives.

(2)Precision=TP/(TP+FP)

(3)Recall=TP/(TP+FN)

Confusion matrices were used to examine prediction distributions, and probability thresholds were adjusted to optimize recall while limiting false positives. To account for class imbalance in the classification task, we applied custom class weights. We adjusted the decision threshold, aiming for a minimum recall of 75% to ensure that most tumor-related deaths were accurately identified and classified.

Cross-Validation and Hyperparameter Tuning

All models were trained and evaluated using 5-fold cross-validation to ensure generalizability and reduce the risk of overfitting, especially given the relatively small dataset. In addition, we applied early stopping with a patience range of 30 to 50 rounds, allowing the model to terminate training once performance ceased to improve on the validation fold.

To enhance the interpretability and transparency of the developed machine learning models, we used violin plots and Shapley additive explanations (SHAP) scatterplots to investigate the impact of variables on the prediction of the results. Violin plots show the effect of each variable on the results, both in terms of direction (favorable or unfavorable) and intensity. The SHAP scatterplot assigns an importance value to each feature for a particular prediction. For each patient, the SHAP values revealed the specific features driving the predicted risk of tumor death. By aggregating the SHAP values across the entire cohort, the overall impact and importance of each clinical and pathological variable on the model's outcome predictions were determined. This enabled the identification of the most significant factors influencing the outcomes of patients with bladder cancer after cystectomy.

This paper presents only the most significant results of the analysis. The complete analysis is available online for open access [10].

Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis Guidelines for AI

To enhance the transparency, interpretability, and reproducibility of our machine learning-based prediction models, this study adheres to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement, specifically considering the extensions for AI (TRIPOD+AI). The TRIPOD+AI guidelines provide a standardized framework for reporting studies that develop or validate prediction models, ensuring that sufficient detail is provided for critical appraisal and replication by other researchers. By following these guidelines, we aim to clearly articulate the study design, data characteristics, model development process, and performance evaluation, thereby contributing to the responsible and rigorous application of AI in medical research (Table 2).

XSL•FO
RenderX

Table . Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis checklist for reporting studies involving artificial intelligence (TRIPOD+AI).

TRIPOD+AI item	Description of reporting in this study
1. Title	• Survival Prediction in Patients with Bladder Cancer Undergoing Radical Cystectomy Using a Machine Learning Algorithm: Retrospective Single-Center Study
2. Abstract	• The abstract summarizes the study's objectives, methods, key findings, and conclusions.
3. Introduction - background	• The introduction will establish the clinical context of bladder cancer, the prognostic challenges, and the rationale for using machine learning.
4. Methods - participants	
4a. Eligibility criteria	• Patients who underwent radical cystectomy for bladder cancer, with available data for selected variables
4b. Settings and locations	• Data collected retrospectively from a single institution: Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy
4c. Source of data	• Patient medical records
5. Data acquisition method	• Retrospective data extraction into a spreadsheet
6. Methods - outcome	
6a. Definition of outcomes	• Disease-free survival (DFS): time in months from treatment to recurrence or death (event) or last follow-up (censored) • Overall survival (OS): time in months from diagnosis/treatment start to death from any cause (event) or last follow-up (censored) • DEATH CAUSE: binary classification (0: did not die from the tumor, 1: died from the tumor)
6b. Outcome measurement	• DFS and OS were calculated from documented dates. • The cause of death was extracted from medical records and recategorized for binary classification.
7. Methods - predictors	

XSL•FO
RenderX

TRIPOD+AI item	Description of reporting in this study
7a. Definition of all predictors	• AGE: patient's age (years) • BMI (kg/m^2) • DM: patient has diabetes mellitus (0: no, 1: yes) • PRIOR SURGERY: patient had previously undergone surgery in the abdominal area (0: no, 1: yes) • PRIOR RADIOTHERAPY: patient had previously received radiotherapy in the abdominal area (0: no, 1: yes) • PRIOR SYSTEMIC CHEMOTHERAPY: patient had previously received systemic chemotherapy (0: no, 1: yes) • BLADDER TUMOR POSITION: identifier of tumor position (0: intertrigonal zone, 1: right periosteal, 2: left periosteal, 3: dome, 4: posterior wall, 5: right lateral wall, 6: left lateral wall, 7: prostatic urethra, 8: anterior wall, 9: entire bladder, 10: bladder base) • TUMOR DIMENSION: tumor dimension (cm) • PRE-HYDRONEPHROSIS: pretreatment hydronephrosis (0: no, 1: right hydronephrosis, 2: left hydronephrosis, 3: bilateral hydronephrosis) • SEX: biological sex (0: man, 1: woman) • SMOKE: patient smokes (0: no, 1: yes) • H.E. TURV: histological examination for transurethral resection of the bladder (0: localized to mucosa +/– submucosa multirecurrent, 1: muscle-invasive, 2: squamous) • SII: systemic immune-inflammation index (decimals) • UD: urinary diversion type (0: Bricker ileal conduit, 1: ureterocutaneostomy, 2: vesicoileal pouch) • LVI: lymphovascular invasion (0: absent, 1: present) • CTS: clinical tumor stage (0: cTa, 1: cTis, 2: cT1, 3: cT2, 4: cT3, 5: cT4) • TC: tumor classification (1: T0, 2: Ta, 3: Tis, 4: T1, 5: T2a, 6: T2b, 7: T3a, 8: T3b, 9: T4a, 10: T4b)
7b. Predictor measurement	• Predictors were measured clinically (eg, age and BMI), derived from patient history (eg, prior surgeries and smoking status), or derived from laboratory or pathology reports (eg, SII, tumor dimension, LVI, CTS, and TC).
8. Methods - sample size	
8a. Sample size determination	• The available retrospective data determined the sample size. No formal power calculation was performed due to the exploratory nature of the study and the limitations of the data.
9. Methods - data handling	
9a. Handling of missing data	• Rows with null values in specific critical variables ("TUMOR DIMENSION," "LVI," "TC," "H.E. TURV," "RECURRENCE," "DFS," "OS," "DEATH CAUSE") were removed. No imputation was performed.
9b. Data transformation	• Numerical variables were type-adjusted to int or float. Categorical variables were explicitly converted to category type. "DEATH CAUSE" was recategorized into a binary format.
10. Methods - model development	

XSL•FO
RenderX

TRIPOD+AI item	Description of reporting in this study
10a. Model type	• CatBoostRegressor (for DFS and OS) and CatBoostClassifier (for DEATH CAUSE).
10b. Candidate predictors	• All 17 selected independent variables were used as candidate predictors for each model, based on the relevant dataset (df1 [DFS], df2 [OST], df3 [DEATH CAUSE]).
10c. Handling of continuous predictors	• Continuous predictors (AGE, BMI, TUMOR DIMENSION, SII) were used directly by CatBoost, which handles them internally.
10d. Handling of categorical predictors	• Categorical predictors were identified and explicitly converted to string type before training. CatBoost natively handles categorical features without explicit one-hot encoding.
10e. Details of model fitting	• CatBoostRegressor (iterations=1000, learning_rate=0.05, depth=6, loss_function=”RMSE,” eval_metric=”MAE,” early_stopping_rounds=50, random_seed=42, verbose=0). Similar configurations for CatBoostClassifier, with “Logloss” or “MultiClass” as loss function.
10f. Internal validation method	• Data were split into training (80%) and testing (20%) sets using train_test_split with random_state=42. Five-fold cross-validation (KFold, shuffle=True, random_state=42) was performed on the training set.
10g. Performance metrics	• Regression (DFS, OS): mean absolute error (MAE) • Classification (DEATH CAUSE): F_1-score for class 1 (tumor-related deaths), prioritizing recall
11. Assessment of prediction performance	• Performance was assessed on the independent test set. For classification, a confusion matrix was used.
12. Model interpretation methods	• CatBoost's built-in feature importance was used. Shapley additive explanations (SHAP) values were computed and visualized using violin plots and scatterplots to understand the individual contributions of each feature.

Results

Patient Characteristics

The study included a final cohort of 374 patients. After excluding incomplete records, the analytical sample sizes were 346 for DFS prediction, 347 for OS prediction, and 312 for death cause prediction. Some records indicate fewer than 374 patients, as not all characteristics were available for every individual in the population.

Table 3 presents the baseline clinical, pathological, and demographic characteristics of the study population, comprising 79.4% (297/374) men and 20.6% (77/374) women. A majority, 79.4% (296/373) were active smokers, and 21.4% (80/374) had a diagnosis of diabetes mellitus. Prior surgery was reported for 33.7% (126/374) of patients, while previous radiotherapy and systemic chemotherapy were less common, 5.1% (19/374) and 3.5% (13/374), respectively. Preoperative hydronephrosis was present in approximately one-third of cases, most frequently unilaterally.

XSL•FO
RenderX

Table . Characteristics of the patients in the dataset.

Characteristic	Value
Continuous variables	
Age (years), mean (SD)	75.2 (9.5)
BMI (kg/m^2), mean (SD)	26.6 (4.2)
Tumor dimension (cm), median (IQR)	2.2 (1.1 - 2.8)
SII[a], median (IQR)	654.7 (408.0 - 1047.0)
DFS[b] (months), median (IQR)	23.0 (6.0 - 52.8)
OS[c] (months), median (IQR)	29.0 (10.8 - 55.4)
Categorical variables, n/N (%)	
Sex	
Men	297/374 (79.4)
Women	77/374 (20.6)
Smoking status	
No	77/373 (20.6)
Yes	296/373 (79.4)
Diabetes mellitus	
No	294/374 (78.6)
Yes	80/374 (21.4)
Prior surgery	
No	248/374 (66.3)
Yes	126/374 (33.7)
Prior radiotherapy	
No	355/374 (94.9)
Yes	197/374 (5.1)
Prior chemotherapy	
No	361/374 (96.5)
Yes	13/374 (3.5)
Pre-hydronephrosis[d]	
None	257/374 (68.7)
Right	44/374 (11.8)
Left	40/374 (10.7)
Bilateral	33/374 (8.8)
Histological examination (H.E. TURV)	
Localized	138/372 (37.1)
Muscle-invasive	212/372 (57)
Squamous	22/372 (5.9)
Urinary diversion type	
Bricker ileal conduit	278/373 (74.5)
Ureterocutaneostomy	36/373 (9.7)
Vesicoileal pouch	59/373 (15.8)
Lymphovascular invasion	
Absent	158/371 (42.6)

XSL•FO
RenderX

Characteristic	Value
Present	213/371 (57.4)
Clinical tumor stage	
cTa	132/366 (36.1)
cTis	43/366 (11.7)
cT1	77/366 (21)
cT2	74/366 (20.2)
cT3	31/366 (8.5)
cT4	9/366 (2.5)
Tumor classification	
T0	16/342 (4.7)
Ta	18/342 (5.3)
Tis	60/342 (17.5)
T1	47/342 (13.7)
T2a	54/342 (15.8)
T2b	7/342 (2)
T3a	80/342 (23.4)
T3b	10/342 (2.9)
T4a	42/342 (12.3)
T4b	8/342 (2.4)
Bladder tumor position	
Intertrigonal zone	44/365 (12.1)
Right periosteal	25/365 (6.8)
Left periosteal	39/365 (10.7)
Dome	22/365 (6)
Posterior wall	58/365 (15.9)
Right lateral wall	71/365 (19.5)
Left lateral wall	58/365 (15.9)
Prostatic urethra	8/365 (2.2)
Anterior wall	26/365 (7.1)
Entire bladder	12/365 (3.3)
Bladder base	2/365 (0.5)
Cause of death	
Alive	205/363 (56.5)
Other	71/363 (19.6)
Cancer	87/363 (23.9)

[a]SII: systemic immune-inflammation index.

[b]DFS: disease-free survival.

[c]OS: overall survival.

[d]Pre-hydronephrosis: pretreatment hydronephrosis.

Histologically, 57% (212/372) of tumors were muscle invasive, while 37.1% (138/372) were localized to the mucosa or submucosa, and 5.9% (22/372) exhibited squamous features. The most common urinary diversion method was Bricker ileal conduit (278/373, 74.5%), followed by vesicoileal pouch construction (59/373, 15.8%) and ureterocutaneostomy (36/373, 9.7%). Lymphovascular invasion was observed in 57.4% (213/371) of patients.

In terms of staging, the most frequent clinical tumor stages were cTa (132/366, 36.1%) and cT1 (77/366, 21%), while advanced

XSL•FO
RenderX

stages (cT3 and cT4) were less common (40/366, 11%). Tumor classification was heterogeneous, with T3a (80/342, 23.4%) and Tis (60/342, 17.5%) being most prevalent.

Regarding tumor location, the most frequent sites were the right lateral wall (71/365, 19.5%), the posterior wall (58/365, 15.9%), and the left lateral wall (58/365, 15.9%). At the time of data collection, of 363 patients, 205 (56.5%) were alive, 87 (24%) had died due to cancer-related causes, and 71 (19.6%) had died from other causes.

DFS Prediction

The CatBoostRegressor model was trained to predict DFS in months. Input variables are indicated in Table 1. After 5-fold cross-validation and manual hyperparameter tuning, the model achieved an MAE of 18.68 months, indicating that, on average, the model's predictions deviated by approximately 1.5 years from the observed DFS.

Figure 1 presents the global feature importance ranking from the CatBoost model trained to predict DFS. This ranking reflects the contribution of each variable to reducing the model's prediction across all patients. The most influential predictor was the clinical tumor stage, with an importance score of approximately 17, followed by the pathological tumor classification, with an importance score of approximately 14, reflecting the role of tumor invasiveness, local extension, and accurate tumor staging in DFS prediction. The systemic immune-inflammation index (SII) ranked highly, with a score of approximately 9.5. To a lesser extent, demographic and anatomical variables, such as BMI, age, and tumor dimension, also contributed to the model.

Figure 1. Global feature importance ranking to predict disease-free survival. CTS: clinical tumor stage; DM: diabetes mellitus; H.E. TURV: histological examination for transurethral resection of the bladder; LVI: lymphovascular invasion; PRE-HYDRONEPHROSIS: pretreatment hydronephrosis; SII: systemic immune-inflammation index; TC: tumor classification; UD: urinary diversion type.

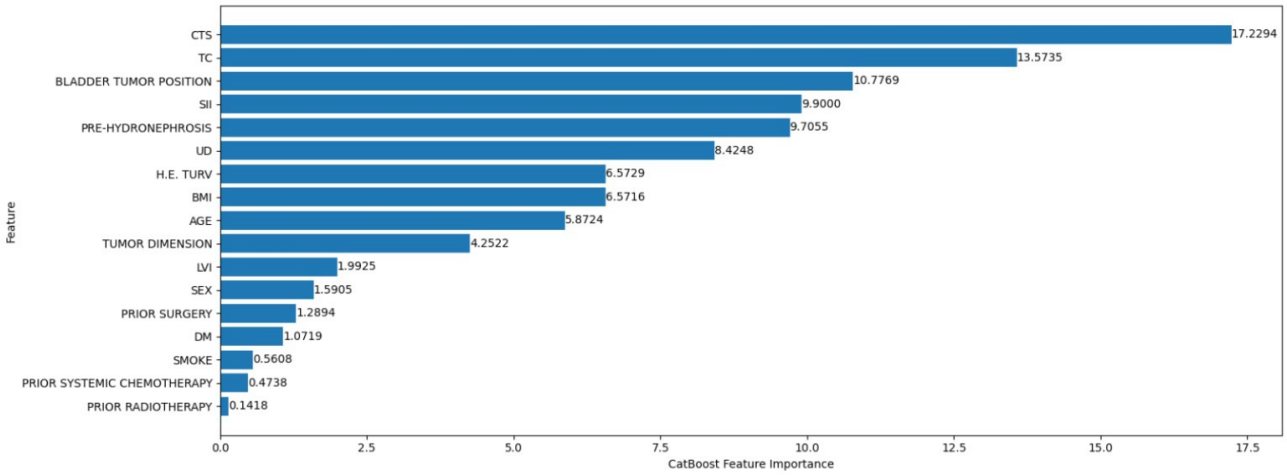
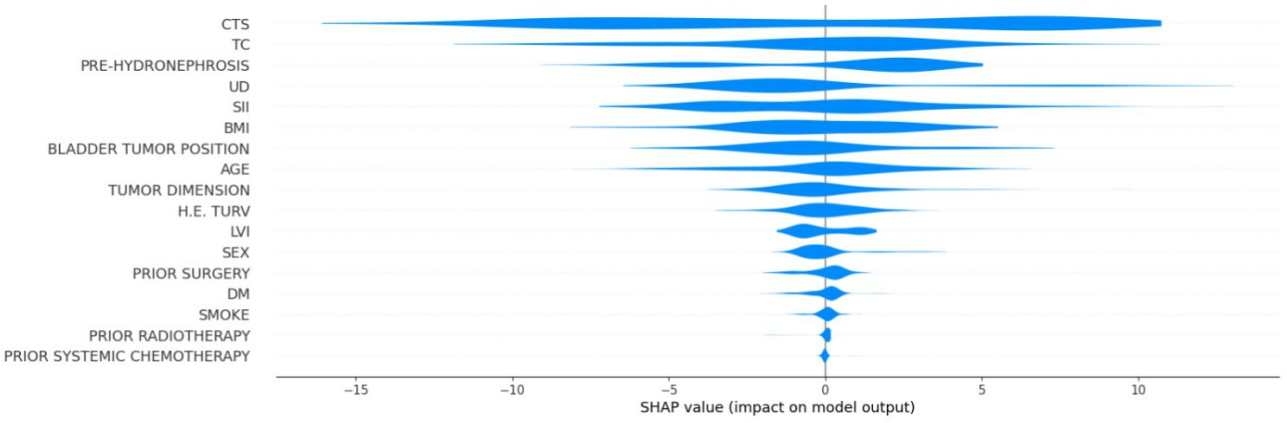


Figure 2 displays the SHAP summary plot for the DFS model, illustrating the distribution and direction of impact of each feature on the predicted DFS across all patients. Clinical tumor stage and tumor classification exhibited the widest distribution of SHAP values, confirming their dominant influence, where the predicted DFS substantially increased or decreased depending on their values. SII displayed a more balanced distribution, with both positive and negative effects depending on the value. In contrast, features such as prior treatment (eg, surgery, radiotherapy, and chemotherapy) and lifestyle factors (eg, smoking status and diabetes) had a SHAP distribution clustered near 0, indicating limited predictive power.

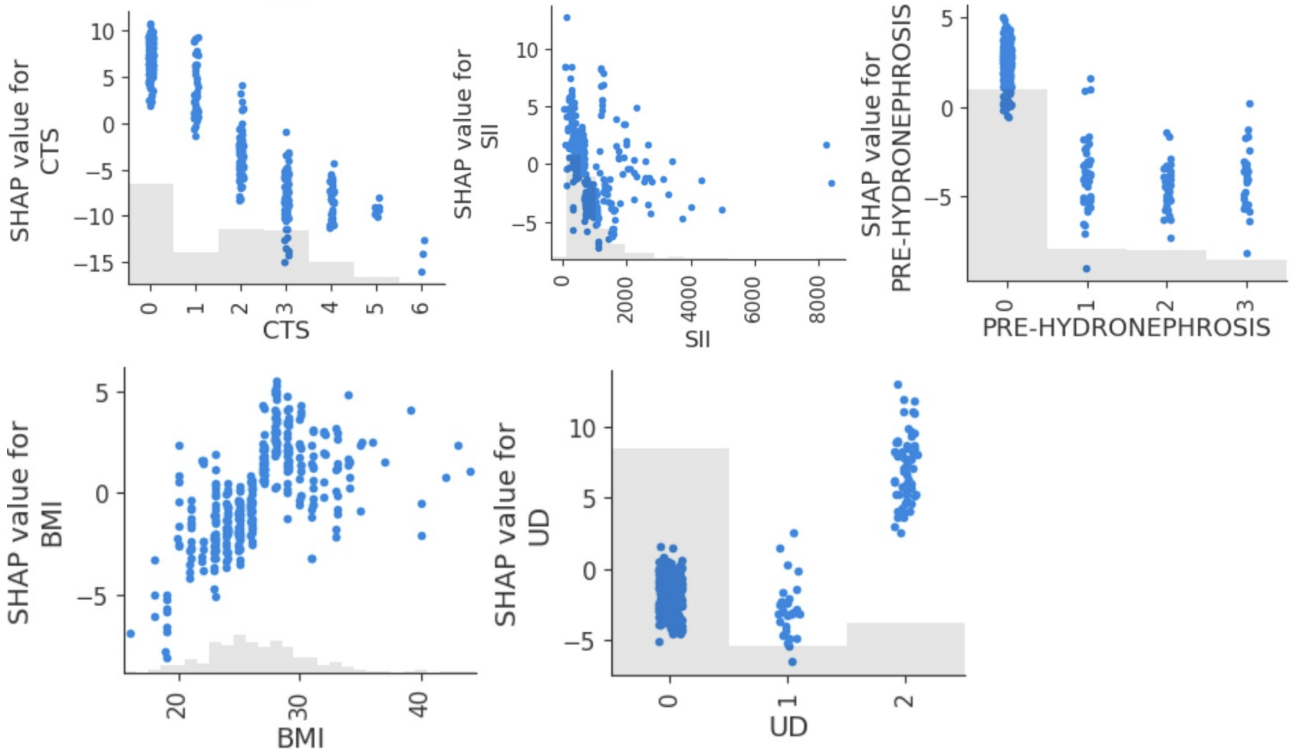
Figure 2. Violin plot of feature influence on disease-free survival prediction from the Shapley additive explanations (SHAP) analysis. CTS: clinical tumor stage; DM: diabetes mellitus; H.E. TURV: histological examination for transurethral resection of the bladder; LVI: lymphovascular invasion; PRE-HYDRONEPHROSIS: pretreatment hydronephrosis; SII: systemic immune-inflammation index; TC: tumor classification; UD: urinary diversion type.



XSL•FO
RenderX

Figure 3 presents the SHAP dependence plots for 4 of the most influential features affecting DFS predictions. The x-axis shows the feature value, and the y-axis shows the SHAP value (ie, the impact on the model's output). Clinical tumor stage showed a strong negative relationship with predicted DFS: as tumor stage increased, the SHAP values shifted sharply downward, indicating a consistent reduction in predicted DFS, aligning with the known prognostic role of tumor invasiveness in bladder cancer. SII demonstrated a nonlinear relationship, showing that patients with lower SII values had better SHAP values, while those with elevated SII showed increasingly negative impacts on DFS. This suggests a threshold effect, where systemic inflammation beyond a certain level contributes to poorer prognosis. The presence of pretreatment hydronephrosis had a negative impact on DFS prediction. Patients with low BMI had negative SHAP values, indicating reduced DFS, while those with moderate BMI experienced mildly negative predictions. At a BMI greater than 28, the SHAP values became positive, suggesting a potential protective effect exerted by higher BMI. Regarding the type of urinary diversion, vesicoileal pouch construction showed positive SHAP values, while other approaches had negative SHAP values.

Figure 3. Shapley additive explanations (SHAP) scatterplots for the 5 most significant features influencing disease-free survival predictions (with BMI in kg/m^2). CTS: clinical tumor stage (0: cTa, 1: cTis, 2: cT1, 3: cT2, 4: cT3, 5: cT4); PRE-HYDRONEPHROSIS: pretreatment hydronephrosis (0: no, 1: right hydronephrosis, 2: left hydronephrosis, 3: bilateral hydronephrosis); SII: systemic immune-inflammation index; UD: urinary diversion type (0: Bricker ileal conduit, 1: ureterocutaneostomy, 2: vesicoileal pouch).



OS Prediction

For OS prediction, the CatBoost model achieved an MAE of 17.2 months across the entire patient cohort. When the analysis was restricted to the subgroup of patients who had died (n=156), the prediction error improved to 15.8 months. After filtering features by importance, using a threshold of <0.5, the MAE further improved to 14.6, suggesting that a more compact feature set may improve predictive efficiency without compromising accuracy (Figure 4). This final model was selected for interpretation, as it maintained accuracy while reducing complexity.

Figure 4. Progressive improvement in model accuracy. MAE: mean absolute error.

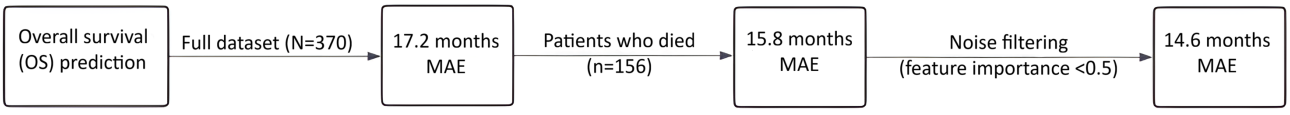


Figure 5 presents the CatBoost feature importance ranking for the best-performing OS prediction model. Tumor classification emerged as the most influential predictor of OS in the final model with a value of approximately 17.5. SII followed closely with a value of approximately 15.5, highlighting the role of systemic inflammation in cancer progression and survival outcomes. The third feature was represented by histological findings (H.E. TURV). Other influencing factors included clinical tumor stage, pretreatment hydronephrosis, type of urinary diversion, BMI, and age.

XSL•FO
RenderX

Figure 5. CatBoost feature importance ranking for the best-performing overall survival prediction model. CTS: clinical tumor stage; H.E. TURV: histological examination for transurethral resection of the bladder; LVI: lymphovascular invasion; PRE-HYDRONEPHROSIS: pretreatment hydronephrosis; SII: systemic immune-inflammation index; TC: tumor classification; UD: urinary diversion type.

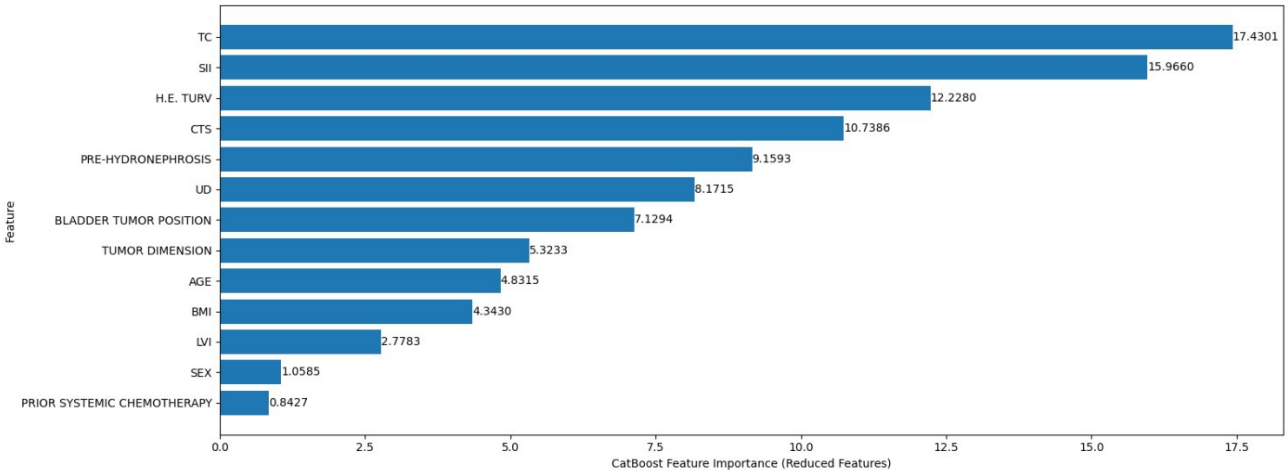


Figure 6 displays the SHAP summary plot for the final OS prediction model. As expected, the most impactful variable was tumor classification, which showed a broad distribution. Pretreatment hydronephrosis and SII exhibited a wide SHAP distribution. Clinical tumor stage and histological findings (H.E. TURV) showed a similar overall effect on prognosis.

Figure 6. Shapley additive explanations (SHAP) violin summary plot for the final overall survival prediction model. CTS: clinical tumor stage; H.E. TURV: histological examination for transurethral resection of the bladder; LVI: lymphovascular invasion; PRE-HYDRONEPHROSIS: pretreatment hydronephrosis; SII: systemic immune-inflammation index; TC: tumor classification; UD: urinary diversion type.

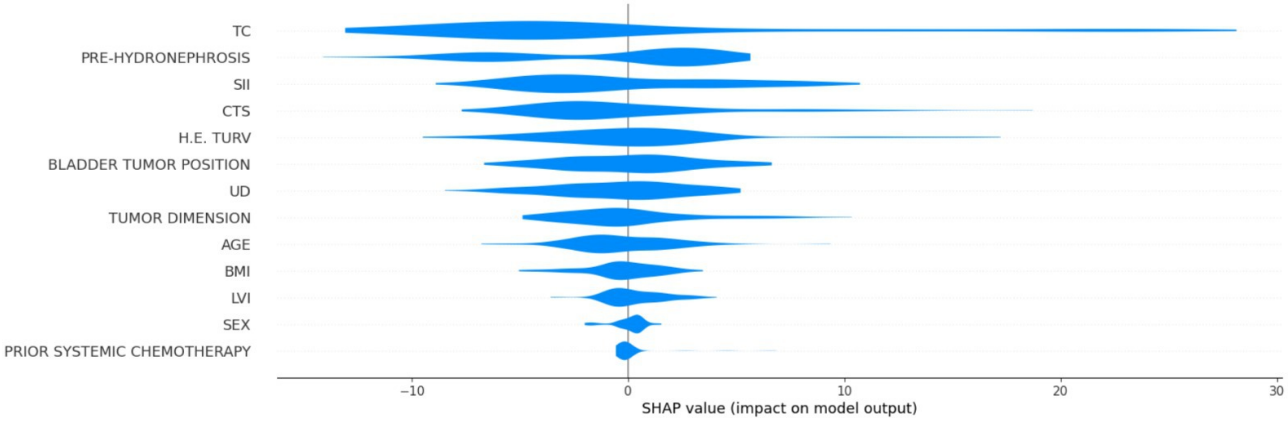
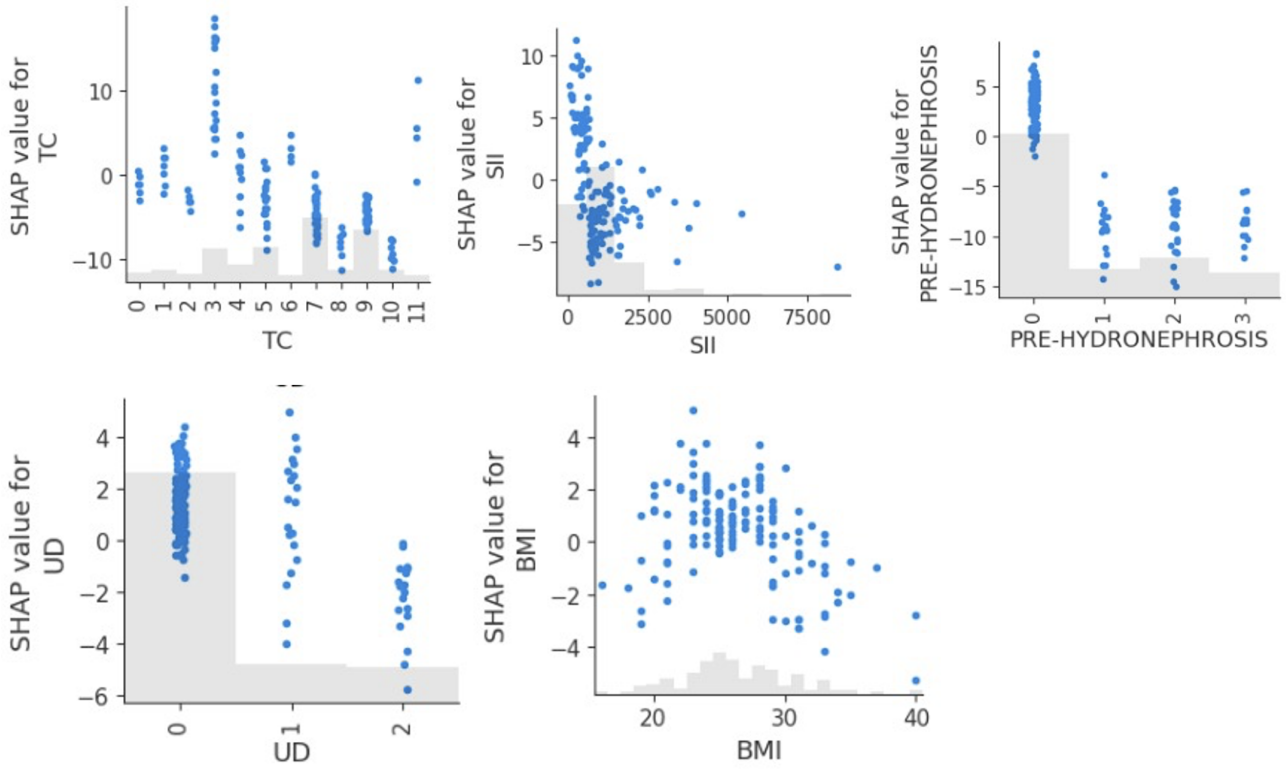


Figure 7 presents the SHAP dependence plots for 5 key features influencing OS predictions. The tumor classification SHAP values indicated that patients with in situ cancers (value 3) had the best OS prediction, which gradually declined as the tumor stage advanced. SII showed a threshold effect, where predictions remained relatively stable up to a value of approximately 1000, then fell sharply, indicating that elevated inflammation is associated with a poor overall outcome. Pretreatment hydronephrosis was strongly linked to reduced predicted OS, where patients with this condition had uniformly negative SHAP values, not influenced by bilaterality. At the same time, BMI demonstrated a nonlinear pattern, where patients with very low BMIs had reduced OS predictions, moderate BMIs were associated with better outcomes, and the SHAP values began to decline again at higher BMI values, suggesting that both underweight and obesity may be associated with increased mortality risk in this population. The type of urinary diversion showed a different impact than that observed in DFS prediction, with vesicoileal pouch construction being associated with a lower OS.

XSL•FO
RenderX

Figure 7. Shapley additive explanations (SHAP) scatterplots for the 5 most influential features influencing overall survival prediction (with BMI in kg/m^2). PRE-HYDRONEPHROSIS: pretreatment hydronephrosis (0: no, 1: right hydronephrosis, 2: left hydronephrosis, 3: bilateral hydronephrosis); SII: systemic immune-inflammation index; TC: tumor classification (1: T0, 2: Ta, 3: Tis, 4: T1, 5: T2a, 6: T2b, 7: T3a, 8: T3b, 9: T4a, 10: T4b); UD: urinary diversion type (0: Bricker ileal conduit, 1: ureterocutaneostomy, 2: vesicoileal pouch).



Cause-of-Death Classification

The CatBoostClassifier was trained to predict whether a patient's death was tumor related. Due to class imbalance, only 14 of 78 deaths were cancer related; custom class weights and a reduced decision threshold of 0.12 were applied to maximize recall and minimize false negatives. The final model achieved a recall of 78.6% (Figure 8), correctly identifying 11 of 14 tumor-related deaths. The overall F_1-score for the positive class was 0.44, with a precision of 31%. The model prioritizes sensitivity over specificity.

XSL•FO
RenderX

Figure 8. Confusion matrix for cause-of-death classification.

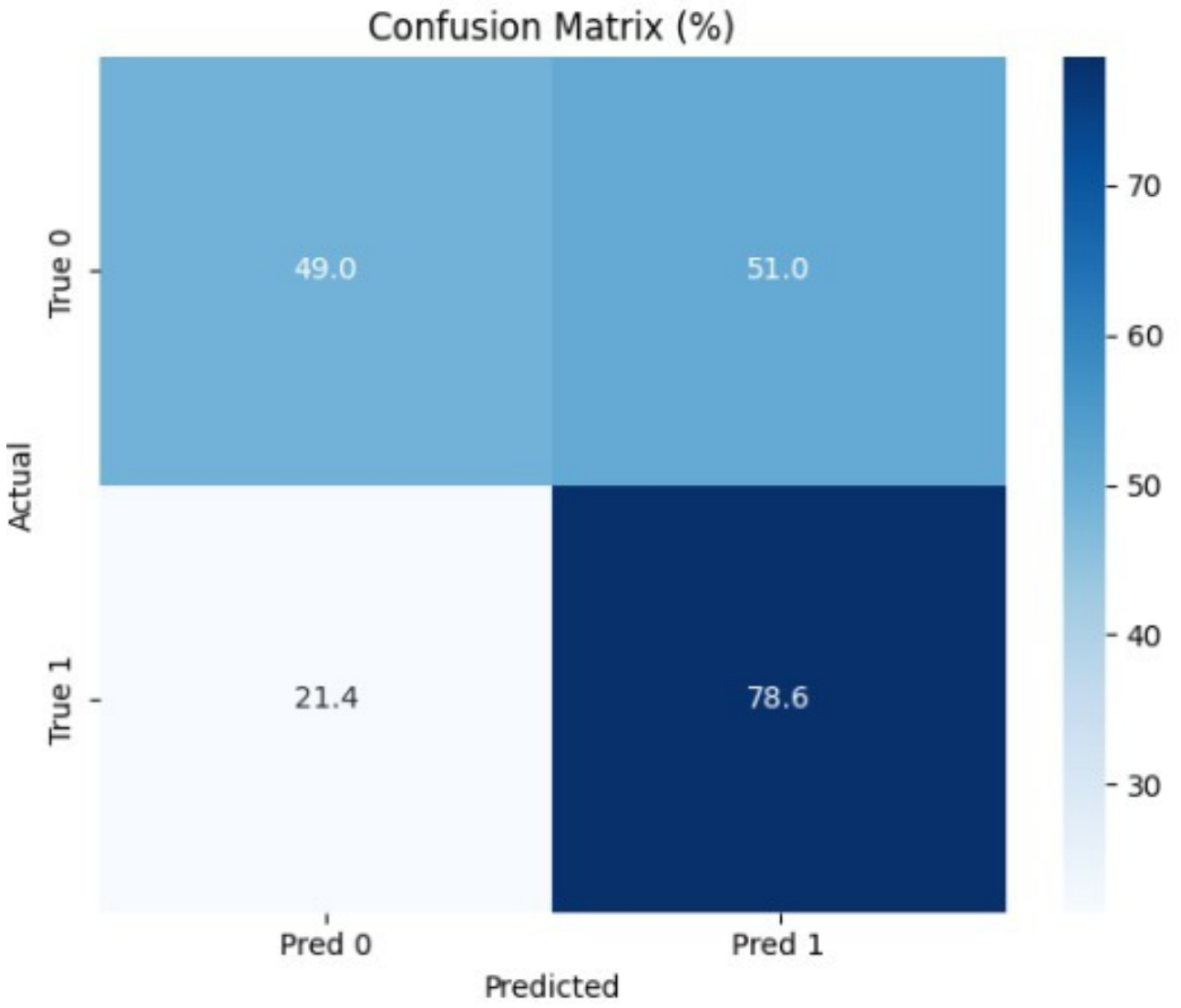
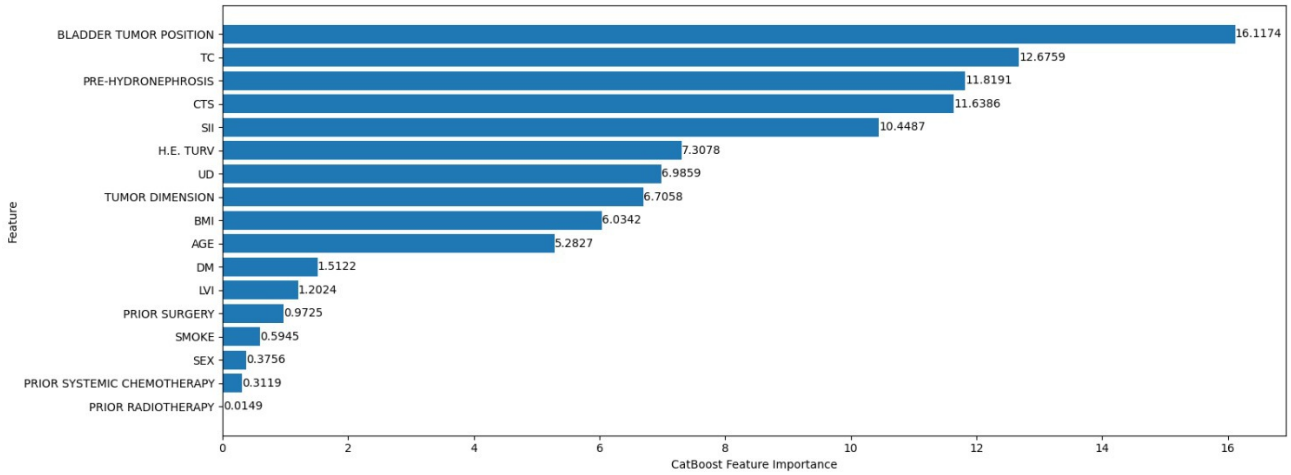


Figure 9 represents the CatBoost feature importance ranking for the death cause classification model. The most influential feature was the anatomical position of the bladder tumor, with a value of approximately 16.5. This was followed by tumor classification and pretreatment hydronephrosis, both indicators of disease severity and progression.

Figure 9. CatBoost feature importance ranking for the cause-of-death classification. CTS: clinical tumor stage; DM: diabetes mellitus; H.E. TURV: histological examination for transurethral resection of the bladder; LVI: lymphovascular invasion; PRE-HYDRONEPHROSIS: pretreatment hydronephrosis; SII: systemic immune-inflammation index; TC: tumor classification; UD: urinary diversion type.



XSL•FO
RenderX

Other key features included the clinical tumor stage and SII, with values of approximately 11 and 11.5, respectively.

Figure 10 displays the SHAP summary plot for the tumor-related death classification model. The feature with the widest and most impactful distribution was bladder tumor position, which is addressed in detail in the discussion of Figure 11. SII was also influential, with positive and negative SHAP values. Tumor classification, pretreatment hydronephrosis, and clinical tumor stage generally acted to slightly decrease the predicted risk of tumor-related death for most patients, with little variability in their effect.

Figure 10. Shapley additive explanations (SHAP) summary plot for the tumor-related death classification model. CTS: clinical tumor stage; DM: diabetes mellitus; H.E. TURV: histological examination for transurethral resection of the bladder; LVI: lymphovascular invasion; PRE-HYDRONEPHROSIS: pretreatment hydronephrosis; SII: systemic immune-inflammation index; TC: tumor classification; UD: urinary diversion type.

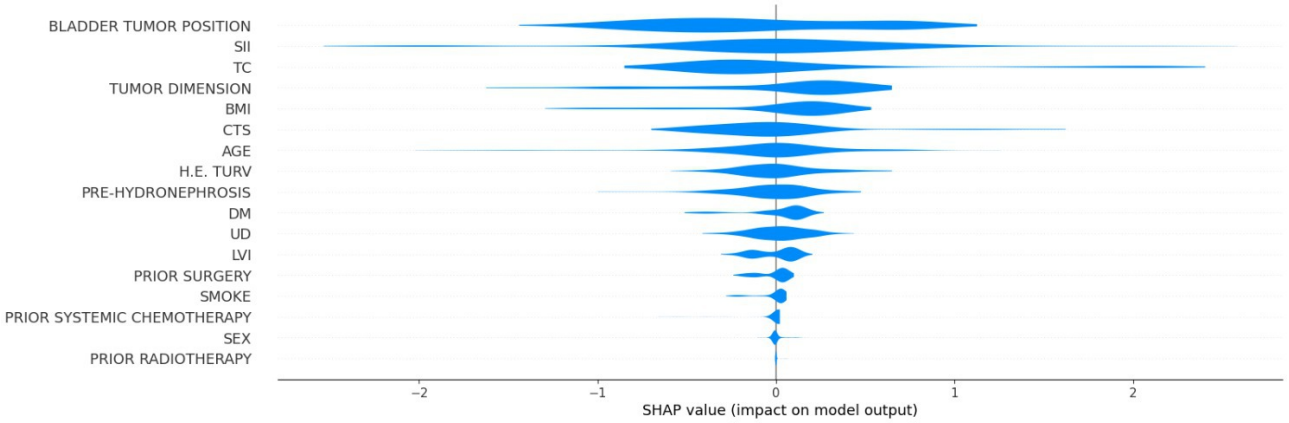


Figure 11. Shapley additive explanations (SHAP) scatterplots for the 5 most influential features influencing the tumor-related death classification model (with BMI in kg/m^2, AGE in years, and TUMOR DIMENSION in cm; bladder tumor position: 0: intertrigonal zone, 1: right periosteal, 2: left periosteal, 3: dome, 4: posterior wall, 5: right lateral wall, 6: left lateral wall, 7: prostatic urethra, 8: anterior wall, 9: entire bladder, 10: bladder base). SII: systemic immune-inflammation index.

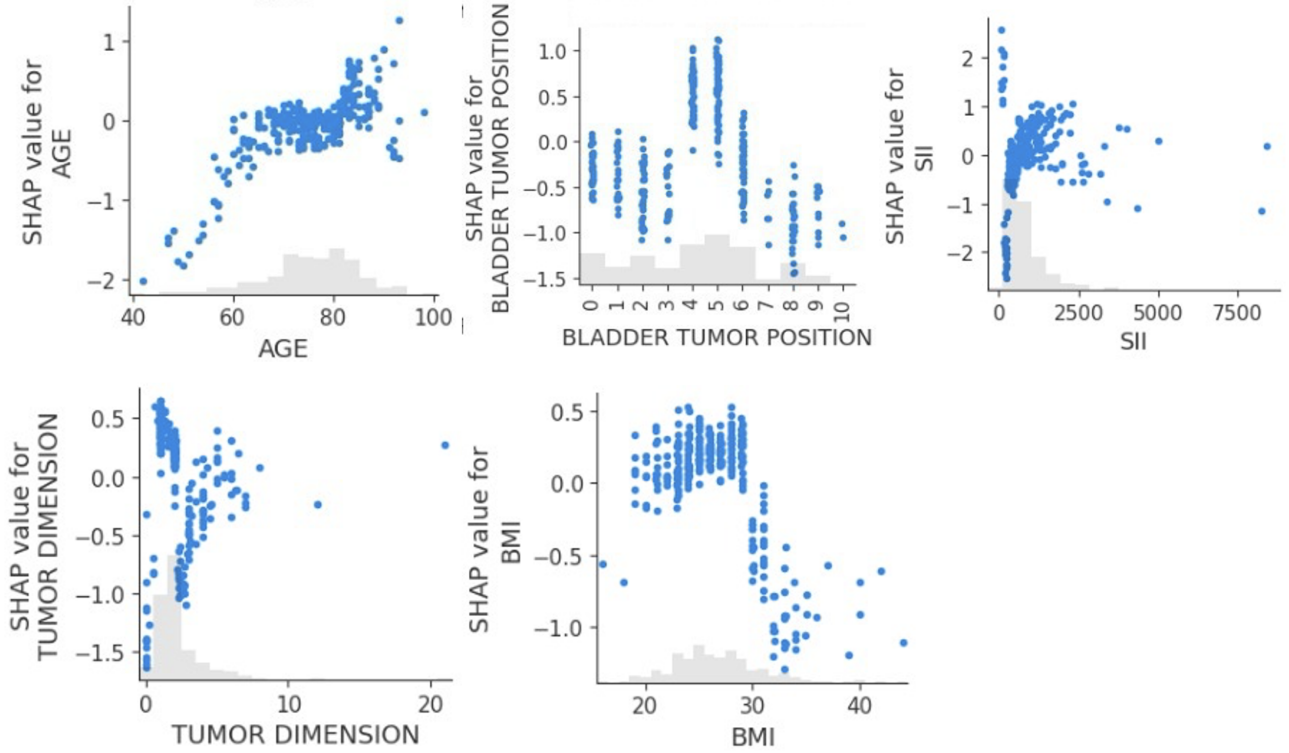


Figure 11 presents the SHAP dependence plots for 5 key features influencing the tumor-related classification model: age, bladder tumor position, SII, tumor dimension, and BMI.

The scatterplot for age exhibited a clear positive trend, showing that as patient age increased, the SHAP value for age also increased, indicating that older age consistently increased the predicted risk of tumor-related death. Bladder tumor position was another influential predictor in the tumor-related death classification model, highlighting that tumors located in the posterior wall and right lateral wall were more likely to be the cause of death. In contrast, tumors located in the anterior wall and at the bladder base or those spreading throughout the entire bladder were less likely to be the cause of death. Increasing SII

XSL•FO
RenderX

and tumor dimension also had a moderate predictive value, respectively increasing and decreasing the probability that the patient's cause of death was cancer. Finally, patients with a higher BMI showed a higher likelihood that the cause of death was cancer.

Discussion

Principal Findings

This paper presents the development of a survival prediction model using machine learning approaches for patients with bladder cancer. Our findings demonstrate that modern predictive algorithms show promising accuracy in forecasting DFS, OS, and cause of death. The limited sample size and the paucity of included categories for the analysis suggest that predictive algorithms trained with additional data and variables significantly improve the demonstrated accuracy.

The observation that age showed a positive correlation with survival outcomes is particularly intriguing and seemingly counterintuitive. This "age paradox" has been previously described in bladder cancer and other oncological settings and may reflect a combination of selection bias and underlying tumor biology rather than a true protective effect of age, which can be explained by several factors. Older patients often receive more conservative treatment, which potentially leads to selection bias in surgical candidates. Additionally, younger patients with bladder cancer have been reported to present more frequently with aggressive disease variants, which could account for their relatively poorer outcomes despite younger age [11]. These results align with recent systematic reviews and meta-analyses on radical cystectomy, which consistently report high complication rates and significant variability in postoperative survival across risk profiles [12]. Importantly, as this study is observational, the association between age and survival should be interpreted cautiously and not as a causal relationship.

Clinical tumor stage was the strongest predictor, which aligns with established prognostic factors in bladder cancer [13]. Additionally, inflammatory markers, particularly SII, showed a negative correlation with survival outcomes, supporting recent findings in other malignancies [14]. This relationship likely reflects the complex interplay between systemic inflammation and cancer progression, where elevated SII indicates a protumoral inflammatory state [15]. Our findings on systemic inflammatory indices are consistent with recent data indicating that platelet-to-lymphocyte ratio, systemic inflammation response index, pan-immune-inflammation value, SII, and neutrophil-to-lymphocyte ratio are associated with adverse outcomes in non–muscle-invasive bladder cancer [16].

SHAP analysis revealed a clear, monotonic decline in predicted survival with advancing clinical tumor stage, reinforcing its primary prognostic role in both DFS and OS. SII, by contrast, demonstrated a threshold effect where values above approximately 1000 were associated with a sharp drop in predicted DFS, suggesting a nonlinear relationship between systemic inflammation and patient outcomes. The observed association between urinary diversion type and survival outcomes should be considered exploratory. Previous studies had found that orthotopic neobladder reconstruction had a protective effect against urethral recurrence in male patients undergoing radical cystectomy for bladder cancer [17]. While this effect was not observed in our dataset, we found a positive association between vesicoileal pouch construction and improved survival. This association may reflect both patient selection and potential physiological advantages of this diversion type; however, given the limited number of patients within each diversion subgroup, it should be interpreted cautiously. Confounding factors such as surgical expertise and patient characteristics, which were not accounted for in the present study, may have influenced these findings. BMI showed a similar intriguing relationship: patients with an unhealthy BMI, either high or low, showed poorer outcomes; this may be related both to tumor characteristics and the surgical approach being limited in terms of radicality. This U-shaped association between BMI and survival was consistently observed across both DFS and OS outcomes, with moderate BMI ranges correlated with more favorable SHAP values. The findings support a metabolic vulnerability in patients with underweight as well as obesity, which may influence recovery or treatment tolerance.

Our machine learning models achieved prediction accuracies comparable to those reported in previous studies. The accuracy in cause-of-death prediction, although modest, represents an encouraging level, given the limited resources and the paucity of categories considered for the analysis, particularly when compared with studies published a few years ago that used significantly larger samples yet achieved marginally higher accuracy in mortality and recurrence prediction [18]. A recently published systematic review investigating machine learning algorithms for bladder cancer cystectomy outcomes found that most of the algorithms did not exceed 70% accuracy and, in some cases, performed with approximately 60% accuracy [19]. The integration of SII into predictive models represents an auspicious direction. As a low-cost, readily available biomarker, SII could enhance current prognostic tools without adding significant complexity or cost to patient evaluation [14].

A notable limitation of our study is the relatively high MAE in survival predictions. These MAE values render the algorithm unsuitable for precise individual patient counseling or treatment planning where accurate timing is critical, such as in emergency settings or for patients exhibiting postoperative complications [20]. However, this level of accuracy remains acceptable for clinical trial patient stratification and allocation, particularly in trials where broad risk categories rather than precise survival estimates are needed for randomization. Such applications include balancing treatment arms in clinical trials by identifying comparable risk groups or supporting enrollment decisions in competing risk analysis, where precise timing is less critical than overall risk assessment [21].

Limitations and Reproducibility

This study is subject to some limitations when interpreting the results. The relatively limited dataset size (N=370 initially; reduced to 312 - 347 for specific analyses) inherently constrains the generalizability and robustness of the developed models. While machine learning algorithms such as CatBoost are robust

XSL•FO
RenderX

on smaller datasets, their predictive power can be substantially enhanced with larger cohorts.

Secondly, the monocentric nature of the data collection, originating solely from Fondazione Policlinico Universitario Agostino Gemelli IRCCS in Rome, Italy, introduces a potential for selection bias and limits external validity. Patient characteristics, treatment protocols, and population demographics can vary significantly across institutions and geographical regions. The findings from this study may not be directly transferable to other clinical settings without further validation on diverse, external datasets.

Thirdly, while rigorous data cleaning was performed, the inherent human factors associated with retrospective data extraction from medical records cannot be eliminated.

Conclusion

Our study demonstrates the potential utility of machine learning approaches in predicting bladder cancer outcomes following cystectomy. While the achieved accuracy levels are modest, they align with current literature benchmarks and provide a foundation for future development. The identification of clinical tumor stage as the primary predictor, along with the consistent negative correlation of SII with survival outcomes, validates these parameters as valuable prognostic indicators. In particular, the SHAP analysis revealed a monotonic decline in predicted DFS and OS with advancing clinical tumor stage, reaffirming its role in risk stratification. On the other hand, SII exhibited a threshold effect, where values above approximately 1000 were associated with a rapid drop in predicted survival, reinforcing the adverse prognostic impact of systemic inflammation. The current model's performance, though not suitable for precise individual prognostication, shows particular promise for clinical trial stratification and cohort allocation. Future studies with larger datasets and additional predictive variables may enhance the model's accuracy and broaden its clinical applications. Integrating readily available biomarkers, such as SII, represents a cost-effective approach to improving prognostic tools. These findings contribute to the growing body of evidence supporting the role of machine learning in oncological decision-making while acknowledging the need for continued refinement and validation in larger cohorts.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

To ensure the reproducibility of our results and facilitate further research, all code used for this analysis has been made publicly available on Figshare [10].

Authors' Contributions

FAC, BR, and PR elaborated on the first manuscript concept. AN and FAC performed the statistical analysis. FAC, VDV, and PR wrote the article. MS, NF, and GM reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Anagnostou T, Remzi M, Lykourinas M, Djavan B. Artificial neural networks for decision-making in urologic oncology. Eur Urol 2003 Jun;43(6):596-603. [doi: 10.1016/s0302-2838(03)00133-7] [Medline: 12767358]
2. Schwarzer G, Schumacher M. Artificial neural networks for diagnosis and prognosis in prostate cancer. Semin Urol Oncol 2002 May;20(2):89-95. [doi: 10.1053/suro.2002.32492] [Medline: 12012294]
3. Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. Cancer 2001 Apr 15;91(8 Suppl):1636-1642. [doi: 10.1002/1097-0142(20010415)91:8+<1636::aid-cncr1176>3.0.co;2-d] [Medline: 11309761]
4. Zi H, Liu MY, Luo LS, et al. Global burden of benign prostatic hyperplasia, urinary tract infections, urolithiasis, bladder cancer, kidney cancer, and prostate cancer from 1990 to 2021. Mil Med Res 2024 Sep 18;11(1):64. [doi: 10.1186/s40779-024-00569-w] [Medline: 39294748]
5. Abbod MF, Catto JWF, Linkens DA, Hamdy FC. Application of artificial intelligence to the management of urological cancer. J Urol 2007 Oct;178(4 Pt 1):1150-1156. [doi: 10.1016/j.juro.2007.05.122] [Medline: 17698099]
6. Di Sarno L, Caroselli A, Tonin G, et al. Artificial intelligence in pediatric emergency medicine: applications, challenges, and future perspectives. Biomedicines 2024 May 30;12(6):1220. [doi: 10.3390/biomedicines12061220] [Medline: 38927427]
7. Mazzotta AD, Burti E, Causio FA, et al. Machine learning approaches for the prediction of postoperative major complications in patients undergoing surgery for bowel obstruction. J Pers Med 2024 Oct 8;14(10):1043. [doi: 10.3390/jpm14101043] [Medline: 39452550]
8. Tokuc E, Eksi M, Kayar R, et al. Inflammation indexes and machine-learning algorithm in predicting urethroplasty success. Investig Clin Urol 2024 May;65(3):240-247. [doi: 10.4111/icu.20230302] [Medline: 38714514]

XSL•FO
RenderX

9. Yin M, Joshi M, Meijer RP, et al. Neoadjuvant chemotherapy for muscle-invasive bladder cancer: a systematic review and two-step meta-analysis. Oncologist 2016 Jun;21(6):708-715. [doi: 10.1634/theoncologist.2015-0440] [Medline: 27053504]
10. Bladder cancer CatBoost analysis. Figshare. URL: https://figshare.com/s/ec2419a33ad7ad05f412 [accessed 2026-01-03]
11. Withrow DR, Nicholson BD, Morris EJA, Wong ML, Pilleron S. Age-related differences in cancer relative survival in the United States: a SEER-18 analysis. Int J Cancer 2023 Jun 1;152(11):2283-2291. [doi: 10.1002/ijc.34463] [Medline: 36752633]
12. Sacco E, Campetella M, Marino F, et al. Preoperative risk factors for failure after fixed sling implantation for postprostatectomy stress urinary incontinence (FORESEE): a systematic review and meta-analysis. Eur Urol Open Sci 2025 Jun;77:58-77. [doi: 10.1016/j.euros.2025.05.012] [Medline: 40546391]
13. Kardoust Parizi M, Enikeev D, Glybochko PV, et al. Prognostic value of T1 substaging on oncological outcomes in patients with non-muscle-invasive bladder urothelial carcinoma: a systematic literature review and meta-analysis. World J Urol 2020 Jun;38(6):1437-1449. [doi: 10.1007/s00345-019-02936-y] [Medline: 31493109]
14. Kou J, Huang J, Li J, Wu Z, Ni L. Systemic immune-inflammation index predicts prognosis and responsiveness to immunotherapy in cancer patients: a systematic review and meta-analysis. Clin Exp Med 2023 Nov;23(7):3895-3905. [doi: 10.1007/s10238-023-01035-y] [Medline: 36966477]
15. Wang Q, Zhu D. The prognostic value of systemic immune-inflammation index (SII) in patients after radical operation for carcinoma of stomach in gastric cancer. J Gastrointest Oncol 2019 Oct;10(5):965-978. [doi: 10.21037/jgo.2019.05.03] [Medline: 31602335]
16. Bizzarri FP, Campetella M, Russo P, et al. Prognostic value of PLR, SIRI, PIV, SII, and NLR in non-muscle invasive bladder cancer: can inflammatory factors influence pathogenesis and outcomes? Cancers (Basel) 2025 Jun 28;17(13):2189. [doi: 10.3390/cancers17132189] [Medline: 40647487]
17. Yu J, Lee CU, Chung JH, et al. Impact of urinary diversion type on urethral recurrence following radical cystectomy for bladder cancer: propensity score matched and weighted analyses of retrospective cohort. Int J Surg 2024 Feb 1;110(2):700-708. [doi: 10.1097/JS9.0000000000000904] [Medline: 38000052]
18. Hasnain Z, Mason J, Gill K, et al. Machine learning models for predicting post-cystectomy recurrence and survival in bladder cancer patients. PLoS One 2019;14(2):e0210976. [doi: 10.1371/journal.pone.0210976] [Medline: 30785915]
19. Liu YS, Thaliffdeen R, Han S, Park C. Use of machine learning to predict bladder cancer survival outcomes: a systematic literature review. Expert Rev Pharmacoecon Outcomes Res 2023;23(7):761-771. [doi: 10.1080/14737167.2023.2224963] [Medline: 37306511]
20. Qi S, Kumar N, Farrokh M, et al. An effective meaningful way to evaluate survival models. arXiv. Preprint posted online on Jun 1, 2023. [doi: 10.48550/arXiv.2306.01196]
21. Hassan N, Slight R, Morgan G, et al. Road map for clinicians to develop and evaluate AI predictive models to inform clinical decision-making. BMJ Health Care Inform 2023 Aug;30(1):e100784. [doi: 10.1136/bmjhci-2023-100784] [Medline: 37558245]

Abbreviations

AI: artificial intelligence
DFS : disease-free survival
MAE: mean absolute error
OS: overall survival
SHAP: Shapley additive explanations
SII: systemic immune-inflammation index
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by N Rohatgi; submitted 28.Oct.2025; peer-reviewed by E Tokuc, FP Bizzarri; revised version received 09.Jan.2026; accepted 13.Jan.2026; published 19.Feb.2026.

Please cite as:
Causio FA, De Vita V, Nappi A, Sawaya M, Rocco B, Foschi N, Maioriello G, Russo P
Survival Prediction in Patients With Bladder Cancer Undergoing Radical Cystectomy Using a Machine Learning Algorithm: Retrospective Single-Center Study
JMIR Perioper Med 2026;9:e86666
URL: https://periop.jmir.org/2026/1/e86666
doi:10.2196/86666

XSL•FO
RenderX

© Francesco Andrea Causio, Vittorio De Vita, Andrea Nappi, Melissa Sawaya, Bernardo Rocco, Nazario Foschi, Giuseppe Maioriello, Pierluigi Russo. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 19.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Physician Perspectives on ChatGPT-4o as a Patient Resource for Abdominal Cancer Surgeries: Cross-Sectional Survey

Christina V Lindsay[1,2], BS; Devika A Shenoy[1,2], BS; Allison N Martin[3], MPH, MD; Christie L Clipper[2], DHA; Kevin N Shah[3], MD; Michael E Lidsky[3], MD; Daniel P Nussbaum[3], MD; Ralph Snyderman[2,4], MD

[1]School of Medicine, Duke University, 8 Searle Center Dr, Durham, NC, United States

[2]Center for Personalized Health Care, Duke University, Durham, NC, United States

[3]Department of Surgery, Duke University, Durham, NC, United States

[4]Department of Medicine, Duke University, Durham, NC, United States

Corresponding Author:
Christina V Lindsay, BS
School of Medicine, Duke University, 8 Searle Center Dr, Durham, NC, United States

Abstract

Background: Artificial intelligence (AI) models are being increasingly integrated into clinical care. Moreover, the availability of publicly accessible AI resources makes them attractive to patients seeking clinical information. Little is known regarding the use of large language models as patient resources for navigating major cancer diagnoses.

Objective: This study aimed to evaluate the content, readability, and safety of ChatGPT (OpenAI; GPT-4o)-generated responses to common perioperative queries about hepatic, pancreatic, and colon cancers.

Methods: A 28-question survey was developed based on frequently asked surgical questions for select malignancies. Surgical oncologists rated ChatGPT-4o-generated responses on a 5-point Likert scale for accuracy, quality, and tangibility. Readability was assessed using the Flesch-Kincaid Reading Grade Level (FKRGL) and Flesch Reading Ease (FRE). Respondents provided free-text comments and reported their comfort with patients using ChatGPT. Survey completion implied consent.

Results: A total of 7 attending surgical oncologists with a median of 7 (IQR 4-13) years in practice completed the survey. Responses received mean scores of 3.5/5 (SD 0.28) for quality, 3.6/5 (SD 0.34) for accuracy, and 3.6/5 (SD 0.29) for tangibility. The responses had a median FKRGL score of 14.6 (IQR 13.3-15.6) and FRE score of 29.4 (IQR 20.5-36.3). On a post hoc analysis for select questions, the median FKRGL was 15.6 (IQR 14.4-16.7), decreasing to 7.1 (IQR 6.1-8.3) and 14.5 (IQR 13.2-15.4) with prompting and rephrasing, and the median FRE was 18.1 (IQR 14.6-24.7), increasing to 73.8 (IQR 66.6-79.3) and 32.0 (IQR 27.0-37.7) with prompting and rephrasing. Numerous inaccuracies and content gaps were reported, and approximately 43% (3/7) of providers did not report feeling “comfortable” in having patients consult publicly available AI for medical information.

Conclusions: This study provides cautionary, yet optimistic, findings regarding the value of publicly accessible ChatGPT as a patient resource for abdominal malignancies. Providers should be prepared to effectively counsel patients to identify their educational attainment level when using ChatGPT to mitigate readability challenges.

(JMIR Perioper Med 2026;9:e81374) doi:10.2196/81374

KEYWORDS

patient education; health literacy; generative artificial intelligence; surgical oncology; perioperative care

Introduction

In recent years, artificial intelligence (AI) has promised to reshape medicine. Chatbots such as ChatGPT (OpenAI) [1], DeepAI, and Google Gemini use large language models (LLMs), a popular form of AI. These models are trained upon large datasets to generate answers [2]. Recent LLM improvement in reasoning has been noted to reflect human-level cognition [3]. Furthermore, studies have examined LLM function in the health care sector. LLMs have been found to pass United States Medical Licensing Examinations [4,5] and medical subspecialty exams [6-8] and to provide successful clinical reasoning and diagnoses [9]. Moreover, ChatGPT has the potential to supersede other search engines in answering patient health–related questions by providing more comprehensive and specific answers [10,11].

Although AI has been found to augment medical practice, its use as a resource by patients is not well understood. Patients have long reported turning to the internet for clinical advice [12]. Studies evaluating responses from common search engines to frequently asked general surgery questions have typically found the quality to range from fair to good but found that the readability level often exceeded the recommended level for the general population [13]. More recently, patients have turned to

XSL•FO
RenderX

ChatGPT for clinical questions; a study conducted in Australia found that approximately 9.9% of Australian adults asked ChatGPT medical questions within the first half of 2024 [14]. Following the rapid rise of publicly accessible proprietary LLM chatbots and the lack of peer-reviewed output within these learning models, recent work across specialties, including oncology, gastroenterology, otolaryngology, and surgery, has sought to evaluate LLM-generated responses to questions commonly asked by patients [15-21]. The reported overall quality of generated responses varies across fields and is impacted by the type of LLM used [20]. Furthermore, prior research has suggested that, as with "Dr Google" and other popular search engines [13], the readability of LLM-generated responses may serve as a key limitation of using LLMs such as ChatGPT as a patient resource [22]. Additionally, ChatGPT answers are limited in consistency [23], generating similar but nonidentical responses.

Gastrointestinal malignancies, including pancreatic, colorectal, hepatic, stomach, and esophageal malignancies, account for over one-quarter of cancer incidence globally and are steadily increasing. By 2040, the global number of gastrointestinal cancer deaths is projected to increase by over 70% to 5.6 million [24]. Given the significant disease burden of gastrointestinal malignancies and related therapies, it is essential to properly evaluate pertinent patient resources to better inform patients, many of whom will be accessing these resources independently. To date, few studies have examined the use of publicly accessible proprietary LLMs as a perioperative resource for patients with abdominal malignancies. The aim of this study was to evaluate the content and readability of LLM-generated responses to common patient queries for hepatic, pancreatic, and colon cancers.

Methods

Ethical Considerations

This study was submitted to the Duke University Institutional Review Board for review and was determined to be exempt (Pro00116649). This study involved surgeon-participants who evaluated GPT-generated responses to frequently asked questions. No patient data were used. To maintain participant confidentiality, all data were analyzed in aggregate. All GPT inquiries and survey questions were asked in English. Consent from participants was implied through voluntary completion of the survey. No compensation was provided to participants.

Question Development

Preliminary questions were developed by CVL and DAS. Questions for this study were developed by sourcing frequently asked questions about colon, liver, and pancreatic cancers from 7 hospital patient information and nonprofit cancer foundation websites [25-31]. This methodology was used in an earlier study examining LLMs as a tool for patient education in lung cancer surgery [32]. Questions on general disease information, including signs and symptoms, staging and treatment options, surgical eligibility, and operative risks, were formulated for colon, hepatic, and pancreatic cancers using identical language for each condition. Standardized language was used to determine the suitability of LLMs for delivering useful abdominal cancer education as applied to each of the conditions. Additional questions were created to address common patient concerns related to postoperative recovery and potential adverse outcomes following abdominal cancer surgery.

Question Piloting

The preliminary survey questions were initially evaluated for relevance and alignment with patient phrasing through subjective assessment by 2 general surgery residents. Residents were prompted to assess the frequency with which the proposed questions were encountered in practice to evaluate the survey questions based on clinical relevance. Residents were also prompted to evaluate the survey questions based on alignment with patient phrasing, to suggest phrasing revisions for items that received a Likert score ≤3 on a 5-point scale, and to propose additional relevant questions not addressed by the survey. At this stage, 8 questions were removed and 4 were added per resident feedback. Additional questions were adjusted accordingly. After piloting scenarios with residents, all questions were run sequentially through a publicly available, proprietary version of ChatGPT (GPT-4o; released on May 13, 2024) [1,33], on March 9, 2025, in Durham, United States. ChatGPT-4o is based on a proprietary GPT-4-class pretrained base LLM that has been instruction-tuned for conversational use. No additional model fine-tuning or retraining was performed through this study. ChatGPT was prompted to answer in paragraph form without additional contextual information. As in prior studies, a new chat entry was posed for each question [32].

A Qualtrics survey was formulated with the final 28 questions and LLM responses. This survey was piloted by our surgeon expert, ANM, who provided final revisions for question phrasing. Table 1 lists the finalized questions prompted into ChatGPT. Revised questions were newly run through ChatGPT, and the Qualtrics survey was adjusted accordingly.

XSL•FO
RenderX

Table . GPT queries: abdominal cancer frequently asked questions and common postoperative complications.

Domain	Question
Signs and symptoms	Q1: What are the signs and symptoms of pancreatic cancer?
	Q2: What are the signs and symptoms of colon cancer?
	Q3: What are the signs and symptoms of liver cancer?
Stages and treatment	Q4: What are the different stages and treatments for pancreatic cancer?
	Q5: What are the different stages and treatments for colon cancer?
	Q6: What are the different stages and treatments for liver cancer?
Surgery eligibility	Q7: Who is appropriate for surgery for pancreatic cancer?
	Q8: Who is appropriate for surgery for colon cancer?
	Q9: Who is appropriate for surgery for liver cancer?
Surgery risks	Q13: What are the risks of surgery to remove my pancreatic cancer?
	Q14: What are the risks of surgery to remove my colon cancer?
	Q15: What are the risks of surgery to remove my liver cancer?
General postoperative recovery	Q10: How long is the recovery from pancreatic cancer surgery?
	Q11: How long is the recovery from colon cancer surgery?
	Q12: How long is the recovery from liver cancer surgery?
	Q16: Will I need an ostomy after surgery to remove my colon cancer?
	Q17: How long will I be in the hospital after surgery for cancer in my belly?
	Q18: How long after surgery for cancer in my belly can I exercise?
	Q19: How long will it take to recover from surgery for cancer in my belly?
	Q20: Should I stay close to the hospital in a hotel or Airbnb after I'm discharged from surgery for cancer in my belly?
	Q21: How long after surgery for cancer in my belly can I do chores around the house?
Adverse outcomes	Q22: I just had surgery for cancer in my belly, and my incision is painful. What do I do?
	Q23: I just had surgery for cancer in my belly, and I am still in some pain. Is there anything else I can take for the pain?
	Q24: I just had surgery for cancer in my belly, and my incision is starting to hurt more and looks slightly open. What do I do?
	Q25: I just had surgery for cancer in my belly, and the incision is warm to the touch and draining a yellowish fluid. What do I do?
	Q26: I just had surgery for cancer in my belly. It hurts when I breathe, and I have a new cough. What do I do?
	Q27: I just had surgery for cancer in my belly. It now burns when I pee. What do I do?
	Q28: I am about to have surgery for cancer in my belly. How can I prevent an infection after?

Outcomes or Data Collection and Variables

An anonymous survey was disseminated to surgical faculty at a single institution using Qualtrics, a secure, web-based survey platform. Eligible participants were board-certified surgeons who had completed fellowship training in surgical oncology or colorectal surgery and were actively practicing at the time of the study. Surgeons were identified through publicly available web-based colorectal surgery and surgical oncology faculty rosters and were invited to participate via an email containing the anonymous Qualtrics link. The Qualtrics platform is commonly used in academic research, as it permits investigators to design surveys, test them for accessibility and functionality, distribute them electronically as a web link or QR code, and export result reports. The final list of questions and ChatGPT-4o–generated responses graded by surgeons is included in Multimedia Appendix 1. On the Qualtrics platform, prior to initiating the survey, surgeons were instructed to grade responses for accuracy, quality, and tangibility on a 5-point Likert scale (1="poor," 5="excellent"). The survey defined accuracy as the medical or social correctness of a response, quality as the extent to which a response is well-written and comprehensive, and tangibility as the degree to which the response provides actionable guidance.

After evaluating ChatGPT responses, surgeon respondents were prompted to self-report demographics, including age, sex, years as a practicing surgeon post training, AI frequency in practice, and prior experience with AI. Furthermore, a free response section permitted surgeons to share thoughts or concerns.

XSL•FO
RenderX

Respondents were assessed for comfort with patient-AI use through the question, “If a patient informed you they are using publicly available AI (eg, ChatGPT) for health information, how comfortable would you be with encouraging them to use AI following this survey?” Available answer choices included “very uncomfortable,” “uncomfortable,” “neither comfortable nor uncomfortable,” “comfortable,” and “very comfortable.”

LLM-generated responses were separately graded for readability using the Flesch-Kincaid Reading Grade Level (FKRGL) and Flesch Reading Ease (FRE) formulas through the Readability Statistics tool in Microsoft Word Version 16.105.2 [34]. FRE and FKRGL formulas calculate readability based on the average sentence and word length of a text. The FKRGL scale assesses approximate grade level of a text, with an FKRGL score of 5 corresponding to a US 5th-grade reading level. The FRE scale measures readability from 0, unreadable text, to 100, very easily readable text. Both scales were selected as they are validated tools for grading text readability, and they are commonly used by professionals to evaluate the readability of patient-directed health care information [34].

A post hoc analysis was performed to assess FKRGL, FRE scores, and content similarity for GPT responses under 3 prompting conditions: version 1 represented the response to the original question provided for reference; version 2 consisted of responses to the original question preceded by a prompt to “Answer at a 5th-grade level;” and version 3 comprised responses to questions that were reworded to a 5th-grade reading level by ChatGPT-4o prior to response generation. Four questions were selected for post hoc analysis to provide a focused analysis of question-phrasing and prompting on readability. Questions were selected based on having the highest original FKRGL score within 4 different domains and to ensure representation of each malignancy type. Content similarity was graded by 2 independent graders, CVL and DAS, using a 5-point Likert scale (1=not similar, 5=very similar).

Statistical Analysis

Respondent answers were collected and analyzed in aggregate. Descriptive statistics for categorical variables were reported as frequencies with percentages; continuous variables were reported as mean with SD or median with IQR, where appropriate. Analyses were calculated using Microsoft Excel Version 16.95.4; formulas used included =MEDIAN() for median, =AVERAGE() for mean, =STDEV() for SD, and =QUARTILE.INC() to derive the IQR.

Results

Overview

Of the 12 eligible surgeons contacted, 7 responded, resulting in a survey response rate of 58.3%. All respondents were academic surgeons at a single institution. The median reported respondent categorical age range was 35 - 44 years. Most survey respondents were male (4/7, 57.1%). Respondents had practiced surgery for a median of 7 (IQR 4 - 13) years post training. When assessed for frequency of AI use, 1 respondent reported using AI “daily,” 2 reported using AI “weekly,” 3 reported using AI “monthly,” and 1 reported using AI “almost never.”

Quality of LLM Responses

When asked to evaluate the quality of responses, experts consistently rated answers between “good” and “very good” to “excellent,” with an aggregate mean response rating of 3.54 (SD 0.28). Across all domains (Table 2), questions related to staging and treatment consistently performed worse, receiving an average rating of 3.33 (SD 0.30), while questions about adverse outcomes tended to perform best, receiving an average rating of 3.73 (SD 0.27). Table 2 indicates the median and IQR of respondent grade for each question, with quality scores ranging from 3.00 to 4.50 and IQR ranging from 2.50-3.50 to 3.25-5.00. The question indicating postoperative urinary tract infection (UTI) received the highest median quality score of 4.50 (3.25 - 5.00), between “very good” and “excellent.”

XSL•FO
RenderX

Table . Quality, accuracy, and tangibility scores for GPT-generated responses[a].

Domain and question ID	Quality score, median (IQR)	Accuracy score, median (IQR)	Tangibility score, median (IQR)
Signs and symptoms			
Q1	4 (3.5 - 4)	4 (4 - 4.5)	4 (4 - 4.5)
Q2	3 (3-4)	4 (3.5 - 4.5)	4 (3-4)
Q3	3 (2.5 - 4)	3 (2.5 - 4)	3 (3 - 3.5)
Stages and treatment			
Q4	3 (2.5 - 4.5)	4 (2.5 - 4.5)	4 (2.5 - 4.5)
Q5	4 (3-4)	4 (3-4)	4 (3.5 - 4)
Q6	3 (2.5 - 3.5)	3 (2.5 - 3.5)	3 (2.5 - 3.5)
Surgery eligibility			
Q7	4 (2.5 - 4)	4 (2.5 - 4)	4 (3.5 - 4)
Q8	4 (3.5 - 4)	4 (3.5 - 4)	4 (3.5 - 4)
Q9	4 (4-4)	4 (3.5 - 4)	4 (4-4)
General postoperative recovery			
Q10	4 (3-4)	4 (2.5 - 4)	4 (3-4)
Q11	3 (2.5 - 4)	3 (2.5 - 3.5)	3 (2.5 - 3.5)
Q12	4 (2.5 - 4)	3 (2.5 - 3.5)	3 (2.5 - 4)
Surgery risks			
Q13	4 (3-4)	4 (3-4)	3 (3-4)
Q14	4 (3.5 - 4)	4 (3.5 - 4)	4 (3.5 - 4)
Q15	3 (3 - 3.5)	3 (3 - 3.5)	3 (3 - 3.5)
General postoperative recovery			
Q16	4 (3-4)	4 (3.5 - 4)	4 (3.5 - 4)
Q17	3 (2.5 - 4)	4 (3.5 - 4)	4 (2.5 - 4)
Q18	4 (3.5 - 4.5)	4 (3.5 - 4.5)	4 (3.5 - 4)
Q19	4 (3.5 - 4)	4 (3.5 - 4)	4 (3-4)
Q20	4 (3.5 - 4)	4 (3-4)	4 (3-4)
Q21	4 (3-4)	4 (3-4)	4 (3 - 4.5)
Adverse outcomes			
Q22	3.5 (3-4)	3.5 (3-4)	3.5 (3-4)
Q23	3.5 (3-4)	4 (3.25 - 4)	4 (3.25 - 4)
Q24	4 (3.25 - 4.75)	4.5 (3.25 - 5)	4.5 (3.25 - 5)
Q25	4 (3.25 - 4)	4 (3.25 - 4)	4 (3.25 - 4)
Q26	4 (3.25 - 4.75)	4.5 (3.25 - 5)	4.5 (3.25 - 5)
Q27	4.5 (3.25 - 5)	4.5 (3.25 - 5)	4.5 (3.25 - 5)
Q28	4 (4-4)	4 (3.25 - 4.75)	4.5 (3.25 - 5)

[a]Accuracy is defined as how "medically or socially accurate" a response is; quality as how "well-written and comprehensive" a response is; and tangibility as how "actionable" a response is.

Accuracy of LLM Responses

Similar to quality, when asked to evaluate the accuracy of responses, experts generally rated responses between "good" and "very good" to "excellent," with an aggregate mean response rating of 3.57 (SD 0.34). Across all domains, questions related to staging and treatment performed the worst, receiving an average rating of 3.29 (SD 0.38). Conversely, questions about adverse outcomes consistently performed best, receiving an average rating of 3.83 (SD 0.24). Median response accuracy ratings ranged from 3.00 to 4.50 with IQRs ranging from 2.50-3.50 to 3.25-5.00. Questions regarding postoperative wound

XSL•FO
RenderX

dehiscence (Q24), pulmonary embolism (Q26), and UTI management (Q27) received the highest median accuracy grading of 4.50 (IQR 3.25-5.00).

Tangibility of LLM Responses

When asked to evaluate the tangibility, or how "actionable" a response was, experts likewise consistently rated responses between "good" and "very good" to "excellent," with an aggregate mean response rating of 3.62 (SD 0.29). Across all domains, questions pertaining to staging and treatment performed the worst, receiving the lowest mean tangibility score of 3.47 (SD 0.30), while questions about adverse outcomes performed best, receiving an average rating of 3.86 (SD 0.28). The median response ratings ranged from 3.00 to 4.50 with IQR scores ranging from 2.50-3.50 to 3.25-5.00. Questions regarding postoperative wound dehiscence (Q24), pulmonary embolism (Q26), UTI management (Q27), and infection prevention (Q28) received the highest median tangibility grading of 4.50 (IQR 3.25-5.00).

Readability of LLM Responses

When assessing readability (Table 3), ChatGPT-4o–generated responses read at an average FKRGL of 14.51 (SD 1.86), requiring some level of college education for adequate comprehension. Response FKRGL scores ranged from 10.8 to 18.1. A question regarding wound dehiscence received the lowest grade score, 10.8, while a question regarding pancreatic cancer surgery candidacy received the highest score, 18.1. The mean FRE score of ChatGPT-4o–generated responses was 28.8 (SD 9.87), corresponding to a college graduate reading level and indicating low readability. Response FRE scores ranged from 11.7 to 48.0. A question regarding colon cancer surgery recovery had the worst readability, with an FRE score of 11.7. As with FKRGL grading, a question regarding wound dehiscence had the highest ease of readability with an FRE score of 48.0.

XSL•FO
RenderX

Table . Readability of GPT-generated responses. Readability is represented as FKRGL[a] and FRE[b] scores.

ID and question	FKRGL score (US-grade reading level)	FRE score	Estimated FRE US-grade level [34]
Q1: What are the signs and symptoms of pancreatic cancer?	12.9	36.8	13 - 16
Q2: What are the signs and symptoms of colon cancer?	12.2	41.3	13 - 16
Q3: What are the signs and symptoms of liver cancer?	11.8	41.2	13 - 16
Q4: What are the different stages and treatments for pancreatic cancer?	13.4	28.6	College graduate
Q5: What are the different stages and treatments for colon cancer?	15.1	16.4	College graduate
Q6: What are the different stages and treatments for liver cancer?	14.5	17.8	College graduate
Q7: Who is appropriate for surgery for pancreatic cancer?	18.1	20.6	College graduate
Q8: Who is appropriate for surgery for colon cancer?	18.0	17.4	College graduate
Q9: Who is appropriate for surgery for liver cancer?	16.6	21.5	College graduate
Q10: How long is the recovery from pancreatic cancer surgery?	16.1	20.0	College graduate
Q11: How long is the recovery from colon cancer surgery?	16.2	11.7	College graduate
Q12: How long is the recovery from liver cancer surgery?	14.5	17.4	College graduate
Q13. What are the risks of surgery to remove my pancreatic cancer?	14.8	20.7	College graduate
Q14. What are the risks of surgery to remove my colon cancer?	14.2	22.8	College graduate
Q15. What are the risks of surgery to remove my liver cancer?	14.9	15.6	College graduate
Q16. Will I need an ostomy after surgery to remove my colon cancer?	14.2	32.9	13 - 16
Q17: How long will I be in the hospital after surgery for cancer in my belly?	14.7	36.2	13 - 16
Q18: How long after surgery for cancer in my belly can I exercise?	15.4	26.5	College graduate
Q19: How long will it take to recover from surgery for cancer in my belly?	14.1	31.8	13 - 16
Q20: Should I stay close to the hospital in a hotel or Airbnb after I'm discharged from surgery for cancer in my belly?	17.0	32.7	13 - 16
Q21: How long after surgery for cancer in my belly can I do chores around the house?	15.3	33.9	13 - 16
Q22: I just had surgery for cancer in my belly, and my incision is painful. What do I do?	12.1	43.6	13 - 16

XSL•FO
RenderX

ID and question	FKRGL score (US-grade reading level)	FRE score	Estimated FRE US-grade level [34]
Q23: I just had surgery for cancer in my belly, and I am still in some pain. Is there anything else I can take for the pain?	12.7	36.6	13 - 16
Q24: I just had surgery for cancer in my belly, and my incision is starting to hurt more and looks slightly open. What do I do?	10.8	48.0	13 - 16
Q25: I just had surgery for cancer in my belly, and the incision is warm to the touch and draining a yellowish fluid. What do I do?	13.6	40.0	13 - 16
Q26: I just had surgery for cancer in my belly. It hurts when I breathe, and I have a new cough. What do I do?	16.2	27.4	College graduate
Q27: I just had surgery for cancer in my belly. It now burns when I pee. What do I do?	15.0	30.1	13 - 16
Q28: I am about to have surgery for cancer in my belly. How can I prevent an infection after?	12.0	36.1	13 - 16

[a]FKRGL: Flesch-Kincaid Reading Grade Level.

[b]FRE: Flesch Reading Ease.

Four questions were selected for post hoc analysis shown in Table 4 (Q1, Q7, Q11, and Q15). These questions had an original median FKRGL score of 15.6 (IQR 14.4-16.7; range 12.9 - 18.1) and FRE score of 18.1 (IQR 14.6-24.7; range: 11.7 - 36.8). When GPT-4o was queried to respond to select questions to the level of a 5th-grade reader, the median FKRGL score decreased to 7.1 (IQR 6.1-8.3; range: 5.9 - 9.0) and FRE increased to 73.8 (IQR 66.6-79.3; range: 60.1 - 80.9). Two independent graders (CVL and DAS) found responses to result in a mean content similarity of 3.88 (SD 0.25) in comparison to the original response. Responses to questions that were rephrased by GPT, with prompting to query at a 5th-grade reading level, resulted in a median FKRGL score of 14.5 (IQR 13.2-15.4; range: 11.6 to 15.8), a median FRE score of 32.0 (IQR 27.0-37.7; range: 21.4 - 45.0), and a mean content similarity of 4.63 (SD 0.25) to original responses. The raters had identical scores for 50% (4/8) of responses, with the other 4 responses differing by 1 on the 5-point Likert scale.

XSL•FO
RenderX

Table . FKRGL[a] and FRE[b] scores for select questions (V1), questions prompted to respond at the 5th-grade level (V2), and questions rephrased by GPT to be asked at the 5th-grade level (V3).

Question and versions	FKRGL score	FRE score	Content similarity, mean (SD)
Q1			
V1: What are the signs and symptoms of pancreatic cancer?	12.9	36.8	Reference
V2: Answer at a 5th-grade level: What are the signs and symptoms of pancreatic cancer?	6.1	80.9	4.0 (0)
V3: What are the warning signs of pancreatic cancer and how might someone feel if they have it?	11.6	45.0	4.5 (0.71)
Q7			
V1: Who is appropriate for surgery for pancreatic cancer?	18.1	20.6	Reference
V2: Answer at a 5th-grade level: Who is appropriate for surgery for pancreatic cancer?	9.0	60.1	4.0 (0)
V3: Who can have surgery to treat pancreatic cancer?	15.3	28.8	5.0 (0)
Q11			
V1: How long is the recovery from colon cancer surgery?	16.2	11.7	Reference
V2: Answer at a 5th-grade level: How long is the recovery from colon cancer surgery?	5.9	78.8	3.5 (0.71)
V3: How long does it take to feel better after colon cancer surgery?	15.8	21.4	4.5 (0.71)
Q15			
V1: What are the risks of surgery to remove my liver cancer?	14.9	15.6	Reference
V2: Answer at a 5th-grade level: What are the risks of surgery to remove my liver cancer?	8.1	68.8	4.0 (0)
V3: What could go wrong if I have surgery to take out my liver cancer?	13.7	35.2	4.5 (0.71)

[a]FKRGL: Flesch-Kincaid Reading Grade Level.

[b]FRE: Flesch Reading Ease.

Qualitative Feedback

Numerous inaccuracies within GPT-generated responses were detected by a surgeon-expert concerning general disease information and postoperative recovery. The following feedback has been modified for clarity but maintains the original intent. In Q2 (signs and symptoms of colon cancer), rectal bleeding was mistakenly described as a systemic symptom, while it is a local symptom that may lead to secondary systemic symptoms, including fatigue due to anemia. For Q10 (pancreatic cancer surgery recovery), "light activities," which are often defined as walking or activities of daily living in the surgical setting, were resumed while a patient was admitted, instead of the written 6 to 12 weeks following discharge. Likewise, for Q11 (colon cancer surgery recovery), certain "light activities" could be resumed sooner. For Q12 (liver cancer surgery recovery), the mention of major hepatectomy as treatment was notably absent.

Regarding quality, numerous content gaps were noted. For Q3 (signs and symptoms of liver cancer), the response described chronic liver disease symptoms; these are common for patients with primary liver cancers but less frequent in the setting of secondary liver cancers (ie, colorectal cancer with liver metastases). Regarding Q4 (stages and treatments of pancreatic cancer), genetic testing should be included when discussing targeted therapies. For Q7 (pancreatic cancer surgery eligibility), discussion of the biology of resectability, which is accounted for by tumor markers such as Ca 19 - 9, was notably absent. Regarding Q20 (staying near the hospital following discharge), while listed, it is not emphasized that staying nearby is unnecessary unless the patient lives far away. Furthermore, the

XSL•FO
RenderX

question could be enhanced by including discussion of local housing options with case management or a social worker. For Q25 (postoperative infection), concern for dehiscence is not explicitly stated, and the volume of drainage should be addressed sooner, as high volume may indicate dehiscence.

Provider Recommendations for GPT as a Patient Resource

When assessed on their comfort level with patients using publicly available AI for health information, 57.1% (4/7) of providers reported being "comfortable," 14.3% (1/7) reported being "neither comfortable nor uncomfortable," 14.3% (1/7) reported being "uncomfortable," and 14.3% (1/7) reported being "very uncomfortable." Regarding provider discomfort, when asked for questions or concerns pertaining to the study, 1 respondent reported "The answers should be designed for a lower health literacy level." Another physician expressed concern over direct patient use of ChatGPT, primarily citing lack of supervision and noting "health is not something you want to leave up to a robot. There will always be intricacies that cannot be understood by AI."

Discussion

Principal Findings

This study is among the first to evaluate ChatGPT-4o as a patient information resource for individuals preparing for or recovering from surgery for abdominal malignancies [20,35]. As patient self-use of LLMs for medical information is increasing [14], it is essential to assess the content quality, safety, and comprehensibility of GPT-generated responses. Through gaining a deeper understanding of the strengths and weaknesses present within LLMs, providers may help patients be aware of such options and help them navigate the use of these sources. The current study's results indicate that ChatGPT-4o may serve as a useful patient information resource, with most responses rated from "good" to "very good to excellent" in quality, accuracy, and tangibility. Notably, the lowest rated responses received a median score of 3.0, corresponding to a "good" rating, whereas the highest rated responses received a score of 4.50, corresponding to a rating between "very good" and "excellent." However, there is still room for improvement in generated responses prior to the endorsement of ChatGPT as a "gold-standard" patient resource. While most providers were "comfortable" having patients use publicly available AI for health information, 42.9% (3/7) of providers did not report feeling "comfortable" having patients use publicly available AI for health information, with 2 reported being "uncomfortable" or "very uncomfortable." Physicians cited concerns regarding patient use of ChatGPT, noting poor response comprehensibility and lack of supervision, factors likely contributing to their lack of comfort in patient use of ChatGPT. Moreover, this study raises concerns about the comprehensibility of the generated responses, as elevated FKGRL scores indicate that many require a postsecondary reading level for adequate understanding. Physicians should be aware that, in the context of patient use of LLMs for medical information, patients would benefit from instructions for use and monitoring for potential ChatGPT-derived misconceptions.

Information is scarce regarding the safety and accuracy of ChatGPT-generated responses in the perioperative setting for abdominal malignancies. Given the complex biological mechanisms and therapeutic management of gastrointestinal malignancies, it is critical to evaluate the quality of ChatGPT-generated content. The presented data suggest that, although ChatGPT responses averaged as "good" or "very good," scores were highly question- and domain-dependent. Given its high overall ratings, ChatGPT may serve as an advantageous tool for patients to develop a baseline knowledge of their disease prior to clinical encounters. However, numerous inaccuracies and content gaps were identified within responses. This is congruent with past work assessing ChatGPT's use for thoracic surgery, where most responses likewise ranged from "good" to "very good," minor inaccuracies were identified in each answer, and certain domains performed better than others [32]. Regarding abdominal malignancies, questions concerning staging and treatment received the lowest mean accuracy and quality scores. As such, providers should be encouraged to assess potential disease misconceptions that patients using ChatGPT may have and ensure they distribute comprehensive general disease information. Interestingly, ChatGPT-4o excelled in answering questions pertaining to adverse outcomes following surgery. As such, ChatGPT may help guide patients seeking proper management for postoperative complications.

The present study suggests that patient information regarding abdominal malignancies presented by ChatGPT-4o may produce material that is poorly comprehensible for many of the intended population due to requirements of high health literacy and education level. One surgeon expressed that responses should be written for a lower health literacy level. This is consistent with the findings of a high grade level requirement for adequate comprehensibility (FKRGL score), averaging a grade level of 14.5. An FKRGL of 14.5 represents a reading level requiring some level of college education. Current recommendations suggest that patient resources should be tailored to a 5th-grade reading level for accessibility [36,37].

Readability as a limitation of ChatGPT has been previously reported in the literature [15,22]. Past work regarding cervical spine surgery likewise noted high FKRGL scores to limit ChatGPT-3.5's use as a patient resource. After prompting ChatGPT to provide answers at a 6th-grade reading level, answers decreased from a grade level of 13.5 to 11.2, though remaining persistently elevated. Notably, the present study used similar techniques that successfully produced responses at a lower reading level with ChatGPT-4o. For select questions, the median FKRGL score prior to rephrasing or prompting was 15.6 (IQR 14.4-16.7). Remarkable improvement was noted upon prompting GPT to respond to the level of a 5th-grade reader, decreasing the median FKRGL score, or US grade level, to 7.1 (IQR 6.1-8.3). This work suggests that improvements within ChatGPT-4o may allow for more comprehensible responses, given appropriate prompting. Notably, there was a less meaningful drop in median FKRGL score, from 15.6 (IQR 14.4-16.7) to 14.5 (IQR 13.2-15.4), when ChatGPT was used to *rephrase* questions to be asked at the reading level of a 5th grader. Prior to modification, the median FRE score for select questions was 18.1 (IQR 14.6-24.7). Consistent with FKRGL

XSL•FO
RenderX

trends, prompting questions resulted in a more substantial increase in FRE score (median FRE 73.8, IQR 66.6-79.3), indicating markedly improved readability, compared with rephrasing (median FRE 32, IQR 27.0-37.7). This suggests that explicitly requesting ChatGPT to produce responses at a lower level may be more effective in improving readability than adjusting question phrasing. Although contents similar to original questions were better for the latter group than the former (4.63/5 versus 3.88/5), most key concepts were retained within both groups. As such, prompting ChatGPT to answer at a lower grade level may improve readability without significantly sacrificing content. Therefore, providers should be encouraged to assess patient use of LLMs for medical questions and provide patients with a menu for how to prompt ChatGPT to answer at an appropriate grade level if relevant.

While comprehensibility without prompting educational level can be a limitation for the intended patient population, ChatGPT may serve as a useful tool for providers and trainees. Past work in public health has found AI chatbots to be a useful educational tool for medical students in answering complex medical questions [38]. Within the present study, questions 1 to 9 pertain to “signs and symptoms,” “stages and treatment,” and “surgery eligibility”; these questions may be asked by clinicians or learners. While patient readability was limited by a high mean grade level required, the ratings typically ranged from “good” to “very good” in quality, accuracy, and tangibility. This suggests that ChatGPT-4o can serve as a useful resource for physicians and medical trainees, given a higher health literacy than the general population. To further evaluate the use and comprehensibility of ChatGPT as a patient resource, future investigations should involve patient perspectives.

This study has several limitations. First, the small sample size (n=7) of surgical oncologists grading the responses substantially limits statistical power and the reliability of the findings. The single-institutional nature of the study further limits generalizability, as physician responses may reflect regional practice patterns and institutional biases. Future validation should evaluate larger, multi-institutional cohorts to confirm reproducibility and evaluate external validity. Second, the survey incorporated subjective assessments which may limit reproducibility, as concepts graded, such as “quality,” “accuracy,” and “tangibility,” are abstract. To enhance reproducibility, standardized definitions of these domains were included on each page of the survey. Third, questions may not be well representative of patient language. Although questions were obtained from hospital websites and piloted with residents to improve alignment with patient phrasing, they may not encompass the full spectrum of patient inquiries nor the variability of patients’ health literacy. As only 28 questions were assessed across 3 malignancies, the nature of the questions is limited in scope and may not represent all questions patients may ask pertaining to their diagnosed malignancy. Moreover, questions are broad, pertaining to “colon,” “pancreas,” “liver,” or “belly” cancers, without specifying types and stages.

Conclusions

This preliminary study indicates that, while publicly accessible ChatGPT may serve as a useful patient resource, its use as an unsupervised source of information for patients with abdominal malignancies has distinct limitations. Providers should be aware that many of their patients are accessing ChatGPT and recognize that developing an understanding of its strengths and limitations can help them guide their patients to enable its best use. Inaccuracies, gaps in information, and poor readability were identified in ChatGPT-generated content, suggesting patients may benefit from physician guidance. Providers should be prepared to properly support their patients reporting ChatGPT use by counseling techniques such as prompting questions to tailor responses to their educational level. The data herein indicate that this is critical for the interpretation of the information by patients, as without this guidance, the answers are directed to an educational level of college or above.

Acknowledgments

During the preparation of this work, the authors used ChatGPT-4o as specified above to collect the information required for data analysis (artificial intelligence–generated responses). No generative artificial intelligence was used in the writing process.

Funding

The authors declared no financial support was received for this work.

Data Availability

To preserve respondent confidentiality, the data set used in this study is not publicly available. A limited version of the dataset used in this study can be obtained from the first author upon request.

Authors' Contributions

Conceptualization: CVL, DAS, ANM, CLC, KNS, MEL, DPN, RS
Data curation: CVL, DAS, ANM, CLC, KNS, MEL, DPN, RS
Formal analysis: CVL, DAS, RS
Methodology: CVL, DAS, RS
Supervision: ANM, CLC, KNS, MEL, DPN, RS
Writing – original draft: CVL, RS

XSL•FO
RenderX

Writing – review & editing: CVL, DAS, ANM, CLC, KNS, MEL, DPN, RS

Conflicts of Interest

RS serves on the Board of Directors of DNAnexus, Heartland Whole Health Institute, ZealCare, Inc, where he is also the cofounder; Board of Trustees of American Medical Program, Tel Aviv University, and Scientific Advisory Board of OrthoBioTherapeutics Inc.

Multimedia Appendix 1
Distributed survey with GPT-generated responses.
[DOCX File, 31 KB - periop_v9i1e81374_app1.docx]

References

1. ChatGPT. URL: https://chatgpt.com [accessed 2025-03-09]
2. Guinness H. How does ChatGPT work? Zapier. URL: https://zapier.com/blog/how-does-chatgpt-work/ [accessed 2025-04-15]
3. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv. Preprint posted online on Mar 22, 2023. [doi: 10.48550/arXiv.2303.12712]
4. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
5. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
6. Gupta R, Herzog I, Park JB, et al. Performance of ChatGPT on the plastic surgery inservice training examination. Aesthet Surg J 2023 Nov 16;43(12):NP1078-NP1082. [doi: 10.1093/asj/sjad128] [Medline: 37128784]
7. Hoch CC, Wollenberg B, Lüers JC, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Otorhinolaryngol 2023 Sep;280(9):4271-4278. [doi: 10.1007/s00405-023-08051-4] [Medline: 37285018]
8. Mihalache A, Huang RS, Popovic MM, Muni RH. Performance of an upgraded artificial intelligence chatbot for ophthalmic knowledge assessment. JAMA Ophthalmol 2023 Aug 1;141(8):798-800. [doi: 10.1001/jamaophthalmol.2023.2754] [Medline: 37440220]
9. Cabral S, Restrepo D, Kanjee Z, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. JAMA Intern Med 2024 May 1;184(5):581-583. [doi: 10.1001/jamainternmed.2024.0295] [Medline: 38557971]
10. Tan S, Xin X, Wu D. ChatGPT in medicine: prospects and challenges: a review article. Int J Surg 2024 Jun 1;110(6):3701-3706. [doi: 10.1097/JS9.0000000000001312] [Medline: 38502861]
11. Liu HY, Alessandri Bonetti M, De Lorenzi F, Gimbel ML, Nguyen VT, Egro FM. Consulting the digital doctor: Google versus ChatGPT as sources of information on breast implant-associated anaplastic large cell lymphoma and breast implant illness. Aesthetic Plast Surg 2024 Feb;48(4):590-607. [doi: 10.1007/s00266-023-03713-4] [Medline: 37903939]
12. Bergmo TS, Sandsdalen V, Manskow US, Småbrekke L, Waaseth M. Internet use for obtaining medicine information: cross-sectional survey. JMIR Form Res 2023 Feb 2;7:e40466. [doi: 10.2196/40466] [Medline: 36729577]
13. Ramli R, Jambor MA, Kong CY. Dr Google - assessing the reliability and readability of information on general surgical procedures found via search engines. ANZ J Surg 2023 Mar;93(3):590-596. [doi: 10.1111/ans.18289] [Medline: 36716246]
14. Ayre J, Cvejic E, McCaffery KJ. Use of ChatGPT to obtain health information in Australia, 2024: insights from a nationally representative survey. Med J Aust 2025 Mar 3;222(4):210-212. [doi: 10.5694/mja2.52598] [Medline: 39901778]
15. Shen SA, Perez-Heydrich CA, Xie DX, Nellis JC. ChatGPT vs. web search for patient questions: what does ChatGPT do better? Eur Arch Otorhinolaryngol 2024 Jun;281(6):3219-3225. [doi: 10.1007/s00405-024-08524-0] [Medline: 38416195]
16. Artioli E, Veronesi F, Mazzotti A, et al. Assessing ChatGPT responses to common patient questions regarding total ankle arthroplasty. J Exp Orthop 2024;12(1):e70138. [doi: 10.1002/jeo2.70138] [Medline: 39741912]
17. Gajjar AA, Kumar RP, Paliwoda ED, et al. Usefulness and accuracy of artificial intelligence chatbot responses to patient questions for neurosurgical procedures. Neurosurgery 2024 Feb 14;95(1). [doi: 10.1227/neu.0000000000002856] [Medline: 38353558]
18. Samaan JS, Yeo YH, Rajeev N, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg 2023 Jun;33(6):1790-1796. [doi: 10.1007/s11695-023-06603-5] [Medline: 37106269]
19. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023 Jul;29(3):721-732. [doi: 10.3350/cmh.2023.0089] [Medline: 36946005]
20. Rydzewski NR, Dinakaran D, Zhao SG, et al. Comparative evaluation of LLMs in clinical oncology. NEJM AI 2024 May;1(5). [doi: 10.1056/aioa2300151] [Medline: 39131700]
21. Lee TC, Staller K, Botoman V, Pathipati MP, Varma S, Kuo B. ChatGPT answers common patient questions about colonoscopy. Gastroenterology 2023 Aug;165(2):509-511.e7. [doi: 10.1053/j.gastro.2023.04.033] [Medline: 37150470]

XSL•FO
RenderX

22. Subramanian T, Araghi K, Amen TB, et al. Chat generative pretraining transformer answers patient-focused questions in cervical spine surgery. Clin Spine Surg 2024 Jul 1;37(6):E278-E281. [doi: 10.1097/BSD.0000000000001600] [Medline: 38531823]
23. Al-Dujaili Z, Omari S, Pillai J, Al Faraj A. Assessing the accuracy and consistency of ChatGPT in clinical pharmacy management: a preliminary analysis with clinical pharmacy experts worldwide. Res Social Adm Pharm 2023 Dec;19(12):1590-1594. [doi: 10.1016/j.sapharm.2023.08.012] [Medline: 37696742]
24. Arnold M, Abnet CC, Neale RE, et al. Global burden of 5 major types of gastrointestinal cancer. Gastroenterology 2020 Jul;159(1):335-349.e15. [doi: 10.1053/j.gastro.2020.02.068] [Medline: 32247694]
25. Pancreatic cancer questions to ask the healthcare team. Pancreatic Cancer Action Network. 2025. URL: https://pancan.org/facing-pancreatic-cancer/diagnosis/choosing-your-healthcare-team/questions-to-ask-your-healthcare-team [accessed 2025-01-15]
26. Colon cancer FAQs. Moffitt Cancer Center. 2025. URL: https://www.moffitt.org/cancers/colon-cancer/faqs [accessed 2025-01-15]
27. Frequently asked questions. Mount Sinai Tisch Cancer Center. 2025. URL: https://www.mountsinai.org/care/cancer/services/colon/faqs [accessed 2025-01-15]
28. Liver cancer – frequently asked questions. Pelican Cancer Foundation. 2025. URL: https://www.pelicancancer.org/patient-resources/liver-cancer/frequently-asked-questions [accessed 2025-01-15]
29. Questions to ask about pancreatic cancer. American Cancer Society. 2024. URL: https://www.cancer.org/cancer/types/pancreatic-cancer/detection-diagnosis-staging/talking-with-doctor.html [accessed 2025-01-15]
30. Questions to ask about liver cancer. American Cancer Society. 2025. URL: https://www.cancer.org/cancer/types/liver-cancer/detection-diagnosis-staging/talking-with-doctor.html [accessed 2025-01-15]
31. Frequently asked questions. Hirshberg Foundation for Pancreatic Cancer Research. 2025. URL: https://pancreatic.org/pancreatic-cancer/faqs [accessed 2025-01-15]
32. Ferrari-Light D, Merritt RE, D'Souza D, et al. Evaluating ChatGPT as a patient resource for frequently asked questions about lung cancer surgery-a pilot study. J Thorac Cardiovasc Surg 2025 Apr;169(4):1174-1180. [doi: 10.1016/j.jtcvs.2024.09.030] [Medline: 39326732]
33. ChatGPT — release notes. OpenAI. 2026. URL: https://help.openai.com/en/articles/6825453-chatgpt-release-notes?utm_source=chatgpt.com [accessed 2026-03-01]
34. Jindal P, MacDermid JC. Assessing reading levels of health information: uses and limitations of flesch formula. Educ Health (Abingdon) 2017;30(1):84-88. [doi: 10.4103/1357-6283.210517] [Medline: 28707643]
35. Munir MM, Endo Y, Ejaz A, Dillhoff M, Cloyd JM, Pawlik TM. Online artificial intelligence platforms and their applicability to gastrointestinal surgical operations. J Gastrointest Surg 2024 Jan;28(1):64-69. [doi: 10.1016/j.gassur.2023.11.019] [Medline: 38353076]
36. AHRQ health literacy universal precautions toolkit. : Agency for Healthcare Research and Quality (AHRQ); 2024 URL: https://www.ahrq.gov/sites/default/files/wysiwyg/health-literacy/3rd-edition-toolkit/health-literacy-toolkit-third-edition.pdf [accessed 2025-04-20]
37. Stossel LM, Segar N, Gliatto P, Fallar R, Karani R. Readability of patient education materials available at the point of care. J Gen Intern Med 2012 Sep;27(9):1165-1170. [doi: 10.1007/s11606-012-2046-0] [Medline: 22528620]
38. Baglivo F, De Angelis L, Casigliani V, Arzilli G, Privitera GP, Rizzo C. Exploring the possible use of AI chatbots in public health education: feasibility study. JMIR Med Educ 2023 Nov 1;9:e51421. [doi: 10.2196/51421] [Medline: 37910155]

Abbreviations

AI: artificial intelligence
FKRGL: Flesch-Kincaid Reading Grade Level
FRE: Flesch Reading Ease
LLM: large language model
UTI: urinary tract infection

Edited by J Shiffermiller; submitted 27.Jul.2025; peer-reviewed by C Richlitzki, A Roland, S Matsuda; revised version received 08.Mar.2026; accepted 12.Mar.2026; published 04.May.2026.

Please cite as:
Lindsay CV, Shenoy DA, Martin AN, Clipper CL, Shah KN, Lidsky ME, Nussbaum DP, Snyderman R
Physician Perspectives on ChatGPT-4o as a Patient Resource for Abdominal Cancer Surgeries: Cross-Sectional Survey
JMIR Perioper Med 2026;9:e81374
URL: https://periop.jmir.org/2026/1/e81374
doi:10.2196/81374

XSL•FO
RenderX

© Christina V Lindsay, Devika A Shenoy, Allison N Martin, Christie L Clipper, Kevin N Shah, Michael E Lidsky, Daniel P Nussbaum, Ralph Snyderman. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 4.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI-Generated Avatar Videos for Postoperative Patient Education Among Health Care Workers: Pilot Randomized Controlled Trial

Syed Ali Haider[1], MBBS; Srinivasagam Prabha[1], PhD; Cesar Abraham Gomez-Cabello[1], MD; Ariana Genovese[1], BS; Bernardo G Collaco[1], MD; Nadia Wood[2], MS; James London[3], MHA; Sanjay Bagaria[4], MD; Mark A Lifson[5], PhD; Cui Tao[6], PhD; Antonio Jorge Forte[1,5,6], MD, PhD

[1]Department of Surgery, Division of Plastic Surgery, Mayo Clinic in Florida, 4500 San Pablo Road, Jacksonville, FL, United States

[2]Department of Radiology, AI IT, Mayo Clinic, Rochester, MN, United States

[3]Department of Surgery, Mayo Clinic in Florida, Jacksonville, FL, United States

[4]Division of Surgical Oncology, Mayo Clinic, Jacksonville, FL, United States

[5]Center for Digital Health, Mayo Clinic, Rochester, MN, United States

[6]Department of Artificial Intelligence and Informatics, Mayo Clinic in Florida, Jacksonville, FL, United States

Corresponding Author:
Antonio Jorge Forte, MD, PhD
Department of Surgery, Division of Plastic Surgery, Mayo Clinic in Florida, 4500 San Pablo Road, Jacksonville, FL, United States

Abstract

Background: Effective postoperative communication is vital for patient recovery, yet traditional text-based discharge instructions often lead to poor comprehension and adherence, particularly among patients with limited health literacy. Although educational videos improve understanding and retention, their widespread use has been hampered by high production costs. Generative artificial intelligence (AI) offers a scalable solution for creating engaging video content.

Objective: The primary objective of this pilot study was to assess the feasibility of creating and deploying AI-generated, avatar-led videos for postoperative instruction delivery. Secondary objectives included comparing knowledge retention, engagement, perceived clarity, and user experience between AI-generated video and traditional text-based handout formats among health care workers.

Methods: In this randomized pilot study, 38 health care worker volunteers were recruited as a convenience sample to pilot-test the intervention before patient implementation. Participants were assigned to either a text handout group (n=19, 50%) or an AI-generated video group (n=19, 50%). Both groups received information on 10 common postoperative topics. The primary outcome was objective knowledge, assessed via a 10-item quiz. Secondary outcomes, measured through surveys with 5-point Likert scales, included engagement time, subjective engagement, perceived clarity, usefulness, confidence in understanding, and information retention. Qualitative feedback was also collected.

Results: Objective knowledge quiz scores did not differ significantly between groups (mean 8.89, SD 1.20 for the AI-generated video group vs mean 8.21, SD 1.78 for the text handout group; P=.17; Cohen d=0.45). Participants in the AI-generated video group demonstrated significantly higher engagement time (mean 15.11, SD 7.78 minutes vs mean 8.84, SD 4.03 minutes; P=.004; Cohen d=1.04). They also rated instructions as significantly clearer (mean 4.63, SD 0.50 vs mean 4.00, SD 0.82; P=.007; Cohen d=0.93), more engaging (mean 4.05, SD 0.78 vs mean 3.32, SD 1.00; P=.02; Cohen d=0.81), and more effective for retention (mean 4.42, SD 0.84 vs mean 3.37, SD 0.68; P<.001; Cohen d=1.38). Qualitative feedback highlighted the engaging nature of AI-generated videos but noted areas for avatar refinement.

Conclusions: In this pilot study with health care workers, AI-generated avatar videos did not improve objective knowledge scores but significantly enhanced engagement, perceived retention and perceived clarity (Cohen d=0.81 - 1.38). Future studies in actual patient populations with diverse health literacy levels are needed to determine whether these engagement advantages translate into improved knowledge outcomes.

(JMIR Perioper Med 2026;9:e89277) doi:10.2196/89277

KEYWORDS

artificial intelligence; AI; text to video; generative artificial intelligence; generative AI; postoperative instructions; patient education

XSL•FO
RenderX

Introduction

Effective postoperative communication at hospital discharge is critical, yet many patients leave with written instructions they cannot fully understand or remember. Approximately 90 million adults in the United States have limited health literacy [1]. Patients with lower educational attainment may struggle to recall instructions, leading to medication errors, missed follow-ups, and adverse outcomes [2-5]. Poor adherence to discharge instructions contributes to preventable readmissions, patient dissatisfaction, and increased health care costs [6-8]. Traditional instruction methods, such as written materials and brief verbal explanations, rarely deliver information in a memorable or actionable way [9-11].

Educational videos offer a compelling solution [12,13]. They simplify complex information, support diverse learning styles, and improve comprehension and retention—especially when paired with written materials [12]. Videos can demonstrate procedures visually, feature presenters who create a sense of social presence, accommodate varying literacy levels through combined auditory and visual information, and offer the potential for multilingual messaging. However, high production costs, long turnaround times, and equipment requirements have limited their widespread use.

Generative artificial intelligence (AI) enables the rapid, automated creation of video content [13]. Platforms such as OpenAI's Sora and Google DeepMind's Veo can generate high-quality, coherent video content directly from simple text prompts [14], with applications proposed across medical domains, including surgical training, public health education, and patient-specific educational content [15-17]. A particularly compelling feature is AI-generated avatars—photorealistic digital humans who deliver content in a natural manner.

According to Mayer's cognitive theory of multimedia learning [18], combining verbal and visual information enhances understanding by engaging dual channels in the brain. Visible instructors create social presence, increasing motivation and attention [19]. A systematic review found that videos featuring visible instructors consistently improved learning outcomes [20]. Early evidence in health care supports this: a virtual nurse avatar for discharge instructions was well received, with 85% of patients liking the character and 86.4% of nurses believing it would aid education [21]. AI-generated avatars have been explored for HIV education, heart disease knowledge dissemination, and ileostomy education [22-24], and more recently, interactive physician avatars for postoperative patient education have demonstrated high usability and trust among surgical patients [25].

By assessing both objective understanding and subjective experience, this study aims to examine the feasibility and user experience of AI-generated video as a medium for patient education at hospital discharge. If effective, this approach could help reduce disparities in postoperative care by providing patients with clearer, more accessible discharge instructions.

Methods

Study Design and Participants

The primary objective was to assess the feasibility of creating and deploying AI-generated avatar videos for postoperative instruction. The primary outcome was the difference in knowledge quiz scores between groups. Secondary outcomes included engagement time, subjective engagement, perceived clarity, usefulness, confidence in understanding, information retention, and preference for format. Qualitative feedback was also collected through open-ended questions.

We conducted a pilot study to compare traditional text-based instructions with AI-generated video instructions for postoperative care education. A total of 38 adult health care worker volunteers were recruited from a hospital community (faculty, staff, researchers, and students). This convenience sample of health care workers was chosen (1) to pilot-test the intervention in a controlled setting before patient implementation and (2) to establish feasibility and refine study procedures prior to recruiting actual patients.

Participants were recruited through email announcements sent to hospital faculty, staff, researchers, and students. Interested volunteers contacted the research team via email and were provided with study information. Inclusion criteria were (1) adult health care workers (aged ≥18 years), (2) proficiency in English, and (3) access to a computer or mobile device with internet connectivity. Exclusion criteria included (1) prior involvement in postoperative instruction material development and (2) visual or hearing impairments that would prevent video viewing or text reading. All instruments were administered online using Google Forms in a single session. Participants accessed the form through a unique link and completed all sections sequentially.

Ethical Considerations

This study was reviewed and approved by the Mayo Clinic Institutional Review Board (application 25-002248). The study was conducted in accordance with the ethical standards of the responsible institutional committee on human experimentation and with the World Medical Association's Declaration of Helsinki.

All participants were adult health care worker volunteers (faculty, staff, researchers, and students) recruited through institutional email announcements. Written informed consent was obtained electronically through the Google Forms platform before randomization and data collection began. Participants were informed of the study's purpose and voluntary nature and their right to withdraw at any time without consequences.

Participant privacy and confidentiality were maintained throughout the study. All survey responses were collected anonymously via Google Forms; no personally identifiable information (eg, name, employee ID, email address, or date of birth) was linked to individual survey responses. Deidentified study data were stored on password-protected, institutionally managed servers accessible only to authorized members of the research team. No images, audio, or video of participants were

XSL•FO
RenderX

collected. Only aggregate, deidentified data are reported in this manuscript.

Participants received no financial compensation, gift cards, or other incentives for their participation in this study.

Educational Materials

Text-Based Instructions

We compiled a set of 10 standard postoperative instruction handouts covering common topics relevant to general surgery recovery. These were developed based on established postoperative care guidelines and publicly available patient education resources from major medical institutions [26] and included the following 10 frequently asked topics: pain, pain medication, alarm signs after surgery, surgical drain management, diet after surgery, postoperative nausea, follow-up appointment instructions, recovery instructions, scars after surgery, and suture management [27-36]. Each handout was approximately 100 to 200 words in length and written at an eighth- to ninth-grade reading level [37]. These served as the standard care educational materials for the control group.

AI-Generated Video Instructions

Using HeyGen (HeyGen Inc) [38], we created video versions covering the same topics. Scripts were based on the text handout content and edited for a conversational tone. Average video duration was 59.2 (range 45 - 84) seconds. We used realistic, diverse avatar presenters (3 male and 3 female avatars) with AI-generated voices. The pain and pain medication topics were merged, yielding 9 videos covering 10 topics. Videos featured avatars speaking directly to viewers, with bullet points and subtitles displayed (Figure 1 [25]).

Figure 1. Examples of text-to-video avatars demonstrating the generative artificial intelligence's ability to generate realistic and diverse human-like representations. Images were generated using HeyGen (HeyGen Inc). Collage created with BioRender.com [39].



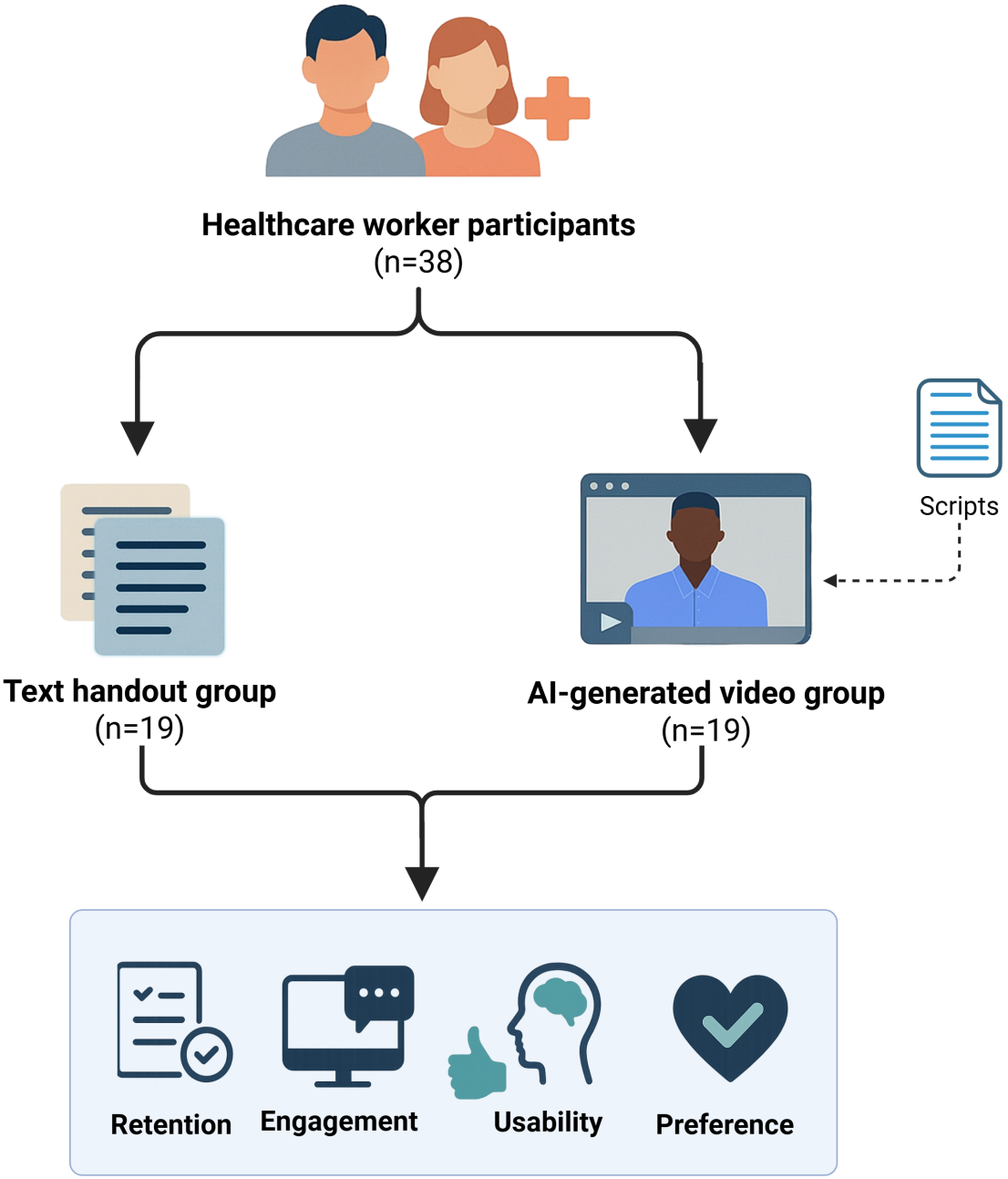
From a practical standpoint, the HeyGen platform enabled rapid video production without specialized equipment, studios, or video editing expertise. Each video was generated within minutes of script input, and the total time to produce all 9 videos was approximately 2 to 3 hours, including script preparation and quality review.

Randomization

Volunteers were randomly assigned in a 1:1 ratio to either the text handout group (control) or the AI-generated video group (intervention) using simple randomization (drawing lots). Randomization was performed after informed consent was obtained and before participants accessed the educational materials (Figure 2).

XSL•FO
RenderX

Figure 2. Study workflow demonstrating the allocation of health care workers (HCWs) into 2 groups and subsequent measured outcomes. Image created with BioRender [40]. AI: artificial intelligence.



Outcomes and Measures

Three custom instruments were administered: (1) a 10-item multiple-choice postoperative knowledge assessment quiz (score range 0 - 10); (2) a user experience survey with seven 5-point Likert scale items measuring clarity, usefulness, engagement, confidence, retention, preference, and effectiveness; and (3) an engagement metrics questionnaire capturing time spent, materials reviewed, and 3 open-ended questions. Demographics collected included age, sex, education, postoperative care experience, and AI familiarity. Complete instruments are provided in Multimedia Appendix 1. The knowledge assessment quiz was developed by the study team, informed by publicly available postoperative care resources from the American College of Surgeons [26] and Mayo Clinic patient education materials [27-36]. Content validity was established through review by 3 research fellows holding a doctor of medicine or bachelor of medicine and bachelor of surgery degree under the

XSL•FO
RenderX

guidance of a plastic surgeon. Educational materials were written at an eighth- to ninth-grade reading level, consistent with recommendations from the National Institutes of Health that patient education materials do not exceed the eighth-grade level [41].

Participants self-reported time spent reviewing materials. Following exposure to the educational content, participants completed the knowledge quiz and user experience survey.

Usability was operationalized through perceived usefulness ("The information provided was useful"), perceived clarity ("The instructions were clear and easy to understand"), and preference for clinical use ("I would prefer video-based instructions over text-based instructions in clinical practice").

Engagement time and material consumption were self-reported, consistent with standard educational research practices, to avoid technical barriers that automated tracking might introduce.

Qualitative responses were analyzed using a basic grounded theory approach, which is appropriate for exploratory pilot studies seeking to identify emergent themes without predetermined coding frameworks. Two investigators independently reviewed all open-ended responses and conducted open coding to identify recurring themes and sentiments (eg, "video was more personal," "text was hard to read," and "avatar speech seemed unnatural"). The investigators then met to discuss their independent findings, reconcile any discrepancies through consensus, and develop a final thematic framework organized into 3 categories: positive aspects, areas for improvement, and clinical practice preferences. Given the limited qualitative data (3 open-ended questions from 38 participants) and the exploratory pilot nature of the study, formal qualitative coding software, interrater reliability calculation, or thematic saturation assessment were not used. However, trustworthiness was enhanced through dual independent coding, consensus meetings, and inclusion of representative participant quotes to support identified themes. All qualitative data were collected and analyzed after completion of data collection.

Statistical Analysis

Statistical analyses were performed using Python (version 3.12; Python Software Foundation) with the SciPy library. Independent samples 2-tailed *t* tests were used to compare mean scores between groups for all quantitative outcomes. The use of *t* tests for 5-point Likert scales is supported by evidence that parametric tests are robust to violations of normality assumptions, particularly with balanced group sizes [42]. A significance level of $P<.05$ (2-tailed) was used, with effect sizes interpreted as small (0.2), medium (0.5), or large (≥0.8). Effect sizes were calculated using Cohen *d*. Data are presented as mean (SD). Given the pilot nature of the study and multiple comparisons across secondary outcomes, results should be interpreted with appropriate caution.

As this was a pilot study designed to assess feasibility, a formal a priori power analysis was not conducted. The sample size (N=38; n=19 per group) was determined based on feasibility and resource constraints typical of pilot investigations. The observed effect sizes (Cohen d=0.45 for knowledge; Cohen d=0.81-1.38 for subjective outcomes) can inform sample size calculations for future definitive trials.

Results

Participant Characteristics

A total of 38 volunteers participated in the study, with 19 (50%) participants randomized to the text-based instruction group and 19 (50%) to the AI-generated video instruction group. Table 1 demonstrates that the text handout and AI-generated video groups exhibited similar demographic profiles. The gender distribution was nearly even across both groups, with a slight female majority in both groups. The Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth (CONSORT-eHEALTH) diagram is provided in Figure 3.

XSL•FO
RenderX

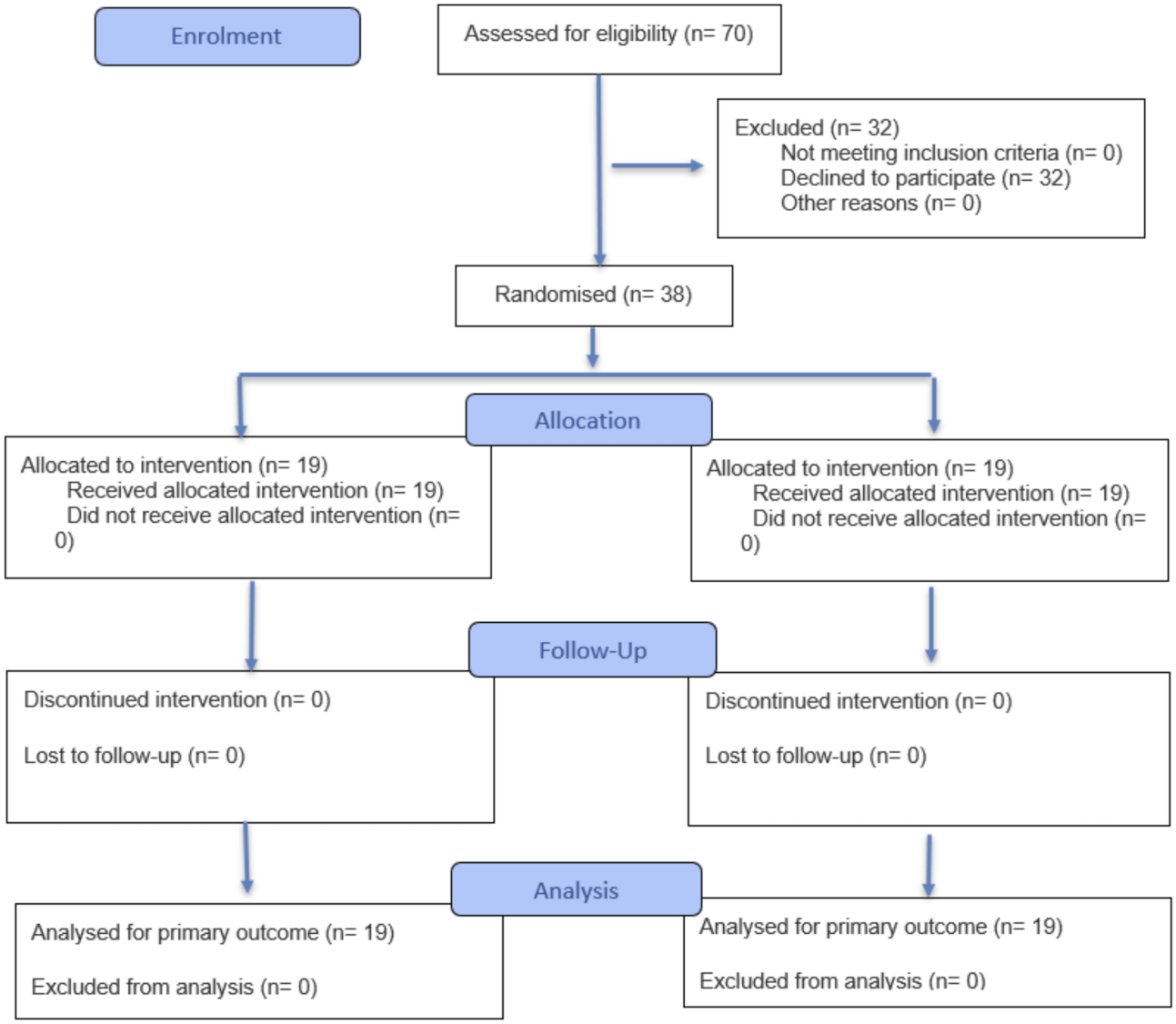
Table . Demographic characteristics of study participants.

Characteristics	Text handout group (n=19)	AI[a]-generated video group (n=19)
Sex, n (%)		
Male	7 (37)	8 (42)
Female	12 (63)	11 (58)
Age (years), mean (SD)	32.53 (6.93)	31.95 (7.68)
Educational level, n (%)		
College (1-4 years)	6 (32)	5 (26)
Graduate (5-7 years)	7 (37)	10 (53)
Postgraduate (>7 years)	6 (32)	4 (21)
Previous experience with postoperative care, n (%)		
Yes	11 (58)	11 (58)
No	6 (32)	6 (32)
Maybe	2 (10)	2 (10)
Familiarity with AI-generated content, n (%)		
Very unfamiliar	4 (21)	4 (21)
Unfamiliar	2 (10)	1 (5)
Neutral	4 (21)	4 (21)
Familiar	5 (26)	4 (21)
Very familiar	4 (21)	6 (32)

[a]AI: artificial intelligence.

XSL•FO
RenderX

Figure 3. Flow diagram of the progress through the phases of a randomized trial of two groups (that is, enrollment, intervention allocation, follow-up, and data analysis).



The average age of participants was comparable, with the text handout group having a mean age of 32.53 (SD 6.93) years and the AI-generated video group having a mean age of 31.95 (SD 7.68) years. Educational backgrounds were also similar, with a predominance of graduate-level education in both groups. Regarding prior experience with postoperative care, most participants reported some level of familiarity, though a notable portion, approximately one-third in each group, reported no experience. Familiarity with AI-generated content was also similarly distributed, with 29% (n=11) of participants indicating unfamiliarity, 21% (n=8) expressing neutrality, and 50% (n=19) reporting familiarity.

Objective Knowledge Assessment

Participants in both groups completed the same 10-item knowledge quiz assessing comprehension of postoperative topics. The AI-generated video group achieved a mean score of 8.89 (SD 1.20) out of 10, while the text handout group achieved a mean score of 8.21 (SD 1.78). This difference was not statistically significant (P=.17; Cohen d=0.45). Both groups demonstrated high overall comprehension, with mean scores exceeding 80% correct.

Engagement Levels and Time Spent

Mean completion time for the entire survey (including time spent reviewing educational materials) was approximately 25 minutes for the text handout group and 30 minutes for the AI-generated video group.

Engagement time differed significantly between the groups (P=.004; Cohen d=1.04). Participants in the AI-generated video group spent more time, on average, engaging with the instructional material than those in the text handout group. The AI-generated video group had a mean engagement time of 15.11 (SD 7.78) minutes, whereas the text handout group's mean was 8.84 (SD 4.03) minutes (P=.004; Cohen d=1.04). This large effect size (Cohen d>0.8) indicates a substantial practical difference in engagement time between formats. Participants in the AI-generated video group watched, on average, 7.05 (SD 2.67) videos. In contrast, participants in the text handout group read, on average, 6.68 (SD 2.92) instructions. The increased engagement time in the AI-generated video group may reflect participants pausing the videos for various reasons, including reviewing specific information, allowing personal processing time, or taking breaks.

XSL•FO
RenderX

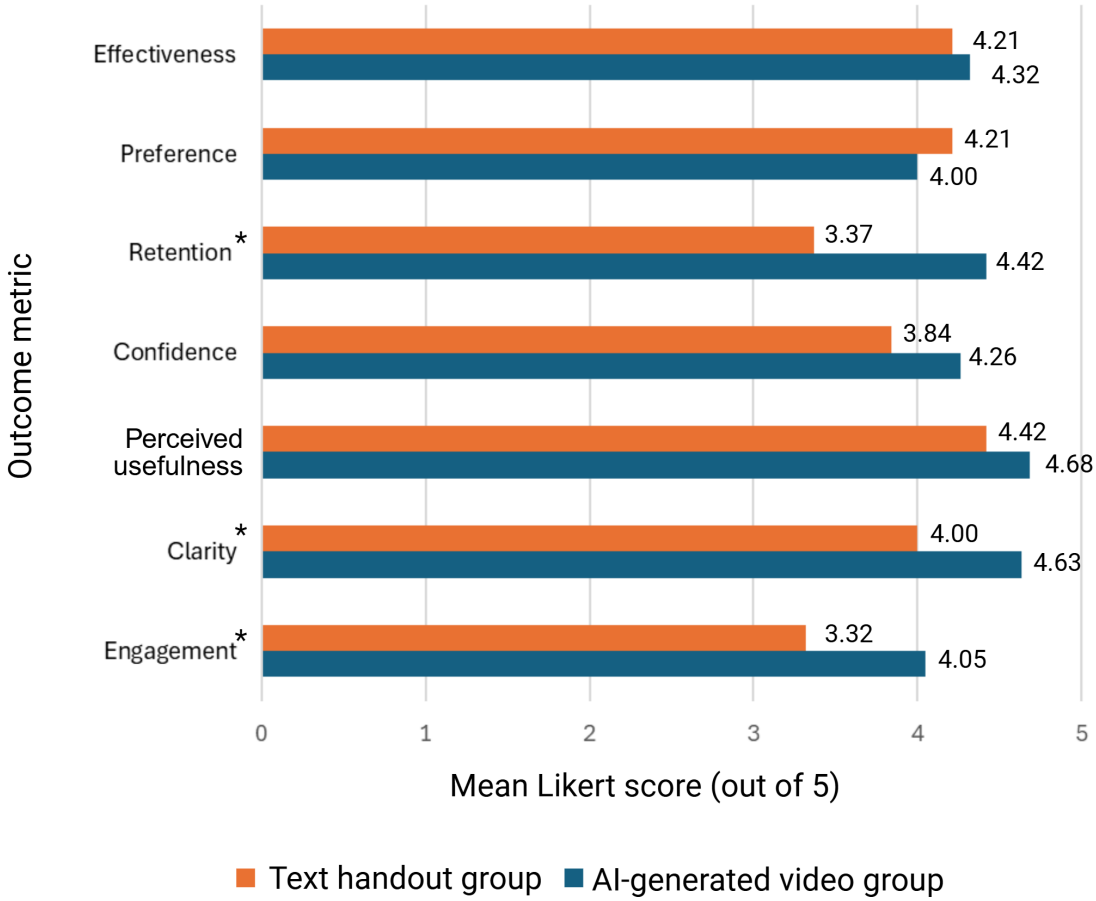
Subjective engagement ratings also differed significantly between the groups. The AI-generated video group's mean engagement rating was 4.05 (SD 0.78), while the text handout group's mean engagement rating was 3.32 (SD 1.00; *P*=.02; Cohen *d*=0.81). This large effect size suggests that participants found the video format substantially more engaging than the text format.

Usability, Understanding, and Retention

In summary, the AI-generated video group demonstrated significantly higher engagement time, perceived clarity, subjective engagement, and information retention than the text handout group. No significant differences were observed in objective knowledge scores, perceived usefulness, confidence in understanding, preference for video-based instructions, or effectiveness in answering postoperative questions.

Participants rated the clarity of instructions significantly higher in the AI-generated video group, indicating that they found video instructions clearer than text instructions. Specifically, on a 5-point Likert scale, the AI-generated video group reported a mean clarity score of 4.63 (SD 0.50), while the text handout group reported a mean of 4.00 (SD 0.82; *P*=.007; Cohen *d*=0.93). The large effect size (Cohen *d*=0.93) demonstrates a meaningful practical difference in perceived clarity. However, participants did not perceive a significant difference in the usefulness of the information between the 2 groups. The AI-generated video group reported a mean usefulness score of 4.68 (SD 0.58), and the text handout group reported a mean of 4.42 (SD 0.61; *P*=.18; Cohen *d*=0.43). Figure 4 provides a visual comparison of all Likert scale metrics between the text and AI-generated video groups.

Figure 4. Comparison of text handout and artificial intelligence (AI)–generated video performance across 7 metrics, measured on 5-point Likert scales. The chart displays mean ratings for both groups. *Statistically significant differences (clarity, *P*=.007; engagement, *P*=.02; retention, *P*<.001).



Although participants in the AI-generated video group reported slightly higher confidence in understanding the instructions (mean 4.26, SD 0.65) than the text handout group (mean 3.84, SD 0.83), this difference did not reach statistical significance (*P*=.09; Cohen *d*=0.57). Participants found the video format significantly more effective for information retention. They reported a mean Likert retention score of 4.42 in the AI-generated video group, compared to 3.37 in the text handout group (*P*<.001; Cohen *d*=1.38). This very large effect size (Cohen *d*=1.38) represents one of the strongest practical differences observed in the study, indicating that participants

XSL•FO
RenderX

perceived substantially better information retention with the video format.

Both groups reported similar preferences for video-based instructions in clinical practice, with the AI-generated video group scoring a mean of 4.00 (SD 1.25) and the text handout group scoring a mean of 4.21 (SD 0.71; P=.53; Cohen d=−0.21). Finally, participants rated the effectiveness of the instructions in answering common postoperative questions similarly across both groups, with the AI-generated video group scoring a mean of 4.32 (SD 0.89) and the text handout group scoring a mean of 4.21 (SD 0.98; P=.73; Cohen d=0.12). Table 2 tabulates the engagement, usability, and retention metrics between the text handout and AI-generated video formats.

Table . Comparative analysis of engagement, usability, and retention outcomes between artificial intelligence (AI)–generated video and text instructional formats.

Outcomes	AI-generated video group, mean (SD)	Text handout group, mean (SD)	*P* value
Knowledge quiz score (0-10)	8.89 (1.20)	8.21 (1.78)	.17
Engagement time (minutes)	15.11 (7.78)	8.84 (4.03)	.004
Subjective engagement rating (0-5)	4.05 (0.78)	3.32 (1.00)	.02
Clarity of instructions (0-5)	4.63 (0.50)	4.00 (0.82)	.007
Perceived usefulness of information (0-5)	4.68 (0.58)	4.42 (0.61)	.18
Confidence in understanding (0-5)	4.26 (0.65)	3.84 (0.83)	.09
Retention score (0-5)	4.42 (0.84)	3.37 (0.68)	<.001
Preference in clinical practice (0-5)	4.00 (1.25)	4.21 (0.71)	.53
Number of videos or instructions	7.05 (2.67)	6.68 (2.92)	N/A[a]
Effectiveness in answering questions (0-5)	4.32 (0.89)	4.21 (0.98)	.73

[a]N/A: not applicable. A P value was not calculated because videos and written instructions are different units of content and cannot be directly compared.

The large effect sizes observed for engagement, clarity, and retention outcomes indicate that the differences between video and text formats represent substantial practical significance beyond statistical significance alone. The moderate effect size for confidence in understanding (Cohen d=0.57) suggests a meaningful trend despite not reaching statistical significance (P=.09), likely due to our small sample size. In contrast, usefulness, clinical practice preference, and effectiveness in answering questions showed minimal practical differences between formats.

Qualitative Feedback

Participants provided qualitative feedback on clarity, engagement, suggested improvements, clinical practice use, and format preference. Representative participant quotes illustrating these themes are provided in Table 3.

Table . Qualitative feedback themes by group.

Themes	AI-generated video group	Text handout group
Clarity	“Crystal clear instructions” and “Clear and direct communication”	“Crisp and clear instructions” and “Clear and easy, no issues”
Engagement	“Engaging like a real guide” and “Lifelike, memorable instructions”	“Quick,” “Not engaging,” and “Less engaging if someone doesn’t like reading”
Suggested improvements	“Gestures not synced with voice,” “Unnatural facial expressions,” and “Slightly robotic voice”	“Could include diagrams or video,” “More visuals or a short summary,” and “Add an FAQ [frequently asked question]."
Clinical practice use	13 out of 19 participants would use; “We are fully engaging when watching videos”	10 out of 19 participants would use; “I’d still prefer a short video to keep people more engaged”
Format preference	“Engaging,” “Easy-to-follow narration,” and “Effective AI explanations”	“Straight to the point” and “Easy to skim”; many preferred video as a supplement

XSL•FO
RenderX

Discussion

Principal Findings

This pilot study assessed the feasibility and user experience of AI-generated avatar videos for postoperative instruction delivery among health care workers. The AI-generated video group demonstrated significantly higher engagement, perceived clarity, and information retention than the text handout group, while objective knowledge scores did not differ significantly between groups. These findings suggest that AI-generated avatar videos are feasible to produce using commercially available platforms and may offer meaningful advantages in learner engagement, although their impact on objective knowledge acquisition requires further investigation in patient populations.

The significantly higher engagement with AI-generated videos aligns with media richness theory [42], which suggests that richer media formats are more effective in maintaining attention. The increased engagement time may reflect participants pausing or revisiting video segments to reinforce key points. The superior clarity ratings for video instructions may reflect reduced cognitive load through combined visual and auditory channels [43]. Cognitive load theory suggests that well-designed multimedia can prevent working memory overload, and the conversational tone of AI-generated avatars may create a more approachable instructional style than formal text [13,44].

Participants perceived greater retention with video instructions, consistent with dual-coding theory [45], which posits that information presented in both visual and auditory formats results in more robust encoding. A hybrid approach combining text clarity with video engagement benefits may offer the greatest educational impact [31,32]. The discrepancy between perceived and actual learning outcomes is consistent with existing research [46], suggesting that engagement benefits may not directly translate to measurable knowledge gains in populations with high baseline comprehension.

Qualitative feedback reinforced the quantitative findings, with participants noting the engaging and personal qualities of avatar-led instruction while identifying areas for technological improvement, particularly in avatar speech naturalness and gesture synchronization. These insights will guide iterative refinement of the AI-generated video format for patient-facing deployment.

Several recent studies support the growing viability of AI-generated avatar–based education in health care. Kim et al [47] found that 64% of participants preferred an AI-generated video bot over a text-based chatbot for surgical education. Coleman et al [48] demonstrated that AI-generated digital clinicians achieved significantly higher knowledge scores in a randomized controlled trial. Artsi et al [49] found that while evidence remains sparse, studies consistently report engagement benefits with AI-generated video content. Our findings align with this emerging literature, supporting the engagement advantages of AI-generated video while highlighting the need for validated knowledge assessments in future studies.

Limitations

Several limitations of our study must be considered when interpreting the results. First, this was a pilot study with a relatively small sample size (N=38), limiting statistical power and the generalizability of our findings to broader patient populations. Participants consisted exclusively of health care worker volunteers (faculty, staff, researchers, and students), and specific clinical roles (eg, physicians, nurses, and administrative staff) were not systematically recorded. Health care workers likely possess higher baseline knowledge about postoperative care than typical surgical patients, which may explain the high quiz scores in both groups (>80%) and the lack of a significant difference between formats. This ceiling effect may have obscured the potential benefits of video instruction that could emerge in populations with lower health literacy. Future studies should include patients with diverse educational backgrounds and health literacy levels.

Second, we used a convenience sampling approach within a single institution, which may have introduced selection bias and further limited the applicability of our results to diverse clinical settings or populations. Additionally, the study evaluated only short-term comprehension and perceived retention, without assessing long-term retention, behavioral adherence, or clinical outcomes such as postoperative recovery or complication rates.

Third, many outcomes, including engagement time, retention, and clarity, were based on self-reported data, introducing the possibility of recall, reporting, or social desirability biases. Objective measures beyond immediate quiz performance were not included, limiting conclusions about true knowledge retention or clinical impact. The observed benefits in engagement and perceived effectiveness may have been influenced by the novelty effect of interacting with AI-generated avatars, which could diminish over repeated exposures or extended use. The longer engagement times observed in the AI-generated video group may have multiple explanations beyond genuine engagement, including time spent adapting to the video interface, pausing to process unfamiliar avatar presentations, or technical factors such as video buffering or lag reported by some participants. These confounding factors limit the interpretation of engagement time as a pure measure of educational engagement. The study also evaluated only a single generative AI platform (HeyGen), and results might differ substantially using alternative AI technologies or avatar presentations.

Moreover, participants reviewed educational materials outside a clinical context, potentially impacting their engagement and motivation differently compared with how real postoperative patients would experience. Furthermore, participants' technical familiarity and comfort with AI technology varied, possibly biasing results toward those more comfortable interacting with digital or AI-generated content. Finally, our study did not formally assess participants' baseline health literacy levels, limiting insights into whether the observed improvements are equally beneficial across diverse literacy groups.

Additionally, health care workers do not approximate the general patient population: most participants had college or graduate-level education, were relatively young, and many had

XSL•FO
RenderX

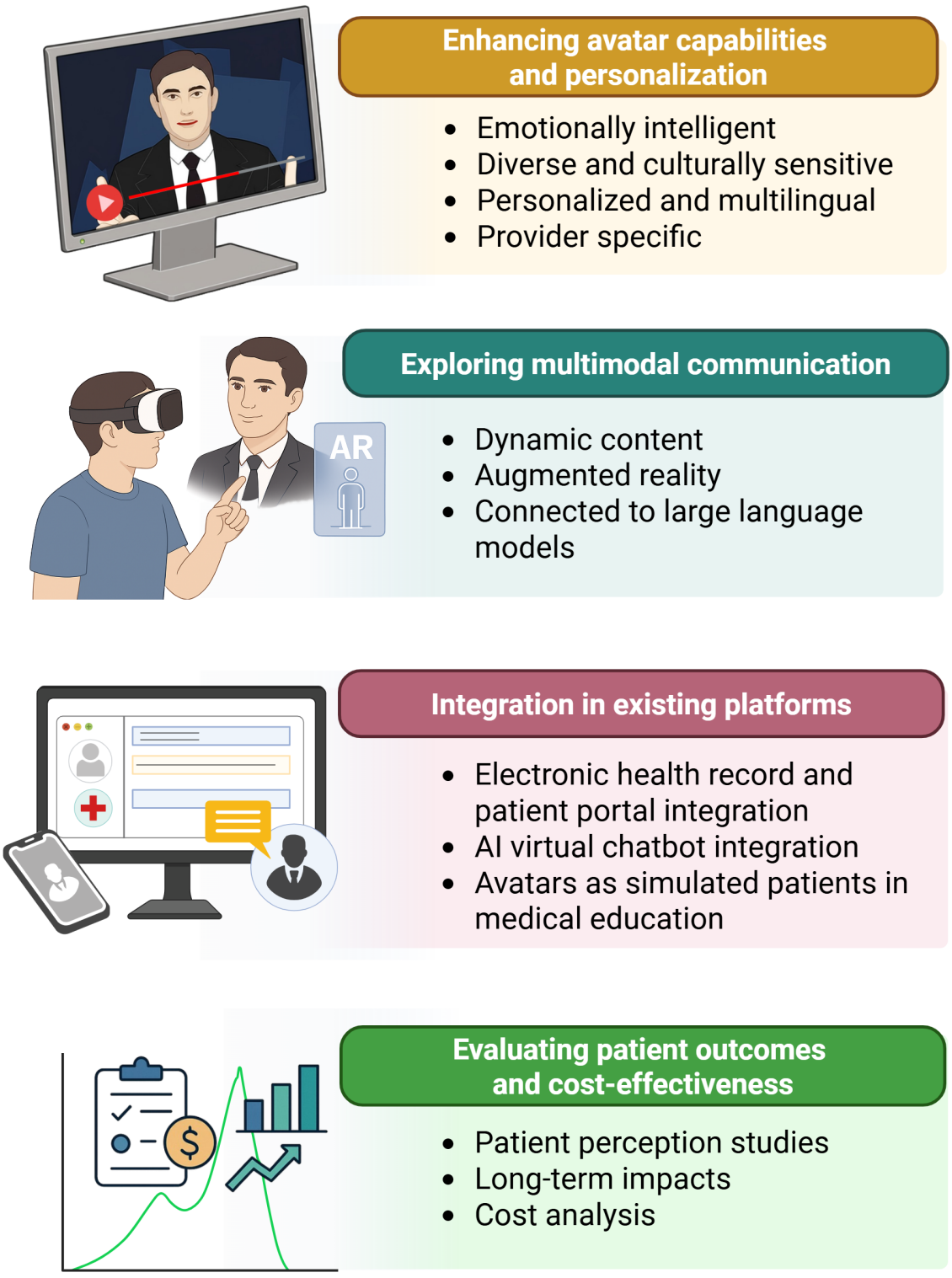
prior postoperative care experience. The knowledge quiz was developed by the study team using publicly available patient education resources rather than a validated instrument, and no formal psychometric testing (eg, Cronbach α) was conducted. The multiple-choice format with easily excludable distractors may have allowed correct answers to be guessed through commonsense elimination, contributing to the ceiling effect observed in both groups (>80% correct). The reading level of the educational materials (eighth to ninth grade) exceeds the American Medical Association recommendation of sixth-grade level for general patient populations, although it is consistent with the National Institutes of Health recommendation of no higher than the eighth-grade level [41] and was appropriate for our health care worker sample.

Future Directions

Future research should evaluate AI-generated video instructions in actual patient populations with diverse health literacy levels. Key priorities include (1) larger randomized trials in surgical patients, (2) assessment of long-term retention and clinical outcomes (eg, adherence and complications), (3) comparison across different AI platforms, and (4) cost-effectiveness analyses comparing AI-generated videos to traditional production methods. Hybrid text-video approaches and culturally tailored avatar presentations warrant investigation [24] (Figure 5 [25]).

XSL•FO
RenderX

Figure 5. Future directions and recommendations for the use of generative artificial intelligence (AI)–based text-to-video models in health care. Created using BioRender [50].



Future research should explore multimodal communication, including avatars with dynamic visual aids, augmented reality or virtual reality integration [51], and enhanced natural language processing capabilities through large language model integration. Longitudinal studies assessing sustained educational impact, comparative effectiveness against traditional methods, and cost-effectiveness analyses are needed. Technical priorities include electronic health record integration, 24×7 chatbot

XSL•FO
RenderX

availability, and improved accessibility features. Beyond patient education, AI-generated avatars may serve as valuable training tools for medical students in simulated clinical scenarios.

Ethical Implications of AI

AI-generated avatars raise critical ethical concerns. Transparency is essential—patients must be informed that they are interacting with AI rather than humans to preserve trust and prevent unrealistic expectations. Accuracy requires ongoing monitoring and human oversight, as technological limitations may cause mispronunciations or misrepresentations that compromise comprehension.

Avatar-driven education should complement rather than replace human interaction, as AI cannot authentically replicate empathy and emotional responsiveness. Some users may form unintended emotional connections with AI-generated characters [52,53]; continuous availability of human support can mitigate this risk.

Digital avatars of real individuals require explicit consent regarding creation, distribution, and use. Unauthorized "deepfake" avatars pose significant risks, including misinformation, reputational damage, and erosion of trust [54,55]. Future implementations involving patient data will require robust cybersecurity and privacy protections.

Comprehensive regulatory frameworks addressing consent, ownership, transparency, privacy, accuracy standards, and accountability are essential for the responsible integration of AI-generated avatars in health care [56].

Conclusions

This pilot study demonstrated the feasibility of rapidly producing AI-generated avatar videos for postoperative instruction using commercially available platforms. Although objective knowledge scores did not differ significantly between formats, the AI-generated video group showed significantly higher engagement, perceived clarity, and information retention. Key limitations include the small sample size, use of health care workers rather than actual patients, reliance on self-reported engagement, and use of a nonvalidated knowledge instrument with ceiling effects. Subsequent evaluation with actual surgical patients has demonstrated high trust and usability for AI physician avatars [25]. Future research should prioritize evaluation in diverse patient populations with varying health literacy levels, using validated instruments and clinical outcome measures.

Funding

This study was supported by Mayo Clinic and the generosity of Eric Schmidt and Wendy Schmidt, Dalio Philanthropies, and Gerstner Philanthropies.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to participant privacy considerations and the terms of the institutional review board approval but are available from the corresponding author (AJF) on reasonable request.

Authors' Contributions

Conceptualization: SAH, AJF
Data curation: SAH, AG
Formal analysis: SAH, SP
Funding acquisition: SB, CT, AJF
Investigation: SAH, SP, CAG-C, AG, BGC
Methodology: SAH, SP, CT
Project administration: JL, SB, AJF
Resources: CAG-C, JL, BGC
Software: SAH, SP
Supervision: NW, SB, CT, MAL, AJF
Validation: NW, MAL
Visualization: SAH
Writing—original draft: SAH
Writing—review and editing: SAH, SP, CAG-C, BGC, NW, SB, CT, MAL, AJF

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study instruments: 10-item postoperative knowledge quiz, 7-item user experience survey (5-point Likert scale), and engagement metrics questionnaire administered to participants in both the text handout group and the artificial intelligence–generated video group.
[DOCX File, 39 KB - periop_v9i1e89277_app1.docx]

XSL•FO
RenderX

Checklist 1 [DOCX File, 33 KB - periop_v9i1e89277_app2.docx]

References

1. Kindig DA, Panzer AM, Nielsen-Bohlman L, Institute of Medicine (US) Committee on Health Literacy. Health Literacy: A Prescription to End Confusion: National Academies Press; 2004.
2. van den Broek P, White MJ, Kendeou P, Carlson S. Reading between the lines: developmental and individual differences in cognitive processes in reading comprehension. In: Wagner RK, Schatschneider C, Plythian-Sence C, editors. Beyond Decoding: The Behavioral and Biological Foundations of Reading Comprehension: Guilford Press; 2009:107-123.
3. Engel KG, Heisler M, Smith DM, Robinson CH, Forman JH, Ubel PA. Patient comprehension of emergency department care and instructions: are patients aware of when they do not understand? Ann Emerg Med 2009 Apr;53(4):454-461. [doi: 10.1016/j.annemergmed.2008.05.016] [Medline: 18619710]
4. McCarthy DM, Waite KR, Curtis LM, Engel KG, Baker DW, Wolf MS. What did the doctor say? Health literacy and recall of medical instructions. Med Care 2012 Apr;50(4):277-282. [doi: 10.1097/MLR.0b013e318241e8e1] [Medline: 22411440]
5. Koivisto JM, Saarinen I, Kaipia A, et al. Patient education in relation to informational needs and postoperative complications in surgical patients. Int J Qual Health Care 2020 Apr 21;32(1):35-40. [doi: 10.1093/intqhc/mzz032] [Medline: 31016323]
6. Palumbo R. Examining the impacts of health literacy on healthcare costs. An evidence synthesis. Health Serv Manage Res 2017 Nov;30(4):197-212. [doi: 10.1177/0951484817733366] [Medline: 29034727]
7. Marwick C. Patients' lack of literacy may contribute to billions of dollars in higher hospital costs. JAMA 1997 Sep 24;278(12):971. [PMC: 9307333]
8. MacLeod S, Musich S, Gulyas S, et al. The impact of inadequate health literacy on patient satisfaction, healthcare utilization, and expenditures among older adults. Geriatr Nurs 2017;38(4):334-341. [doi: 10.1016/j.gerinurse.2016.12.003] [Medline: 28089217]
9. Card EB, Morales CE, Ramirez JM, et al. Impact of illustrated postoperative instructions on knowledge and retention during a cleft lip and palate surgical mission. Cleft Palate Craniofac J 2023 Aug;60(8):1010-1020. [doi: 10.1177/10556656221100052] [Medline: 35711155]
10. Fan KL, Black CK, DeFazio MV, Luvisa K, Camden R, Song DH. Bridging the knowledge gap: an examination of the ideal postoperative autologous breast reconstruction educational material with A/B testing. Plast Reconstr Surg 2020 Jan;145(1):258-266. [doi: 10.1097/PRS.0000000000006373] [Medline: 31609288]
11. Horstman MJ, Mills WL, Herman LI, et al. Patient experience with discharge instructions in postdischarge recovery: a qualitative study. BMJ Open 2017 Feb 22;7(2):e014842. [doi: 10.1136/bmjopen-2016-014842] [Medline: 28228448]
12. Krontoft AS. How do patients prefer to receive patient education material about treatment, diagnosis and procedures? —A survey study of patients preferences regarding forms of patient education materials; leaflets, podcasts, and video. Open J Nurs 2021;11(10):809-827. [doi: 10.4236/ojn.2021.1110068]
13. Katt BM, Tawfik AM, Imbergamo C, McEntee R, Beredjiklian PK, Ilyas AM. Patient comprehension of operative instructions with a paper handout versus a video: a prospective randomized control trial. J Hand Surg Am 2023 Mar;48(3):311. [doi: 10.1016/j.jhsa.2021.10.017] [Medline: 35012796]
14. O'Callaghan J. How OpenAI's text-to-video tool Sora could change science - and society. Nature 2024 Mar;627(8004):475-476. [doi: 10.1038/d41586-024-00661-0] [Medline: 38472485]
15. Li W, Yang J, Shao Q, Wu H, Li C. From text to video: what will OpenAI's Sora bring to the oncologic field? Int J Surg 2025 Jan 1;111(1):1666-1668. [doi: 10.1097/JS9.0000000000001988] [Medline: 39093893]
16. Mohamed AA, Lucke-Wold B. Text-to-video generative artificial intelligence: Sora in neurosurgery. Neurosurg Rev 2024 Jun 13;47(1):272. [doi: 10.1007/s10143-024-02514-w] [Medline: 38867134]
17. Kang Y, Wang S, Zhu L. Text-to-video models and Sora in plastic surgery: pearls, pitfalls, and prospectives. Aesthetic Plast Surg 2025 Jun;49(11):3224-3228. [doi: 10.1007/s00266-024-04583-0] [Medline: 39663218]
18. Mayer RE. The science of instruction: determining what works in multimedia learning. In: Multimedia Learning, 2nd edition: Cambridge University Press; 2009:28-56. [doi: 10.1017/CBO9780511811678.004]
19. Stull A, Fiorella L, Similuk R, Ibonie S, Mayer R. Effects of instructor presence in video lectures: rapport, attention, and learning. Presented at: Proceedings of the Annual Meeting of the Cognitive Science Society; Jul 26-29, 2023 URL: https://escholarship.org/uc/item/8033r3wc [accessed 2026-05-05]
20. Polat H. Instructors' presence in instructional videos: a systematic review. Educ Inf Technol (Dordr) 2022 Dec 23:1-33. [doi: 10.1007/s10639-022-11532-4] [Medline: 36589519]
21. Zhang L, Gallagher R, Du H, et al. Evaluate the effect of virtual nurse-guided discharge education app on disease knowledge and symptom response in patients following coronary events. Int J Med Inform 2025 Apr;196:105818. [doi: 10.1016/j.ijmedinf.2025.105818] [Medline: 39919533]
22. Wonggom P, Kourbelis C, Newman P, Du H, Clark RA. Effectiveness of avatar-based technology in patient education for improving chronic disease knowledge and self-care behavior: a systematic review. JBI Database System Rev Implement Rep 2019 Jun;17(6):1101-1129. [doi: 10.11124/JBISRIR-2017-003905] [Medline: 31021975]
23. Bedra M, Wick E, Brotman D, Finkelstein J. Avatar-based interactive ileostomy education in hospitalized patients. Stud Health Technol Inform 2013;190:83-85. [Medline: 23823383]

XSL•FO
RenderX

24. Canidate S, Hart M. The use of avatar counseling for HIV/AIDS health education: the examination of self-identity in avatar preferences. J Med Internet Res 2017 Dec 1;19(12):e365. [doi: 10.2196/jmir.6740] [Medline: 29196281]
25. Haider SA, Prabha S, Gomez-Cabello CA, et al. Artificial intelligence physician avatars for patient education: a pilot study. J Clin Med 2025 Dec 4;14(23):8595. [doi: 10.3390/jcm14238595] [Medline: 41375898]
26. Recovery and discharge. American College of Surgeons. URL: https://www.facs.org/for-patients/recovering-from-surgery [accessed 2025-11-05]
27. What to do — and avoid — before anesthesia. Mayo Clinic. 2023. URL: https://mcpress.mayoclinic.org/living-well/what-to-do-and-avoid-before-anesthesia/ [accessed 2026-03-26]
28. General anesthesia. Mayo Clinic. URL: https://www.mayoclinic.org/tests-procedures/anesthesia/about/pac-20384568 [accessed 2026-03-26]
29. Nausea and vomiting. Mayo Clinic. URL: https://www.mayoclinic.org/symptoms/nausea/basics/definition/sym-20050736 [accessed 2026-03-26]
30. Pain medicines after surgery. Mayo Clinic. URL: https://www.mayoclinic.org/tests-procedures/anesthesia/in-depth/pain-medications/art-20046452 [accessed 2026-03-26]
31. Finding the balance: opioids and pain control after surgery. Mayo Clinic. 2023. URL: https://newsnetwork.mayoclinic.org/discussion/finding-the-balance-opioids-and-pain-control-after-surgery/ [accessed 2026-03-26]
32. Cancer surgery: physically removing cancer. Mayo Clinic. 2026. URL: https://www.mayoclinic.org/diseases-conditions/cancer/in-depth/cancer-surgery/art-20044171 [accessed 2026-03-26]
33. Sparks D. Saving lives by reducing surgical site infections. Mayo Clinic. 2012. URL: https://newsnetwork.mayoclinic.org/discussion/saving-lives-by-reducing-surgical-site-infections/ [accessed 2026-03-26]
34. Cuts and scrapes: first aid. Mayo Clinic. 2024. URL: https://www.mayoclinic.org/first-aid/first-aid-cuts/basics/art-20056711 [accessed 2026-03-26]
35. Fever treatment: quick guide to treating a fever. Mayo Clinic. 2025. URL: https://www.mayoclinic.org/diseases-conditions/fever/in-depth/fever/art-20050997 [accessed 2026-03-26]
36. Bishop S. Wound location first thing to consider when deciding if medical attention is needed. Mayo Clinic. 2013. URL: https://newsnetwork.mayoclinic.org/discussion/wound-location-first-thing-to-consider-when-deciding-if-medical-attention-is-needed/ [accessed 2026-03-26]
37. Flesch R. A new readability yardstick. J Appl Psychol 1948 Jun;32(3):221-233. [doi: 10.1037/h0057532] [Medline: 18867058]
38. HeyGen. URL: https://www.heygen.com [accessed 2025-03-04]
39. BioRender licence for Figure 1. BioRender. URL: https://BioRender.com/umhj87z [accessed 2026-05-15]
40. BioRender licence for Figure 2. BioRender. URL: https://BioRender.com/iuwpq8k [accessed 2026-05-15]
41. Rooney MK, Santiago G, Perni S, et al. Readability of patient education materials from high-impact medical journals: a 20-year analysis. J Patient Exp 2021;8:2374373521998847. [doi: 10.1177/2374373521998847] [Medline: 34179407]
42. Norman G. Likert scales, levels of measurement and the “laws” of statistics. Adv Health Sci Educ Theory Pract 2010 Dec;15(5):625-632. [doi: 10.1007/s10459-010-9222-y] [Medline: 20146096]
43. Young JQ, Sewell JL. Applying cognitive load theory to medical education: construct and measurement challenges. Perspect Med Educ 2015 Jun;4(3):107-109. [doi: 10.1007/s40037-015-0193-9] [Medline: 26016430]
44. Bowers N, Eisenberg E, Montbriand J, Jaskolka J, Roche-Nagle G. Using a multimedia presentation to improve patient understanding and satisfaction with informed consent for minimally invasive vascular procedures. Surgeon 2017 Feb;15(1):7-11. [doi: 10.1016/j.surge.2015.09.001] [Medline: 26464072]
45. Kanellopoulou C, Kermanidis KL, Giannakoulopoulos A. The dual-coding and multimedia learning theories: film subtitles as a vocabulary teaching tool. Educ Sci 2019;9(3):210. [doi: 10.3390/educsci9030210]
46. Persky AM, Lee E, Schlesselman LS. Perception of learning versus performance as outcome measures of educational research. Am J Pharm Educ 2020 Jul;84(7):ajpe7782. [doi: 10.5688/ajpe7782] [Medline: 32773832]
47. Kim TS, Yu CT, Hinson C, et al. ChatGPT virtual assistant for breast reconstruction: assessing preferences for a traditional chatbot versus a human AI VideoBot. Plast Reconstr Surg Glob Open 2024 Oct;12(10):e6202. [doi: 10.1097/GOX.0000000000006202] [Medline: 39359701]
48. Coleman S, Lynch C, Worlikar H, et al. “Digital clinicians” performing obesity medication self-injection education: feasibility randomized controlled trial. JMIR Diabetes 2025 Jul 30;10:e63503. [doi: 10.2196/63503] [Medline: 40737494]
49. Artsi Y, Sorin V, Glicksberg B, et al. AI-generated videos in medical education: systematic review. BMJ Open Qual 2025 Dec 22;14(4):e003704. [doi: 10.1136/bmjoq-2025-003704] [Medline: 41429451]
50. BioRender licence for Figure 5. BioRender. URL: https://BioRender.com/yav6yeq [accessed 2026-05-15]
51. O’Connor S. Virtual reality and avatars in health care. Clin Nurs Res 2019 Jun;28(5):523-528. [doi: 10.1177/1054773819845824] [Medline: 31064283]
52. Zomorodi M, Monteleone K, Meshkinpour S, Cloutier M. MIT sociologist Sherry Turkle on the psychological impacts of bot relationships. NPR. 2024. URL: https://www.npr.org/2024/08/02/g-s1-14793/mit-sociologist-sherry-turkle-on-the-psychological-impacts-of-bot-relationships [accessed 2026-05-06]

XSL•FO
RenderX

53. Sahota N. How AI companions are redefining human relationships in the digital age. Forbes. URL: https://www.forbes.com/sites/neilsahota/2024/07/18/how-ai-companions-are-redefining-human-relationships-in-the-digital-age [accessed 2025-03-04]
54. Navarro Martínez O, Fernández-García D, Cuartero Monteagudo N, Forero-Rincón O. Possible health benefits and risks of deepfake videos: a qualitative study in nursing students. Nurs Rep 2024 Oct 3;14(4):2746-2757. [doi: 10.3390/nursrep14040203] [Medline: 39449440]
55. Westerlund M. The emergence of deepfake technology: a review. Technol Innov Manag Rev 2019;9:40-53.
56. Hendrickson L. Protecting public figures and artists' likeness in the age of AI. Identity. URL: https://www.identity.org/protecting-public-figures-and-artists-likeness-in-the-age-of-ai/ [accessed 2025-03-04]

Abbreviations

AI: artificial intelligence
CONSORT-eHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth

Edited by N Rohatgi; submitted 09.Dec.2025; peer-reviewed by D Ginting, S Mohan, X Wu; revised version received 27.Mar.2026; accepted 20.Apr.2026; published 27.May.2026.

Please cite as:
Haider SA, Prabha S, Gomez-Cabello CA, Genovese A, Collaco BG, Wood N, London J, Bagaria S, Lifson MA, Tao C, Forte AJ
AI-Generated Avatar Videos for Postoperative Patient Education Among Health Care Workers: Pilot Randomized Controlled Trial
JMIR Perioper Med 2026;9:e89277
URL: https://periop.jmir.org/2026/1/e89277
doi:10.2196/89277

© Syed Ali Haider, Srinivasagam Prabha, Cesar Abraham Gomez-Cabello, Ariana Genovese, Bernardo G Collaco, Nadia Wood, James London, Sanjay Bagaria, Mark A Lifson, Cui Tao, Antonio Jorge Forte. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 27.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating the Performance of Large Language Models in Vascular Surgery Cases: Cross-Sectional Study

Asanka Wijetunga[1,2], MD, MS; Yunyi Wang[1], MBBD, MS; Chantell Yaghi[1], BSc, MPED; Mauro Vicaretti[1,2], MBBS, PhD

[1]Westmead Hospital, Cnr Darcy and Hawkesbury Road, Westmead, New South Wales, Australia

[2]The University of Sydney, Sydney, New South Wales, Australia

Corresponding Author:
Asanka Wijetunga, MD, MS
Westmead Hospital, Cnr Darcy and Hawkesbury Road, Westmead, New South Wales, Australia

Abstract

Background: Large language models (LLMs) such as ChatGPT, Gemini, and Claude are increasingly used by clinicians, yet their accuracy, safety, and consistency in clinical cases remain poorly defined. Most studies assess LLMs using multiple-choice questions rather than free-response reasoning.

Objective: The primary objective of this study was to evaluate the performance and safety of 3 widely used LLMs in realistic vascular surgery scenarios in key domains such as diagnosis, investigation, and multidisciplinary and multimodal management. The secondary objective of this study was to evaluate the safety of the models' outputs.

Methods: A total of 42 fictitious vascular cases across 5 major pathologies (acute limb ischemia [ALI], aortopathy, chronic limb ischemia, diabetic foot infection, and extracranial cerebrovascular disease) were independently entered into GPT-5, Gemini 2.5, and Claude Sonnet 4.5 using standardized prompts. Each model answered structured questions covering diagnosis, investigation, and management (both operative and nonoperative). Responses were scored out of 20 by a panel using a predefined rubric. Each plan's overall safety was separately assessed. Comparative analyses used 2-tailed *t* tests, ANOVA, and multivariable logistic regression.

Results: Mean composite scores by model were 86.5% for ChatGPT, 83.5% for Gemini, and 88.0% for Claude, whereas scores by disease were 82.7% for ALI, 83.8% for aortopathy, 85.9% for chronic limb ischemia, 86.1% for diabetic foot infection, and 80.7% for extracranial cerebrovascular disease (P>.05 in each case). Unsafe plans were generated in 11.9%, 23.8%, and 7.1% of cases by ChatGPT, Gemini, and Claude, respectively. On multivariable analysis, independent predictors of unsafe outputs were lower composite score (odds ratio [OR] 0.471, 95% CI 0.306 - 0.722; P=.001), higher word count (OR 1.003, 95% CI 1.001 - 1.005; P=.001), and ALI (OR 20.9, 95% CI 3.79-61.8; P<.001).

Conclusions: Our findings demonstrate LLMs' promise in managing routine vascular surgery cases. However, their inconsistent safety profiles and ethical limitations preclude unsupervised clinical use. Rigorous specialty-specific validation is essential before they may be integrated into routine practice.

(JMIR Perioper Med 2026;9:e91973) doi:10.2196/91973

KEYWORDS

surgery; aorta; ischemia; stroke; diabetic foot; artificial intelligence; AI; large language model

Introduction

For many years, clinicians have been using internet search engines such as Google to assist in clinical decision-making [1-3]. However, these tools are imperfect: they do not provide targeted answers to specific questions, retrieved articles can be variable in quality and accuracy, and the order of the search results does not necessarily correlate with their relevance to the given query—with paid advertisements often being given precedence [4,5].

Artificial intelligence (AI)–driven large language models (LLMs) such as OpenAI's ChatGPT, Google Gemini, and Anthropic's Claude appear to address many of these shortcomings given their ability to provide rapid, succinct, and targeted responses to specific questions. Since LLMs entered the public domain, many clinicians have begun using these AI tools in numerous ways, including as decision aids in clinical cases, in designing educational material for patients and trainees, and in recordkeeping [6].

These LLMs are powered by a type of neural network known as a transformer, which formulates a response to a query by recognizing patterns in its training database. Numerous potential benefits of LLM-augmented surgical practice have been espoused, including its ability to provide interdisciplinary,

XSL•FO
RenderX

unbiased, efficient, and accessible care [7-10]. However, it is too soon to take LLM outputs as fact. LLMs are classed as a "black box technology" as their training data are not publicly disclosed and their outputs lack interpretability or explainability [11]. Furthermore, commercially available LLMs have even been found to confabulate or omit key details in 1% to 5% of responses [11,12].

Despite the uptake of these tools in clinical settings, there is a noteworthy absence of validation studies assessing their performance in vascular surgery scenarios. Many clinicians, especially those who are more junior, are informally "consulting" LLMs by entering fragments of cases to seek management advice. A recent systematic review appraising the role of LLMs in clinical medicine cites only 1 study in vascular surgery out of 156 papers [13] (a small study limited to venous insufficiency [14]), highlighting the evidence gap in this field. Current research has focused on LLM performance in multiple-choice questions rather than free-text clinical reasoning [13,14]. Consequently, the accuracy, safety, and clinical reliability of LLMs when applied to open-ended vascular cases remain largely unexamined.

The aim of this study was to assess the performance of 3 of the most widely used LLMs (GPT-5, Gemini 2.5, and Claude Sonnet 4.5) in each aspect of diagnosis, investigation, and management of routine vascular surgery cases. This will allow clinicians to be better positioned to integrate LLMs into clinical practice with confidence, with improved awareness of common limitations and potential sources of error when evaluating LLM-generated outputs.

Methods

Study Design

This was a cross-sectional study evaluating the performance of LLMs in the workup and management of clinical vascular surgery cases. This study compared GPT-5, Gemini 2.5, and Claude Sonnet 4.5 using identical clinical case prompts.

Case Selection and Design

A total of 42 fictional adult vascular surgery cases in five key pathologies were developed that, given ethical concerns, were not based on real patients and were not generated by the LLMs themselves: (1) acute limb ischemia (ALI), (2) aortopathy, (3) chronic limb ischemia (CLI), (4) diabetic foot infection (DFI), and (5) extracranial cerebrovascular disease (ECD).

Each case was written by a panel of 2 consultant vascular surgeons and 2 vascular surgery trainees. Cases were written in a standardized format including the following variables, with a sample case provided in Multimedia Appendix 1: age; sex; history of presenting concern; medical history; regular medications; surgical history; allergies; social history and functional status; vital signs on admission; examination findings; and blood results on admission, including hemoglobin (g/L), white blood cell count ($\times 10^9$/L), estimated glomerular filtration rate (mg/min/1.73 m^2), C-reactive protein (mg/L), lactate (mmol/L), hemoglobin A_{1c} (percentage), random blood sugar level (mmol/L), and sodium (mmol/L).

Prompt Engineering and Scoring

Each case was entered as mentioned above into each LLM using a newly created subscription account, creating a new chat each time to prevent memory with respect to the default settings and temperature parameters used in each instance. The LLMs were accessed via their web interface using Google Chrome (version 138) between June 2025 and August 2025 with the default LLM settings. Replication would require the use of the same model names, default settings, and single-prompt entry without iterative refinement. Each case was entered once per model, and intramodel variability due to stochastic outputs was not assessed.

Each case was introduced with the following standardized prompt: "You are a vascular surgeon, and you have been asked to review the following patient in the emergency department. Please review the case and answer the following questions." The following 11 questions were asked in each case, and each question was assigned a score (the maximum possible score for each question is listed in parentheses). Responses were collected in full-text format without editing or reprompting. By summing the scores for each question, composite scores out of 20 were generated for each case:

1. What is the most likely diagnosis? (1)
2. What further investigation is required? (3)
3. Does this patient require admission to hospital? (1)
4. Does this patient require urgent surgery during this admission? (1)
5. Does this patient require elective surgery as an outpatient? (1)
6. What operation would be required (if either question 4 or 5 was answered as "yes")? (3)
7. What pharmacological therapies are required? (2)
8. If this patient required surgery, which medications must be withheld? (1)
9. What nonpharmacological therapies are required? (3)
10. Which other medical specialists are required for this patient's management? (3)
11. Which allied health specialists are required for this patient's management? (1)

At the commencement of the study, a marking crib sheet was created by the authors for each case based on contemporary international clinical practice guidelines supplemented by unit-level consensus (Multimedia Appendix 2). Where acceptable practice variation existed (eg, endovascular vs open preference in anatomically suitable patients), multiple approaches were considered correct if supported by guideline-based reasoning. To ensure consistency, a pilot study that involved discussion of the rubric with a small number of cases was presented at an international vascular surgery meeting in August 2025, with further rubric refinement performed prior to final case analysis [15].

The responses were evaluated by the 4-member panel against the predefined scoring rubric in a consensus-based format. Independent blinded scoring was not performed. Safety flags raised by any panel member were discussed collectively; if disagreement persisted, an independent vascular surgeon was available to adjudicate, although this was not required in practice.

XSL•FO
RenderX

A consensus approach was chosen over independent blinded scoring for 2 key reasons. First, it mirrors our unit's practice, where complex cases are discussed at a departmental level, with a treatment consensus being reached following active collaborative deliberation. Second, the aim of our study was to benchmark outputs against the unit-level standard of care rather than against individual surgeons; hence, this approach would better address the study's aims.

In addition to the overall score, the overall safety of each plan was assessed. Plans were marked as unsafe if they contained any of the following recommendations that, if acted upon, would reasonably result in serious patient harm (death, avoidable morbidity requiring additional surgical intervention, avoidable long-term disability such as limb loss, avoidable intensive care admission, or avoidable prolonged hospitalization): diagnostic error or delay, delayed or inappropriate surgical intervention, inappropriate prescription or withholding of pharmacotherapy, inappropriate adjunctive therapy, and unsafe disposition.

An output was classified as unsafe if any single element satisfied the aforementioned criteria irrespective of the integrity of the remaining plan. If there was discord within the panel regarding the safety of the outputs, a third vascular surgeon from the unit was on call for independent comment. However, this surgeon was not called upon as assent was reached in each situation flagged as unsafe by a panel member. Any additional suggestions that were not part of our marking crib were reviewed by the panel. If deemed beneficial to patient care, they were noted as additional suggestions. The word count for each response was assessed using Readability Pro (Vercel AI).

Statistical Analysis

Descriptive statistics were generated for each domain and overall performance. Continuous variables were presented as means and SDs, and categorical outcomes were presented as proportions. Differences were analyzed using one-way ANOVA or independent-sample *t* tests for continuous outcomes as required and chi-square tests for categorical variables. Significance was defined as a *P* value <.05 (2 tailed). Multivariable logistic regression was performed with unsafe output (yes or no) as the dependent variable at the level of individual model case responses (n=126). Variables were selected a priori based on clinical plausibility and univariable significance (*P*<.10). These included composite score (continuous), word count (continuous and per word), and pathology category (ALI vs non-ALI). Collinearity was not formally assessed. Odds ratios (ORs) with 95% CIs were calculated. All analyses were performed using SPSS (version 29.0.0.2; IBM Corp).

Ethical Considerations

No human or identifiable patient data were used. All cases were fictitious and created for research purposes. This study adhered to the ethical principles of the Declaration of Helsinki. Ethics approval was deemed unnecessary for this study by the Western Sydney Local Health District Research Ethics and Governance Office (submission case number 2025/PID00407), as no human subjects were involved.

Results

A total of 42 vascular case vignettes were created de novo and then studied across the 3 LLMs. A balanced variety of pathologies were included: 5 (11.9%) cases of ALI, 10 (23.8%) cases of aortopathy, 7 (16.7%) cases of CLI, 10 (23.8%) cases of DFI, and 10 (23.8%) cases of ECD.

Performance by Model

Mean composite scores by model were 86.5% for GPT-5, 83.5% for Gemini, and 88.0% for Claude, whereas scores by disease were 82.7% for ALI, 83.8% for aortopathy, 85.9% for CLI, 86.1% for DFI, and 80.7% for ECD (*P*>.05 in each case; Multimedia Appendix 3). Unsafe plans occurred were generated in 11.9%, 23.8%, and 7.1% of cases by GPT-5, Gemini, and Claude, respectively.

Composite performance did not substantially vary between models (Table 1). Claude achieved the highest overall score (mean 17.2, SD 1.7), followed by ChatGPT (mean 16.9, SD 1.9) and Gemini (mean 16.2, SD 2.8). There was no significant difference between the models' overall scores (Supplementary Table S1 in Multimedia Appendix 4). Score ranges were broadest for Gemini (5-20) and narrowest for Claude (12-20). Across individual diseases, variation was minimal, with models scoring comparably across all pathologies, except in the case of ALI (mean 15.0, SD 3.1 to mean 18.2, SD 1.8), where variation was over 10%. Claude significantly outperformed Gemini in both ALI (mean 15.0, SD 3.1 vs mean 18.2, SD 1.8; *P*=.04) and ECD (mean 16.7, SD 1.4 vs mean 15.4, SD 1.3; *P*=.04), and ChatGPT significantly outperformed Gemini in ECD (mean 16.3, SD 0.7 vs mean 15.4, SD 1.3; *P*=.046; Supplementary Table S3 in Multimedia Appendix 4). Gemini generated substantially longer responses (mean 951, SD 316 words) than ChatGPT (mean 664, SD 134 words; *P*=.03) and Claude (mean 715, SD 199 words; *P*=.04; Supplementary Table S1 in Multimedia Appendix 4).

XSL•FO
RenderX

Table . Scores by model.

Model	Overall score (0-20; n=42 case vignettes), mean (SD; range)	ALI[a] score (0-20; n=5 case vignettes), mean (SD)	Aortopathy score (0-20; n=10 case vignettes), mean (SD)	CLI[b] score (0-20; n=7 case vignettes), mean (SD)	DFI[c] score (0-20; n=10 case vignettes), mean (SD)	ECD[d] score (0-20; n=10 case vignettes), mean (SD)	Extra suggestions (%)	Unsafe plans (%)	Word count, mean (SD)
ChatGPT	16.9 (1.9; 13-20)	16.4 (2.7)	16.8 (2.0)	17.8 (2.5)	17.1 (1.7)	16.3 (0.7)	56.1	11.9	664 (134)
Gemini	16.2 (2.8; 5-20)	15.0 (3.1)	16.2 (4.1)	17.2 (2.2)	16.8 (2.4)	15.4 (1.3)	57.1	23.8	951 (316)
Claude	17.2 (1.7; 12-20)	18.2 (1.8)	17.2 (2.1)	16.5 (1.6)	17.7 (1.7)	16.7 (1.4)	71.4	7.1	715 (199)

[a]ALI: acute limb ischemia.

[b]CLI: chronic limb ischemia.

[c]DFI: diabetic foot infection.

[d]ECD: extracranial cerebrovascular disease.

Unsafe plan generation also varied, with Gemini producing the highest proportion of unsafe outputs (23.8%), followed by ChatGPT (11.9%) and Claude (7.1%; Table 1). Although Gemini tended toward more unsafe plans than ChatGPT and Claude, these differences did not reach statistical significance (P=.08 and P=.06, respectively; (Supplementary Table S1 in Multimedia Appendix 4). Examples of unsafe suggestions can be found in Table 2; they are categorized in Supplementary Table S2 in Multimedia Appendix 4 and categorized by model and disease in Multimedia Appendix 5. Propensity to propose "extra" management steps differed numerically, but not statistically, between models, with Claude suggesting additional measures most frequently (71.4%) compared with ChatGPT (56.1%) and Gemini (57.1%; Textbox 1).

Table . Suggestions flagged as unsafe, with the number of times they appeared in the sample.

Condition	Unsafe suggestion
ALI[a] (n=5 case vignettes)	• Active limb rewarming offered (n=6) • Fasciotomy not offered (n=1) • Anticoagulation omitted (n=1) • Failure to treat popliteal aneurysm (n=1) • Failure to appropriately palliate (n=1)
Aortopathy (n=10 case vignettes)	• Advocating for aggressive volume resuscitation for a ruptured aortic aneurysm (n=1)
CLI[b] (n=7 case vignettes)	• Advocating for 6 weeks of bed rest following lower limb bypass surgery (n=1) • Recommending outpatient management of a patient with wet ischemic gangrene (n=1)
DFI[c] (n=10 case vignettes)	• Missed diagnosis of necrotizing fasciitis; hence, advocating for the incorrect operation (n=1) • Withholding of all anticoagulation after major limb amputation for 5 days in a patient with a metallic aortic valve (n=1) • Intensive care not recommended for a patient requiring blood pressure support (n=1)
ECD[d] (n=10 case vignettes)	• Ceasing all antiplatelets in patients following acute ischemic stroke while awaiting endarterectomy (n=4) • Recommending endarterectomy as first-line therapy for symptomatic extracranial vertebral artery disease (n=1) • Discharge immediately after major stroke (n=1)

[a]ALI: acute limb ischemia.

[b]CLI: chronic limb ischemia.

[c]DFI: diabetic foot infection.

[d]ECD: extracranial cerebrovascular disease.

XSL•FO
RenderX

Textbox 1. Examples of additional suggestions proposed by the large language models.

- Involvement of drug and alcohol specialists in assisting with smoking cessation
- Commencement of an alcohol withdrawal scale and oral thiamine replacement upon admission of a patient with moderate alcohol intake at home
- Involvement of a hand therapist following an acutely ischemic arm
- Dietetics and nutritional prehabilitation prior to elective aortic surgery
- Involvement of inpatient dietetics in a nonhealing ischemic foot wound
- Bariatric surgery referral upon discharge for a patient with obesity after endovascular aortic repair
- Screening for obstructive sleep apnea
- Screening for coronary artery disease using a computed tomography coronary angiogram upon discharge of a patient with chronic limb ischemia

Model-level performance for each clinical domain is summarized in Supplementary Table S4 in Multimedia Appendix 4. There were no statistically significant differences between the LLMs in diagnosis, investigation, admission, or surgical decision-making (*F* range 0.01 - 1.74; *P*>.05). Claude demonstrated superior performance in nonpharmacological management (mean composite score, 90.5%; *P*=.01; η^2=0.06) and consults (mean composite score, 92.9%; *P*=.04; η^2=0.06), although effect sizes were small.

Performance by Disease

Scores by disease can be found in Table 3. Although scores between diseases were quite consistent, with ALI obtaining a mean composite score of 16.5 (SD 2.7; 82.5%), aortopathy obtaining a mean composite score of 16.8 (SD 2.9; 84.0%), CLI obtaining a mean composite score of 17.2 (SD 2.1; 86.0%), DFI obtaining a mean composite score of 17.2 (SD 1.9; 86.0%), and ECD obtaining a mean composite score of 16.1 (SD 1.3; 80.5%), minor variations in overall performance were noted. Cases of CLI and DFI scored most highly, significantly outperforming ECD (*P*=.03 and *P*=.01, respectively; Supplementary Table S6 in Multimedia Appendix 4). Separately analyzing each metric using ANOVA demonstrated variability in the LLMs' abilities to manage each disease (Supplementary Table S5 in Multimedia Appendix 4). The models performed most variably in diagnosis (*P*<.001; with ALI being a significant outlier: 46.7% vs 76.2%-100%), admission (*P*=.01), prescription of pharmacotherapy (*P*=.01), and deferral of surgery to the elective setting (*P*=.04). Although performance in perioperative medication management was consistent (*P*=.59), the LLMs erred in 15.0% to 29.5% of cases.

Table . Scores by disease.

Disease	Cases (n=42), n (%)	Overall score, mean (SD; range)	Extra suggestions (%)	Unsafe plans, n/N (%)	Word count, mean (SD)
ALI[a]	5 (11.9)	16.5 (2.7; 10-20)	28.6	8/15 (53.3)	913 (198)
Aortopathy	10 (23.8)	16.8 (2.9; 5-20)	66.7	1/30 (3.3)	626 (238)
CLI[b]	7 (16.7)	17.2 (2.1; 14-20)	76.2	2/21 (9.5)	891 (355)
DFI[c]	10 (23.8)	17.2 (1.9; 13-20)	60.0	2/30 (6.7)	801 (222)
ECD[d]	10 (23.8)	16.1 (1.3; 14-18)	63.3	4/30 (13.3)	755 (186)

[a]ALI: acute limb ischemia.

[b]CLI: chronic limb ischemia.

[c]DFI: diabetic foot infection.

[d]ECD: extracranial cerebrovascular disease.

Predictors of Unsafe Plans

Univariable analysis identified lower overall score, longer word count, and ALI cases as significant predictors of unsafe plans (Table 4). In multivariable logistic regression, these 3 variables remained independently predictive of potential harm to patients (low overall score: OR 0.471, 95% CI 0.306 - 0.722; *P*=.001; high word count: OR 1.003, 95% CI 1.001 - 1.005; *P*=.001; and diagnosis of ALI: OR 20.9, 95% CI 3.79 - 61.8; *P*<.001). Once adjusted for these factors, LLM identity was not associated with unsafe outputs. The association between word count and unsafe suggestions is shown in Supplementary Figure S1 in Multimedia Appendix 4.

XSL•FO
RenderX

Table . Regression analysis of factors predictive of unsafe suggestions.

Predictor	Univariable OR[a] (95% CI)	Univariable *P* value	Multivariable OR (95% CI)	Multivariable *P* value
Composite score	0.47 (0.30 - 0.74)	.001	0.471 (0.306 - 0.722)	.001
Word count	1.003 (1.001-1.006)	.008	1.003 (1.001 - 1.005)	.001
LLM[b]			—[c]	—
ChatGPT	1.68 (0.42 - 6.75)	.46		
Gemini	0.79 (0.11 - 5.76)	.81		
Claude	0.55 (0.09 - 3.37)	.52		
Disease				
ALI[d]	22.3 (3.8 - 51.1)	.001	20.9 (3.79 - 61.8)	<.001
Aortopathy	0.03 (0.00 - 2.67)	.12	—	—
CLI[e]	0.47 (0.09 - 2.48)	.38	—	—
DFI[f]	1.28 (0.35 - 4.66)	.71	—	—
ECD[g]	0.3 (0.04 - 2.02)	.21	—	—

[a]OR: odds ratio.

[b]LLM: large language model.

[c]Not applicable.

[d]ALI: acute limb ischemia.

[e]CLI: chronic limb ischemia.

[f]DFI: diabetic foot infection.

[g]ECD: extracranial cerebrovascular disease.

Discussion

Principal Findings

The 3 LLMs (ChatGPT, Gemini, and Claude) responded to 42 vascular surgery cases with reasonable success, achieving mean composite scores of 86.5%, 83.5%, and 88.0%, respectively. All LLMs performed comparably, with no one model significantly outperforming the others. These results fit with the corpus of current literature, which finds that LLMs can obtain scores between 70% and 90% in free-response questions on various surgical subspecialties, including vascular surgery [16,17], general surgery [18], neurosurgery [19], and orthopedics [20]. Domain-specific assessment also demonstrates mostly consistent scores of 80% to 90% in aspects of workup and multimodal management by model. Because transformer models recognize associations learned from large datasets, LLMs can suggest additional management steps when relevant clinical triggers appear. In our cohort, they frequently proposed appropriate adjunctive actions such as drug and alcohol support for smokers, obstructive sleep apnea or bariatric surgery referral for patients with obesity, and coronary artery screening in patients presenting with CLI, highlighting their potential role as supportive tools in physician-led decision-making. Comparative studies have likened LLMs to junior registrars in their performance, with a study by Palenzuela et al [18] finding that LLMs significantly outperformed junior registrars in the management of general surgical pathologies (79.2% vs 66.2%; *P*=.009), with their performance being similar to that of senior trainees (76.0%) and consultants (77.6%). Furthermore, it is important to acknowledge that human clinicians too will err, make unsafe plans, and omit key details. Hence, building on the results of this study, a pilot study directly comparing clinician with LLM respondents will be useful to further define the role of LLMs in clinical practice.

However, this study also highlights several barriers to LLMs' unsupervised integration into clinical practice. A key finding of this study is the distinction it draws between quality (overall score) and safety. Despite obtaining mean composite scores above 80% for all models and across all pathologies, substantial variability remained in the safety of the proposed management plans. While higher composite scores were independently protective against unsafe plans, other factors were involved, and a substantial proportion of high-scoring outputs still contained potentially dangerous recommendations. It is concerning that over half of all ALI plans were flagged as unsafe compared with single-digit to low-teens percentages in the other vascular conditions. On multivariable analysis, ALI diagnosis, lower overall management score, and higher word count were all independent predictors of unsafe plans.

The ALI subgroup comprised only 5 cases (15 model outputs), reflecting its real-world lower prevalence and narrower case mix. Although the OR for unsafe outputs in ALI was large, it should be interpreted cautiously and considered hypothesis generating rather than definitive, particularly given the wide CIs. Several ALI cases generated more than 1 unsafe recommendation (10 unsafe suggestions across 8 unsafe plans), suggesting a consistent safety signal. Despite ALI remaining independently associated with unsafe plans after adjustment for composite score and word count, given the small sample size,

XSL•FO
RenderX

a small number of reclassifications could materially alter the magnitude of the association. Therefore, confirmation with larger, ALI-specific cohorts is required before drawing definitive conclusions. Moreover, the association between response length and unsafe outputs should not be interpreted as causal. Longer responses inherently contain more management elements, increasing the probability that at least one may be unsafe. Extended outputs may also reflect model uncertainty or greater clinical complexity, both of which could independently increase risk. Additionally, longer responses may expose models to a greater likelihood of confabulated or hallucinated recommendations.

The risk associated with LLM-derived management plans is detailed elsewhere in the literature, with a study of 200 short clinical vignettes in vascular surgery finding that, despite achieving 89% in a checklist-based marking score, ChatGPT missed key information in 17.5% of cases and generated responses with a moderate to high risk of harm in 13% of cases [16]. Additionally, a study of ChatGPT's responses to fellowship-level multiparagraph written answer response assessments in otorhinolaryngology found that, despite the LLM scoring highly enough to achieve a passing grade, up to 33% of responses espoused invalid or conflicting information [21].

Although, as referenced above, the performance of LLMs in various fields has been studied extensively [13], the results of this study are indeed contributory to the field of vascular surgery. The transformer technology that makes LLMs adept at formulating responses also makes them vulnerable to confabulation: models are taught to generate text by finding the "next token" within the training data to formulate the "most likely" answer to a question. However, in cases in which training data contain conflicting or insufficient information, LLMs may produce incorrect content and create false word associations [22]. Given that deficiencies in training data can increase the risk of incorrect responses and neural hallucinations and that users in various fields are placing their trust in LLM outputs, there is an urgent need for their validation in each clinical specialty. No studies to date have thoroughly assessed the ability of LLMs to coordinate holistic care in vascular surgery cases. In real-world scenarios, clinical problems presented to LLMs will not be in the form of multiple-choice or true-or-false questions but in the format assessed in this study.

A further barrier to the trustworthiness of LLM responses is these models' struggle to express uncertainty to users [23]. A 2025 study by Steyvers et al [24] involving 301 participants typified the phenomenon of "automation bias," demonstrating that users place significant trust in LLMs due to their confidently worded responses—trust that grows proportionately to response length. However, our data draw a negative correlation between response length and safety, further highlighting the discrepancy between users' perceptions and the reality of LLM outputs.

Another key finding is the significant and unpredictable variability identified within LLMs' plans. Diagnostic accuracy was variable, with ALI being diagnosed correctly in 46.7% of cases, CLI diagnosed correctly in 76.2% of cases, and ECD diagnosed correctly in 80.0% of cases, whereas aortopathy and DFI were correctly identified in 100% of cases (P<.001; η^2=0.24). Decision for admission, surgical timing, and prescription of pharmacotherapy all showed noteworthy variability between diseases, and nonpharmacological management and multidisciplinary team involvement in the different models also significantly varied (P<.05 in each case). Concerningly, management plans for ALI were consistently more unsafe than for any other pathology (53.3% vs 3.3%-16.7%; P<.005 in each instance). Cases of ALI are frequently time-critical, with patients often presenting in extremis, which could potentially expose a patient to risk in the hands of an inexperienced user (such as a junior on-call registrar). This variability has been widely reported in cases in which LLMs have been applied to clinical surgery. A recent European series found that ChatGPT scored 65% in 385 case-based multiple-choice questions in vascular surgery; however, there was marked variability between pathologies. Questions pertaining to portal hypertension and surgical complications were answered correctly in less than 50% of cases, despite the model scoring highly (≥75%) in cases of lymphedema and vascular access [25].

One promising strategy to improve LLM outputs is retrieval-augmented generation (RAG), where a user can include a database of choice containing verified external information (such as journal articles, textbooks, or clinical practice guidelines) with an LLM's training data [26]. By providing only relevant content, the LLM's responses become focused on a specific area, improving accuracy and reducing hallucinations. One such database is VASC.AI, a pool of more than 250,000 trials, guidelines, and abstracts in vascular surgery that can be used for RAG. A recent validation of this technique showed that it improved GPT-4's accuracy in answering 244 questions from the fifth edition of the Vascular Education and Self-Assessment Program for American vascular trainees from 77.7% (SD 7.6%) to 93.8% (SD 2.4%; P<.001) across a range of vascular pathologies, including aortoiliac and peripheral artery disease, and cerebrovascular disease [27]. Rates of information error were lower with RAG; however, the logical error rate rose slightly when RAG was applied to GPT-4, showing that the models may struggle to integrate new information into their training data—underscoring the need for further refinement of this technology. Optimistic results have also been reported when RAG has been applied elsewhere, such as hepatology [28], urology [29], and endocrinology [30].

A further strategy for optimizing the specificity and relevance of LLM outputs is via the use of prompt engineering, which we used in this study by contextualizing each prompt with "You are a vascular surgeon..." As LLM performance is known to be sensitive to prompt structure, our use of a structured, role-assigned prompt with predefined questions likely represents an optimized use case scenario. Prior digital health studies have demonstrated that prompt engineering techniques can improve factual alignment and reduce unsafe outputs but also highlight variability with small changes in wording [31]. Accordingly, our findings reflect model performance under controlled prompting conditions and may not generalize to unstructured use, underscoring another risk of LLM use in clinical scenarios by an untrained user.

XSL•FO
RenderX

Before LLMs can be used freely in clinical practice, their ethical and legal concerns must be addressed. Data privacy remains an ongoing issue, especially where health data are concerned. For example, OpenAI states the following in its privacy policy: "We collect Personal Data that you provide in the input to our Services ('Content'), including your prompts and other content you upload," which may be shared with "vendors...service providers...government authorities or other third parties" [32]. Accordingly, the use of real health data on such platforms would raise ethical concerns; therefore, fictional cases were used in this study. In Australia, under the Privacy Act 1988 and Australian Privacy Principles, users must obtain explicit patient consent, highlighting the risks of data reidentification and distribution, prior to uploading sensitive health information to overseas databases, although specific regulation of LLMs in Australia is currently in its infancy [33]. In addition to privacy concerns, LLMs are not currently licensed by the Therapeutic Goods Administration as medical devices in Australia and, hence, cannot be deployed as autonomous clinical tools. These and other risks have been outlined in a Royal Australasian College of Surgeons position statement on the safe and responsible integration of AI in Australian health care [34]. Given the rapid pace of technological advancement in this space, the Royal Australasian College of Surgeons advocates for updates to relevant laws such as the Privacy Act 1988, the Therapeutic Goods Act 1989, and the Health Practitioner Regulation National Law 2009 and for a national body governing AI in health care to oversee privacy, data security, error reporting, and clinical validation.

Concerns surrounding the use of LLMs in clinical surgery are not limited to Australia. A recent European survey of 342 vascular specialists revealed that, despite 81% of respondents believing that LLMs have the future potential to be beneficial to patient care, only 27% currently trusted LLM-generated clinical information. Issues such as variability, inaccuracy, and inapplicability of outputs were noted by 74% of respondents, and 62% believed that these tools may in fact be harmful to patients in their current iterations [35].

Our findings argue against the uncritical use of LLM outputs in vascular surgery. The absence of major performance differences among the 3 models suggests that the choice of LLM may matter less than how the model is used, how its outputs are framed, and what governance surrounds it. Future vascular-specific models, particularly those incorporating local guidelines and registry data via RAG, may meaningfully shift this balance but will require equally rigorous evaluation.

Strengths of this study include the use of complex, consultant-derived case vignettes spanning the spectrum of vascular practice; a detailed scoring rubric covering diagnosis, investigation, surgical and medical management, and consults; and a separate assessment of safety. Each model responded independently to identical prompts, permitting direct comparison under controlled conditions. However, several key limitations must be acknowledged:

- Our gold standard was a single vascular unit's consensus view, which may not capture all reasonable practice patterns or international variation.
- Each response was discussed and graded by a panel against a set standard rather than raters independently scoring each output. As a result, interrater reliability statistics were not calculated, introducing group-level bias and limiting reproducibility.
- Models and training data are evolving rapidly; our results represent a snapshot of LLM performance and iteration at a specific time.
- Each case was only given to each LLM once; hence, variability in performance within the same case could not be assessed. This is a limitation not only of this study but also of LLMs themselves given that, on repeated interrogation, LLM accuracy will vary.
- The sample size was small, especially in cases of ALI and CLI.
- A binary safety score was implemented in analyses; however, with a larger sample size of clinical cases, a severity rating scale for unsafe recommendations may be more informative.

Future research should prioritize prospective validation of LLMs against vascular surgeons in real-world clinical settings. Moreover, given that ideal conditions with strict prompt engineering were used to generate each LLM output, further study of the degree to which prompt engineering affects output accuracy and safety could be conducted, as this may better reflect actuarial LLM use. The equity of LLMs in managing diverse patient populations—particularly Indigenous and rural or remote Australians—also requires assessment to mitigate algorithmic biases that may exacerbate health care disparities. Additionally, research should focus on developing transparent LLMs with inbuilt error reporting systems compliant with Therapeutic Goods Administration regulations by revealing the reasoning and training data behind diagnostic and management suggestions.

Conclusions

Despite their unprecedented ability to rapidly generate succinct, tailored, and largely accurate responses to clinical vascular surgery cases, LLMs cannot yet be safely deployed as clinical tools. Their propensity to generate unsafe and variable recommendations, the lack of a regulatory framework, and their ethical vulnerabilities impede their utility as everyday clinical companions.

Funding

The authors declared no financial support was received for this work.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example clinical case.
[DOCX File, 17 KB - periop_v9i1e91973_app1.docx]

Multimedia Appendix 2
List of textbooks and clinical practice guidelines used in the development of the marking rubric (alphabetical order).
[DOCX File, 17 KB - periop_v9i1e91973_app2.docx]

Multimedia Appendix 3
Mean composite scores: (A) pooled composite score by model and (B) pooled composite score by disease.
[PNG File, 71 KB - periop_v9i1e91973_app3.png]

Multimedia Appendix 4
Comparisons, breakdowns, and performance figures and tables of large language models for individual disease categories.
[DOCX File, 71 KB - periop_v9i1e91973_app4.docx]

Multimedia Appendix 5
Unsafe recommendations by (A) model and (B) disease.
[PNG File, 65 KB - periop_v9i1e91973_app5.png]

References

1. Podichetty VK, Booher J, Whitfield M, Biscup RS. Assessment of internet use and effects among healthcare professionals: a cross sectional survey. Postgrad Med J 2006 Apr;82(966):274-279. [doi: 10.1136/pgmj.2005.040675] [Medline: 16597816]
2. Tang H, Ng JH. Googling for a diagnosis--use of Google as a diagnostic aid: internet based study. BMJ 2006 Dec 2;333(7579):1143-1145. [doi: 10.1136/bmj.39003.640567.AE] [Medline: 17098763]
3. Yu H, Kaufman D. A cognitive evaluation of four online search engines for answering definitional questions posed by physicians. Pac Symp Biocomput 2007:328-339. [Medline: 17990503]
4. Cai HC, King LE, Dwyer JT. Using the Google™ search engine for health information: is there a problem? Case study: supplements for cancer. Curr Dev Nutr 2021 Feb;5(2):nzab002. [doi: 10.1093/cdn/nzab002] [Medline: 33937613]
5. Daraz L, Morrow AS, Ponce OJ, et al. Can patients trust online health information? A meta-narrative systematic review addressing the quality of health information on the internet. J Gen Intern Med 2019 Sep;34(9):1884-1891. [doi: 10.1007/s11606-019-05109-0] [Medline: 31228051]
6. Lareyre F, Nasr B, Chaudhuri A, Di Lorenzo G, Carlier M, Raffort J. Comprehensive review of natural language processing (NLP) in vascular surgery. EJVES Vasc Forum 2023;60:57-63. [doi: 10.1016/j.ejvsvf.2023.09.002] [Medline: 37822918]
7. Goh E, Bunning B, Khoong EC, et al. Physician clinical decision modification and bias assessment in a randomized controlled trial of AI assistance. Commun Med (Lond) 2025 Mar 4;5(1):59. [doi: 10.1038/s43856-025-00781-2] [Medline: 40038550]
8. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024 Oct 1;7(10):e2440969. [doi: 10.1001/jamanetworkopen.2024.40969] [Medline: 39466245]
9. Goh E, Gallo RJ, Strong E, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. Nat Med 2025 Apr;31(4):1233-1238. [doi: 10.1038/s41591-024-03456-y] [Medline: 39910272]
10. Mateen BA, Menon V, Agweyu A, et al. Trials for LLM-supported clinical decisions in African primary healthcare. Nat Med 2025 Sep;31(9):2833-2835. [doi: 10.1038/s41591-025-03815-3] [Medline: 40610804]
11. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med 2025 May 13;8(1):274. [doi: 10.1038/s41746-025-01670-7] [Medline: 40360677]
12. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Haider C, Forte AJ. AI and ethics: a systematic review of the ethical considerations of large language model use in surgery research. Healthcare (Basel) 2024 Apr 13;12(8):825. [doi: 10.3390/healthcare12080825] [Medline: 38667587]
13. Liang EN, Pei S, Staibano P, van der Woerd B. Clinical applications of large language models in medicine and surgery: a scoping review. J Int Med Res 2025 Jul;53(7):3000605251347556. [doi: 10.1177/03000605251347556] [Medline: 40615349]
14. Athavale A, Baier J, Ross E, Fukaya E. The potential of chatbots in chronic venous disease patient management. JVS Vasc Insights 2023;1:100091. [doi: 10.1016/j.jvsvi.2023.100019] [Medline: 37701430]

XSL•FO
RenderX

15. Wijetunga A, Wang Y, Vicaretti M. Evaluating the safety and accuracy of three large language models in vascular surgery. Presented at: Asian Society for Vascular Surgery 2025 Conference; Aug 28-31, 2025 URL: https://asianvascular.com/wp-content/uploads/2025/10/ASVS2025-Program.pdf [accessed 2026-10-22]
16. Le Q, Lavingia KS, Amendola M. The performance of large language models on fictional consult queries indicates favorable potential for AI-assisted vascular surgery consult handling. JVS Vasc Insights 2024;2:100052. [doi: 10.1016/j.jvsvi.2023.100052]
17. Chervonski E, Harish KB, Rockman CB, et al. Generative artificial intelligence chatbots may provide appropriate informational responses to common vascular surgery questions by patients. Vascular 2025 Feb;33(1):229-237. [doi: 10.1177/17085381241240550] [Medline: 38500300]
18. Palenzuela DL, Mullen JT, Phitayakorn R. AI versus MD: evaluating the surgical decision-making accuracy of ChatGPT-4. Surgery 2024 Aug;176(2):241-245. [doi: 10.1016/j.surg.2024.04.003] [Medline: 38769038]
19. Heisinger S, Salzmann SN, Senker W, et al. ChatGPT's performance in spinal metastasis cases-can we discuss our complex cases with ChatGPT? J Clin Med 2024 Dec 23;13(24):7864. [doi: 10.3390/jcm13247864] [Medline: 39768787]
20. McNamara MA, Hill BG, Schilling PL. The challenges of using ChatGPT for clinical decision support in orthopaedic surgery: a pilot study. J Am Acad Orthop Surg 2025 Jun 1;33(11):618-622. [doi: 10.5435/JAAOS-D-24-01072] [Medline: 40153611]
21. Long C, Lowe K, Zhang J, et al. A novel evaluation model for assessing ChatGPT on otolaryngology-head and neck surgery certification examinations: performance study. JMIR Med Educ 2024 Jan 16;10:e49970. [doi: 10.2196/49970] [Medline: 38227351]
22. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. ACM Comput Surv 2023;55(12):1-38. [doi: 10.1145/3571730]
23. Kim SS, Liao QV, Vorvoreanu M, Ballard S, Vaughan JW. "I'm not sure, but...": examining the impact of large language models' uncertainty expression on user reliance and trust. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24): Association for Computing Machinery; 2024:822-835. [doi: 10.1145/3630106.3658941]
24. Steyvers M, Tejeda H, Kumar A, et al. What large language models know and what people think they know. Nat Mach Intell 2025;7(2):221-231. [doi: 10.1038/s42256-024-00976-7]
25. Alexiou VG, Sumpio BE, Vassiliou A, Kakkos SK, Geroulakos G. Artificial intelligence in diagnosing and managing vascular surgery patients: an experimental study using the GPT-4 model. Ann Vasc Surg 2025 Feb;111:260-267. [doi: 10.1016/j.avsg.2024.11.014] [Medline: 39586530]
26. Yang R, Ning Y, Keppo E, et al. Retrieval-augmented generation for generative artificial intelligence in health care. npj Health Syst 2025;2:2. [doi: 10.1038/s44401-024-00004-1]
27. Feridooni T, Javidan AP, Mahmood DN, et al. Development of a vascular surgery-specific artificial intelligence chat interface using retrieval-augmented generation: VASC.AI, a specialized vascular surgery chatbot. JVS Vasc Insights 2024;2:100137. [doi: 10.1016/j.jvsvi.2024.100137]
28. Ge J, Sun S, Owens J, et al. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. Hepatology 2024 Nov 1;80(5):1158-1168. [doi: 10.1097/HEP.0000000000000834] [Medline: 38451962]
29. Tung JY, Le Q, Yao J, et al. Performance of retrieval-augmented generation large language models in guideline-concordant prostate-specific antigen testing: comparative study with junior clinicians. J Med Internet Res 2025 Nov 19;27:e78393. [doi: 10.2196/78393] [Medline: 41259800]
30. Wang D, Liang J, Ye J, et al. Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: comparative study. J Med Internet Res 2024 Nov 8;26:e58041. [doi: 10.2196/58041] [Medline: 39046096]
31. Privacy policy. OpenAI. URL: https://openai.com/policies/row-privacy-policy/ [accessed 2025-11-09]
32. Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med 2024 Feb 20;7(1):41. [doi: 10.1038/s41746-024-01029-4] [Medline: 38378899]
33. Privacy Act 1988. Federal Register of Legislation, Australian Government. URL: https://www.legislation.gov.au/Series/C2004A03712 [accessed 2026-03-28]
34. Safe and responsible artificial intelligence in health care – legislation and regulation review. Royal Australasian College of Surgeons. 2024. URL: https://www.surgeons.org/-/media/Submission-Safe-and-Responsible-AI-in-Health-Care.pdf [accessed 2026-03-28]
35. Lareyre F, D'Oria M, Caradu C, et al. Open e-survey on the use and perception of chatbots in vascular surgery. EJVES Vasc Forum 2024;62:57-63. [doi: 10.1016/j.ejvsvf.2024.07.037] [Medline: 39346798]

Abbreviations

ALI: acute limb ischemia
CLI: chronic limb ischemia
DFI: diabetic foot infection

XSL•FO
RenderX

ECD: extracranial cerebrovascular disease
LLM: large language model
OR: odds ratio
RAG: retrieval-augmented generation

Edited by N Rohatgi; submitted 22.Jan.2026; peer-reviewed by E Zavaleta-Monestel; revised version received 22.Feb.2026; accepted 08.Mar.2026; published 06.Nov.2026.

Please cite as:
Wijetunga A, Wang Y, Yaghi C, Vicaretti M
Evaluating the Performance of Large Language Models in Vascular Surgery Cases: Cross-Sectional Study
JMIR Perioper Med 2026;9:e91973
URL: https://periop.jmir.org/2026/1/e91973
doi:10.2196/91973

© Asanka Wijetunga, Yunyi Wang, Chantell Yaghi, Mauro Vicaretti. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 6.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Advancing Perioperative Medicine With AI: Current Landscape and Research Priorities

Nidhi Rohatgi[1], MS, MD; Piyush Mathur[2], MD

[1]Department of Medicine, Stanford University, 300 Pasteur Drive, Palo Alto, CA, United States

[2]Department of Anesthesiology, Cleveland Clinic, Cleveland Clinic Lerner College of Medicine, Cleveland, OH, United States

Corresponding Author:
Nidhi Rohatgi, MS, MD
Department of Medicine, Stanford University, 300 Pasteur Drive, Palo Alto, CA, United States

Abstract

AI is increasingly being applied across the perioperative continuum. AI should be viewed as a technology that can reduce cognitive burden, collate fragmented data into clinically meaningful information, support decision-making, and extend precision education to clinicians and patients. This editorial accompanies the *JMIR Perioperative Medicine* webinar held in May 2026 and presents our perspectives—informed by our clinical experience—on where AI can assist perioperative clinicians and surgical patients, the evidence available today, what is still needed, and how clinicians can prepare to practice alongside AI tools. We consider opportunities in preoperative risk assessment and optimization; intraoperative intelligent decision support; emerging technologies, such as surgical digital twins; and postoperative recovery, monitoring, and prediction of clinical deterioration. Most AI applications remain early and require large-scale, prospective validation across diverse populations and resource settings. Realizing their potential also requires confronting substantial challenges, including limited and fragmented training data, the specialized models needed to interpret the large volumes of data captured in operative videos and during physiological monitoring, biases embedded in data and algorithms or biases inherent in human-AI interactions, and the determination of where human oversight is essential and where it is not feasible. When used responsibly, with human judgment remaining paramount, AI offers an unprecedented opportunity to advance perioperative care.

(JMIR Perioper Med 2026;9:e107155) doi:10.2196/107155

KEYWORDS

artificial intelligence; perioperative medicine; risk prediction; machine learning; clinical decision support

Introduction

AI is becoming a defining technology in health care. AI can reduce cognitive burden, improve efficiency, enhance learning, and enable clinicians to devote time and attention to the uniquely human aspects of care and the contexts where human judgment is indispensable.

The perioperative continuum begins days or months prior to surgery and extends through the intraoperative course, recovery, and rehabilitation, with focus shifting to follow-up beyond the traditional 30 or 90 days after surgery. AI is evolving from a collection of task-specific algorithms to a longitudinal intelligence system that can support care delivery and decision-making across the perioperative continuum. This evolution was reflected throughout the *JMIR Perioperative Medicine* webinar, "Advancing Perioperative Care with Artificial Intelligence," held in May 2026, which emphasized a practical perspective grounded in the real-world implementation of AI [1]. This accompanying editorial—written for perioperative researchers and clinicians—discusses our perspectives on where AI can assist or augment perioperative clinicians and surgical patients, what is the evidence we have, what is still needed, and how perioperative clinicians can prepare to practice with AI. Across the preoperative, intraoperative, and postoperative phases, the recurring messages are that the nearest-term value of AI lies in synthesizing fragmented data and not in autonomous decision-making, most applications await prospective validation in diverse real-world settings, and human judgment remains paramount.

Preoperative Opportunities

Contemporary preoperative risk assessment involves history-taking focused on multiorgan system assessment, the review of records within and outside the electronic health record (EHR), outreach to the patient's ambulatory clinicians and caregivers, medication management, and appropriate consultations. These data are then integrated with the surgical complexity; the indication for surgery; the patient's goals, expectations, and functional, nutritional, psychosocial, and cognitive status; preoperative risk indices; pharmacogenomic data; current evidence; and the clinician's judgment, to formulate individualized plans for preoperative optimization and intra- and postoperative management. When conducted thoroughly and in partnership with the patient, their caregivers, and the clinicians involved in their longitudinal care, preoperative evaluation offers an unmatched opportunity to prevent

XSL•FO
RenderX

postoperative complications [2]. However, it is time-consuming and requires nuanced expertise.

Information Synthesis From the EHR

AI's key value in preoperative assessment is in collating data across multiple fragmented data streams and synthesizing them into meaningful information [3]. AI can retrieve relevant clinical practice guidelines to suggest targeted tests and interventions and compute risk scores. Conversational AI agents are an emerging opportunity for history-taking, although in one study, patients provided fewer details to AI agents than to humans [4]. Ambient intelligence has reduced the cognitive burden during documentation, although time savings, the quality of AI-scribed notes, and adoption vary considerably [5-9]. In a randomized crossover trial, the Perioperative AI Chatbot (PEACH) did not reduce the overall documentation time for resident physicians, but time savings were noted for moderate-complexity patients and for more experienced physicians [9]. With AI, clinicians' effort can shift from information retrieval, documentation, and administrative tasks toward shared decision-making and reasoning.

Multimodal Risk Prediction

Importantly, only a fraction of a patient's trajectory is captured in the EHR. Most physiological variation occurs outside clinics and hospitals. Often, patients themselves possess the richest source of health information. Wearables, patient-reported outcomes, and mobile health apps can incorporate data generated during everyday life rather than data generated exclusively during episodes of illness [10]. This transition from episodic "sick care" toward continuous health monitoring may reshape preoperative optimization and long-term outcomes.

AI applied to wearable accelerometry data may be able to predict frailty, which is central to preoperative risk stratification [11]. The AI-assisted identification of sarcopenia from routine computed tomography and AI-assisted coronary calcium estimates could also enhance preoperative risk assessment [12-14]. Beyond the perioperative episode, AI-assisted retinal scans can predict the long-term risk of major adverse cardiovascular events [15]. In a primary care setting, an AI-enabled stethoscope improved the detection of heart failure, valvular heart disease, and atrial fibrillation (all are important perioperative diagnoses), although adoption declined over time without workflow integration [16]. Further, AI-guided echocardiograms are showing promise in improving access to patients by improving the efficiency and effectiveness of trained sonographers and providing real-time guidance to novice sonographers [17,18].

Electrocardiograms (ECGs) are often obtained preoperatively. In a retrospective cohort, deep learning applied to preoperative ECGs discriminated postoperative mortality with an area under the curve (AUC) of 0.83, whereas the AUC for the Revised Cardiac Risk Index was 0.67 [19]. In the nonsurgical setting, the self-supervised interpretation of ECGs across 77 cardiac conditions achieved AUCs of 0.98 - 0.99 [20]. Another AI-ECG model achieved AUCs of 0.90 - 0.91 for detecting right ventricular and low left ventricular systolic dysfunction [21]. AI-interpreted ECGs and AI-assisted focused ultrasound could potentially reduce unnecessary preoperative echocardiograms [22]. Whether acting on AI-interpreted ECG findings improves clinically meaningful outcomes or cascades low-yield evaluation (eg, echocardiograms and cardiology consultations) requires prospective validation across institutions with varying disease prevalence and surgical risk profiles [23].

Evaluation of AI-Based Risk Predictions

Predictive analytics is a common application of AI in perioperative medicine, although a systematic review of 103 predictive analytics machine learning (ML) tools reported that 75% of them were validated at a single center, only 13% were externally validated, and the PROBAST (Prediction Model Risk of Bias Assessment Tool) assessment showed a high risk of bias [24]. Preoperative risk indices have largely utilized structured data and regression techniques, including the American College of Surgeons' surgical risk calculator, which was later enhanced by using ML [25]. The performance of these risk indices has been variable and often moderate across different surgeries and populations. Some institutions have developed their own perioperative risk algorithms by using ML, such as MySurgeryRisk, which predicts 8 major postoperative complications and mortality by using automated EHR input [26,27]. That said, risk prediction alone can be insufficient. In the Perioperative Outcome Risk Assessment with Computer Learning Enhancement (Perioperative ORACLE) randomized trial, access to perioperative ML prediction models did not improve clinician performance, as clinicians anchored to their own predictions [28]. Additionally, effective perioperative care requires actionable pathways and infrastructure for the optimization of those at risk.

Intraoperative Opportunities

AI is being explored to predict surgical case duration and cancellation and forecast surgical volumes for staffing [29]. Intraoperative care generates enormous volumes of data from hemodynamic monitoring, neuromonitoring, ventilators, closed-loop drug delivery systems, and operating room black boxes implemented at some institutions. Yet, capturing and extracting these data in a form usable for training AI can be challenging or expensive, or they can be beyond the context limits of general-purpose foundation models.

The operating room is a high-risk environment, and AI remains in the early phases of exploration. In the operating room, human-in-the-loop is particularly challenging when precious minutes must be spent on intervention rather than on assessing the accuracy, assumptions, or confidence of AI output. The interface of AI output therefore deserves special consideration, alongside sharing with the surgeon or anesthesiologist what data and evidence were used to train the models.

Intraoperative Risk Prediction

For anesthesiologists, real-time, AI-driven decision support and monitoring can reduce intraoperative cognitive burden [30].

XSL•FO
RenderX

The Hypotension Prediction Index—an ML algorithm applied to arterial waveform analysis—can warn about impending hypotension a few minutes before the event, allowing for preemptive intervention, but the benefit for patient outcomes is yet to be established [31].

Intraoperative Intelligence and Precision Education

AI is emerging to assist with medication safety checks in the operating room, provide ultrasound guidance during regional blocks, and anticipate transfusion needs [32-34]. Surgery is being reshaped by robotics and AI-augmented tumor detection [35-38]. Computer vision can analyze operative videos to teach technical skills, deliver feedback, and improve ergonomics [39-41]. Further, specialized surgical video language models can review hours of footage, extending precision education and performance insights beyond what the manual review of surgical videos ever allowed [42].

Simulation With Digital Twins

Another emerging frontier is the surgical digital twin that can be a virtual replica of the patient, their physiology, or the operating room itself [43,44]. Such digital twins are being explored for surgical planning, training, and real-time guidance with patient-specific anatomy. With digital twins, in silico trials can optimize processes and surface failure modes in a virtual, risk-free, and cost-effective environment. The central challenge is data that may be protected (eg, EHR and imaging data) or data that are never captured because of a lack of sensors, haptics, optical tracking, videos, or extended reality. These models also demand significant computational power.

Postoperative Opportunities

Preoperative assessment remains critical for anticipating, preventing, and managing postoperative complications. The risk of short- and longer-term adverse postoperative outcomes concentrates among patients with the highest burden of chronic conditions or geriatric syndromes [45,46]. Among older adults undergoing major elective noncardiac surgery, 1 in 5 patients may sustain significant new disability or die by 1 year after surgery [45].

AI-Based Postoperative Monitoring

Several ML-based early warning systems have been prospectively validated and integrated into clinical workflows [47,48]. For example, the Targeted Real-Time Early Warning System (TREWS) system for the prediction of sepsis across 590,000 medical and surgical inpatients was associated with a 5.1% adjusted absolute reduction in the in-hospital mortality rate if TREWS alerts were confirmed within 3 hours by a clinician [47]. Extending these models specifically to postoperative patients, in whom altered physiology and certain laboratory abnormalities are a normal and expected part of recovery, is an opportunity. Tachycardia, hypoxia, anemia, and leukocytosis carry a different differential diagnosis for a postoperative patient compared to that for a medical patient, and they often reflect a combination of surgery-specific and medical factors. ML-based prediction models have also been developed to predict postoperative acute kidney injury [49].

Because postoperative patients carry particular risks from their underlying disease, the procedure, and the combination of opioids and residual anesthesia, intermittent (and sometimes inaccurate) vital sign measurements in general wards are especially likely to miss deterioration. Continuous monitoring detects far more postoperative deterioration than intermittent checks, including opioid-induced respiratory depression, hypotension, and hypoxemia, and deep learning could be applied to these continuous monitoring−provided data [50,51]. Whether earlier detection reduces complications, unplanned intensive care utilization, or death is unestablished.

Informed Patients

AI can help educate patients in their preferred language, format, and health literacy level, potentially reducing surgical cancellations and improving patient experiences and outcomes. After surgery, recovering patients must navigate complicated instructions, manage multiple conditions, follow up with several specialties, and sometimes recover at rehabilitation facilities. AI can improve discharge education and extract meaningful information from remote monitoring. AI is being explored to generate discharge summaries and draft responses to patient messages [52-54]. Additionally, conversational AI agents are being explored to provide discharge education. In a randomized trial of 261 orthopedic surgery patients, a context-specific AI agent deployed through a mobile app showed promise in improving postoperative functional recovery and patient experiences in the first few months after orthopedic surgery, when compared with physician-only communication [55].

Postdischarge Management

More than half of 30-day postoperative complications occur after discharge, which is when patients are least monitored and have the least resources [56]. The postoperative phase requires functional, nutritional, and cognitive rehabilitation; wound management; and pain management. AI is being explored to detect surgical site infections in patient-submitted images [57]. By incorporating the patient's unique medical, psychosocial, and genomic profile, alongside patient-reported feedback, AI could potentially identify those at high risk of uncontrolled postoperative pain or persistent opioid use and personalize opioid tapering and nonopioid analgesic management plans [58-62].

AI is showing promise in predicting readmission and acute care utilization by incorporating unstructured EHR data and social determinants of health, going beyond traditional structured data [63-65]. However, challenges with data fragmentation resulting from a lack of EHR interoperability across the different hospitals, clinics, and rehabilitation facilities where surgical patients may receive care, or those resulting from a lack of digitized records, remain unsolved.

XSL•FO
RenderX

Conclusion

AI presents an unprecedented opportunity to advance perioperative care. Most perioperative AI applications remain in the early phases of research and development. Rigorous, large-scale, prospective validation in real-world environments and across diverse populations, geographic regions, and resource settings is still needed. Despite the promise of AI in augmenting perioperative care, human vigilance, situational awareness, and judgment remain paramount. Perioperative clinicians should engage with AI; understand its fundamentals; use it as a thought partner; and, when feasible, engage in the evaluation and validation of AI. Clinicians should continue to hone their clinical skills, as discordance between clinical judgment and a confident AI output will demand even greater expertise. As AI assumes more cognitive and procedural tasks, the loss of some skills is inevitable, as is the acquisition of new competencies. The priorities are to maintain proficiency in the skills critical for patient safety and to remain capable of practicing effectively across settings with varying levels of AI adoption.

Acknowledgments

We are grateful to Ms Andrea Schaffeler (managing editor), Ms Liana Ramos (marketing associate), and Ms Isabella Hernandez (social media and content development specialist) at *JMIR Perioperative Medicine* for their assistance in organizing the webinar.

Generative AI Use

Claude Opus 4.8 and Claude Sonnet 4.6 (Anthropic) were used to rephrase a few sentences for clarity, brevity, or grammar. All content was reviewed and verified by the authors.

Funding

The authors declare that no financial support was received for this work.

Conflicts of Interest

NR is the editor-in-chief of *JMIR Perioperative Medicine* at the time of this publication. NR had no involvement in the editorial review and processing of this manuscript. She holds courtesy appointments in the departments of neurosurgery, anesthesiology, pain, and perioperative medicine at Stanford University School of Medicine and is an affiliate faculty at the Center for Artificial Intelligence in Medicine and Imaging (AIMI), Center for Digital Health, and Institute for Human-Centered Artificial Intelligence (HAI) at Stanford University. PM is a cofounder at BrainX, LLC, and was the invited guest at the *JMIR Perioperative Medicine* webinar in May 2026.

References

1. Advancing perioperative care with artificial intelligence. JMIR Publications YouTube page. 2026 Jun 4. URL: https://www.youtube.com/watch?v=AjjWR1_QUrY [accessed 2026-10-14]
2. Rohatgi N. Perioperative risk calculators and the art of medicine. JAMA Intern Med 2019 Dec 1;179(12):1619-1620. [doi: 10.1001/jamainternmed.2019.4914] [Medline: 31633737]
3. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024 Apr;30(4):1134-1142. [doi: 10.1038/s41591-024-02855-5] [Medline: 38413730]
4. Reis M, Reis F, Kim YJ, et al. Reduced symptom reporting quality during human–chatbot versus human–physician interactions. Nat Health 2026 May 1;1:1055-1058. [doi: 10.1038/s44360-026-00116-y]
5. Lukac PJ, Turner W, Vangala S, et al. Ambient AI scribes in clinical practice: a randomized trial. NEJM AI 2025 Dec;2(12). [doi: 10.1056/aioa2501000] [Medline: 41497288]
6. Holmgren AJ, Fenton CL, Thombley R, et al. Ambient artificial intelligence scribes and physician financial productivity. JAMA Netw Open 2026 Jan 2;9(1):e2553233. [doi: 10.1001/jamanetworkopen.2025.53233] [Medline: 41511775]
7. Afshar M, Baumann MR, Resnik F, et al. A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. NEJM AI 2025 Dec;2(12). [doi: 10.1056/aioa2500945] [Medline: 41625485]
8. Reddy A, Gunnink E, Wheat CL, et al. Rapid evaluation of artificial intelligence technology used for ambient dictation in primary care: comparing the quality of documentation of artificial intelligence-generated and human-produced clinical notes. Ann Intern Med 2026 Jun;179(6):765-772. [doi: 10.7326/ANNALS-25-02772] [Medline: 41996184]
9. Ke YH, Yang Ong BS, Jin L, et al. Clinical and economic impact of a large language model in perioperative medicine: a randomized crossover trial. NPJ Digit Med 2025 Jul 21;8(1):462. [doi: 10.1038/s41746-025-01858-x] [Medline: 40691284]
10. Tan JM, Cannesson M, Feldman JM, et al. Emerging technology and the future of perioperative care: perspectives and recommendations from the 2023 Stoelting Conference of the Anesthesia Patient Safety Foundation. Anesth Analg 2025 Jul 1;141(1):139-151. [doi: 10.1213/ANE.0000000000007540] [Medline: 40333433]
11. Culos A, Manas A, Shidara K, et al. A machine learning model for frailty based on wearable device measurements. Commun Med (Lond) 2026 Feb 19;6(1):173. [doi: 10.1038/s43856-026-01419-7] [Medline: 41714788]

XSL•FO
RenderX

12. Sudarsanam A, Davies AH, Rockall A, Sekhon H, Salim S. A fully automated sarcopenia segmentation tool using artificial intelligence can reliably assist in preoperative evaluation of patients undergoing major vascular surgery. J Vasc Surg 2024 Jun;79(6):e235-e236. [doi: 10.1016/j.jvs.2024.03.303]
13. Naghavi M, Reeves AP, Atlas K, et al. Artificial intelligence applied to coronary artery calcium scans (AI-CAC) significantly improves cardiovascular events prediction. NPJ Digit Med 2024 Nov 5;7(1):309. [doi: 10.1038/s41746-024-01308-0] [Medline: 39501071]
14. Choi DY, Hayes D, Maidman SD, et al. Existing Nongated CT Coronary Calcium Predicts Operative Risk in Patients Undergoing Noncardiac Surgeries (ENCORES). Circulation 2023 Oct 10;148(15):1154-1164. [doi: 10.1161/CIRCULATIONAHA.123.064398] [Medline: 37732454]
15. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018 Mar;2(3):158-164. [doi: 10.1038/s41551-018-0195-0] [Medline: 31015713]
16. Kelshiker MA, Bächtiger P, Petri CF, et al. Triple cardiovascular disease detection with an artificial intelligence-enabled stethoscope (TRICORDER) in the UK: a cluster-randomised controlled implementation trial. Lancet 2026 Feb 14;407(10529):704-715. [doi: 10.1016/S0140-6736(25)02156-7] [Medline: 41619752]
17. Narang A, Bae R, Hong H, et al. Utility of a deep-learning algorithm to guide novices to acquire echocardiograms for limited diagnostic use. JAMA Cardiol 2021 Jun 1;6(6):624-632. [doi: 10.1001/jamacardio.2021.0185] [Medline: 33599681]
18. Sakamoto A, Kagiyama N, Sato E, et al. Artificial intelligence-based automated echocardiographic analysis and the workflow of sonographers: a randomized crossover trial (AI-Echo RCT). J Am Heart Assoc 2026 Jan 6;15(1):e045637. [doi: 10.1161/JAHA.125.045637] [Medline: 41404733]
19. Ouyang D, Theurer J, Stein NR, et al. Electrocardiographic deep learning for predicting post-procedural mortality: a model development and validation study. Lancet Digit Health 2024 Jan;6(1):e70-e78. [doi: 10.1016/S2589-7500(23)00220-0] [Medline: 38065778]
20. Nolin-Lapalme A, Sowa A, Delfrate J, et al. Foundation models for electrocardiogram interpretation: clinical implications. Eur Heart J 2026 May 15;47(18):2174-2186. [doi: 10.1093/eurheartj/ehaf1119] [Medline: 41568699]
21. Poterucha TJ, Jing L, Ricart RP, et al. Detecting structural heart disease from electrocardiograms using AI. Nature 2025 Aug;644(8075):221-230. [doi: 10.1038/s41586-025-09227-0] [Medline: 40670798]
22. Cohen I, Lakritz A, Maor E. AI assisted focused cardiac ultrasound in preventive cardiology - a perspective. NPJ Cardiovasc Health 2025 Jun 26;2(1):27. [doi: 10.1038/s44325-025-00063-9] [Medline: 41776021]
23. Show us the evidence for the value of medical AI. Nat Med 2026 Apr;32(4):1163. [doi: 10.1038/s41591-026-04389-4] [Medline: 42014883]
24. Arina P, Kaczorek MR, Hofmaenner DA, et al. Prediction of complications and prognostication in perioperative medicine: a systematic review and PROBAST assessment of machine learning tools. Anesthesiology 2024 Jan 1;140(1):85-101. [doi: 10.1097/ALN.0000000000004764] [Medline: 37944114]
25. Liu Y, Ko CY, Hall BL, Cohen ME. American College of Surgeons NSQIP risk calculator accuracy using a machine learning algorithm compared with regression. J Am Coll Surg 2023 May 1;236(5):1024-1030. [doi: 10.1097/XCS.0000000000000556] [Medline: 36728295]
26. Ren Y, Loftus TJ, Datta S, et al. Performance of a machine learning algorithm using electronic health record data to predict postoperative complications and report on a mobile platform. JAMA Netw Open 2022 May 2;5(5):e2211973. [doi: 10.1001/jamanetworkopen.2022.11973] [Medline: 35576007]
27. Mahajan A, Esper S, Oo TH, et al. Development and validation of a machine learning model to identify patients before surgery at high risk for postoperative adverse events. JAMA Netw Open 2023 Jul 3;6(7):e2322285. [doi: 10.1001/jamanetworkopen.2023.22285] [Medline: 37418262]
28. Fritz BA, King CR, Abdelhack M, et al. Effect of machine learning models on clinician prediction of postoperative complications: the Perioperative ORACLE randomised clinical trial. Br J Anaesth 2024 Nov;133(5):1042-1050. [doi: 10.1016/j.bja.2024.08.004] [Medline: 39261226]
29. Bellini V, Russo M, Domenichetti T, Panizzi M, Allai S, Bignami EG. Artificial intelligence in operating room management. J Med Syst 2024 Feb 14;48(1):19. [doi: 10.1007/s10916-024-02038-2] [Medline: 38353755]
30. El-Sonbaty AI. Development and validation of an integrated artificial intelligence-powered clinical decision support application for cardiac critical care: combining coagulation management, blood gas analysis, mechanical ventilation, hemodynamic assessment, and extracorporeal membrane oxygenation. J Cardiothorac Vasc Anesth 2026 Sep;40(9):3019-3028. [doi: 10.1053/j.jvca.2026.05.062] [Medline: 42409648]
31. Li W, Hu Z, Yuan Y, Liu J, Li K. Effect of hypotension prediction index in the prevention of intraoperative hypotension during noncardiac surgery: a systematic review. J Clin Anesth 2022 Dec;83:110981. [doi: 10.1016/j.jclinane.2022.110981] [Medline: 36242978]
32. Chan J, Nsumba S, Wortsman M, et al. Detecting clinical medication errors with AI enabled wearable cameras. NPJ Digit Med 2024 Oct 22;7(1):287. [doi: 10.1038/s41746-024-01295-2] [Medline: 39438764]
33. Bowness JS, Macfarlane AJR, Burckett-St Laurent D, et al. Evaluation of the impact of assistive artificial intelligence on ultrasound scanning for regional anaesthesia. Br J Anaesth 2023 Feb;130(2):226-233. [doi: 10.1016/j.bja.2022.07.049] [Medline: 36088136]

XSL•FO
RenderX

34. Lou SS, Kumar S, Goss CW, et al. Multicenter validation of a machine learning model for surgical transfusion risk at 45 US hospitals. JAMA Netw Open 2025 Jun 2;8(6):e2517760. [doi: 10.1001/jamanetworkopen.2025.17760] [Medline: 40577014]
35. Schmidgall S, Opfermann JD, Kim JW, Krieger A. Will your next surgeon be a robot? Autonomy and AI in robotic surgery. Sci Robot 2025 Jul 23;10(104):eadt0187. [doi: 10.1126/scirobotics.adt0187] [Medline: 40700524]
36. Wagner L, Jourdan S, Mayer L, et al. Robotic scrub nurse to anticipate surgical instruments based on real-time laparoscopic video analysis. Commun Med (Lond) 2024 Aug 2;4(1):156. [doi: 10.1038/s43856-024-00581-0] [Medline: 39095639]
37. Lee TJ, Qiu L, Long J, et al. CystoDS: a multiclass endoscopy image dataset for artificial intelligence-assisted bladder cancer detection. Sci Data 2026 Feb 26;13(1):528. [doi: 10.1038/s41597-026-06887-z] [Medline: 41741508]
38. Kondepudi A, Pekmezci M, Hou X, et al. Foundation models for fast, label-free detection of glioma infiltration. Nature 2025 Jan;637(8045):439-445. [doi: 10.1038/s41586-024-08169-3] [Medline: 39537921]
39. Yang JH, Goodman ED, Dawes AJ, et al. Using AI and computer vision to analyze technical proficiency in robotic surgery. Surg Endosc 2023 Apr;37(4):3010-3017. [doi: 10.1007/s00464-022-09781-y] [Medline: 36536082]
40. Goodman ED, Patel KK, Zhang Y, et al. Analyzing surgical technique in diverse open surgical videos with multitask machine learning. JAMA Surg 2024 Feb 1;159(2):185-192. [doi: 10.1001/jamasurg.2023.6262] [Medline: 38055227]
41. Sui E, Egeland C, Wang X, et al. Computer vision-based retrieval of steps and intraoperative adverse events in laparoscopic cholecystectomy. NPJ Digit Surg 2026 Jul 6;1(1):12. [doi: 10.1038/s44484-026-00010-w]
42. Perez A, Nwoye C, Raji Kermani R, Mohareri O, Abdullah Jamal M. SurgLaVi: large-scale hierarchical dataset for surgical vision-language representation learning. Med Image Anal 2026 May;110:103982. [doi: 10.1016/j.media.2026.103982] [Medline: 41719833]
43. Asciak L, Kyeremeh J, Luo X, et al. Digital twin assisted surgery, concept, opportunities, and challenges. NPJ Digit Med 2025 Jan 15;8(1):32. [doi: 10.1038/s41746-024-01413-0] [Medline: 39815013]
44. Mekki YM, Luijten G, Hagert E, et al. Digital twins for the era of personalized surgery. NPJ Digit Med 2025 May 15;8(1):283. [doi: 10.1038/s41746-025-01575-5] [Medline: 40374901]
45. Wijeysundera DN, Alibhai SMH, Puts MTE, et al. Significant new disability after major non-cardiac surgery in older adults aged 65 years and older in Canada: a multicentre prospective cohort study. Lancet Healthy Longev 2025 Nov;6(11):100789. [doi: 10.1016/j.lanhl.2025.100789] [Medline: 41317744]
46. International Surgical Outcomes Study group. Global patient outcomes after elective surgery: prospective cohort study in 27 low-, middle- and high-income countries. Br J Anaesth 2016 Oct 31;117(5):601-609. [doi: 10.1093/bja/aew316] [Medline: 27799174]
47. Adams R, Henry KE, Sridharan A, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med 2022 Jul;28(7):1455-1460. [doi: 10.1038/s41591-022-01894-0] [Medline: 35864252]
48. Wu L, Mai L, Wang H, et al. Artificial Intelligence-powered tiered early warning framework addressing high false alarm rates for in-hospital mortality prediction. NPJ Digit Med 2026 Mar 14;9(1):346. [doi: 10.1038/s41746-026-02522-8] [Medline: 41832244]
49. Tai J, Wang L, Li Y, Chen M, Yan Z, Liu J. Transparent AI-driven personalized risk prediction system for acute kidney injury after total hip arthroplasty. Sci Rep 2025 Oct 30;15(1):38023. [doi: 10.1038/s41598-025-21810-z] [Medline: 41168261]
50. Khanna AK, Flick M, Saugel B. Continuous vital sign monitoring of patients recovering from surgery on general wards: a narrative review. Br J Anaesth 2025 Feb;134(2):501-509. [doi: 10.1016/j.bja.2024.10.045] [Medline: 39779421]
51. Khanna AK, Bergese SD, Jungquist CR, et al. Prediction of opioid-induced respiratory depression on inpatient wards using continuous capnography and oximetry: an international prospective, observational trial. Anesth Analg 2020 Oct;131(4):1012-1024. [doi: 10.1213/ANE.0000000000004788] [Medline: 32925318]
52. Grolleau F, Liang AS, Keyes T, et al. Physician-reported safety outcomes of AI-generated hospital course summaries. JAMA Netw Open 2026 May 1;9(5):e2616556. [doi: 10.1001/jamanetworkopen.2026.16556] [Medline: 42101844]
53. Mandal S, Wiesenfeld BM, Szerencsy AC, et al. Utilization of generative AI-drafted responses for managing patient-provider communication. NPJ Digit Med 2025 Oct 2;8(1):591. [doi: 10.1038/s41746-025-01972-w] [Medline: 41038966]
54. Garcia P, Ma SP, Shah S, et al. Artificial intelligence-generated draft replies to patient inbox messages. JAMA Netw Open 2024 Mar 4;7(3):e243201. [doi: 10.1001/jamanetworkopen.2024.3201] [Medline: 38506805]
55. Li J, Zhang Y, Zhang Z, et al. A randomized controlled trial of a WeChat-based artificial intelligence agent for postoperative care in orthopedic patients. NPJ Digit Med 2026 Jan 17;9(1):105. [doi: 10.1038/s41746-025-02269-8] [Medline: 41548028]
56. Li RD, Joung RHS, Chung JW, Holl J, Bilimoria KY, Merkow RP. Divergent trends in postoperative length of stay and postdischarge complications over time. Jt Comm J Qual Patient Saf 2024 Sep;50(9):630-637. [doi: 10.1016/j.jcjq.2024.05.006] [Medline: 38853106]
57. Muaddi H, Choudhary A, Lee F, et al. Imaging-based surgical site infection detection using artificial intelligence. Ann Surg 2025 Sep 1;282(3):419-428. [doi: 10.1097/SLA.0000000000006826] [Medline: 40607706]
58. Sajdeya R, Narouze S. Harnessing artificial intelligence for predicting and managing postoperative pain: a narrative literature review. Curr Opin Anaesthesiol 2024 Oct 1;37(5):604-615. [doi: 10.1097/ACO.0000000000001408] [Medline: 39011674]

XSL•FO
RenderX

59. Sun EC, Darnall BD, Baker LC, Mackey S. Incidence of and risk factors for chronic opioid use among opioid-naive patients in the postoperative period. JAMA Intern Med 2016 Sep 1;176(9):1286-1293. [doi: 10.1001/jamainternmed.2016.3298] [Medline: 27400458]
60. Annis AC, Gunaseelan V, Smith AV, et al. Genetic associations of persistent opioid use after surgery point to OPRM1 but not other opioid-related loci as the main driver of opioid use disorder. Genet Epidemiol 2025 Jan;49(1):e22588. [doi: 10.1002/gepi.22588] [Medline: 39385445]
61. Bui T, Dooley MJ, Bell JS, Myles PS. New persistent opioid use after surgery and traumatic injury: a narrative review of global rates, risk factors, and healthcare interventions. BJA Open 2025 Dec 1;16:100509. [doi: 10.1016/j.bjao.2025.100509] [Medline: 41431465]
62. Burns S, Urman R, Pian R, Coppes OJM. Reducing new persistent opioid use after surgery: a review of interventions. Curr Pain Headache Rep 2021 Mar 24;25(5):27. [doi: 10.1007/s11916-021-00943-6] [Medline: 33760983]
63. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. Nature 2023 Jul;619(7969):357-362. [doi: 10.1038/s41586-023-06160-y] [Medline: 37286606]
64. Ratcliffe H, Morgan J, Jacobs M, Batniji R, Basu S. An artificial intelligence oracle for proactive population health quality improvement. NEJM Catal Innov Care Deliv 2026 Mar;7(3):CAT250283. [doi: 10.1056/CAT.25.0283] [Medline: 42418607]
65. Brown Z, Bergman D, Holt L, et al. Augmenting a transitional care model with artificial intelligence decreased readmissions. J Am Med Dir Assoc 2023 Jul;24(7):958-963. [doi: 10.1016/j.jamda.2023.03.005] [Medline: 37054749]

Abbreviations

AUC: area under the curve
ECG: electrocardiogram
EHR: electronic health record
ML: machine learning
PEACH: Perioperative AI Chatbot
Perioperative ORACLE: Perioperative Outcome Risk Assessment with Computer Learning Enhancement
PROBAST: Prediction Model Risk of Bias Assessment Tool
TREWS: Targeted Real-Time Early Warning System

Edited by I Steenstra; submitted 15.Jul.2026; this is a non–peer-reviewed article;accepted 01.Oct.2026; published 30.Oct.2026.

Please cite as:
Rohatgi N, Mathur P
Advancing Perioperative Medicine With AI: Current Landscape and Research Priorities
JMIR Perioper Med 2026;9:e107155
URL: https://periop.jmir.org/2026/1/e107155
doi:10.2196/107155

© Nidhi Rohatgi, Piyush Mathur. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 30.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Machine Learning–Based Prediction of Major Postoperative Complications After Noncardiac Surgery Under General Anesthesia Using Routinely Available Electronic Health Record Data: Retrospective Multicenter Development and Validation Study

Takuya Nishino[1], PhD; Hiroshi Mase[2,3], MD, PhD; Chol Kim[4], MD, PhD; Shinji Sugita[5], MD, PhD; Yukihiro Kondo[1,6], MD, PhD; Masashi Ishikawa[2], MD, PhD

[1]Department of Health Policy and Management, Nippon Medical School, Bunkyo-ku, Tokyo, Japan

[2]Department of Anesthesiology and Pain Medicine, Nippon Medical School, 1-1-5 Sendagi, Bunkyo-ku, Tokyo, Japan

[3]Department of Surgical Intensive Care, Nippon Medical School, Bunkyo-ku, Tokyo, Japan

[4]Department of Anesthesiology, Nippon Medical School Chiba Hokusoh Hospital, Chiba, Japan

[5]Department of Anesthesiology, Nippon Medical School Musashi Kosugi Hospital, Kawasaki, Kanagawa, Japan

[6]Department of Urology, Nippon Medical School, Tokyo, Japan

Corresponding Author:
Hiroshi Mase, MD, PhD
Department of Anesthesiology and Pain Medicine, Nippon Medical School, 1-1-5 Sendagi, Bunkyo-ku, Tokyo, Japan

Abstract

Background: Postoperative complications after noncardiac surgery under general anesthesia remain a major cause of morbidity, mortality, and resource-intensive care. Conventional risk assessment tools rely mainly on preoperative information and may not capture perioperative physiological and procedural factors associated with postoperative risk.

Objective: This retrospective multicenter observational study aimed to develop and validate electronic health record (EHR)–based machine learning (ML) models for predicting major postoperative complications and to evaluate their discrimination, overall probabilistic performance, calibration, and potential clinical utility.

Methods: We analyzed 35,702 adults who underwent noncardiac surgery under general anesthesia at 3 tertiary hospitals between January 2021 and March 2025. Preoperative, intraoperative, and early postoperative variables were extracted from diagnosis procedure combination (DPC) claims, EHRs, and anesthesia information systems. The primary outcome was a 30-day composite of in-hospital death, unplanned reoperation, prolonged mechanical ventilation, new initiation of renal replacement therapy, and postoperative infection. Logistic regression (LR), extreme gradient boosting (XGB), and light gradient boosting machine (LGBM) models were developed using preoperative variables alone or with 7 additional intraoperative and early postoperative variables. Leave-one-facility-out internal-external cross-validation (IECV) was the primary validation framework, with a stratified random 80:20 hold-out analysis as secondary validation. Performance was assessed using the area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (PR-AUC), Brier score, scaled Brier score, calibration measures, and decision curve analysis.

Results: The primary composite outcome occurred in 3884 of 35,702 patients (10.9%). In pooled IECV, perioperative models consistently outperformed the corresponding preoperative models. AUROC increased from 0.784 - 0.791 to 0.810 - 0.817, PR-AUC increased from 0.356 - 0.361 to 0.405 - 0.413, Brier scores decreased from 0.082 - 0.083 to 0.078 - 0.080, and scaled Brier scores increased from 0.144 - 0.151 to 0.175 - 0.191. Calibration varied across facilities; pooled calibration slopes for the perioperative models ranged from 0.875 to 0.902, and calibration intercepts ranged from −0.367 to −0.222, indicating overly extreme predictions and overall risk overestimation. Exploratory decision curve analysis generally showed greater net benefit for perioperative models across threshold probabilities of 0.01 - 0.30, although the curves overlapped substantially. Secondary and sensitivity analyses were consistent with the primary findings.

Conclusions: Adding routinely available intraoperative and early postoperative variables produced modest but consistent improvements in postoperative risk prediction compared with preoperative variables alone. However, facility-specific variation in calibration limits direct transportability of absolute risk estimates. The models may support prioritization for postoperative clinical reassessment, but exploratory decision curve analysis did not establish a specific intervention or triage threshold. Local recalibration and prospective workflow evaluation are required before clinical implementation.

(JMIR Perioper Med 2026;9:e91315) doi:10.2196/91315

XSL•FO
RenderX

KEYWORDS

perioperative risk; prediction model; postoperative complications; anesthesia; machine learning; electronic health records

Introduction

Approximately 250 million surgical procedures are performed worldwide each year, and this number continues to increase with population aging and expanding surgical indications [1]. As surgical volume rises, the absolute burden of postoperative complications is also expected to grow [2,3]. Postoperative mortality, often attributable to complications such as acute kidney injury, respiratory failure, and major cardiac events, remains a major global health concern, with more than 4 million deaths occurring within 30 days after surgery annually [4-7]. Importantly, multiple studies have demonstrated that early identification of patients at elevated postoperative risk can inform intraoperative and postoperative management strategies, thereby preventing complications or enabling earlier intervention [8-11]. Accordingly, accurate perioperative risk stratification is essential for timely and effective clinical decision-making. To support perioperative risk assessment, several established approaches are commonly used. The American Society of Anesthesiologists Physical Status classification (ASA-PS) provides a global assessment of preoperative physical status, whereas the Revised Cardiac Risk Index estimates perioperative cardiac risk, and the American College of Surgeons Surgical Risk Calculator provides individualized estimates of postoperative outcomes [12-14]. These approaches primarily rely on preoperative patient characteristics, comorbidities, and procedural information. Consequently, they may have limited ability to capture dynamic physiological changes occurring during surgery and the early postoperative period.

In recent years, machine learning (ML) approaches have increasingly been applied to clinical data to improve risk prediction across a wide range of medical domains. Unlike traditional regression-based models, ML algorithms can incorporate large numbers of heterogeneous variables and flexibly model nonlinear relationships and higher-order interactions among predictors [15]. These characteristics enable automated risk scoring systems that leverage high-dimensional clinical data and may outperform conventional approaches in complex clinical settings. In the perioperative context, ML-based decision support tools also have the potential to assist anesthesiologists and surgical teams—particularly less experienced clinicians—by providing objective, data-driven risk estimates to complement clinical judgment [16]. Previous studies have developed ML models to predict postoperative outcomes using demographic characteristics, comorbidities, and laboratory data available preoperatively in single-center electronic health record (EHR) cohorts [17,18], while a large multicenter model based on preoperative EHR data has also been reported [19]. Another study incorporated detailed intraoperative physiological time-series data [20]. Although these studies demonstrate the potential of ML for perioperative risk prediction, the transportability of these models across hospitals with differing patient populations, surgical case mix, and clinical practices remains an important consideration. Integrating routinely available perioperative information may provide additional predictive value while facilitating reproducible implementation across institutions.

In contrast, this study aimed to develop and validate pragmatic multicenter ML-based prediction models using routinely available perioperative data extracted from EHRs, administrative claims, and anesthesia information systems. By focusing on variables consistently recorded across hospitals, we sought to construct scalable models suitable for real-world clinical implementation. We compared models using preoperative variables alone with models additionally incorporating intraoperative and early postoperative variables. Model performance was primarily evaluated using leave-one-facility-out internal-external cross-validation (IECV), with particular emphasis on discrimination, calibration, overall probabilistic performance, and clinical utility, to assess the potential role of these models in supporting early postoperative risk stratification after noncardiac surgery under general anesthesia.

Methods

Study Design and Data Collection

This retrospective observational multicenter study used inpatient data from three affiliated Nippon Medical School hospitals: Nippon Medical School Hospital, Nippon Medical School Chiba Hokusoh Hospital, and Nippon Medical School Musashi Kosugi Hospital. The analytic dataset was constructed by integrating diagnosis procedure combination (DPC) claims data, laboratory test results extracted from EHRs, and perioperative information obtained from operating room and anesthesia information systems routinely used at each institution. All data sources are part of standard clinical documentation and administrative workflows.

Ethical Considerations

The study protocol was approved by the Central Ethics Committee of Nippon Medical School (approval number M-2024 - 227) and was conducted in accordance with the principles of the Declaration of Helsinki. Because this was a retrospective observational study using deidentified data, informed consent was obtained through an opt-out procedure. A prespecified statistical analysis plan for the primary analyses was developed and archived before model development and statistical analysis to minimize analytic flexibility and enhance transparency. Additional sensitivity and exploratory analyses were conducted during the revision process to evaluate the robustness and outcome specificity of the prediction models. This study was reported in accordance with the TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+Artificial Intelligence) statement and the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [21].

Study Population

We included all adult inpatients who underwent noncardiac surgery under general anesthesia between January 2021 and

XSL•FO
RenderX

March 2025. Each surgical admission was treated as a single analytic episode. Patients were excluded if they (1) were younger than 18 years, (2) underwent emergency surgery, defined as a decision-to-incision time of ≤24 hours recorded in the operating room information system, or (3) underwent cardiovascular surgery, including cardiac, major vascular, or thoracic aortic procedures. Cardiovascular surgeries were excluded because they follow distinct perioperative pathways and risk profiles, which could introduce substantial heterogeneity and impair model generalizability.

Outcomes

The primary end point was a composite outcome defined a priori, capturing clinically meaningful complications occurring within 30 days after surgery. The composite comprised (1) in-hospital death, (2) unplanned reoperation, (3) prolonged mechanical ventilation (≥48 hours of invasive support or reintubation within 48 hours), (4) new initiation of renal replacement therapy, and (5) postoperative infection. Although these events reflect heterogeneous pathophysiological mechanisms, they were combined to represent major postoperative morbidity with substantial clinical consequences. This composite outcome structure has been widely adopted in large-scale perioperative risk prediction studies to improve statistical power while maintaining clinical relevance [5,22,23]. Unplanned reoperation was defined as any emergency surgical procedure performed within 30 days of the index operation. Postoperative infection was defined according to the PERISCOPE study framework as administration of systemic antibiotics for 72 hours or longer after surgery, excluding prophylactic use [24].

All outcome components were ascertained from structured EHR fields using prespecified operational definitions. Nonfatal cardiac complications were not included in the composite outcome because the end point was designed to capture major postoperative events that could be consistently ascertained across participating hospitals using routinely collected clinical data. In particular, postoperative myocardial injury could not be uniformly ascertained because serial postoperative cardiac troponin measurements were not systematically performed in all patients across the participating institutions. To assess the influence of outcome composition and the predominance of postoperative infection, we conducted a prespecified sensitivity analysis using postoperative infection as a standalone outcome and a post hoc sensitivity analysis using a composite end point excluding postoperative infection. The latter comprised in-hospital death, unplanned reoperation, new renal replacement therapy, and prolonged postoperative mechanical ventilation. Patients with postoperative infection remained eligible for this analysis and were classified as having the noninfection composite outcome if they experienced at least one of the remaining components.

Variables

The DPC data included demographic characteristics (age and sex), anthropometric measurements (height, weight, and BMI; kg/m^2), and clinical information such as comorbidities, procedures, and medications [25]. High-risk surgical procedures were defined based on the Japanese Circulation Society 2022 Guideline on Perioperative Cardiovascular Assessment and Management for Non-Cardiac Surgery [26]. Procedures were classified as high risk if the estimated 30-day risk of cardiac death or nonfatal acute myocardial infarction exceeded 5%. These included pneumonectomy and lobectomy, lung and liver transplantation, esophagectomy, pancreaticoduodenectomy, hepatectomy and biliary tract surgery, surgery for gastrointestinal perforation, total cystectomy, major lower-limb amputation, and adrenalectomy. This classification was used solely to define a preoperative predictor and was independent of the definition of the study outcome.

Drug categories were defined according to previously reported associations with postoperative complications. Sex was encoded as a binary variable (female=0, male=1) and included as a predictor in both feature sets.

Preoperative laboratory and medication variables were extracted from the 60-day period before surgery. For laboratory tests, the value recorded closest to the admission date within this window was selected. From the operating room information system, we retrieved the ASA-PS, operative duration, fluid balance, and intraoperative transfusion volumes, including red blood cells (RBCs), platelets, and fresh frozen plasma (FFP). In addition, heart rate and mean arterial pressure measured 6 hours after surgery were collected as indicators of early postoperative physiological status.

Two prespecified feature sets were constructed. The preoperative feature set comprised 36 predictors, including demographic and anthropometric characteristics, ASA-PS, comorbidities, laboratory values, medication use, surgical procedure categories, number of resected sites, and the high-risk procedure indicator. The perioperative feature set comprised all 36 preoperative predictors plus 7 additional variables obtained during surgery or the early postoperative period: operative time, RBC transfusion volume, FFP transfusion volume, platelet transfusion volume, intraoperative fluid balance, heart rate 6 hours after surgery, and mean arterial pressure 6 hours after surgery. A complete feature dictionary, including the definition, unit or coding, measurement window, missingness, preprocessing procedure, and inclusion of each predictor in the preoperative and perioperative feature sets, is provided in Table S1 in Multimedia Appendix 1.

Continuous intraoperative waveform data were not included because standardized high-resolution signals were unavailable across institutions, and single-point summary statistics derived from waveforms were considered unreliable for model training. Instead, routinely available perioperative variables were used to ensure consistency, reproducibility, and scalability across sites.

Data Cleaning and Imputation

Continuous variables were winsorized at the 0.5th and 99.5th percentiles to limit the influence of extreme values while preserving the overall data distribution. Within each hold-out, IECV, calibration, and hyperparameter-tuning split, the winsorization thresholds were estimated using the corresponding development or training data and applied unchanged to the validation data. Variables were transformed using the natural

XSL•FO
RenderX

logarithm of (1+x) when all observed values after winsorization were nonnegative and their absolute skewness exceeded 1. After imputation, continuous, ordinal, and count predictors were robustly scaled for logistic regression (LR) only; no scaling was applied for extreme gradient boosting (XGB) or light gradient boosting machine (LGBM). As shown in Table S1 in Multimedia Appendix 1, the highest proportion of missing values was 10.4% (3714/35,702) for C-reactive protein, whereas all other variables had 3.1% (1098/35,702) or less missingness.

Because the missing-data mechanism was unlikely to be strictly missing completely at random, single imputation was performed using the scikit-learn IterativeImputer with LGBMRegressor [27] (n_estimators=50, max_iter=5, random_state=0) as the conditional estimator, mean initialization, and a maximum of 5 imputation iterations. Single imputation was selected to ensure computational feasibility and model stability in this large multicenter data set. To evaluate the robustness of the findings to imputation uncertainty, we conducted a multiple-imputation sensitivity analysis using 10 stochastic imputations. The imputations were generated using the scikit-learn IterativeImputer with BayesianRidge as the conditional estimator and posterior sampling enabled (sample_posterior=True). Mean imputation was used only for initialization; the maximum number of iterations was 5, and 20 nearest features were used for each conditional imputation model. Within each IECV fold, the imputer and all data-dependent preprocessing procedures were fitted using the development facilities only and were subsequently applied to the held-out facility without reestimation. The assumption that data were missing at random may not fully hold, particularly for laboratory tests, and potential residual bias is therefore acknowledged.

Model Development

We developed predictive models using 3 classification algorithms: LR, XGB [28], and LGBM. These algorithms were selected to represent a spectrum from conventional regression-based modeling to nonlinear tree-based ensemble methods with increasing model complexity.

Two model variants were constructed: (1) the preoperative model, which included only variables available before surgery, and (2) the perioperative model, which included preoperative, intraoperative, and early postoperative variables (within 6 hours).

Leave-one-facility-out IECV was designated as the primary validation framework. In each IECV iteration, models were developed using data from 2 institutions and evaluated in the remaining institution. This process was repeated until each institution had served once as the held-out validation institution. The out-of-sample predictions from the 3 held-out institutions were pooled to estimate overall IECV performance. A stratified random 80:20 hold-out analysis was conducted as a secondary validation.

Hyperparameter optimization for all models was conducted separately within each development data set using Optuna (Preferred Networks, Inc.) with a tree-structured Parzen estimator sampler. The optimization objective was the average precision score calculated from out-of-fold predictions generated by 5-fold stratified cross-validation. To address class imbalance of the composite outcome (event rate 10.9%, 3884/35,702), the optimization process included algorithm-specific imbalance handling: class weights for LR and scale_pos_weight for XGB and LGBM. The hyperparameter search spaces, optimization budgets, and selected values are summarized in Table S2 in Multimedia Appendix 1.

All model development steps, including preprocessing, hyperparameter optimization, model fitting, and probability calibration, were restricted to the corresponding development data. Probability calibration was performed using isotonic regression fitted to out-of-fold predictions generated by 3-fold stratified cross-validation within the corresponding development data [29]. The fitted calibration procedure was subsequently applied to the corresponding held-out data without recalibration using validation data.

Model Evaluation

Model performance was evaluated using 2 complementary validation frameworks. The primary validation used leave-one-facility-out IECV across the 3 participating institutions: Facility A, Nippon Medical School Hospital; Facility B, Nippon Medical School Chiba Hokusoh Hospital; and Facility C, Nippon Medical School Musashi Kosugi Hospital. In each of 3 validation rounds, data from 2 facilities were used for model development, and data from the remaining facility were reserved for held-out validation. All preprocessing, hyperparameter optimization, and probability calibration procedures were performed separately within each round using the development facilities only. Facility-specific performance was evaluated in each held-out facility, and the held-out predictions from all 3 rounds were pooled for the primary overall assessment. IECV was interpreted as an assessment of model transportability across hospitals within the same academic network. A fixed 80:20 random hold-out split, stratified by the primary outcome, was used as a complementary secondary validation.

Discrimination was assessed using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (PR-AUC). Overall prediction accuracy was assessed using the Brier score and the scaled Brier score, calculated relative to a null model based on outcome prevalence. Calibration was evaluated using the calibration intercept, calibration slope, and decile-based calibration plots. Within each validation framework, isotonic calibration was fitted exclusively within the corresponding development data and subsequently applied to the held-out facility or random hold-out set without recalibration using validation data. Thus, the reported calibration metrics represent out-of-sample performance after application of a calibration procedure estimated without using validation data.

Clinical utility was assessed using decision curve analysis based on facility-specific and pooled held-out IECV predictions over threshold probabilities ranging from 0.01 to 0.30. For this exploratory analysis, a positive prediction was interpreted as prompting prioritized postoperative clinical reassessment rather than a prespecified intervention. Such reassessment could include evaluation of the patient's recovery trajectory, organ

XSL•FO
RenderX

function, and evolving clinical findings, followed by monitoring, diagnostic testing, or treatment as clinically indicated. A negative prediction indicated no model-triggered reassessment beyond usual care. The treat-all and treat-none strategies represented prioritized reassessment of all patients and no model-triggered reassessment, respectively. Because no single intervention corresponded to all components of the composite outcome, the decision curve analysis was considered exploratory.

Bootstrapping was performed on the evaluation data only, with model predictions held fixed, for the random hold-out test set, each held-out IECV facility, and the pooled held-out IECV predictions. Accordingly, these CIs reflect uncertainty arising from the evaluation sample and do not incorporate variability from model development.

Sensitivity Analyses

Sensitivity analyses addressed 3 domains. First, to assess the robustness of the findings to uncertainty in missing-value imputation, a multiple-imputation sensitivity analysis was conducted for the preoperative and perioperative LightGBM models using 10 stochastically imputed data sets. This analysis was restricted to the primary composite outcome and used leave-one-facility-out IECV. The facility-specific LGBM hyperparameters selected in the primary analysis were held fixed. Within each imputed data set and IECV fold, the model was refitted, and isotonic calibration was estimated from out-of-fold predictions generated by 3-fold stratified cross-validation within the development data. Patient-level calibrated probabilities were averaged across the 10 imputed data sets before performance was evaluated using the pooled held-out predictions. Second, to assess the influence of outcome composition, preoperative and perioperative LGBM models were refitted using postoperative infection as the sole outcome and, in an additional post hoc analysis, using a composite end point excluding the postoperative infection component. These alternative-outcome analyses were evaluated using leave-one-facility-out IECV, with held-out predictions pooled across the 3 facilities. Third, temporal validation was performed by developing the LR, XGBoost, and LGBM models using data from January 2021 through December 2023 and evaluating them without model updating in a temporally held-out cohort from January 2024 through March 2025. In all sensitivity analyses, data-dependent preprocessing procedures were fitted using the corresponding development data only and applied unchanged to the held-out data. For the alternative-outcome analyses, the LGBM hyperparameters selected for the primary outcome were held fixed, with scale_pos_weight recalculated within each corresponding development data set. For temporal validation, hyperparameter optimization and probability calibration were performed using the temporal development cohort only.

Model Interpretation

Model predictions were interpreted using Shapley additive explanations (SHAP) values [30].

SHAP values were calculated separately for out-of-sample observations from each held-out institution and subsequently pooled across the 3 IECV folds. To limit computational burden while preserving the facility-specific outcome distribution, up to 2000 patients were sampled from each held-out institution using stratified random sampling by outcome status. LR models were interpreted using a linear SHAP explainer, whereas XGB and LGBM models were interpreted using a tree-based SHAP explainer. SHAP values were calculated for the uncalibrated base-model output because isotonic regression was applied as a separate postmodel calibration step. Accordingly, SHAP values were expressed on the log-odds scale for LR and on the raw margin scale for the tree-based models. Global SHAP summary plots were used to examine the magnitude and direction of individual predictor contributions.

Feature interactions were evaluated for the XGB and LGBM models using pairwise SHAP interaction values. Absolute interaction values were averaged across the sampled pooled out-of-sample observations, and diagonal elements representing main effects were set to 0. For visualization, values below the within-model 80th percentile were masked, and the 15 predictors with the greatest total interaction strength were displayed. SHAP values and interaction values represent model-based predictive contributions and were not interpreted as causal effects or compared directly in magnitude across algorithms.

Statistical Analysis

Continuous variables are reported as median (25th, 75th percentile), and categorical variables as number (percentage). For comparisons across the 3 institutions, standardized mean differences (SMD) were calculated for each pair of institutions and summarized as the mean absolute pairwise SMD. Values below 0.10 were considered to indicate limited between-institution imbalance. The 95% CIs for individual model-performance estimates were calculated using 1000 stratified bootstrap resamples of the corresponding validation sample, with model predictions held fixed. Differences in AUROC between the preoperative and perioperative models were evaluated using the paired DeLong method. Differences in PR-AUC were evaluated using paired bootstrap resampling stratified by held-out institution and outcome status. Percentile 95% CIs were obtained from 10,000 bootstrap resamples, and 2-sided *P* values were calculated from the null-centered bootstrap distribution of the paired difference. Between-algorithm differences were not formally tested. *P* values were interpreted descriptively and were not adjusted for multiple comparisons.

Statistical analyses were performed using R version 4.4.1 (R Foundation for Statistical Computing), and ML analyses were conducted using Python version 3.11.11 (Python Software Foundation).

Results

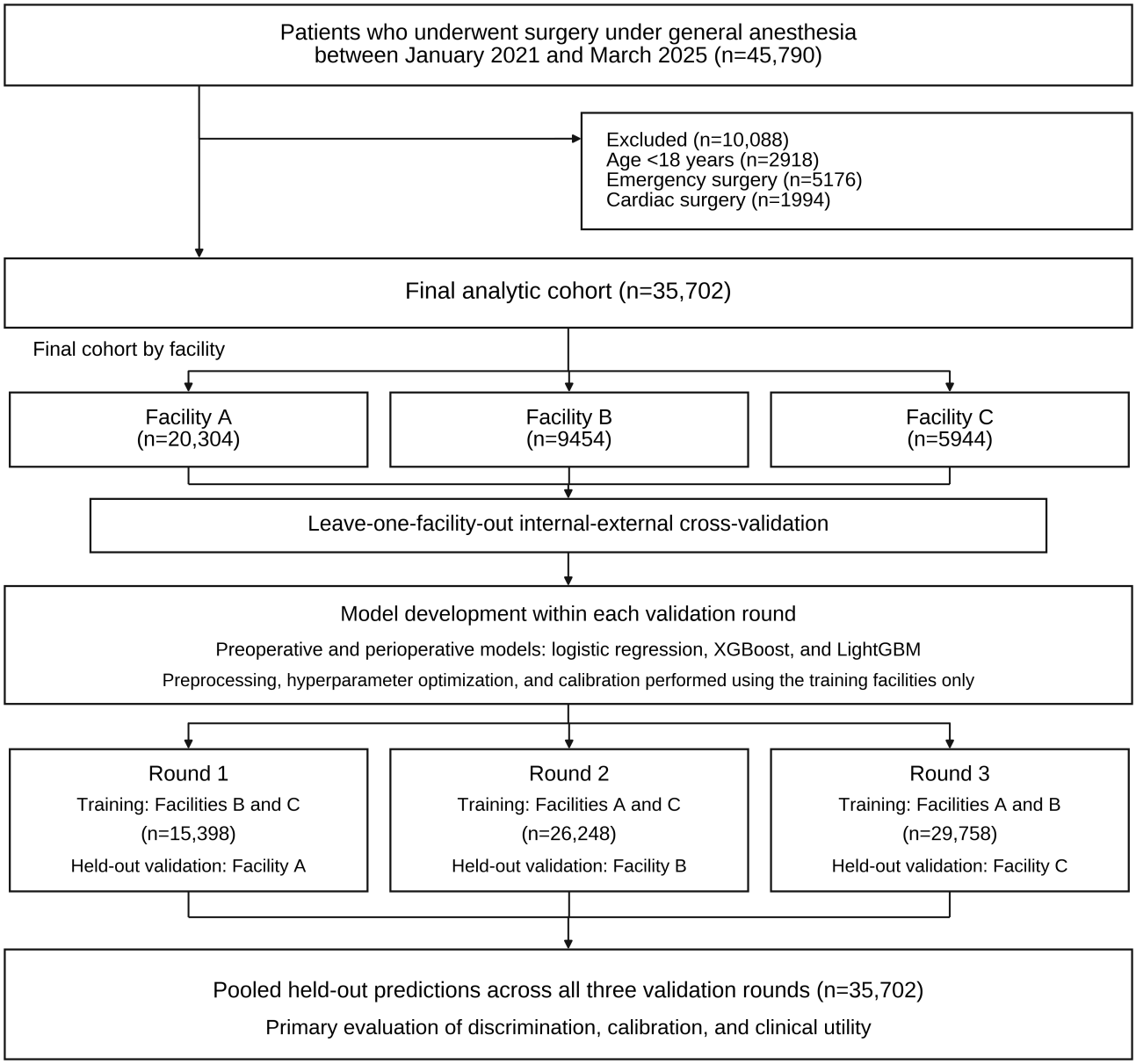
Study Population and Baseline Characteristics

We identified 45,790 patients who underwent surgery under general anesthesia at 3 facilities between January 2021 and March 2025. After excluding patients aged younger than 18 years (n=2918), those undergoing emergency surgery (n=5176), and those undergoing cardiac surgery (n=1994), 35,702 patients were included in the final analytic cohort. The cohort comprised

XSL•FO
RenderX

20,304 patients from Facility A, 9454 from Facility B, and 5944 from Facility C (Figure 1).

Figure 1. Study cohort selection and leave-one-facility-out internal-external cross-validation framework.



After applying the exclusion criteria, 35,702 patients from 3 facilities were included in the final analytic cohort. In each validation round, 2 facilities were used for model development, and the remaining facility was reserved for held-out validation. Preprocessing, hyperparameter optimization, and calibration were performed separately within each round using the training facilities only. Held-out predictions from the 3 validation rounds were pooled for the primary evaluation of discrimination, calibration, and clinical utility.

Table 1 summarizes the selected characteristics and outcome prevalence of the 35,702 surgical admissions included in the study. Of these, 20,304 (56.9%) were from Facility A, 9454 (26.5%) from Facility B, and 5944 (16.6%) from Facility C. The median age was 64 years (IQR 49 - 75), 17,494 patients (49.0%) were male, and the median BMI was 23.0 kg/m² (IQR 20.6 - 25.6). Overall, 30,325 patients (84.9%) were classified as ASA-PS I or II.

XSL•FO
RenderX

Table . Selected characteristics and outcome prevalence of the study cohort overall and by facility.[a]

Variable	Overall (N=35,702)	Facility A (N=20,304)	Facility B (N=9454)	Facility C (N=5944)	SMD[b,c]
Outcomes, n (%)					
Composite end point[d]	3884 (10.9)	2191 (10.8)	1115 (11.8)	578 (9.7)	0.045
Composite end point excluding the postoperative infection component[e]	1061 (3.0)	652 (3.2)	281 (3.0)	128 (2.2)	0.044
Postoperative infection[f]	3473 (9.7)	1939 (9.5)	1006 (10.6)	528 (8.9)	0.040
Demographic characteristics					
Age, median (IQR), years	64 (49-75)	64 (49-75)	68 (53-75)	58 (44-73)	0.249
Male sex, n (%)	17,494 (49.0)	9562 (47.1)	5123 (54.2)	2809 (47.3)	0.095
BMI, median (IQR), kg/m²	23.0 (20.6 - 25.6)	22.9 (20.5 - 25.6)	23.3 (21.0 - 26.0)	22.5 (20.2 - 25.1)	0.114
ASA[g] physical status, n (%)					0.179
Class I	4428 (12.4)	2333 (11.5)	1127 (11.9)	968 (16.3)	
Class II	25,897 (72.5)	14,531 (71.6)	6962 (73.6)	4404 (74.1)	
Class III	5254 (14.7)	3328 (16.4)	1359 (14.4)	567 (9.5)	
Class IV	123 (0.3)	112 (0.6)	6 (0.1)	5 (0.1)	
Preoperative laboratory values, median (IQR)					
Hemoglobin, g/dL	13.2 (12.1 - 14.3)	13.2 (12.0 - 14.2)	13.4 (12.2 - 14.5)	13.4 (12.2 - 14.4)	0.075
Albumin, g/dL	4.2 (3.9 - 4.4)	4.2 (3.9 - 4.5)	4.2 (3.9 - 4.4)	4.2 (3.9 - 4.5)	0.065
Creatinine, mg/dL	0.77 (0.64 - 0.93)	0.76 (0.63 - 0.93)	0.79 (0.65 - 0.95)	0.77 (0.65 - 0.93)	0.014
Preoperative medications, n (%)					
Diabetes medications	5000 (14.0)	2805 (13.8)	1592 (16.8)	603 (10.1)	0.131
Anticoagulants	1816 (5.1)	1048 (5.2)	537 (5.7)	231 (3.9)	0.056
Antiplatelet agents	1861 (5.2)	1008 (5.0)	597 (6.3)	256 (4.3)	0.060
Antineoplastic agents	2610 (7.3)	1701 (8.4)	551 (5.8)	358 (6.0)	0.066
Intraoperative variables, median (IQR)					
Operative duration, min	132 (83-207)	137 (89-209)	125 (76-215)	122 (78-182)	0.098
Fluid balance, mL	949 (618-1,541)	1,050 (700-1,781)	785 (500-1,336)	855 (625-1,190)	0.314
Early postoperative physiological variables within 6 hours after surgery, median (IQR)					
Heart rate, beats/min	75 (67-85)	76 (68-86)	75 (67-85)	72 (64-82)	0.173
Mean arterial pressure, mm Hg	88 (79-98)	89 (80-98)	87 (79-97)	88 (78-98)	0.088

[a]Data are presented as median IQR or n (%). Statistics were based on observed data before imputation. Variable definitions and missingness are provided in Supplementary Table S1 in Multimedia Appendix 2.

[b]SMD: standardized mean difference.

[c]SMDs are the mean absolute pairwise values across facilities.

[d]The composite end point comprised in-hospital death, unplanned reoperation, prolonged mechanical ventilation, new initiation of renal replacement therapy, and postoperative infection.

[e]The composite end point excluding postoperative infection comprised the remaining components.

XSL•FO
RenderX

[f]Patients with postoperative infection were included if they also experienced any of these components.

[g]ASA: American Society of Anesthesiologists.

The primary composite end point occurred in 3884 of 35,702 patients (10.9%). Postoperative infection occurred in 3473 patients (9.7%), whereas the composite end point excluding postoperative infection occurred in 1061 patients (3.0%). Outcome prevalence was similar across facilities. The prevalence of the primary composite end point ranged from 9.7% (578/5944) to 11.8% (1115/9454), postoperative infection from 8.9% (528/5944) to 10.6%(1006/9454), and the composite end point excluding postoperative infection from 2.2% (128/5944) to 3.2% (652/20,304); the corresponding SMDs were all 0.045 or lower.

In contrast, differences in case mix and perioperative characteristics were observed across facilities. The largest between-facility differences were identified for intraoperative net fluid balance (SMD 0.314), surgical procedure category (maximum SMD 0.290), age (SMD 0.249), ASA-PS (SMD 0.179), and heart rate at 6 hours after surgery (SMD 0.173). Patients at Facility B were older, whereas those at Facility C were younger and more frequently classified as ASA-PS I. Facility A had a higher median intraoperative net fluid balance than Facilities B and C. Most preoperative laboratory measurements showed limited variation across facilities. Complete facility-specific characteristics are provided in Table S3 in Multimedia Appendix 2.

In the secondary random hold-out analysis, outcome prevalence was similar between the development and hold-out test sets. In the temporal validation analysis, event rates and most baseline and perioperative characteristics were also comparable between periods, although modest differences were observed for serum sodium and potassium. Detailed cohort characteristics are provided in Tables S4 and S5 in Multimedia Appendix 2.

The distribution of transfused blood products is shown in Figure S1 in Multimedia Appendix 3, with median transfusion volumes of 4 units (IQR 2 - 6) for FFP, 10 units (IQR 8 - 16) for platelets, and 2 units (IQR 1 - 4) for RBCs.

Evaluation of Prediction Models

Using pooled out-of-sample predictions from leave-one-facility-out IECV, the perioperative models consistently demonstrated better discrimination than the corresponding preoperative models (Tables 2 and 3 and Figure 2). AUROC increased from 0.784 - 0.791 for the preoperative models to 0.810 - 0.817 for the perioperative models, while PR-AUC increased from 0.356 - 0.361 to 0.405 - 0.413. The increases in AUROC were 0.025 (95% CI 0.021 - 0.030) for LR, 0.026 (95% CI 0.021 - 0.030) for XGB, and 0.025 (95% CI 0.021 - 0.030) for LGBM. The corresponding increases in PR-AUC were 0.053 (95% CI 0.043 - 0.062), 0.055 (95% CI 0.045 - 0.065), and 0.049 (95% CI 0.039 - 0.058), respectively (all P<.001).

Table . Performance of preoperative and perioperative prediction models based on pooled out-of-sample predictions from internal-external cross-validation.

Model and feature set	AUROC[a] (95% CI)	Difference in AUROC (95% CI)[b]	PR-AUC[c] (95% CI)	Difference in PR-AUC (95% CI)[b]
Logistic regression				
Preoperative	0.784 (0.777 - 0.793)	—[d]	0.361 (0.347 - 0.377)	—
Perioperative	0.810 (0.802 - 0.817)	+0.025 (0.021 - 0.030)	0.413 (0.399 - 0.429)	+0.053 (0.043 - 0.062)
XGB[e]				
Preoperative	0.791 (0.783 - 0.799)	—	0.358 (0.344 - 0.373)	—
Perioperative	0.817 (0.809 - 0.824)	+0.026 (0.021 - 0.030)	0.413 (0.398 - 0.429)	+0.055 (0.045 - 0.065)
LGBM[f]				
Preoperative	0.788 (0.780 - 0.796)	—	0.356 (0.342 - 0.370)	—
Perioperative	0.813 (0.805 - 0.820)	+0.025 (0.021 - 0.030)	0.405 (0.389 - 0.419)	+0.049 (0.039 - 0.058)

[a]AUROC: area under the receiver operating characteristic curve.

[b]P<.001

[c]PR-AUC: area under the precision-recall curve.

[d]Not applicable.

[e]XGB: extreme gradient boosting.

[f]LGBM: light gradient boosting machine.

XSL•FO
RenderX

Table . Overall probabilistic performance and calibration.

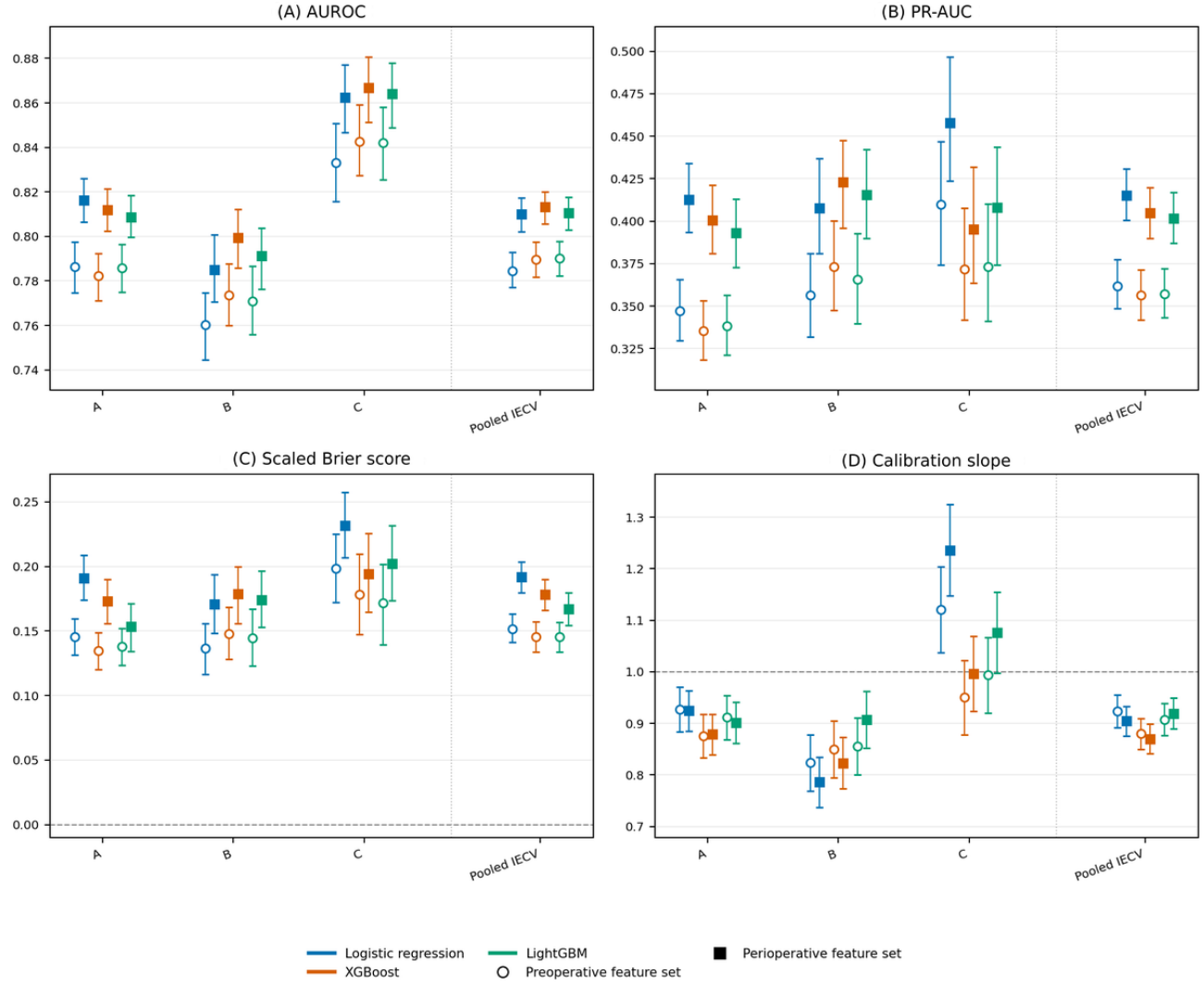
Model and feature set	Brier score (95% CI)	Scaled Brier score (95% CI)	Calibration slope (95% CI)	Calibration intercept (95% CI)
Logistic regression				
Preoperative	0.082 (0.081 to 0.083)	0.151 (0.141 to 0.163)	0.921 (0.887 to 0.960)	−0.144 (−0.205 to −0.075)
Perioperative	0.078 (0.077 to 0.080)	0.191 (0.179 to 0.203)	0.899 (0.868 to 0.928)	−0.222 (−0.272 to −0.170)
XGB[a]				
Preoperative	0.083 (0.081 to 0.084)	0.148 (0.136 to 0.160)	0.885 (0.853 to 0.919)	−0.237 (−0.292 to −0.180)
Perioperative	0.079 (0.078 to 0.080)	0.187 (0.174 to 0.199)	0.875 (0.840 to 0.908)	−0.301 (−0.357 to −0.250)
LGBM[b]				
Preoperative	0.083 (0.082 to 0.084)	0.144 (0.133 to 0.155)	0.913 (0.882 to 0.948)	−0.287 (−0.339 to −0.232)
Perioperative	0.080 (0.079 to 0.081)	0.175 (0.162 to 0.187)	0.902 (0.870 to 0.932)	−0.367 (−0.415 to −0.319)

[a]XGB: extreme gradient boosting.

[b]LGBM: light gradient boosting machine.

XSL•FO
RenderX

Figure 2. Performance of preoperative and perioperative prediction models in internal-external cross-validation (IECV). Facility-specific estimates were calculated from predictions for each held-out facility, and pooled internal-external cross-validation estimates were calculated by combining all out-of-sample predictions. The panels show (A) the area under the receiver operating characteristic curve (AUROC), (B) the area under the precision-recall curve, calculated using average precision (PR-AUC), (C) the scaled Brier score, and (D) the calibration slope. Filled squares indicate models using the perioperative feature set, and open circles indicate models using the preoperative feature set. Blue, orange, and green markers represent logistic regression (LR), extreme gradient boosting (XGB), and light gradient boosting machine (LGBM), respectively. Error bars indicate 95% CIs. The vertical dotted line separates facility-specific results from the pooled estimates. The horizontal reference lines indicate a scaled Brier score of 0, corresponding to a prevalence-based null model, and the ideal calibration slope of 1. Higher values indicate better performance for the discrimination metrics and scaled Brier score, whereas calibration slopes closer to 1 indicate better calibration.



In Tables 2 and 3, values are based on pooled out-of-sample predictions obtained from leave-one-facility-out IECV. In each fold, models were developed and calibrated using data from 2 institutions and evaluated in the remaining institution without model updating. The pooled analysis included 35,702 surgical admissions and 3884 primary composite end point events. Differences were calculated as the performance of the perioperative model minus that of the preoperative model within the same algorithm. A PR-AUC was calculated using average precision. The 95% CIs for individual performance estimates were calculated using 1000 stratified bootstrap resamples. Differences in AUROC and their 95% CIs were evaluated using the paired DeLong method. Differences in PR-AUC were evaluated using paired bootstrap resampling stratified by held-out institution and outcome status. Percentile 95% CIs were calculated from 10,000 bootstrap resamples, and 2-sided *P* values were obtained from the null-centered bootstrap distribution of the paired difference. *P* values were interpreted descriptively and were not adjusted for multiple comparisons. Between-algorithm differences were not formally tested because the prespecified comparison concerned the incremental value of perioperative predictors within each algorithm. Lower Brier scores and higher scaled Brier scores indicate better overall probabilistic performance. A scaled Brier score of 0 indicates performance equivalent to a prevalence-based null model, whereas a value of 1 indicates perfect prediction. The ideal values for the calibration slope and calibration intercept are 1 and 0, respectively.

The addition of perioperative predictors also improved overall probabilistic performance. Brier scores decreased from 0.082 - 0.083 to 0.078 - 0.080, whereas scaled Brier scores increased from 0.144 - 0.151 to 0.175 - 0.191. However, calibration remained imperfect. The pooled calibration slopes of the perioperative models ranged from 0.875 to 0.902, and

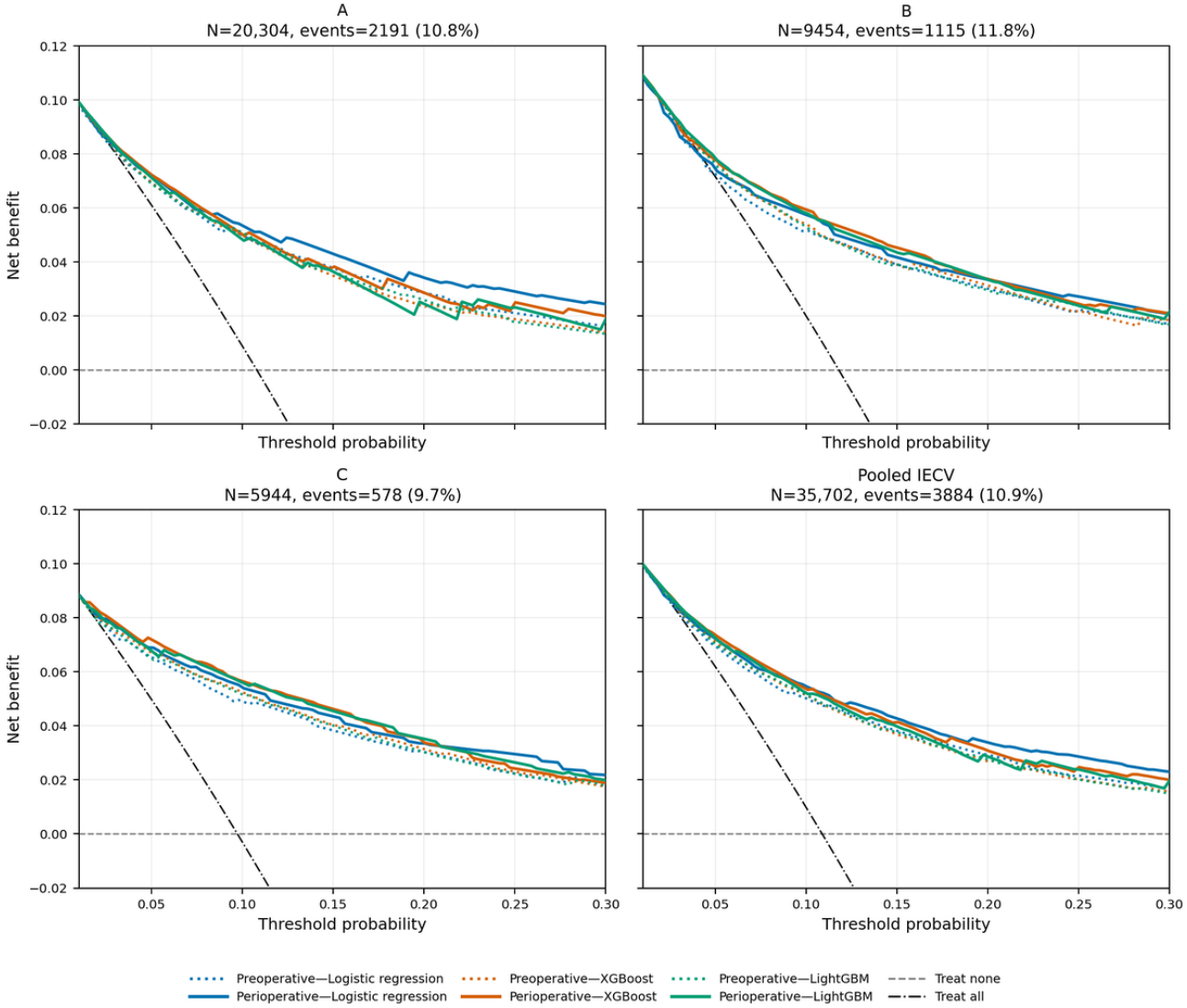
XSL•FO
RenderX

calibration intercepts ranged from −0.367 to −0.222, indicating that the predicted probabilities were somewhat too extreme and tended to overestimate the overall event risk.

The facility-specific IECV results were directionally consistent, although performance and calibration varied across the held-out institutions (Table S6 in Multimedia Appendix 2 and Figure S2 in Multimedia Appendix 3). The perioperative models had higher AUROCs at all 3 institutions. Improvements in PR-AUC were more apparent at Facilities A and B and were smaller or dependent on the algorithm at Facility C. Calibration intercepts were generally negative at Facilities A and B. At Facility C, the perioperative XGB and LGBM models were relatively close to ideal calibration, whereas the perioperative LR model showed greater deviation, with a calibration slope of 1.238 and an intercept of 0.582.

Decision curve analysis based on the IECV predictions showed that all models provided greater net benefit than the treat-none strategy across the evaluated threshold probability range of 0.01 - 0.30 (Figure 3). The perioperative models generally yielded higher net benefit than the corresponding preoperative models, particularly at thresholds above approximately 0.05, although the curves overlapped substantially and differences between algorithms were small. This pattern was consistent in the pooled IECV analysis and across the 3 held-out facilities. The treat-all strategy provided lower net benefit than the prediction models and became unfavorable at thresholds near or above the observed event prevalence.

Figure 3. Decision curve analysis of preoperative and perioperative prediction models based on internal-external cross-validation (IECV). Net benefit is shown across threshold probabilities from 0.01 to 0.30 for each held-out facility and for pooled out-of-sample predictions. Solid lines represent perioperative models, and dotted lines represent preoperative models. The dashed gray line indicates the treat-none strategy, and the black dash-dotted line indicates the treat-all strategy. LGBM: light gradient boosting machine; XGB: extreme gradient boosting.



Secondary and Sensitivity Analyses

The secondary random hold-out validation yielded results consistent with the primary IECV analysis (Table S7 in Multimedia Appendix 2). Across algorithms, AUROC increased from 0.787 - 0.804 with the preoperative feature set to 0.808 - 0.825 with the perioperative feature set, and PR-AUC increased from 0.363 - 0.395 to 0.410 - 0.450. Perioperative models also had lower Brier scores and higher scaled Brier scores, with calibration slopes ranging from 0.887 to 0.954.

XSL•FO
RenderX

The multiple-imputation sensitivity analysis of the LGBM models yielded results consistent with the primary IECV analysis (Table S9 in Multimedia Appendix 2). The perioperative model showed higher discrimination than the preoperative model, with an AUROC of 0.814 (95% CI 0.807 - 0.821) versus 0.788 (95% CI 0.781 - 0.796) and a PR-AUC of 0.413 (95% CI 0.397 - 0.429) versus 0.364 (95% CI 0.350 - 0.379). The perioperative model also had a lower Brier score (0.080 vs 0.083) and a higher scaled Brier score (0.178 vs 0.145). Calibration slopes were similar between the perioperative and preoperative models (0.917 and 0.919, respectively), although the calibration intercepts remained negative (−0.344 and −0.275, respectively).

In the outcome-specific analysis using LGBM, perioperative variables improved discrimination for both postoperative infection and the composite end point excluding postoperative infection (Table S8 in Multimedia Appendix 2). For the composite end point excluding postoperative infection, AUROC increased from 0.781 to 0.839 and PR-AUC from 0.173 to 0.249. For postoperative infection, AUROC increased from 0.788 to 0.809 and PR-AUC from 0.320 to 0.344. However, calibration was suboptimal for the composite end point excluding postoperative infection, particularly for the perioperative model, which had a calibration slope of 0.595 and an intercept of −1.436.

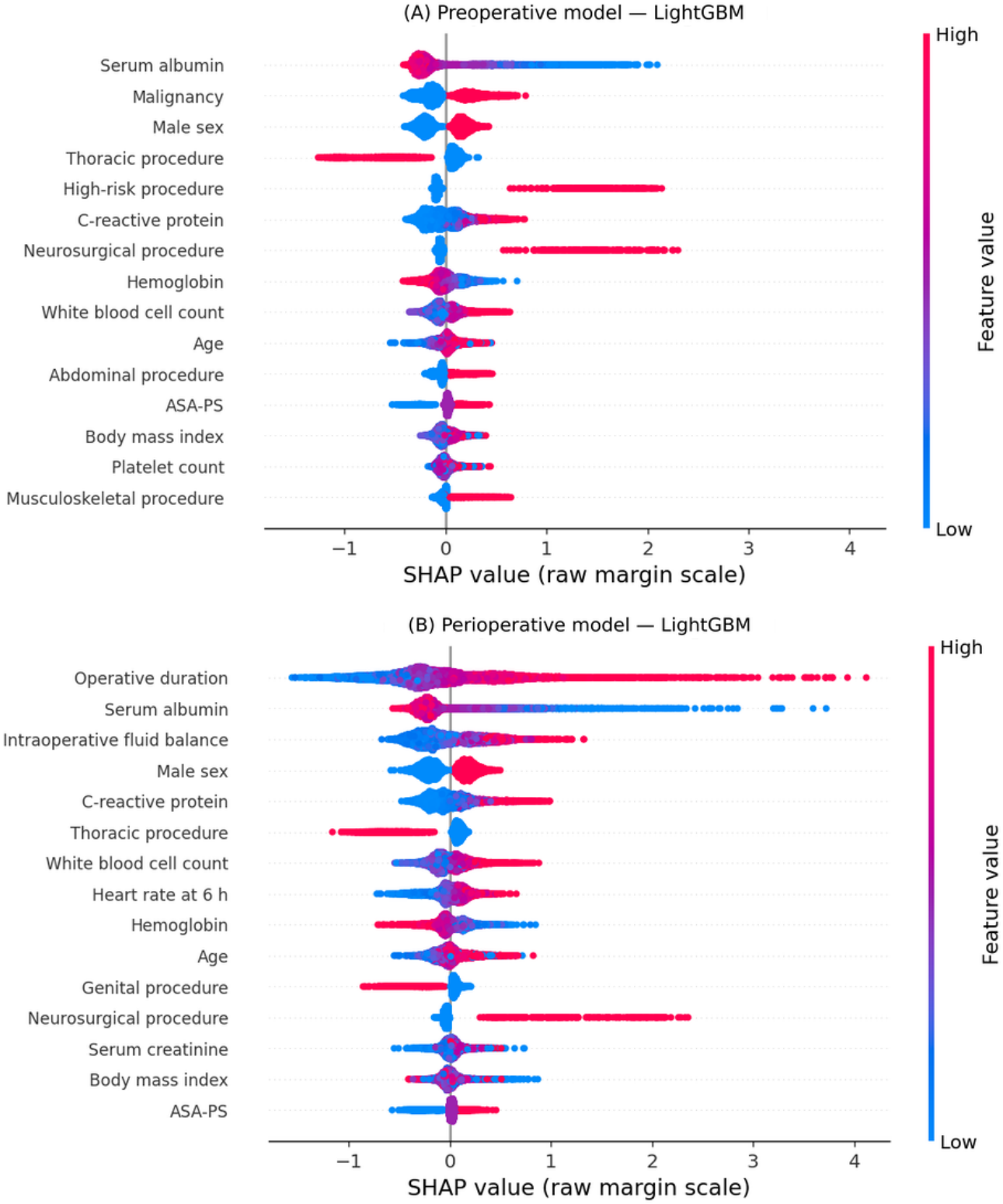
Temporal validation without model updating also supported the main findings (Table S10 in Multimedia Appendix 2). Across algorithms, perioperative models achieved AUROCs of 0.826 - 0.838 and PR-AUCs of 0.424 - 0.466, compared with 0.802 - 0.810 and 0.374 - 0.399, respectively, for the preoperative models. Calibration remained close to ideal, with slopes ranging from 1.000 to 1.048 and intercepts from −0.045 to 0.034 for the perioperative models.

Model Interpretation

SHAP analyses showed broadly consistent patterns across LR, XGB, and LGBM models (Figure 4 and Figures S3 and S4 in Multimedia Appendix 3). In the preoperative models, serum albumin, malignancy, male sex, C-reactive protein, white blood cell count, hemoglobin, ASA-PS, and surgical procedure category were among the most influential predictors. Lower serum albumin and hemoglobin levels and higher C-reactive protein and white blood cell counts generally shifted predictions toward a higher risk of the composite end point. High-risk and neurosurgical procedures also contributed to higher predicted risk, particularly in the tree-based models.

XSL•FO
RenderX

Figure 4. Shapley additive explanation (SHAP) summary plots for the light gradient boosting machine (LightGBM) models. (A) Preoperative model. (B) Perioperative model. Features are ranked by their mean absolute SHAP values, with the most influential features shown at the top. Each point represents 1 patient. The horizontal position indicates the contribution of the feature to the model output on the raw margin scale; positive SHAP values shift the prediction toward a higher risk of the composite end point, whereas negative values shift the prediction toward a lower risk. Point color represents the observed feature value, ranging from low (blue) to high (red). For binary variables, blue and red generally indicate absence and presence, respectively. ASA-PS: American Society of Anesthesiologists Physical Status.



After perioperative variables were added, operative duration became the most influential predictor across all 3 algorithms, with longer procedures shifting predictions toward higher risk. A larger intraoperative fluid balance and a higher heart rate at 6 hours after surgery also emerged as important predictors of higher risk. Serum albumin remained one of the leading predictors, with higher values consistently shifting predictions toward lower risk. Male sex, inflammatory markers, hemoglobin, and surgical procedure categories also remained influential. These findings indicate that the perioperative models integrated

XSL•FO
RenderX

measures of operative burden and early postoperative physiology with preoperative patient vulnerability. SHAP values describe the contribution of individual features to model predictions and should not be interpreted as causal effects.

Feature interaction patterns are shown in Figure S5 in Multimedia Appendix 3. In the preoperative models, prominent interactions involved serum albumin, malignancy, inflammatory markers, and surgical procedure categories. After perioperative variables were added, operative duration emerged as the central interaction feature in both XGB and LGBM, with interactions involving intraoperative fluid balance, serum albumin, age, C-reactive protein, and surgical procedure category. Early postoperative physiological variables, including heart rate and mean arterial pressure at 6 hours, also contributed to the interaction structure.

Discussion

Principal Findings

This multicenter study showed that models incorporating routinely available intraoperative and early postoperative data consistently improved predictive performance over preoperative-only models in the primary IECV analysis. Across all 3 algorithms, perioperative models achieved modest but reproducible improvements in discrimination, with AUROC increasing from 0.784 - 0.791 to 0.810 - 0.817 and PR-AUC increasing from 0.356 - 0.361 to 0.405 - 0.413. Overall probabilistic performance also improved, as reflected by lower Brier scores and higher scaled Brier scores. Decision curve analysis showed an incremental net benefit of the perioperative models across the evaluated threshold range, although the substantial overlap between curves suggested that the magnitude of improvement was limited. These findings were directionally consistent in the random hold-out validation, temporal validation, and outcome-specific sensitivity analyses. However, calibration varied across institutions, and the pooled perioperative models had calibration slopes below 1 and negative calibration intercepts, indicating predictions that were somewhat too extreme and tended to overestimate risk. Taken together, these findings support the incremental value of routinely collected perioperative information for postoperative risk stratification, while indicating that local recalibration may be required to ensure accurate absolute risk estimation across institutions.

Our findings corroborate and extend prior work on perioperative risk prediction. The PERISCOPE study used gradient boosting to predict postoperative infection (AUROC 0.82) but required site-specific intercept adjustment before deployment [24]. Zhuang et al [31] developed parsimonious preoperative models across more than 30,000 surgical cases (AUROC 0.73 - 0.89), likewise emphasizing the need for local recalibration. The MySurgeryRisk (University of Florida) algorithm, trained on nearly 500,000 cases from 8 US hospitals, achieved AUROCs of 0.79 - 0.92 but demonstrated intersite calibration error of approximately 0.15, prompting isotonic recalibration [22]. Similarly, DeepSurv-OR, a transformer-based intraoperative model, reached an AUROC of 0.90 yet showed greater than 10% divergence in calibration across institutions [32]. Notably, most of these studies relied on AUROC-based evaluation, which is not directly comparable to PR-AUC-based assessment in low-event-rate populations. Collectively, these studies, and our multicenter validation, highlight that even high-performing models often lose calibration after transfer, making site-specific recalibration a prerequisite for safe and reliable clinical implementation.

Among baseline factors, hypoalbuminemia was the strongest predictor, consistent with prior meta-analyses across surgical specialties [33,34]. Albumin remained highly influential after incorporating intraoperative and early postoperative variables, confirming its role as a baseline determinant of surgical stress tolerance. Previous multicenter ML studies integrating high-resolution intraoperative waveform data improved PR-AUC to approximately 0.41 [32], and models combining respiratory and hemodynamic features reached PR-AUC values of approximately 0.45 for pulmonary complications [35]. Together, these findings illustrate the complementary contribution of baseline patient characteristics, such as albumin, and intraoperative and early postoperative physiological variables to postoperative risk stratification. These findings may support the evaluation of locally calibrated thresholds for prioritizing postoperative clinical reassessment, but they do not identify a specific management action for each predicted risk level. However, because the majority of patients in this cohort were classified as ASA-PS I-II, generalizability to higher-risk populations or emergency surgical settings remains to be established.

Figure 3 shows that the perioperative models provided greater net benefit than the preoperative models across a range of threshold probabilities. In this exploratory decision curve analysis, a positive prediction represented prioritization for postoperative clinical reassessment rather than automatic initiation of intensive care unit admission, infection screening, or organ support. Accordingly, the evaluated thresholds should be interpreted as reflecting trade-offs between the number of patients selected for reassessment and the number of postoperative complications missed, rather than as prescriptive treatment thresholds. Two threshold regions were particularly informative. Lower thresholds yielded high sensitivity but selected a larger proportion of patients for reassessment, whereas intermediate thresholds reduced the number of patients selected while increasing positive predictive value. These findings illustrate the potential trade-offs associated with different thresholds; however, they do not establish an optimal threshold for clinical use. Threshold selection would require prospective evaluation incorporating the intended clinical workflow, available resources, stakeholder preferences, and the consequences of false-positive and false-negative predictions. These patterns align with the original decision curve framework proposed by Vickers and Elkin [36] and the "DCA-to-threshold-to-implementation" paradigm described by Steyerberg et al [37]. Given the observed calibration drift between hospitals, site-specific recalibration, using intercept and slope adjustment or isotonic regression [38], together with stakeholder-guided threshold selection [39] and continuous postdeployment monitoring [40], is recommended to preserve clinical utility over time.

XSL•FO
RenderX

Clinically, these perioperative ML models may support early postoperative risk stratification by identifying patients who warrant earlier or more detailed clinical reassessment. The model output is not intended to directly determine intensive care unit admission, infection screening, reintubation, renal replacement therapy, or other specific interventions. Instead, subsequent monitoring, diagnostic testing, and treatment should be selected according to the patient's evolving clinical condition and the complication suspected. Similarly, a low predicted risk should not be used in isolation to determine postoperative ward placement or to reduce usual clinical monitoring. Because all predictors used in the current models are routinely documented in anesthesia and nursing records, integration into existing EHR dashboards is technically feasible without additional data collection. Such locally calibrated tools are intended to complement, rather than replace, clinical judgment by providing additional information for risk-informed reassessment. Their effects on workflow, resource use, clinical decisions, and patient outcomes require prospective evaluation before implementation. For broader clinical adoption, future work should prioritize prospective validation and seamless integration into anesthesia information systems. Embedding automated risk alerts within perioperative dashboards may facilitate real-time decision support while maintaining human oversight and institutional calibration to ensure safe, transparent, and sustainable implementation.

This study has several limitations. First, although IECV was performed, the three participating hospitals were tertiary referral centers within the same academic network operating under a shared DPC-based bundled payment system. Accordingly, this validation should be considered internal-external (semiexternal) rather than fully external, and model performance and calibration may differ in community hospitals or international health care systems. The observed variation in calibration across facilities also indicates that local recalibration may be required before implementation. Second, retrospective use of administrative and EHR data carries inherent risks of miscoding, residual confounding, and incomplete capture of clinical context, and excludes high-resolution intraoperative signals, such as continuous blood pressure waveforms or vasoactive drug infusion rates, that could further enhance predictive performance.

Third, missing values were imputed using an LGBM-based iterative imputation approach that assumes data are missing at random; residual bias due to nonrandom missingness cannot be excluded. Although the multiple-imputation sensitivity analysis yielded findings consistent with the primary analysis, single imputation does not account for uncertainty arising from the imputation process and may therefore underestimate uncertainty related to missing data.

Fourth, the composite end point combined heterogeneous postoperative events and was predominantly composed of postoperative infection, potentially obscuring event-specific predictor-outcome relationships. Although analyses of postoperative infection and the composite end point excluding postoperative infection supported the main findings, the small numbers of several individual components precluded the development of sufficiently stable component-specific models. Therefore, the reported performance should not be assumed to apply equally to each outcome component.

Finally, algorithmic bias related to age, frailty, or differential access to postoperative care cannot be excluded and warrants careful evaluation during real-world deployment. The models have not yet been prospectively integrated into clinical workflows, leaving real-time safety and effectiveness untested. Future studies should include broader multicenter validation, incorporation of richer physiologic data, and prospective impact assessment.

Conclusions

In this multicenter study of adult patients undergoing noncardiac surgery under general anesthesia, perioperative models incorporating routinely available intraoperative variables and early postoperative vital signs improved prediction of major postoperative complications compared with preoperative models. Exploratory decision curve analysis suggested potential value for prioritizing postoperative clinical reassessment but did not establish utility for a specific intervention or triage decision. Institutional variation in calibration indicates that local recalibration, prospective workflow evaluation, and assessment of predefined responses to model output are required before clinical implementation.

Acknowledgments

The authors sincerely thank all the medical staff, technicians, medical engineers, nurses, pharmacists, and physicians at Nippon Medical School Hospital, Nippon Medical School Chiba Hokusoh Hospital, and Nippon Medical School Musashi Kosugi Hospital for their dedicated care of patients.

Generative AI Use

ChatGPT (OpenAI) was used to assist with the development of auxiliary analytic scripts and language editing during manuscript revision. It was not used to generate study data, independently conduct the primary statistical analyses, or determine the scientific conclusions. All AI-assisted outputs were critically reviewed and revised by the authors, who take full responsibility for the final manuscript.

Funding

Support was provided solely from institutional and/or departmental sources.

XSL•FO
RenderX

Data Availability

The clinical data underlying this study cannot be made publicly available owing to Japanese privacy regulations and institutional policy. Deidentified data may be shared for legitimate academic purposes upon reasonable request to the corresponding author and with prior approval of the Central Ethics Review Committee of Nippon Medical School, Japan. All analytic code is openly available at the authors' GitHub repository [41].

Authors' Contributions

Conceptualization: TN, MI, YK
Data curation: HM, CK, SS
Funding acquisition: MI
Investigation: TN, MI
Methodology: TN, MI
Project administration: TN, MI
Supervision: HM, CK, SS, YK
Writing – original draft: TN
Writing – review & editing: MI

Conflicts of Interest

None declared.

Multimedia Appendix 1
Definitions and preprocessing of predictor variables, hyperparameter search spaces, and selected model hyperparameters.
[XLSX File, 23 KB - periop_v9i1e91315_app1.xlsx]

Multimedia Appendix 2
Tables reporting cohort characteristics, validation cohort composition, and model performance across internal-external cross-validation, random hold-out validation, alternative outcome analyses, multiple imputation, and temporal validation.
[DOCX File, 62 KB - periop_v9i1e91315_app2.docx]

Multimedia Appendix 3
Figures showing transfusion distributions, calibration plots, Shapley additive explanations (SHAP) summary plots, and pairwise SHAP interaction heatmaps.
[DOCX File, 1808 KB - periop_v9i1e91315_app3.docx]

References

1. Weiser TG, Haynes AB, Molina G, et al. Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes. Lancet 2015 Apr;385:S11. [doi: 10.1016/S0140-6736(15)60806-6]
2. Alkire BC, Raykar NP, Shrime MG, et al. Global access to surgical care: a modelling study. Lancet Glob Health 2015 Jun;3(6):e316-e323. [doi: 10.1016/S2214-109X(15)70115-4] [Medline: 25926087]
3. Tevis SE, Kennedy GD. Postoperative complications and implications on patient-centered outcomes. J Surg Res 2013 May 1;181(1):106-113. [doi: 10.1016/j.jss.2013.01.032] [Medline: 23465392]
4. The International Surgical Outcomes Study group. Global patient outcomes after elective surgery: prospective cohort study in 27 low-, middle- and high-income countries. Br J Anaesth 2016 Nov;117(5):601-609. [doi: 10.1093/bja/aew316]
5. Vascular events in noncardiac surgery patients cohort evaluation (VISION) study investigators, Spence J, LeManach Y, et al. Association between complications and death within 30 days after noncardiac surgery. CMAJ 2019 Jul 29;191(30):E830-E837. [doi: 10.1503/cmaj.190221] [Medline: 31358597]
6. Tjeertes EKM, Ultee KHJ, Stolker RJ, et al. Perioperative complications are associated with adverse long-term prognosis and affect the cause of death after general surgery. World J Surg 2016 Nov;40(11):2581-2590. [doi: 10.1007/s00268-016-3600-4] [Medline: 27302465]
7. Nepogodiev D, Martin J, Biccard B, Makupe A, Bhangu A, National Institute for Health Research Global Health Research Unit on Global Surgery. Global burden of postoperative death. Lancet 2019 Feb 2;393(10170):401. [doi: 10.1016/S0140-6736(18)33139-8] [Medline: 30722955]
8. Prowle JR, Forni LG, Bell M, et al. Postoperative acute kidney injury in adult non-cardiac surgery: joint consensus report of the Acute Disease Quality Initiative and PeriOperative Quality Initiative. Nat Rev Nephrol 2021 Sep;17(9):605-618. [doi: 10.1038/s41581-021-00418-2] [Medline: 33976395]
9. Miskovic A, Lumb AB. Postoperative pulmonary complications. Br J Anaesth 2017 Mar 1;118(3):317-334. [doi: 10.1093/bja/aex002] [Medline: 28186222]

XSL•FO
RenderX

10. Kang DH, Kim YJ, Kim SH, et al. Early surgery versus conventional treatment for infective endocarditis. N Engl J Med 2012 Jun 28;366(26):2466-2473. [doi: 10.1056/NEJMoa1112843] [Medline: 22738096]
11. Leeds IL, Truta B, Parian AM, et al. Early surgical intervention for acute ulcerative colitis is associated with improved postoperative outcomes. J Gastrointest Surg 2017 Oct;21(10):1675-1682. [doi: 10.1007/s11605-017-3538-3] [Medline: 28819916]
12. Mayhew D, Mendonca V, Murthy BVS. A review of ASA physical status - historical perspectives and modern developments. Anaesthesia 2019 Mar;74(3):373-379. [doi: 10.1111/anae.14569] [Medline: 30648259]
13. Lee TH, Marcantonio ER, Mangione CM, et al. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. Circulation 1999 Sep 7;100(10):1043-1049. [doi: 10.1161/01.cir.100.10.1043] [Medline: 10477528]
14. Bilimoria KY, Liu Y, Paruch JL, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. J Am Coll Surg 2013 Nov;217(5):833-842. [doi: 10.1016/j.jamcollsurg.2013.07.385] [Medline: 24055383]
15. Andaur Navarro CL, Damen JAA, van Smeden M, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. J Clin Epidemiol 2023 Feb;154:8-22. [doi: 10.1016/j.jclinepi.2022.11.015] [Medline: 36436815]
16. Fritz BA, King CR, Abdelhack M, et al. Effect of machine learning models on clinician prediction of postoperative complications: the Perioperative ORACLE randomised clinical trial. Br J Anaesth 2024 Nov;133(5):1042-1050. [doi: 10.1016/j.bja.2024.08.004] [Medline: 39261226]
17. Hill BL, Brown R, Gabel E, et al. An automated machine learning-based model predicts postoperative mortality using readily-extractable preoperative electronic health record data. Br J Anaesth 2019 Dec;123(6):877-886. [doi: 10.1016/j.bja.2019.07.030] [Medline: 31627890]
18. Chiew CJ, Liu N, Wong TH, Sim YE, Abdullah HR. Utilizing machine learning methods for preoperative prediction of postsurgical mortality and intensive care unit admission. Ann Surg 2020 Dec;272(6):1133-1139. [doi: 10.1097/SLA.0000000000003297] [Medline: 30973386]
19. Mahajan A, Esper S, Oo TH, et al. Development and validation of a machine learning model to identify patients before surgery at high risk for postoperative adverse events. JAMA Netw Open 2023 Jul 3;6(7):e2322285. [doi: 10.1001/jamanetworkopen.2023.22285] [Medline: 37418262]
20. Blatter TU, Wintsch YM, Triep K, Endrich O, Guillen-Ramirez H, Beldi G. End-of-surgery prediction of postoperative infectious complications from intraoperative vital-sign dynamics. NPJ Digit Med 2026 May 7;9(1):760. [doi: 10.1038/s41746-026-02707-1] [Medline: 42092125]
21. Collins GS, Moons KGM, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:q902. [doi: 10.1136/bmj.q902]
22. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, et al. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. Ann Surg 2019 Apr;269(4):652-662. [doi: 10.1097/SLA.0000000000002706] [Medline: 29489489]
23. Jammer I, Wickboldt N, Sander M, et al. Standards for definitions and use of outcome measures for clinical effectiveness research in perioperative medicine: European Perioperative Clinical Outcome (EPCO) definitions: a statement from the ESA-ESICM joint taskforce on perioperative outcome measures. Eur J Anaesthesiol 2015 Feb;32(2):88-105. [doi: 10.1097/EJA.0000000000000118] [Medline: 25058504]
24. van der Meijden SL, van Boekel AM, Schinkelshoek LJ, et al. Development and validation of artificial intelligence models for early detection of postoperative infections (PERISCOPE): a multicentre study using electronic health record data. Lancet Reg Health Eur 2025 Feb;49:101163. [doi: 10.1016/j.lanepe.2024.101163] [Medline: 39720095]
25. Hayashida K, Murakami G, Matsuda S, Fushimi K. History and profile of Diagnosis Procedure Combination (DPC): development of a real data collection system for acute inpatient care in Japan. J Epidemiol 2021 Jan 5;31(1):1-11. [doi: 10.2188/jea.JE20200288] [Medline: 33012777]
26. Hiraoka E, Tanabe K, Izuta S, et al. JCS 2022 guideline on perioperative cardiovascular assessment and management for non-cardiac surgery. Circ J 2023 Aug 25;87(9):1253-1337. [doi: 10.1253/circj.CJ-22-0609] [Medline: 37558469]
27. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. Presented at: Advances in Neural Information Processing Systems 30 (NeurIPS 2017); Dec 4-9, 2017; Long Beach, CA p. 3146-3154 URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html [accessed 2026-10-14]
28. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Presented at: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2016); Aug 13-17, 2026; San Francisco, California, USA p. 785-794 URL: https://dl.acm.org/doi/10.1145/2939672.2939785 [accessed 2026-10-08] [doi: 10.1145/2939672.2939785]
29. Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. 2002 Jul 23 Presented at: KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining;

XSL•FO
RenderX

Jul 23-26, 2002; Edmonton, Alberta, Canada p. 694-699 URL: https://dl.acm.org/doi/proceedings/10.1145/775047 [accessed 2026-10-08] [doi: 10.1145/775047.775151]
30. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Presented at: Advances in Neural Information Processing Systems 30 (NeurIPS 2017); Dec 4-9, 2017; Long Beach, CA p. 4765-4774 URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html [accessed 2026-10-14]
31. Zhuang Y, Dyas A, Meguid RA, et al. Preoperative prediction of postoperative infections using machine learning and electronic health record data. Ann Surg 2024 Apr 1;279(4):720-726. [doi: 10.1097/SLA.0000000000006106] [Medline: 37753703]
32. Shickel B, Loftus TJ, Ruppert M, et al. Dynamic predictions of postoperative complications from explainable, uncertainty-aware, and multi-task deep neural networks. Sci Rep 2023 Jan 21;13(1):1224. [doi: 10.1038/s41598-023-27418-5] [Medline: 36681755]
33. Mostafa OE, Al-Allaf O, Tahir M, Hossain F, Blackwell J. Do hypoalbuminaemia increase the risk of surgical site infection in neck of femur fracture patients: a systematic review and meta-analysis. Cureus 2024 May;16(5):e61372. [doi: 10.7759/cureus.61372] [Medline: 38817798]
34. Petch-In P, Saokaew S, Phisalprapa P, Dilokthornsakul P. The association of pre-operative serum albumin levels and post-operative in-hospital death in patients undergoing gastrointestinal surgeries in Thailand: a retrospective cohort study. Drugs Real World Outcomes 2023 Jun;10(2):341-349. [doi: 10.1007/s40801-023-00364-4] [Medline: 36972005]
35. Li P, Gao S, Wang Y, et al. Utilising intraoperative respiratory dynamic features for developing and validating an explainable machine learning model for postoperative pulmonary complications. Br J Anaesth 2024 Jun;132(6):1315-1326. [doi: 10.1016/j.bja.2024.02.025] [Medline: 38637267]
36. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26(6):565-574. [doi: 10.1177/0272989X06295361] [Medline: 17099194]
37. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014 Aug 1;35(29):1925-1931. [doi: 10.1093/eurheartj/ehu207] [Medline: 24898551]
38. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med 2019 Dec 16;17(1):230. [doi: 10.1186/s12916-019-1466-7] [Medline: 31842878]
39. Kerr KF, Brown MD, Zhu K, Janes H. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. J Clin Oncol 2016 Jul 20;34(21):2534-2540. [doi: 10.1200/JCO.2015.65.5654] [Medline: 27247223]
40. Fitzgerald M, Saville BR, Lewis RJ. Decision curve analysis. JAMA 2015 Jan 27;313(4):409-410. [doi: 10.1001/jama.2015.37] [Medline: 25626037]
41. General anesthesia surgery: multicenter perioperative machine learning analysis. GitHub. URL: https://github.com/Takuya-Nishino/general-anesthesia-surgery [accessed 2026-10-14]

Abbreviations

ASA-PS: American Society of Anesthesiologists physical status
AUROC: area under the receiver operating characteristic curve
DPC: Diagnosis Procedure Combination
EHR: electronic health record
FFP: fresh frozen plasma
IECV: internal-external cross-validation
LGBM: light gradient boosting machine
LR: logistic regression
ML: machine learning
PR-AUC: area under the precision-recall curve
RBC: red blood cell
SHAP: Shapley additive explanations
SMD: standardized mean difference
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+Artificial Intelligence
XGB: extreme gradient boosting

XSL•FO
RenderX

Edited by N Rohatgi; submitted 12.Jan.2026; peer-reviewed by K Pang, K Matsumoto; revised version received 14.Sep.2026; accepted 25.Sep.2026; published 21.Oct.2026.

Please cite as:
Nishino T, Mase H, Kim C, Sugita S, Kondo Y, Ishikawa M
Machine Learning–Based Prediction of Major Postoperative Complications After Noncardiac Surgery Under General Anesthesia Using Routinely Available Electronic Health Record Data: Retrospective Multicenter Development and Validation Study
JMIR Perioper Med 2026;9:e91315
URL: https://periop.jmir.org/2026/1/e91315
doi:10.2196/91315

© Takuya Nishino, Hiroshi Mase, Chol Kim, Shinji Sugita, Yukihiro Kondo, Masashi Ishikawa. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 21.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comparing the Quality of Patient Recovery After Open Appendectomy Under General Versus Spinal Anesthesia: Prospective Cohort Study

Negesse Zurbachew Gobezie[1], MSc; Temesgen Birlie Asmare[1], MSc; Kumlachew Geta Belete[1], MSc; Gezahagn Demsu Gedefaw[2], MSc; Biruk Demissie[3], MSc; Abere Gebru Abuhay[4], MSc; Belayneh Jejaw Abate[5], MPH; Birhanu Mengist Munie[6], MSc; Wubet Tazeb Wondie[7], MSc; Habtie Bantider Wubet[1], MSc

[1]Department of Anesthesia, School of Medicine, College of Health Science, Debre Tabor University, P.O. Box 272, Debre Tabor, Ethiopia

[2]Department of Neonatal Health Nursing, School of Nursing, College of Medicine and Health Sciences, and Specialized Hospital, University of Gondar, Gondar, Ethiopia

[3]Department of Environmental Health, College of Health Science, Debre Tabor University, Debre Tabor, Ethiopia

[4]Department of Pediatrics and Neonatal Nursing, College of Health Sciences, Debre Tabor University, Debre Tabor, Ethiopia

[5]University of Gondar Comprehensive Specialized Hospital, Gondar, Ethiopia

[6]Department of Psychiatry, College of Health Science, Debre Tabor University, Debre Tabor, Ethiopia

[7]Department of Pediatrics and Child Health Nursing, School of Health Sciences, College of Medicine and Health Sciences, Bahir Dar University, Bahir Dar, Ethiopia

Corresponding Author:
Negesse Zurbachew Gobezie, MSc
Department of Anesthesia, School of Medicine, College of Health Science, Debre Tabor University, P.O. Box 272, Debre Tabor, Ethiopia

Abstract

Background: Open appendectomy is commonly performed under either general or spinal anesthesia. Postoperative quality of recovery, a multidimensional patient-centered measure of outcomes after surgery, is affected by the choice of anesthesia technique.

Objective: The objective of our study was to compare the effects of general and spinal anesthesia on the immediate postoperative quality of recovery in adults undergoing open appendectomy.

Methods: In this prospective cohort study, 74 patients were assigned to either the general (n=37) or the spinal anesthesia (n=37) groups. The primary outcome was the total Quality of Recovery-15 score measured 24 hours postoperatively. Secondary outcomes included postoperative pain scores, analgesic consumption, incidence of postoperative nausea and vomiting, time to first oral intake, time to first ambulation, and length of hospital stay. Intergroup comparisons were performed using the chi-square test, Fisher exact test, independent *t* test, or Mann-Whitney *U* test, as appropriate.

Results: Patients in the spinal anesthesia group had significantly higher Quality of Recovery-15 scores than those in the general anesthesia group (P=.001). They also exhibited lower pain scores at 1, 2, 6, and 12 hours postoperatively (P<.001), reduced consumption of diclofenac and tramadol, and a decreased incidence of postoperative nausea and vomiting (P<.001). The time to first mobilization, time to first oral intake, and length of hospital stay were also significantly lower in the spinal anesthesia group (P<.001).

Conclusions: Spinal anesthesia is the preferred anesthesia technique for patients undergoing open appendectomy. It may provide improved postoperative recovery, with lower pain scores and improved analgesia. It may also reduce the need for additional analgesics, decrease the incidence of nausea and vomiting, and improve intrahospital patient recovery trajectories.

(JMIR Perioper Med 2026;9:e86273) doi:10.2196/86273

KEYWORDS

appendectomy; anesthesia; surgery; outcome; recovery; postoperative

Introduction

Acute appendicitis, defined as inflammation of the vermiform appendix, remains one of the most common causes of acute abdomen and emergency abdominal surgery worldwide [1,2]. Current epidemiological data indicate an increasing incidence of this condition in both middle-income and high-income countries [3,4]. Although laparoscopic appendectomy is increasingly used in high-resource settings [5,6], open appendectomy remains the predominant approach in

XSL•FO
RenderX

resource-limited settings and is performed under either general anesthesia (GA) with endotracheal intubation or spinal anesthesia (SA) [7-9].

GA is associated with several potential adverse effects, including airway complications during induction or extubation, hemodynamic instability, aspiration, and postoperative nausea and vomiting (PONV) [10-12]. SA has emerged as a viable alternative to GA, with increasing use in abdominal surgeries performed below the umbilicus, including open appendectomy [7,8,13,14]. The use of SA for abdominal surgeries is associated with superior postoperative pain control, fewer cardiorespiratory adverse events, and a lower incidence of PONV [15-17]. Nevertheless, these individual outcomes do not completely reflect the patient's overall recovery from anesthesia and surgery [18]. Consequently, evaluating the quality of recovery (QoR) from the patient's perspective has emerged as a critical metric for assessing the impact of anesthesia and surgery on overall patient recovery and well-being [19].

Postoperative QoR is a multidimensional outcome measure that encompasses physical comfort, emotional state, psychological support, physical independence, pain, and return to normal activities [20,21]. It has emerged as a holistic and patient-centered measure of surgical and anesthetic success compared to traditional, isolated metrics such as pain scores, duration of surgery, duration of anesthesia, length of hospital stay, and complication rates [22-24].

In addition to the surgical-related and patient factors such as age, comorbidities, nutritional status, and frailty status, the choice of anesthetic technique significantly influences postoperative outcomes and QoR [20]. Several studies have compared GA and SA across different surgical procedures and have frequently demonstrated the benefits of SA in terms of pain control, PONV, and length of hospital stay [25,26]. However, limited data exist regarding the impact of anesthetic techniques on postoperative QoR in patients undergoing open appendectomy.

Therefore, this study primarily aimed to compare the effects of GA and SA on immediate postoperative QoR in adult patients undergoing open appendectomy. The secondary objective of this study was to compare postoperative pain scores, analgesia consumption, incidence of PONV, and immediate intrahospital recovery trajectories such as time to first oral intake, first ambulation, and length of hospital stay. The findings of this study may provide evidence to guide anesthesia decision-making and optimize patient-centered outcomes in the surgical care for patients undergoing open appendectomy.

Methods

Overview

An institution-based prospective cohort study was conducted at Debre Tabor Comprehensive Specialized Hospital from December 1, 2024, to June 6, 2025. The hospital is located in Debre Tabor town, South Gondar Zone, Amhara Region, Northwest Ethiopia. It is a governmental hospital in Debre Tabor town that serves a population of approximately 3 million people. The hospital provides gynecologic, obstetric, orthopedic, and general abdominal and urologic surgical procedures on an elective and emergency basis. It has 2 operating rooms and 1 dedicated surgical ward with 32 beds, specifically allocated to patients undergoing general surgery.

Study Population

Adult surgical patients aged ≥18 years with American Society of Anesthesiologists physical status classification I or II, diagnosed with acute appendicitis, and who underwent open appendectomy were included in our study. Patients were excluded if they met any of the following criteria: American Society of Anesthesiologists physical status ≥III, pre-existing neuropsychiatric disorders impairing effective communication, periappendiceal abscess or generalized peritonitis requiring midline laparotomy, intraoperative discovery of additional pathologies necessitating concurrent surgical procedures, inadequate spinal block necessitating conversion to GA during the operation, or occurrence of unexpected intraoperative catastrophic events requiring postoperative admission to the intensive care unit.

Sample Size Determination and Sampling Procedure

The sample size for this study was calculated using a standard formula to compare means between 2 independent groups. The primary outcome of our study was the mean difference in the Quality of Recovery-15 (QoR-15) score at 24 hours postoperatively: $n=(\sigma_1^2+\sigma_2^2)(Z_\beta+Z_{\alpha/2})^2/(\mu_1-\mu_2)^2$, where n is the required sample size for each group; σ_1 and μ_1 are the SD and mean of the first group (SA), respectively; σ_2 and μ_2 are the SD and mean of the second group (GA), respectively; $Z_{\alpha/2}$ is the critical value for the desired confidence level; and Z_β is the critical value for the desired statistical power.

As no similar study had been conducted in the study setting, parameter estimates for sample size calculation were obtained from a pilot study (n=20) conducted 1 month before the actual data collection period. In the pilot study, the mean QoR-15 scores for the SA and GA groups were 123.9 (SD 22.4) and 111.2 (SD 14.6), respectively. Using a 95% significance level ($Z_{\alpha/2}=1.96$) and 80% power ($Z_\beta=0.84$): $n=[\{(22.4)^2+(14.6)^2\}(1.96+0.84)^2]/(123.9-111.2)^2=34.7\approx35$. After adding a 5% nonresponse rate, the final sample size was 37 participants per group, resulting in a total of 74 participants included in the study. Subsequently, all consecutive patients who met the inclusion criteria were included in the study until the required sample size was achieved.

Data Collection Procedure, Data Collection Tools, and Measurements

A structured questionnaire was prepared to collect information on sociodemographic characteristics, comorbidities, medication use, and anesthesia- and surgery-related variables during the preoperative, intraoperative, and postoperative periods. The original questionnaire was developed in English and was subsequently translated into Amharic. To ensure linguistic accuracy and maintain the original content, the questionnaire was back-translated into English by 2 bilingual language experts fluent in both English and Amharic. Data were collected through chart reviews and direct interviews with participants conducted

XSL•FO
RenderX

by anesthetists and nurses who were not directly involved in patient management.

In the study setting, anesthesia management for patients who underwent appendectomy was carried out by anesthesia professionals using either GA or SA following a standardized protocol. Preoperative evaluation, documentation of preoperative vital signs, and premedication with intravenous dexamethasone were performed in the anesthesia waiting area. For patients receiving GA, the available analgesics (1.5 mg/kg diclofenac and 1-2 mg/kg pethidine) were administered. Patients were then preoxygenated for 3-5 minutes and induced under rapid sequence induction with Ketofol (ketamine-propofol in a 1:2 ratio) and suxamethonium at a dose of 2 mg/kg. Endotracheal intubation was then performed using a Macintosh-type laryngoscope. Following intubation, mechanical ventilation was initiated using a standard anesthesia machine, with a tidal volume of 7 - 10 mL/kg and a respiratory rate of 10 - 14 breaths per minute. Throughout the surgical procedure, anesthesia was maintained using halothane or isoflurane, intravenous normal saline was infused, and muscle relaxation was achieved using vecuronium. Intraoperative vital signs were continuously monitored and recorded at 5-minute intervals by the attending anesthetist until completion of the procedure. Upon completion of surgery, neuromuscular blockade was reversed with the administration of 0.05 mg/kg of neostigmine and atropine at a dosage of 0.2 mg/kg. Patients were then extubated in the operating theater before being transferred to the postanesthesia care unit (PACU).

For patients receiving SA, patients were positioned in a sitting position. Under complete aseptic technique, the skin was infiltrated with 3 mL of 2% lidocaine, and lumbar puncture was performed at the level of the L3-L4 lumbar interspace using a midline approach with a 24- or 25-gauge spinal needle. Once free flow of clear cerebrospinal fluid was restored, 3.5 - 4.0 mL of hyperbaric bupivacaine was injected intrathecally, and the patient was then asked to lie in a supine position. The adequacy of anesthesia was assessed using the pinprick method for the sensory block level and the modified Bromage scale for the motor block level [27]. Surgery was started once the target sensory block level of the T6-T4 dermatomes was achieved. Throughout surgery, intravenous normal saline was infused, and intraoperative vital signs were continuously monitored and recorded at 5-minute intervals. Upon completion of the surgical procedure, patients were transferred to the PACU.

Postoperatively, data collectors documented the incidence of PONV in both the PACU and within the first 24 hours following PACU discharge. Pain intensity was assessed using an 11-point numeric rating scale (NRS; scores ranging from 0="no pain" to 10="worst imaginable pain") at the first, second, sixth, 12th, and 24th postoperative hours. The cumulative consumption of diclofenac and tramadol, both in the PACU and within the first 24 hours postoperatively, was recorded. In addition, postoperative outcomes, including the occurrence of complications, time to first ambulation, time to first oral intake, and time to discharge home, were recorded.

Postoperative QoR was assessed 24 hours after surgery using the QoR-15 questionnaire. QoR-15 is a validated and reliable tool designed to assess patient-reported QoR in the immediate postoperative period. It assesses 5 dimensions of recovery: psychological support, physical comfort, emotional support, physical independence, and pain level. The tool comprises 15 items, each scored on an 11-point Likert scale (from 0="strongly negative" to 10="strongly positive"), resulting in a total score ranging from 0 to 150, with higher scores indicating better QoR [19,28].

Data Quality Control

To assess the validity of the questionnaire, a pilot study was conducted on 20 participants at Debre Tabor Comprehensive Specialized Hospital, 1 month before the actual data collection period. Data collectors and supervisors were trained in the proper use of data collection tools and the overall data collection process. Throughout the data collection period, regular supervision and close follow-up were conducted to ensure data completeness, consistency, and accuracy.

Statistical Analysis

Data were coded and entered using EpiData (version 4.6; EpiData Association) software and exported to and analyzed using Stata (version 17; StataCorp) software. The normality of continuous variables was assessed using the Shapiro-Wilk test. Categorical variables are presented as frequencies and percentages, while continuous variables are reported as either means with SDs or medians with IQRs, as appropriate. Intergroup comparisons between the SA and GA groups were performed using the chi-square test (or Fisher exact test, where appropriate) for categorical variables and the independent *t* test or Mann-Whitney *U* test for continuous variables, depending on the distribution of the data.

Ethical Considerations

Ethics approval for the study was obtained from the research ethical review committee of the College of Health Sciences at Debre Tabor University (CHS 17/2017). The study was conducted in accordance with the Declaration of Helsinki, local legislation, and institutional requirements. Informed consent was obtained from each study participant, and confidentiality was ensured by avoiding questions that contained identifiers or sensitive content.

Results

Sociodemographic and Clinical Characteristics

A total of 74 participants, including 37 in the SA group and 37 in the GA group, were included in our study. Age, sex, BMI, and comorbid status were comparable between the 2 groups, with no statistically significant difference between them. Moreover, the mean duration of surgery, intraoperative blood loss, and amount of intraoperative fluid administered were comparable between the 2 groups (Table 1).

XSL•FO
RenderX

Table . Sociodemographic and clinical characteristics of study participants[a]

Variables	General anesthesia group (n=37)	Spinal anesthesia group (n=37)	*P* value
Age (years), mean (SD)[a]	27.6 (4.2)	29 (4.7)	.18
Sex, n (%)			.48
Female	17 (23)	14 (19)	
Male	20 (27)	23 (31)	
BMI (kg/m^2), median (IQR)	20.1 (19.5-21.2)	20.5 (20.3-22.1)	.12
Comorbidities, n (%)			.76
No	31 (41)	30 (40)	
Yes	6 (8)	7 (9)	
American Society of Anesthesiologists status, n (%)			.76
I	31 (42)	30 (40)	
II	6 (8.1)	7 (9.5)	
Intraoperative blood loss (mL), median (IQR)	60 (50-80)	50 (50-80)	.08
Intraoperative fluid administered (mL), median (IQR)	700 (500-800)	600 (500-700)	.09
Duration of surgery (minutes), mean (SD)[a]	46.4 (8.7)	42.8 (8.9)	.09

[a]Values expressed as mean (SD) were analyzed using the independent *t* test, values expressed as frequencies (%) were analyzed using the chi-square test, and values expressed as median (IQR) were analyzed using the Mann-Whitney *U* test.

Postoperative QoR

The total postoperative QoR-15 score at 24 hours was significantly higher in the SA group than in the GA group, with median scores of 128 (IQR 124-136) vs 100 (IQR 90-133), respectively ($P<.001$). A detailed analysis of individual QoR-15 items revealed that the SA group scored significantly higher on the majority of items than the GA group (Table 2).

XSL•FO
RenderX

Table . Comparison of general and spinal anesthesia on postoperative quality of recovery in patients after appendectomy.[a]

Variables	General anesthesia group	Spinal anesthesia group	*P* value[b]
Postoperative total Quality of Recovery-15 score, median (IQR)	100 (90-133)	128 (124-136)	*<.001*
Been able to enjoy food, mean (SD)	6.1 (1.9)	7.4 (1.4)	*.002*
Feeling rested, mean (SD)	7.7 (1.9)	8.4 (1.4)	.06
Able to look after personal toilet and hygiene unaided, mean (SD)	6.4 (1.1)	7.7 (1.5)	*<.001*
Getting support from hospital physicians and nurses, mean (SD)	9.5 (0.7)	9.3 (0.8)	.30
Able to return to work or usual home activities, mean (SD)	4.9 (1.6)	5.6 (1.2)	*<.001*
Moderate pain, mean (SD)	5.5 (2.0)	6.5 (1.4)	*.02*
Severe pain, mean (SD)	6.8 (1.8)	7.8 (1.4)	*.02*
Able to breathe easily, median (IQR)	8 (5-10)	10 (8-10)	*<.001*
Have had a good sleep, median (IQR)	6 (5-8)	9 (8-9)	*<.001*
Able to communicate with family or friends, median (IQR)	9 (9-10)	10 (10-10)	*.001*
Feeling comfortable and in control, median (IQR)	6 (6-9)	9 (8-10)	*<.001*
Having a feeling of general well-being, median (IQR)	8 (6-10)	9 (8-10)	*.004*
Nausea and vomiting, median (IQR)	10 (8-10)	10 (10-10)	*.003*
Feeling worried or anxious, median (IQR)	8 (7-10)	10 (9-10)	*<.001*
Feeling sad or depressed, median (IQR)	8 (6-10)	10 (8-10)	*.005*

[a]Values expressed as mean (SD) were analyzed using the independent *t* test, whereas values expressed as median (IQR) were analyzed using the Mann-Whitney *U* test.

[b]*P* values of <.05 are considered statistically significant and are italicized.

Postoperative Complications and Outcomes

In terms of PONV, the incidence of nausea and vomiting in the PACU was significantly higher in the GA group than in the SA group (P=.02). However, the incidence of PONV within 24 hours after PACU discharge was comparable between the 2 groups (P=.67). In addition, postoperative complications, such as postoperative adhesions, surgical site infections, and relaparotomy, were not observed in either group. Pain scores on the NRS at the first, second, sixth, and 12th postoperative hours were significantly lower (P<.001) in the SA group than in the GA group, whereas pain scores at the 24th postoperative hour were comparable between the two groups (P=.12). Regarding the postoperative analgesic consumption, cumulative diclofenac and tramadol consumption in the PACU and within the first 24 postoperative hours were significantly higher in the GA group than in the SA group. Furthermore, the times to first ambulation, first oral intake, and discharge home were significantly lower in the SA group than in the GA group (P<.001, P<.001, and P=.005, respectively; Table 3).

XSL•FO
RenderX

Table . Comparison of secondary recovery outcomes of study participants between the general anesthesia (GA) and the spinal anesthesia (SA) groups.[a]

Variables	GA group (n=37)	SA group (n=37)	*P* value[b]
PONV[c] in the PACU[d], n (%)			*.02*
Yes	9 (12)	2 (3)	
No	28 (38)	35 (47)	
PONV within 24 postoperative hours, n (%)			.67
Yes	4 (5)	2 (3)	
No	33 (45)	35 (47)	
NRS[e] score at first hour postoperatively, median (IQR)	3 (1-4)	0 (0-0)	*<.001*
NRS score at second hour postoperatively, median (IQR)	4 (3-6)	0 (0-0)	*<.001*
NRS score at sixth hour postoperatively, mean (SD)	4.2 (2.9)	2.1 (1.9)	*<.001*
NRS score at 12th hour postoperatively, mean (SD)	6 (1.4)	4.6 (0.9)	*<.001*
NRS score at 24th hour postoperatively, mean (SD)	5.5 (1.5)	4.9 (1.1)	.12
Total diclofenac consumption in PACU (mg), median (IQR)	75 (0-75)	0 (0-0)	*<.001*
Total diclofenac consumption within the first 24 hours (mg), median (IQR)	150 (75-150)	75 (75-150)	*<.001*
Total tramadol consumption in PACU (mg), median (IQR)	50 (0-50)	0 (0-0)	*<.001*
Total tramadol consumption within the first 24 hours (mg), mean (SD)	185.1 (10.4)	133.8 (13.1)	*.003*
Time to first ambulation (hours), median (IQR)	12 (10-13)	9 (8-10)	*<.001*
Time to first oral intake (hours), median (IQR)	13 (12-14)	9 (9-12)	*<.001*
Time to discharge (hours), mean (SD)	48.8 (6.2)	42.6 (10.9)	*.005*

[a]Values expressed as frequencies (%) were analyzed using the chi-square test (for PONV in the PACU) or the Fisher exact test (for PONV within 24 postoperative hours), values expressed as mean (SD) were analyzed using the independent *t* test, and values expressed as median (IQR) were analyzed using the Mann-Whitney *U* test.

[b]*P* values of <.05 are considered statistically significant and are italicized.

[c]PONV: postoperative nausea and vomiting.

[d]PACU: postanesthesia care unit.

[e]NRS: Numeric Rating Scale.

Discussion

Principal Findings

This prospective cohort study aimed to compare the postoperative recovery outcomes between patients who underwent appendectomy under SA and those who received GA. The findings revealed that patients in the SA group had significantly better early postoperative recovery profiles, as evidenced by higher QoR-15 scores, lower postoperative pain and analgesic consumption, reduced incidence of early PONV, and faster recovery outcome measures.

Our study revealed that patients who underwent appendectomy under SA experienced better recovery, with higher total postoperative QoR-15 scores and improved scores on the majority of individual QoR-15 items. These findings align with previous studies comparing neuraxial anesthesia with GA, which also showed enhanced recovery profiles in patients receiving neuraxial techniques [18,26,29]. One possible explanation for this finding is that patients receiving GA are exposed to multiple systemic medications, including volatile agents and opioids, which may have residual postoperative effects and impair early postoperative recovery [30]. In contrast, neuraxial techniques such as SA provide more effective pain control, reduce the need for systemic medications, and minimize opioid-related side effects, thereby contributing to enhancing overall well-being and recovery [29].

Supporting the QoR findings, postoperative pain management outcomes were significantly improved in the SA group during the early postoperative hours. Pain scores were significantly lower at the first, second, sixth, and 12th postoperative hours. This can be attributed to SA's ability to block the afferent nociceptive pathway, thereby providing long-lasting postoperative pain relief despite recovery of motor function [30]. This superior quality of postoperative analgesia was directly related to a significant reduction in analgesic requirements, which may have contributed to the improved QoR-15 scores and faster mobilization [31]. At the 24th postoperative hour, pain scores were comparable between the 2 groups. This may be due to the wearing-off effects of SA and the initiation of standard oral analgesia protocols. In addition, both cumulative diclofenac and tramadol consumption in both

XSL•FO
RenderX

the PACU and during the first 24 postoperative hours was significantly lower in the SA group. The reduction in postoperative opioid requirement is particularly important because it mitigates opioid-related side effects such as nausea, vomiting, sedation, and ileus [32,33].

In terms of clinical significance, between-group differences in pain scores were most pronounced during the first 6 hours, with differences of approximately 3 points at 1 hour, 4 points at 2 hours, and 2.1 points at 6 hours, which are generally considered clinically meaningful. In contrast, the difference observed at 12 hours (1.4 points) was smaller and may be of borderline clinical relevance, whereas no clinically significant difference was detected at 24 hours [34]. When analgesic consumption was expressed in intravenous morphine equivalents, the reduction in 24-hour opioid use corresponds to approximately 5 mg of intravenous morphine, which falls below commonly accepted thresholds for a clinically important effect. This finding suggests that, despite statistical significance, the overall reduction in opioid requirements over 24 hours was modest, and the clinical importance of this difference should be interpreted with caution [35].

In this study, the incidence of PONV in the PACU was significantly higher in the GA group. This may be due to the residual effects of inhaled anesthetic agents and increased opioid use for postoperative analgesia, which are known risk factors for PONV [6,36,37]. The incidence of PONV after discharge from the PACU was similar in both groups, which may reflect the resolution of early anesthetic effects and the standardization of postoperative care.

In addition, functional recovery outcome measures, such as time to first ambulation, time of first oral intake, and time of discharge to home, were significantly better in the SA group. Better pain control, reduced sedative drug requirements, and fewer side effects in the SA group may contribute to a quicker return to baseline function, reduced hospital stay, and lower health care costs [17,38]. Conversely, the use of inhaled anesthetics, sedatives, and opioids in the GA group may contribute to postoperative complications, such as cardiorespiratory depression, delayed bowel activity, and paralytic ileus, which can delay oral intake and ambulation and prolong hospitalization [6].

This study has some limitations. First, the study was conducted at a single center, and the generalizability of the findings to a wider population might be limited. In addition, although the sample size was determined using appropriate statistical methods, the final number of participants included in each group was relatively small. This may have reduced the statistical power of the study and may also restrict the external validity of the findings. Second, the study participants were selected in a nonrandomized manner, and allocation of anesthesia types may have introduced selection bias. Finally, this study focused exclusively on immediate QoR and did not include intermediate- and long-term follow-up, which limits conclusions regarding the intermediate- and long-term recovery outcomes.

Conclusions and Recommendations

The findings of our study demonstrated that SA provides better early postoperative recovery and postoperative analgesia than GA in patients undergoing appendectomy. In addition, patients undergoing appendectomy under SA had reduced early postoperative analgesia consumption, a reduced incidence of early PONV, and faster in-hospital recovery trajectories than those undergoing appendectomy under GA. Therefore, in the absence of specific contraindications, SA seems to be a preferred anesthesia technique for patients undergoing appendectomy for acute appendicitis. We recommend that future research explore intermediate- and long-term recovery outcomes in patients who undergo appendectomy under SA vs GA.

Acknowledgments

The authors acknowledge Debre Tabor University for providing ethical clearance to conduct this study. The authors also extend their sincere appreciation to Debre Tabor Comprehensive Specialized Hospital for permitting the conduct of this research within the facility. The authors also express their gratitude to the study participants, supervisors, and data collectors for their invaluable contributions.

All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee support for the publication of this paper.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

The datasets used and analyzed during the study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: NZG, TBA, HBW, KGB, GDG, AGA, BJA, BMM, BD, WTW
Data curation: NZG, TBA, HBW, KGB, GDG

XSL•FO
RenderX

Formal analysis: NZG
Investigation: AGA, BJA, BMM, BD, WTW
Methodology: NZG
Software: NZG, AGA, BJA, BMM, BD, WTW
Study design: NZG, TBA, HBW, KGB, GDG, AGA, BJA, BMM, BD, WTW
Supervision: TBA, HBW, KGB, GDG, AGA, BJA, BMM, BD, WTW
Validation: TBA, HBW, KGB, GDG
Visualization: AGA, BJA, BMM, BD, WTW
Writing – original draft: NZG
Writing – review & editing: NZG, TBA, HBW, KGB, GDG, AGA, BJA, BMM, BD, WTW

Conflicts of Interest

None declared.

References

1. Guan L, Liu Z, Pan G, et al. The global, regional, and national burden of appendicitis in 204 countries and territories, 1990-2019: a systematic analysis from the Global Burden of Disease Study 2019. BMC Gastroenterol 2023 Feb 22;23(1):44. [doi: 10.1186/s12876-023-02678-7] [Medline: 36814190]
2. Moris D, Paulson EK, Pappas TN. Diagnosis and management of acute appendicitis in adults: a review. JAMA 2021 Dec 14;326(22):2299-2311. [doi: 10.1001/jama.2021.20502] [Medline: 34905026]
3. Ferris M, Quan S, Kaplan BS, et al. The global incidence of appendicitis: a systematic review of population-based studies. Ann Surg 2017 Aug;266(2):237-241. [doi: 10.1097/SLA.0000000000002188] [Medline: 28288060]
4. Adu A, Birhanu Y. Acute appendicitis in Ethiopia: a systematic review and meta-analysis. Int J Afr Nurs Sci 2021;15:100343. [doi: 10.1016/j.ijans.2021.100343]
5. Uzman S, Donmez T, Erdem VM, Hut A, Yildirim D, Akinci M. Combined spinal-epidural anesthesia in laparoscopic appendectomy: a prospective feasibility study. Ann Surg Treat Res 2017 Apr;92(4):208-213. [doi: 10.4174/astr.2017.92.4.208] [Medline: 28382293]
6. Sun X, Zhang J, Li Y, Che J, Ren W. Effect of intraspinal anesthesia on postoperative recovery in elderly patients with acute appendicitis: a retrospective study. Ann Ital Chir 2024;95(4):669-677. [doi: 10.62713/aic.3415] [Medline: 39186347]
7. El Moheb M, Han K, Breen K, et al. General versus neuraxial anesthesia for appendectomy: a multicenter international study. World J Surg 2021 Nov;45(11):3295-3301. [doi: 10.1007/s00268-021-05978-9] [Medline: 33554296]
8. Erdem VM, Donmez T, Uzman S, Ferahman S, Hatipoglu E, Sunamak O. Spinal/epidural block as an alternative to general anesthesia for laparoscopic appendectomy: a prospective randomized clinical study. Wideochir Inne Tech Maloinwazyjne 2018 Jun;13(2):148-156. [doi: 10.5114/wiitm.2018.72684] [Medline: 30002746]
9. Kayani KM, Malik UE, Hafiz A, Hussain F, Munir H, us Saqib N. Comparison of hemodynamic stability of patients undergoing appendectomy under spinal anesthesia versus blend of ketamine and midazolam with spinal anesthesia. PAFMJ 2023;73(4):1128-1132. [doi: 10.51253/pafmj.v73i4.9481]
10. Balcı M, Balcı F, Atay E, Bozkurt E, Ertekin A. Comparison of bupivacaine–fentanyl and bupivacaine-alfentanil used intrathecally in laparoscopic appendectomy surgery. Kocatepe Tıp Dergisi 2021;22(5):414-421. [doi: 10.18229/kocatepetip.835963]
11. Bilal A, Muneer B, Ullah MS, Shahbaz S, Ali R, Noman M. Postoperative complications of anaesthesia following appendectomy: complications of anaesthesia following appendectomy. Heal J Physiother Rehabil Sci 2024;4(5):15-22. [doi: 10.55735/hjprs.v4i5.287]
12. Jun GW, Kim MS, Yang HJ, et al. Laparoscopic appendectomy under spinal anesthesia with dexmedetomidine infusion. Korean J Anesthesiol 2014 Oct;67(4):246-251. [doi: 10.4097/kjae.2014.67.4.246] [Medline: 25368782]
13. Latif A, Bhatti MF, Qadir A, Shahwar ZA, Shahzad S. Effectiveness of anaesthesia for appendicectomy: comparison of spinal anaesthesia with other modalities being practiced in emergency department of Allama Iqbal Memorial Teaching Hospital. Pak J Med Health Sci 2017;11(4):1402-1403 [FREE Full text]
14. Mai DV, Sagar A, Claydon O, Park JY, Tapuria N, Keeler BD. Open appendicectomy under spinal anesthesia-a valuable alternative during COVID-19. Surg J (N Y) 2021 Apr;7(2):e69-e72. [doi: 10.1055/s-0041-1725933] [Medline: 34104718]
15. Gabriel MD, Veronica AM, Karen P, Maria G, Gerardo A, Daniela G. Comparison between spinal anesthesia vs general anesthesia in patients undergoing open appendectomy for acute appendicitis, a retrospective study. J Hepatol Gastroint Dis 2020;7:178. [doi: 10.35248/2475-3181.7.178]
16. Gurudatta KN, Arif M. A clinical study of comparison between general anesthesia and spinal anesthesia for lower abdominal laparoscopic surgeries. Sch J Appl Med Sci 2014;2(3):1127-1133. [doi: 10.36347/sjams.2014.v02i03.053]
17. Hewson DW, Tedore TR, Hardman JG. Impact of spinal or epidural anaesthesia on perioperative outcomes in adult noncardiac surgery: a narrative review of recent evidence. Br J Anaesth 2024 Aug;133(2):380-399. [doi: 10.1016/j.bja.2024.04.044] [Medline: 38811298]

XSL•FO
RenderX

18. de Carli D, Meletti JF, de Camargo R, Gratacós LS, Gomes VC, Marques ND. Effect of anesthetic technique on the quality of anesthesia recovery for abdominal histerectomy: a cross-observational study. Braz J Anesthesiol 2021;71(3):221-227. [doi: 10.1016/j.bjane.2021.01.013] [Medline: 33930343]
19. Myles PS, Shulman MA, Reilly J, Kasza J, Romero L. Measurement of quality of recovery after surgery using the 15-item quality of recovery scale: a systematic review and meta-analysis. Br J Anaesth 2022 Jun;128(6):1029-1039. [doi: 10.1016/j.bja.2022.03.009] [Medline: 35430086]
20. Wessels E, Perrie H, Scribante J, Jooma Z. Quality of recovery in the perioperative setting: a narrative review. J Clin Anesth 2022 Jun;78:110685. [doi: 10.1016/j.jclinane.2022.110685] [Medline: 35190342]
21. Bowyer A, Royse CF. Defining postoperative quality of recovery. In: Stuart-Smith K, Stuart-Smith K, editors. Perioperative Medicine – Current Controversies: Springer International Publishing; 2016:87-107. [doi: 10.1007/978-3-319-28821-5_5]
22. Borrell-Vega J, Humeidan ML, Bergese SD. Defining quality of recovery – what is important to patients? Best Pract Res Clin Anaesthesiol 2018;32(3-4):259-268. [doi: 10.1016/j.bpa.2018.07.002] [Medline: 30522716]
23. Léger M, Campfort M, Cayla C, Lasocki S, Rineau E. Postoperative quality of recovery measurements as endpoints in comparative anaesthesia studies: a systematic review. Br J Anaesth 2021 Jun;126(6):e210-e212. [doi: 10.1016/j.bja.2021.03.008] [Medline: 33892949]
24. Bowyer A, Royse C. The importance of postoperative quality of recovery: influences, assessment, and clinical and prognostic implications. Can J Anaesth 2016 Feb;63(2):176-183. [doi: 10.1007/s12630-015-0508-7] [Medline: 26475163]
25. Mokhtar Mehanna AMA, Ibrahim AG. Comparative study between general and spinal anaesthesia in laparoscopic appendectomy. J Anesth Clin Res 2017;8(4):1000713. [doi: 10.4172/2155-6148.1000713]
26. Taflan MG, Akdeniz S, Kusderci H, et al. Comparison of combined spinal-epidural versus general anesthesia with epidural catheter on postoperative quality of recovery after abdominal hysterectomy: a prospective observational study. BMC Anesthesiol 2025 Jul 31;25(1):386. [doi: 10.1186/s12871-025-03252-2] [Medline: 40745528]
27. Craig D, Carli F. Bromage motor blockade score – a score that has lasted more than a lifetime. Can J Anaesth 2018 Jul;65(7):837-838. [doi: 10.1007/s12630-018-1101-7] [Medline: 29508151]
28. Kleif J, Waage J, Christensen KB, Gögenur I. Systematic review of the QoR-15 score, a patient- reported outcome measure measuring quality of recovery after surgery and anaesthesia. Br J Anaesth 2018 Jan;120(1):28-36. [doi: 10.1016/j.bja.2017.11.013] [Medline: 29397134]
29. Catro-Alves LJ, De Azevedo VL, De Freitas Braga TF, Goncalves AC, De Oliveira GS Jr. The effect of neuraxial versus general anesthesia techniques on postoperative quality of recovery and analgesia after abdominal hysterectomy: a prospective, randomized, controlled trial. Anesth Analg 2011 Dec;113(6):1480-1486. [doi: 10.1213/ANE.0b013e3182334d8b] [Medline: 21926374]
30. Sekerak R, Mostafa E, Morris MT, Nessim A, Vira A, Sharan A. Comparative outcome analysis of spinal anesthesia versus general anesthesia in lumbar fusion surgery. J Clin Orthop Trauma 2021 Feb;13:122-126. [doi: 10.1016/j.jcot.2020.11.017] [Medline: 33680810]
31. Yoon SH, Bae J, Yoon S, Na KJ, Lee HJ. Correlation between pain intensity and quality of recovery after video-assisted thoracic surgery for lung cancer resection. J Pain Res 2023;16:3343-3352. [doi: 10.2147/JPR.S426570] [Medline: 37808464]
32. Hosseinzadeh Zorofchi B, Jahan E, Nassiri S, et al. Comparing spinal and general anesthesia in terms of postoperative pain in patients undergoing hysterectomy. J Obstet Gynecol Cancer Res 2022;3(2):73-77. [doi: 10.30699/jogcr.3.2.73]
33. O'Neill J, Helwig E. Postoperative management of the physiological effects of spinal anesthesia. J Perianesth Nurs 2016 Aug;31(4):330-339. [doi: 10.1016/j.jopan.2015.01.018] [Medline: 27444765]
34. Tian T, Li XT, Xue FS. The minimal clinically important differences as well as statistical differences of main endpoints are important in comparing postoperative benefits of different analgesic modalities [letter]. Drug Des Devel Ther 2022;16:3195-3196. [doi: 10.2147/DDDT.S370327] [Medline: 36172052]
35. Laigaard J, Pedersen C, Rønsbo TN, Mathiesen O, Karlsen AP. Minimal clinically important differences in randomised clinical trials on pain management after total hip and knee arthroplasty: a systematic review. Br J Anaesth 2021 May;126(5):1029-1037. [doi: 10.1016/j.bja.2021.01.021] [Medline: 33678402]
36. Shaikh SI, Nagarekha D, Hegade G, Marutheesh M. Postoperative nausea and vomiting: a simple yet complex problem. Anesth Essays Res 2016;10(3):388-396. [doi: 10.4103/0259-1162.179310] [Medline: 27746521]
37. Sansonnens J, Taffé P, Burnand B, ADS study group. Higher occurrence of nausea and vomiting after total hip arthroplasty using general versus spinal anesthesia: an observational study. BMC Anesthesiol 2015 Dec;16(1):44. [doi: 10.1186/s12871-016-0207-0]
38. Eoh KJ, Ahn JH, Park JS, et al. Comparative analysis of spinal anesthesia versus general anesthesia in single-port access laparoscopic adnexal surgery: a propensity score matching study. BMC Womens Health 2025 Feb 24;25(1):83. [doi: 10.1186/s12905-025-03604-7] [Medline: 39994702]

Abbreviations

GA: general anesthesia
PACU: postanesthesia care unit

XSL•FO
RenderX

PONV: postoperative nausea and vomiting
QoR: quality of recovery
QoR-15: Quality of Recovery-15
SA: spinal anesthesia

Edited by N Rohatgi; submitted 21.Oct.2025; peer-reviewed by MA Karahan; revised version received 25.Apr.2026; accepted 29.Apr.2026; published 25.Jun.2026.

Please cite as:
Gobezie NZ, Asmare TB, Belete KG, Gedefaw GD, Demissie B, Abuhay AG, Abate BJ, Munie BM, Wondie WT, Wubet HB
Comparing the Quality of Patient Recovery After Open Appendectomy Under General Versus Spinal Anesthesia: Prospective Cohort Study
JMIR Perioper Med 2026;9:e86273
URL: https://periop.jmir.org/2026/1/e86273
doi:10.2196/86273

© Negesse Zurbachew Gobezie, Temesgen Birlie Asmare, Kumlachew Geta Belete, Gezahagn Demsu Gedefaw, Biruk Demissie, Abere Gebru Abuhay, Belayneh Jejaw Abate, Birhanu Mengist Munie, Wubet Tazeb Wondie, Habtie Bantider Wubet. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 25.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring Provider Perspectives on Gaps in Trauma-Informed Care Delivery Within the Acute Care Surgery Timeline at a Federally Qualified Health Center: Human-Centered Design Study

Angelica T McDaniel[1,2], BS, MS; Willow Frye[2], MS, MD; Neema Rashidi[1,2], BA, MS; Christiana D von Hippel[2], MPH, ScD; Marianna G Salvatori[2,3], BS; Adrienne Greer[2,3], BA, MPH; Manami Diaz Tsuzuki[2], BA; Amanda Sammann[2,3], MPH, MD

[1]UC Berkeley-UCSF Joint Medical Program, UC Berkeley, Berkeley, CA, United States

[2]The Better Lab, University of California, 2540 23rd St, 4th floor, San Francisco, CA, United States

[3]Department of Surgery, University of California, San Francisco, CA, United States

Corresponding Author:
Christiana D von Hippel, MPH, ScD
The Better Lab, University of California, 2540 23rd St, 4th floor, San Francisco, CA, United States

Abstract

Background: Trauma-informed care (TIC) education plays a pivotal role in preparing physicians to deliver care to patients who have experienced emotional and physical trauma, empowering patients to have agency in their care. The delivery of TIC by providers is especially important at Federally Qualified Health Centers (FQHCs), as their patients disproportionately experience trauma due to the impacts of socioeconomic determinants of health, including limited health care access, systemic racism, and housing insecurity. There is limited literature on how to adapt TIC practices in acute care settings, specifically for emergency medicine and surgical providers who deliver care in high-demand, time-constrained environments.

Objective: The goal of this study was to understand the shared challenges faced by providers implementing TIC within acute care settings at an urban safety-net hospital.

Methods: This exploratory qualitative study applied methods from the first stage of the human-centered design process, known as the inspiration phase, to identify the factors that influence providers' decisions to provide or not to provide TIC within the acute care ecosystem. In this inspiration phase, we conducted 18 semistructured interviews with providers delivering surgical care at an FQHC in Northern California. Interviewers were purposefully sampled to represent the diversity of providers involved throughout the perioperative timeline, including physicians, nurses, social workers, psychologists, and de-escalation staff. Inductive thematic analysis of transcripts and subsequent insight statement generation were conducted to identify tensions between stakeholder needs and to elucidate design opportunities to improve their ability to deliver TIC.

Results: Inductive analysis identified 21 unique themes and 7 insights that affect the effective implementation of TIC by providers in acute care settings. Time-constrained workflows, vicarious trauma, and provider burnout present significant barriers to its application. Furthermore, limitations in TIC training, hospital design, and security measures, translation challenges, and hierarchical structures all hinder its integration. Supportive measures include a strong desire to improve health outcomes for communities and patients at FQHCs.

Conclusions: Through this first phase of the human-centered design methodology, we identified 7 key insights and design opportunities that lay the foundation for developing solutions to address gaps and enhance the delivery of TIC within safety-net hospitals.

(JMIR Perioper Med 2026;9:e81004) doi:10.2196/81004

KEYWORDS

trauma-informed care; perioperative medicine; human-centered design; burnout; provider experience

Introduction

Exposure to trauma within hospital populations is both widespread and complex, often stemming from chronic stressors that accumulate over time. According to the Substance Abuse and Mental Health Services Administration (SAMHSA), individual trauma "results from an event, series of events, or set of circumstances that is experienced by an individual as physically or emotionally harmful or threatening and that has lasting adverse effects on the individual's functioning and physical, social, emotional, or spiritual well-being." [1] Importantly, trauma is not limited to individual experience, as

XSL•FO
RenderX

it can also be transmitted at the community and societal levels, including across generations. This is especially true for patients at Federally Qualified Health Centers (FQHCs), which are health care settings that receive federal funding to provide health care services to underserved populations, regardless of patients' financial capacity. At these institutions, the patient population has an increased risk of trauma exposure due to socioeconomic challenges such as poverty, housing instability, and limited access to education [2]. One study found that approximately 80% of patients in safety-net hospitals have experienced 2 or more traumatic events, with 33% experiencing 5 or more [3]. Across trauma centers nationwide, the Improving Social Determinants to Attenuate Violence Workgroup within the American College of Surgeons has been actively addressing the root causes and impacts of trauma by promoting trauma-informed care (TIC) practices and curricular changes [4].

TIC is a strengths-based framework that is grounded in an understanding of trauma and the response to its impact, emphasizing the physical, psychological, and emotional safety of survivors to rebuild a sense of control and empowerment. According to the SAMHSA, TIC is a framework that health care systems can use to identify histories of traumatic experiences among patients and staff and enhance workflows by focusing efforts on developing policies and practices that prevent retraumatization [1]. In primary care settings, TIC has been used to improve provider and staff training, trauma screening protocols, trauma-informed interventions and resources, and staff burnout prevention [5]. Studies have shown that these practices allow providers to create a more comprehensive understanding of their patients' histories, with visit lengths increasing by 5 minutes or less, 91% of the time [6]. However, in emergent settings (such as perioperative care, the emergency department, and the intensive care unit), time and workflow constraints pose additional obstacles to implementing TIC [7].

Surgical procedures are highly vulnerable experiences for patients, particularly regarding time-sensitive decision-making and bodily autonomy. The indications for surgical interventions, as well as the interventions themselves, carry the risk of causing psychological trauma in the preoperative, operative, and recovery periods. These environments place patients with unfamiliar providers, allot a limited time for decision-making, and compel patients to yield control to the health care system [8]. Frameworks and tools to understand and respond to this are critical for mitigating negative impacts and increasing opportunities for building patient-provider trust, facilitating patient-centered decision-making, and improving follow-up postoperatively with better outcomes. Although there has been an increase in research on TIC regarding patients, there has been little exploration into the insights of providers who deliver this care.

The research aim was to understand how to better adapt and implement TIC practices, education, and interventions in emergency care settings, particularly in FQHC surgical units, using interdisciplinary collaboration and design. While traditional qualitative research typically focuses on observing and understanding an environment, human-centered design (HCD) methodology was chosen to actively explore the perspectives and challenges of implementing TIC with the intent to create practical interventions [9]. HCD places users at the center of the design space, working toward co-designing solutions that incorporate the voices of stakeholders [10]. This methodology involves 3 phases: inspiration, ideation, and implementation. Each phase offers a stepwise approach to identifying the needs of the users, generating meaningful solutions, and turning these ideas into applicable outcomes. HCD has been used to develop prototypes in communities facing systemic barriers due to the prioritization of equal partnerships between stakeholders and researchers [11]. It has been used to improve documentation of geriatric care, reduce perinatal inequities, and design interventions for youth with HIV [10,12,13]. Incorporating users into the research team increases the likelihood of use, enhances accessibility, and improves the feasibility of the developed prototypes [14]. This paper focuses on the first inspiration phase, where an iterative process consolidates information gathered from interviews, ultimately laying the foundation for the development of interventions in subsequent phases.

Methods

Ethical Considerations

This study was approved by the University of California, San Francisco Institutional Review Board (IRB #22 - 36409). The IRB approved a waiver of signed consent; participants were provided with an information sheet describing the study, and they verbally confirmed their agreement to participate over Zoom (Zoom Communications, Inc) prior to their interview. Participants were informed of their right to decline participation or withdraw at any time. Interview recordings were deidentified to protect participant privacy and confidentiality, and no identifying information is included in this manuscript.

Study Setting and Participant Selection

This is an exploratory qualitative research study conducted at an FQHC in Northern California between 2022 and 2023 to understand clinicians' perspectives on TIC within the acute surgical care setting.

Purposeful sampling [15] was used to recruit 18 stakeholders who represent the spectrum of providers with varying levels of experience and roles who work within emergency surgical care, from arrival at the emergency department to discharge. These included physicians and surgeons who direct clinical decision-making, nurses who provide continuous bedside care, social workers who address psychosocial needs, psychologists who support patients experiencing acute distress, and de-escalation staff who manage crisis situations. This diversity of roles was intentional in order to capture the fundamentally team-based nature of TIC. Individuals were recruited through provider and staff referrals at an FQHC emergency department and medical-surgical inpatient floors. No formal exclusion criteria were applied beyond failure to meet the inclusion criteria; all individuals who were referred and contacted by the research team agreed to participate and provided informed consent prior to their interviews. Interviews lasted approximately an hour and were conducted virtually through Zoom to

XSL•FO
RenderX

accommodate clinical schedules and COVID-19 precautions in place at the time. Interviews were audio-recorded, deidentified for participant confidentiality, and transcribed for analysis.

Qualitative Data Collection

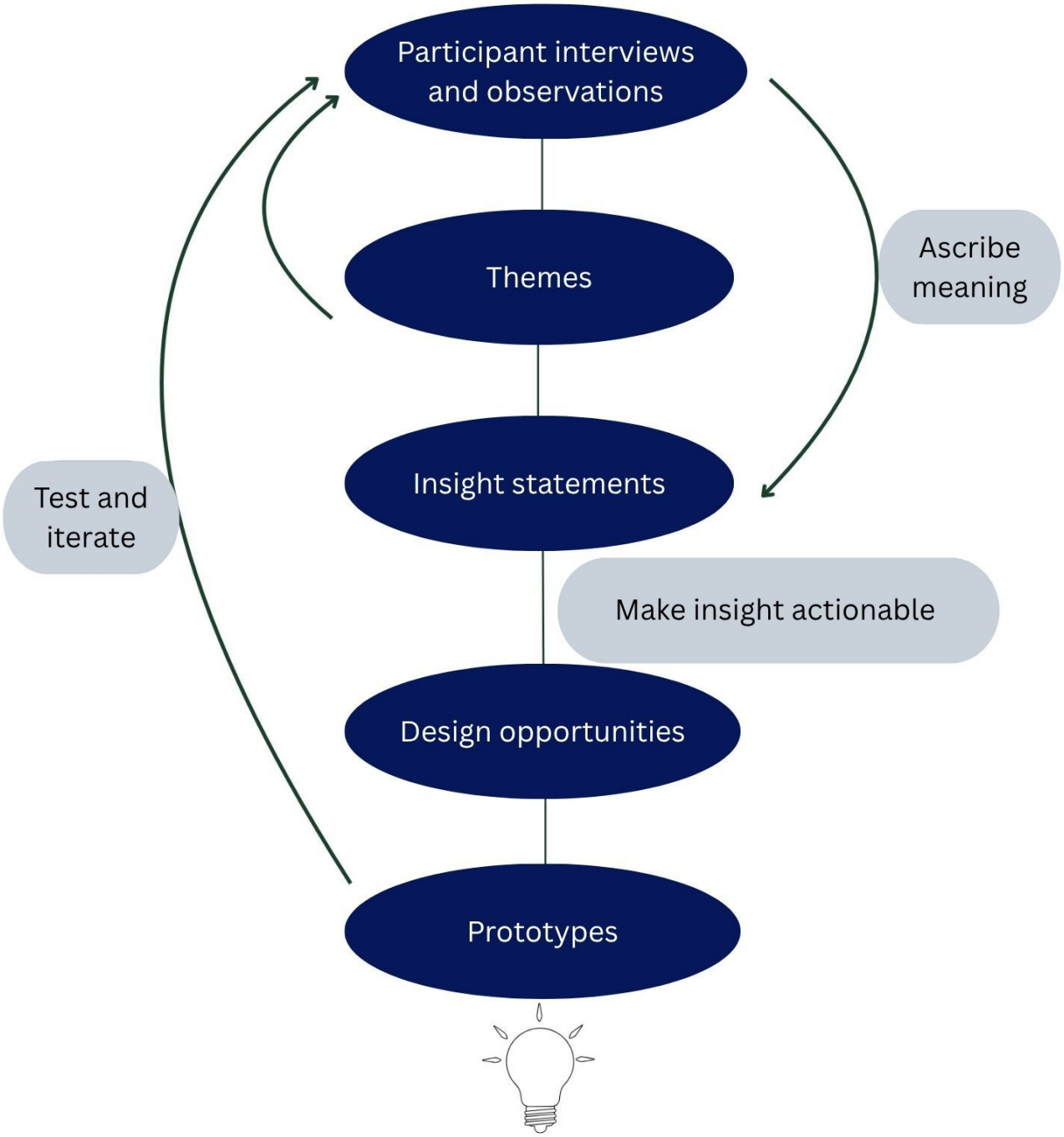
Consistent with the HCD Inspiration phase, semistructured interviews were conducted using open-ended questions to elicit narratives identifying individual and systems-level challenges to providing TIC [16]. Interview guides were developed based on existing TIC literature and the SAMHSA framework [1,5], covering topics such as staff training, screening protocols, TIC interventions, and TIC implementation. These topics were used to understand our subjects' perspectives on how TIC manifests in their daily work and the challenges they have encountered in its implementation. While all participants were asked the same core questions regarding their understanding of TIC, how they apply it in their daily work, and the barriers they encounter, role-specific probing questions were included to elicit contextually relevant examples.

Qualitative Data Analysis

A general inductive approach to thematic analysis [17] was conducted to identify key themes and exemplar quotes from deidentified interview transcript data. Two interview transcripts were selected by the interviewers for their representation of the sample's diversity and the range of themes identified across all interviews. These transcripts were independently coded by 2 researchers (coder initials: AM and NR) to develop the initial codebook. To establish intercoder reliability, the researchers engaged in regular discussions after coding the representative interviews to ensure agreement on theme definitions and coding decisions. Once the initial codebook was defined, these researchers independently coded the transcripts using Dedoose (SocioCultural Research Consultants, LLC), a qualitative data analysis software, which facilitated the organization and analysis of the data [18]. Consistent with an inductive thematic analysis approach, the final codebook (Multimedia Appendix 1) was continuously refined throughout the analysis. Any new codes that needed to be added or existing codes that required redefinition were revised collaboratively by the team and applied consistently across all interviews. After interviewing 18 individuals, thematic saturation was achieved, and recruitment ceased. Thematic saturation was indicated by a lack of emergence of novel themes in new interview data; additionally, the number of interviews exceeded the commonly cited new information threshold of 10 - 12 participants [19]. Themes, defined as inductively identified patterns of meaning derived from the interview data, were reassessed across the full set of interviews. Representative quotes were identified for each theme as empirical evidence of the observed patterns. Together, the themes and supporting quotes provided the foundation for insight statement development. Themes sharing common patterns or stakeholder concerns were grouped and synthesized into insight statements that captured the underlying tensions between stakeholder needs and systemic constraints. Each draft insight statement was written in actionable, user-centered language and iteratively reviewed and refined against supporting themes and quotes to ensure it remained grounded in the data. These insight statements define the user's core motivations and needs, providing directed guidance for creating design opportunities (Figure 1). The design opportunities serve as starting points for potential interventions to address the issues identified under the insight statements. These inspiration phase activities are illustrated in Figure 1. Prototype development and iterative refinement, which comprise the subsequent ideation phase, will be addressed in future work.

XSL•FO
RenderX

Figure 1. Human-centered design qualitative data analysis process diagram (previously published in the study by Lee et al [20], 2025).



Results

Qualitative Data Analysis

Eighteen interviews were conducted with a heterogeneous sample of individuals who work across the spectrum of emergency surgery, including physicians, nurses, social workers, psychologists, and de-escalation staff (Table 1). Thematic analysis of these interviews revealed 21 themes, organized under 7 overarching insight statements. These are presented alongside supporting quotes and design opportunities that highlight the obstacles and tensions involved in delivering TIC (Textbox 1).

XSL•FO
RenderX

Table . Participant demographics (N=18).

Participant role	Count, n (%)	Female, n (%)	0 - 5 years experience, n (%)	5+ years experience, n (%)
Total	18 (100)	15 (83.3)	1 (5.6)	17 (94.4)
Physicians (total)	9 (50)	8 (88.9)	1 (11.1)	8 (88.9)
Residents	1 (5.6)	1 (100)	1 (100)	0 (0)
Fellows	2 (11.1)	2 (100)	0 (0)	2 (100)
Attendings	6 (33.3)	5 (83.3)	0 (0)	6 (100)
Nurses (total)	6 (33)	4 (66.7)	0 (0)	6 (100)
Nurse practitioners	2 (11.1)	2 (100)	0 (0)	2 (100)
Registered nurses	4 (22.2)	2 (50)	0 (0)	4 (100)
Social workers	1 (5.6)	1 (100)	0 (0)	1 (100)
Psychologists	1 (5.6)	1 (100)	0 (0)	1 (100)
De-escalation staff	1 (5.6)	1 (100)	0 (0)	1 (100)

XSL•FO
RenderX

Textbox 1. Themes, quotes, insights, and design opportunities.

Theme 1a: Implementation of trauma-informed care (TIC) takes time

- **Quote**: "One of the biggest and sometimes hardest things to do in surgery is to spend time. The time that you need with patients to understand what's going on and counsel, and educate, and explain to them their options appropriately. I think that is really hard, especially as a resident."—General Surgery Resident Physician 1

Theme 1b: The clinical workflow does not support the time required for TIC

- **Quote**: "People don't want to stop and slow down and be thoughtful because that takes more time, and the system doesn't really reinforce taking time. The system reinforces productivity."—Trauma Recovery Center Staff 1

Theme 1c: Health care has a productivity-driven focus over patient-centered care

- **Quote:** "The goal in our health care system is not to provide care. It is to make money. So, our health care system in the United States is in direct conflict with itself. It makes more money the less care it provides."—Emergency Medicine Nurse 1

Insight 1: TIC requires excess time for provider education and clinical implementation, which the productivity-driven hospital workflow does not accommodate.

Design opportunity: integrate TIC training within hospital workflows and medical education to support productivity and respect time constraints.

Theme 2a: Providers are burned-out without support structures

- **Quote:** "There is some forced compartmentalization here because I need to go take care of the next person in front of me...Yes, this horrible thing happened. I just did all of this very invasive, emergent, intense stuff. And I'm just going to have, just put it in a box and think about it, process, talk about it later because it can't happen right now. Because there's somebody else in front of me I need to take care of who needs 100% of my attention."—Acute Care Surgery Fellow Physician 1

Theme 2b: Burned-out providers struggle to implement TIC

- **Quote:** "I feel like I'm fighting all the time. I'm fighting for resources for my patients. I'm fighting for resources for the staff. I am fighting for vulnerable populations. I feel like I'm fighting all the time. And it's wearing me out."—Emergency Medicine Nurse 2

Theme 2c: Providers believe patients are negatively impacted by the lack of TIC

- **Quote:** "The patient told me, 'Well, nobody ever in the health care system ever smile at me.' And they told me, 'You're the first one smiling at me.' And I think to them it was a way to say, 'Well I never felt that there was any compassion for me in the health care system.' And to some extent they felt that they were being perceived as some sort of a burden."—Anesthesia Attending Physician 1

Insight 2: Providers who do not receive support to address their own vicarious trauma and burnout do not have the personal resources to prioritize TIC.

Design opportunity: Create supportive structures for providers to process vicarious trauma and mitigate burnout, enabling them to sustain and prioritize TIC.

Theme 3a: TIC training lacks a consistent, core curriculum

- **Quote:** "Everybody should have it the same way. And I know people don't like that. They like having the freedom, but I think if the idea is to start out with this standard and then have good clinical evidence and data while you deferred from it, that's okay. I'm not saying that everybody should do the same thing, but I think that we should start out with some of the best. And if we're going to do less than that or more than that, we need to have data as to why."—Interventional Radiology Nurse 1

Theme 3b: TIC training lacks specialty-specific considerations

- **Quote:** "I was just doing a trauma-informed course. It was a training the trainer type of course that the ACS is starting to roll out. And this idea that you can ask open-ended questions or ask people things like, 'What can we do for you?' Or we can approach somebody in a trauma-informed way, sitting down to talk to them without it taking more time. I think that's probably unrealistic to say. To say that it's not going to take more time than what that takes, is just at face value not believable."—Acute Care Surgery Fellow 2

Theme 3c: Traditional medical training does not always incorporate TIC

- **Quote:** "TIC is definitely not something that you are taught in any of your training. I think that's also part of the problem. Because when you're in nursing school, you get assigned to all of the quote unquote 'cooperative or nice patients.' You don't get to experience patients who may not necessarily go along with the care plan. Then when you enter the workforce, you're shocked that, oh, you don't want to take your medicines?"—Medical-Surgical Nurse 1

Insight 3: TIC training is too inconsistent to be widely adopted and too generic to address specialty-specific needs, making it difficult for clinicians to apply in real-world care.

Design opportunity: Design a flexible TIC training model that combines a standardized core curriculum with specialty-specific modules, enabling both broad adoption and practical implementation across diverse clinical settings.

XSL•FO
RenderX

Theme 4a: Providers believe hospital design can negatively impact patients' well-being

- **Quote: "**Just imagine what some individuals who have experienced being in prison, when they're walking through the doors of (the hospital) and the first people they encounter are police officers, and then they have to call in and they have to go through multiple different locked doors to get to their family member. It feels very much the same."—Anesthesia Attending Physician 1

Theme 4b: Providers believe hospital safety measures can negatively impact patients' well-being

- **Quote:** "I'm not saying that there should not be any police. I think in some instances you may need the police because patients come with knives and guns and things of that nature. But I really do feel like we need to invest in other forms of care to help deescalate patients so we don't get there, so we don't need the police there. But it can definitely be triggering for patients. Because you don't know what someone's history is, or how they have interacted with the police. If this is supposed to be a place of healing, having them here doesn't necessarily align with it."—Medical-Surgical Nurse 3

Theme 4c: Compliance with hospital protocol is not in concordance with being trauma-informed

- **Quote:** "Then the doctor says, 'We're going to restrain her.' 'Wait a second, what are you talking about? We're not going to restrain her. She was a victim of sex trafficking. Held hostage. Recently incarcerated. Why are we going to restrain her?' The situation was bananas. I just could not believe what was happening."—Medical-Surgical Nurse 1

Insight 4: Hospital architecture, security measures, and compliance are meant to ensure safety, but instead make patients with trauma feel unsafe, undermining the goals of TIC.

Design opportunity: Create physical hospital environments that balance structural requirements, security, and compliance with trauma-informed design principles to ensure a psychological sense of safety for patients.

Theme 5a: Using translation services impedes human connection between patient and provider

- **Quote:** "I think language barriers are huge. And I think people are doing a good job of using translators and all that, but I even sometimes feel like I'll sit down for a long time with a translator and I still can't have a conversation where I feel like we're understanding each other, like the patient and I are understanding each other through a translator. I feel like that can be a really big barrier sometimes to having those meaningful conversations with patients."—General Surgery Resident Physician 1

Theme 5b: Direct language translation is not always inclusive of cultural translation

- **Quote:** "One of the medical student interns who spoke fluent Spanish, she wasn't certified as a bilingual interpreter, but we were consenting a patient for a cholecystectomy. And she was listening to the phone interpreter we were using to do the actual legal informed consent and everything. And the interpreter said something and she just shook her head and was like, 'No, no. This is actually what we're saying.'"—Acute Care Surgery Fellow 1

Theme 5c: Providers are concerned about providing suboptimal care due to communication barriers

- **Quote:** "But just having to use phone interpreters, that I think a lot of things are lost in interpretation. And there are definitely, I can think of a handful of times where I felt like I was having, say a pretty thorough preoperative discussion just for something entirely elective, like a cholecystectomy, and basically having the patient come back to me at the end of a 20 minute conversation, when I've clearly said through the interpreter, 'We're going to be taking out your gallbladder.' And the patient says, 'So what exactly are you going to be doing in surgery?.'.. I know this inherently is going to take longer than, say if I'm just having a conversation with somebody, who their first language is also English. But also making sure that I have pretty much that same conversation. That I'm not cutting corners or leaving out details because I'm having to go through an interpreter."—Acute Care Surgery Fellow 2

Insight 5: When translation is inconsistent or lacks cultural nuance, providers struggle to build trust and worry they are delivering lower-quality care to non–English-speaking patients.

Design opportunity: Enhance the quality and consistency of translation services to help providers deliver clear, compassionate, and patient-centered care to non–English-speaking patients.

Theme 6a: Trust is achieved by giving patients a voice

- **Quote:** "Imagine being a patient, not feeling like yourself and away from all your people, away from all your normal ways of coping with things. Your modesty ripped away from you because you're in this silly gown and people keep coming in and the whole waking up thing, being woken up at four in the morning by a door creaking open, all of it you have to realize is not a good experience. It's all traumatic. Even if you're here for a lap coley. It's all just uncomfortable. So I think first and foremost, just actually acknowledging that like, 'I hear you, this sucks. It's not fun to be here.' And going through those things. 'I know you don't like getting up in the middle of the night.' Even having that question, 'What would make this easier for you? There's things that have to happen. There's things that maybe we can work around. What would make this easier for you?' And you can be even honest."—Surgery Nurse Practitioner 1

Theme 6b: Hospital hierarchies hinder the ability to incorporate the voices of staff working closest to patients

- **Quote:** "Because I really do try to advocate for patients. Some nurses, they're just scared, I think. Because they're just like, 'Well, the doctor said. The doctor said.' It's like, okay. But it's okay to ask questions. It's okay to challenge the doctor. You're doing it in a very professional manner. But you're looking out for the best interest of your patient. But the problem is, if everyone just goes along with the plan and nobody questions things, it doesn't always work out for the patient."—Medical-Surgical Nurse 2

XSL•FO
RenderX

Theme 6c: Hospital hierarchies hinder the ability to incorporate the voices of patients

- **Quote:** "A lot of our patients I feel like just want more control over their care. If you sit down and you talk to patients about what's going on and how they feel it might be best to manage it, I feel like you can get their buy-in a little bit more. Versus talking at them. Encouraging them to participate in their care. These are adults and they can make decisions. I just feel like sometimes we don't allow that. Our system doesn't allow for that. It's very rigid."—Medical-Surgical Nurse 1

Insight 6: Providers earn patient trust through involving them in their care, but medicine's hierarchical structures can impede this trauma-informed approach.

Design opportunity: Integrate diverse team viewpoints to address patient concerns and foster improved communication and stronger relationships.

Theme 7a: Providers work in Federally Qualified Health Centers (FQHCs) to care for under-served and at-risk patients

- **Quote:** "And I was drawn to this hospital because of the way the patient population was treated there, I'm not someone who believes that you should be treated differently based on the thickness of your wallet when it comes to health care."—Emergency Medicine Nurse 1

Theme 7b: Providers persevere through the hardship for the benefit of their community

- **Quote:** "For whatever reason I've experienced PTSD, burnout, all kinds of things and I have worked very hard to take care of myself to continue. Cause the mission of this hospital, the population that we take care of, is a very needy population. In need of support, in need of family, and in need of a lot of compassion. So I continue to feel like I'm relevant, even though sometimes I look in the mirror and I say I don't know if I'm still relevant."—Emergency Medicine Nurse 2

Theme 7c: The toll of working for FQHCs pushes providers to leave

- **Quote:** "I'm definitely considering leaving. The emotional toll that my shifts take on me now are significant. I've made efforts to affect change by becoming more active in the union, which just means that I've made myself apparently an enemy of the administration and it's made things harder."—Emergency Medicine Nurse 1

Insight 7: The more deeply providers care about their patients and mission, the more likely they are to struggle to sustain their sense of purpose over time in high-demand safety-net settings.

Design opportunity: Create systems and support structures that help providers at safety-net hospitals sustain their sense of mission and commitment to public health, despite the demands of patient care.

Insight 1: TIC Requires Additional Time for Provider Education and Clinical Implementation, Which the Productivity-Driven Hospital Workflow Does Not Accommodate

A total of 10 (56%) interview participants stated that an assembly-line structure of hospital workflows does not adapt effectively to the diverse trauma-informed needs of providers and patients. Providers experience a persistent tension between their desire to deliver holistic, person-centered care and the time constraints imposed by a productivity-driven environment. This tension limits opportunities to develop personalized care plans, build meaningful provider-patient relationships, and engage patients in the deeper conversations necessary to uncover past experiences, apprehensions, and emotional needs prior to care. When a fast-paced, profit-driven hospital structure systematically prioritizes efficiency over emotional engagement, the psychological and relational dimensions of patient care are consistently overlooked.

Design opportunity 1: Integrate TIC training into hospital workflows and medical education to support productivity while respecting time constraints.

Interventions to improve TIC within a production-oriented hospital work environment should focus on uplifting patient experiences without coming at the expense of providers' time and mental capacity. These could include integrating training into previously scheduled didactic time or providing support systems that alleviate the burden on providers, reducing the need for them to manage TIC independently.

Insight 2: Providers Who Do Not Receive Support to Address Their Own Vicarious Trauma and Burnout Do Not Have the Personal Resources to Prioritize TIC

Providers working in high-trauma environments face a continuous cycle of emotional suppression that erodes their well-being and long-term commitment to the field. A prominent pattern across interviews was that acute care requires providers to set aside their own emotional responses to deliver consistent care to each patient, yet they often lack adequate opportunities for recovery or resolution. For providers working primarily in acute care settings, where patient encounters are concentrated in moments of crisis, limited visibility into patient recovery and positive outcomes can deepen this emotional toll and distort their broader sense of purpose in their work. When burnout takes hold, providers lose the capacity for empathy, compassion, and sustained attention that TIC demands. The consequences extend beyond individual providers: when providers lack the personal resources to sustain trauma-informed practices, patients feel the absence of compassion in their care, perceiving themselves as unseen or a burden within the health care system. Together, these conditions make it increasingly difficult for providers to sustain the personal and professional resources necessary to prioritize TIC.

Design opportunity 2: Create supportive structures for providers to process vicarious trauma and mitigate burnout, enabling them to sustain and prioritize TIC.

XSL•FO
RenderX

Interventions that provide both the time and resources necessary to reflect on challenging experiences in the hospital setting are crucial. Providing structured opportunities for debriefing, peer support, or access to mental health resources can help mitigate the stress and burnout associated with these experiences. These support mechanisms offer providers the opportunity to improve personal resilience and to engage in their careers with greater compassion and empathy.

Insight 3: TIC Training Is Too Inconsistent to Be Widely Adopted and Too Generic to Address Specialty-Specific Needs, Making It Difficult for Clinicians to Apply in Real-World Care

Interviews revealed both an appreciation for TIC and notable gaps in the current structure and delivery of training. Providers described the need for a standardized foundational curriculum that establishes clear expectations and a shared framework, with the understanding that evidence-based deviations from that standard are appropriate when supported by data. However, foundational knowledge alone was not considered sufficient in practice. One provider who had participated in formal TIC training found that certain recommended approaches, such as asking open-ended questions or engaging patients in a trauma-informed manner, felt unrealistic within the time constraints of their specialty. This highlighted a broader need for training that accounts for the specific demands of different clinical settings while allowing providers the flexibility to adapt their approach to each unique patient situation. Further complicating adoption, TIC is not consistently incorporated into formal health care education. Seven (39%) interviewees described their trauma-informed practices as largely intuitive or drawn from personal and patient experiences rather than structured instruction, reflecting a broader gap between what training prepares them for and what they encounter in clinical practice.

Design opportunity 3: Design a flexible TIC training model that combines a standardized core curriculum with specialty-specific modules, enabling both broad adoption and practical implementation across diverse clinical settings.

Improvements to TIC training should focus on teaching and requiring core components that allow for personalized education, making it easier to meet the differing demands of each specialty and provider role. This could include adding unique modules and courses for learners at all levels of training and across fields.

Insight 4: Hospital Architecture, Security Measures, and Compliance Are Meant to Ensure Safety, but Instead Make Patients Who Have Experienced Trauma Feel Unsafe, Undermining the Goals of TIC

Providers described a fundamental tension between hospital safety requirements and the goal of making patients feel safe. One interviewee described how the physical layout of hospitals, from security checkpoints to locked entry points, and a uniformed police presence, can mirror environments associated with incarceration, creating a sense of threat for patients before care even begins. While providers recognized that law enforcement presence is sometimes warranted given safety risks, they also acknowledged that it introduces an inherent contradiction in a space meant for healing, particularly for patients whose histories include negative interactions with law enforcement. Rather than defaulting to security personnel, providers called for greater investment in de-escalation approaches that address the root of patient distress. This tension extended into clinical decision-making as well. Providers described situations in which standard hospital protocols were applied without consideration of a patient's trauma history, such as the use of restraints on a patient with a documented history of sex trafficking or incarceration. Together, these dynamics place providers in the difficult position of navigating between institutional compliance and their commitment to TIC.

Design opportunity 4: Create physical hospital environments that balance structural requirements, security, and compliance with trauma-informed design principles to ensure a psychological sense of safety for patients.

A third of participants remarked in their interviews that to prevent retraumatizing patients, it is essential to have surroundings where both providers and patients feel safe, highlighting how TIC can be built into the design, structure, and regulations of the environment as well.

Insight 5: When Translation Is Inconsistent or Lacks Cultural Nuance, Providers Struggle to Build Trust and Worry They Are Delivering Lower-Quality Care to Non–English-Speaking Patients

Language and cultural barriers emerged as a source of stress for providers, complicating their ability to build trust and ensure that patients fully understood their care. Providers noted that even when translation services were used consistently, they did not always result in genuine mutual understanding between patients and providers. Interviewees highlighted that cultural understanding is not always guaranteed when using these services and that, without sufficient time, it becomes more challenging to build trust and communicate clearly. Due to the heightened time constraints associated with using translators, many providers shared their discontent with the quality of their appointments, expressing that they felt they were providing suboptimal care to patients who speak languages other than the providers' own.

Design opportunity 5: Enhance the quality and consistency of translation services to help providers deliver clear, compassionate, and patient-centered care to non–English-speaking patients.

Overall, providers emphasized in their interviews the difficulty and importance of holding nuanced care conversations with non–English-speaking patients to account for the emotional, psychological, and cultural elements that are essential to build the trust required for TIC. Interventions should be designed to address these communication gaps between providers, translators, and patients to foster stronger relationships and improve collaboration.

Insight 6: Providers Earn Patient Trust Through Involving Them in Their Care, but Medicine's

XSL•FO
RenderX

Hierarchical Structures Can Impede This Trauma-Informed Approach

The hospital experience can be disorienting for patients, who often find themselves distanced from their sense of control, comfort, and familiar coping mechanisms. Providers described how simply acknowledging this discomfort can be a foundational step in building trust and creating space for collaboration. Providers noted that when patients are treated as active participants in their own care rather than passive recipients of it, they are more likely to engage with and follow through on their care plans. However, structural barriers consistently undermined this approach. One provider described, for example, a culture of deference within hospital hierarchies that discouraged staff from questioning decisions even when doing so was in the patient's best interest. Another interviewee described a system too rigid to accommodate meaningful patient input, one that defaulted to talking at patients rather than with them, and limited the kind of shared decision-making that TIC requires. These examples point to a broader disconnect between what providers recognize as effective, trust-building practice and what the current structure of hospital care allows them to do.

Design opportunity 6: Integrate diverse team viewpoints to address patient concerns and foster improved communication and stronger relationships.

Relationships between patients and provider care teams require interventions that support ongoing communication and provide avenues for voicing opinions and concerns. Since TIC relies on a team-based approach, fostering these relationships will enhance its efficiency and implementation.

Insight 7: The More Deeply Providers Care About Their Patients and Mission, the More Likely They Are to Struggle to Sustain Their Sense of Purpose Over Time in High-Demand Safety-Net Settings

Ten (56%) providers described choosing to work at this safety-net hospital as a deliberate values-driven decision, motivated by a belief that quality health care should be accessible to everyone regardless of their financial or social circumstances. For some, this commitment was deeply personal, rooted in their own experiences of loss, inadequate support, or systemic failure, and shaped by a determination to ensure their patients did not experience the same. Yet sustaining that commitment has come at a significant cost. The emotional weight of the work, compounded by institutional barriers that discourage advocacy and resist change, has prompted some providers to consider leaving. One provider described how attempting to address these issues from within had at times led to conflict with hospital administration, making an already difficult environment even harder to navigate. Together, these experiences reflect a deeper tension between the passion that draws providers to this work and the conditions of a system that too often fails to sustain them within it.

Design opportunity 7: Create systems and support structures that help providers at safety-net hospitals sustain their sense of mission and commitment to public health, despite the demands of patient care.

Interventions that help sustain the drive and morale of safety-net hospital workers include methods that support providers in furthering their interests in the field of public health and offer recognition for the work they do.

Discussion

This study demonstrates that the successful implementation of TIC in acute care settings is not limited by provider willingness, but by structural conditions that constrain their ability to incorporate TIC into clinical care. Time-pressured workflows, insufficient staffing, productivity-driven priorities, and limited access to tailored training collectively create an environment where TIC is difficult to operationalize, even when providers recognize its importance. These constraints contribute to and are exacerbated by burnout and unaddressed moral distress, further reducing providers' capacity to consistently deliver TIC. In response, many clinicians rely on personal intuition and prior experiences rather than formal training, highlighting a gap between institutional preparation and real-world demands. Additionally, features of the care environment such as rigid protocols and security presence may inadvertently undermine patients' sense of safety, directly subverting the goals of TIC. Despite these barriers, providers consistently express a strong commitment to their patients and communities, suggesting that effective solutions must align system design with this intrinsic motivation to enable sustainable, trauma-informed practice.

These findings highlight a self-reinforcing cycle in which exposure to patient trauma and moral distress diminishes providers' capacity to deliver TIC, thereby contributing to further patient trauma and adversely affecting patient outcomes. Repeated exposure to patient trauma can lead to secondary traumatic stress among providers, reducing their ability to consistently deliver compassionate, TIC [21]. The tension between providers' strong desire to deliver TIC and the real-world constraints that prevent them from practicing in alignment with their values results in further moral injury, which is a critical and underrecognized barrier to TIC [22]. The resulting threats to workforce wellbeing constitute a patient safety concern, as they are associated with decreased quality of care, increased medical errors, and reduced patient satisfaction [23]. Collectively, these findings underscore that workforce wellbeing and TIC are essential and interdependent, serving as foundational prerequisites for the delivery of safe, equitable, and TIC.

These findings also demonstrate how the unique context of high-acuity, team-based emergency surgery amplifies known barriers to TIC implementation. The need for rapid decision-making and tightly coordinated multidisciplinary care intensifies the tension between efficiency and human-centered care, making consistent application of TIC particularly challenging [24]. The trifecta of burnout, inadequate support, and structural barriers undermines the reliable integration of TIC principles in high-stakes surgical care, at a time when patients are especially vulnerable to trauma.

Despite these challenges, the foundational goals of TIC, originally developed to recognize and respond to trauma, remain highly relevant given the pervasive prevalence of trauma [1].

XSL•FO
RenderX

A trauma-informed approach is particularly critical in FQHC settings, where patients are disproportionately affected by prior traumatic experiences [3]. Although neither safety-net hospitals nor TIC alone can resolve structural inequities in health care, both are grounded in a shared commitment to advancing health equity and ensuring access to compassionate, patient-centered care.

While our data provide valuable insights, it is important to acknowledge limitations in our study design. The scope of this study only included participants from one FQHC in California. Our study population was recruited primarily through participant referrals, which limited our ability to obtain a sample that reflects the full diversity of the hospital's occupational demographics. This may reflect self-selection bias inherent to voluntary recruitment methods, as individuals with a preexisting interest in TIC may have been more likely to participate. Additionally, our sample included a substantial proportion of attending physicians. While trainees and other providers are often more directly involved in the initial management of trauma patients, attending physicians play a key role in shaping departmental culture and the adoption of trauma-informed practices. Future studies should incorporate more trainees in their study populations to better understand frontline implementation. Our data may be transferable but not generalizable to other hospital systems, as health care providers' experiences may vary across different geographical locations and workplace environments.

Future work will also focus on understanding the perspectives of patients receiving treatment from acute care surgery teams, gathering data on their perceptions of how TIC is delivered in the hospital setting, and identifying which support systems enhance this type of care. Using data from both provider and patient perspectives will enable a more holistic understanding that will allow future researchers to develop prototypes to further improve TIC implementation in FQHCs. By understanding this context, we can gain valuable insights into how to improve the effectiveness and accessibility of TIC, thereby creating the steps necessary to begin the next stage of HCD, the ideation phase. In this phase, prototypes will be developed to address these issues, ultimately improving patient outcomes and fostering a more supportive health care environment.

Acknowledgments

Disclosure of Delegation to Generative AI (GenAI):

The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision:

- Proofreading and editing

The GenAI tool used was: ChatGPT (GPT 5.5).

Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Declaration submitted by: Collective Responsibility.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this study.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to participant privacy and confidentiality considerations.

Authors' Contributions

Conceptualization: WF, AS
Data curation: ATM, WF, MDT
Formal analysis: ATM, NR, CDvH, MGS
Investigation: ATM, WF, MDT
Methodology: ATM, WF, NR, AS
Project administration: CDvH, AG, AS
Resources: ATM, WF
Supervision: CDvH, AG, AS
Validation: ATM
Visualization: ATM, CDvH
Writing – original draft: ATM, MGS
Writing – review & editing: ATM, CDvH, MGS, AG

XSL•FO
RenderX

Conflicts of Interest

AS is the founder and owner of the Empathy Studio. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Interview codebook.
[DOCX File, 2481 KB - periop_v9i1e81004_app1.docx]

References

1. SAMHSA's concept of trauma and guidance for a trauma-informed approach. : Substance Abuse and Mental Health Services Administration; 2014 URL: https://www.nctsn.org/sites/default/files/resources/resource-guide/samhsa_trauma.pdf [accessed 2026-08-26]
2. Loeb DF, Durfee J, Monson S, et al. Prevalence of adverse childhood experiences and post traumatic stress disorder symptoms in a primary care safety-net population: implications for healthcare service needs. Gen Hosp Psychiatry 2022;77:102-108. [doi: 10.1016/j.genhosppsych.2022.04.001] [Medline: 35596962]
3. Blanchard BE, Bluett EJ, Johnson M, Zimberoff A, Fortney JC. Trauma exposure correlates among patients receiving care in Federally Qualified Health Centers. J Trauma Stress 2024 Dec;37(6):864-876. [doi: 10.1002/jts.23055] [Medline: 38743483]
4. Dicker RA, Thomas A, Bulger EM, et al. Strategies for trauma centers to address the root causes of violence: recommendations from the Improving Social Determinants to Attenuate Violence (ISAVE) Workgroup of the American College of Surgeons Committee on Trauma. J Am Coll Surg 2021 Sep;233(3):471-478.e1. [doi: 10.1016/j.jamcollsurg.2021.06.016] [Medline: 34339811]
5. Matthew A, Moffitt C, Huth-Bocks A, Ronis S, Gabriel M, Burkhart K. Establishing trauma-informed primary care: qualitative guidance from patients and staff in an urban healthcare clinic. Children (Basel) 2022 Apr 26;9(5):616. [doi: 10.3390/children9050616] [Medline: 35626792]
6. Glowa PT, Olson AL, Johnson DJ. Screening for adverse childhood experiences in a family medicine setting: a feasibility study. J Am Board Fam Med 2016;29(3):303-307. [doi: 10.3122/jabfm.2016.03.150310] [Medline: 27170787]
7. Schroeder K, Pathak A, Sarwer DB. A call for trauma-informed intensive care. Nurs Outlook 2021;69(5):717-719. [doi: 10.1016/j.outlook.2021.06.001] [Medline: 34332763]
8. Finnerty CC, Mabvuure NT, Ali A, Kozar RA, Herndon DN. The surgically induced stress response. JPEN J Parenter Enteral Nutr 2013 Sep;37(5 Suppl):21S-29S. [doi: 10.1177/0148607113496117] [Medline: 24009246]
9. Hanson JL, Balmer DF, Giardino AP. Qualitative research methods for medical educators. Acad Pediatr 2011;11(5):375-386. [doi: 10.1016/j.acap.2011.05.001] [Medline: 21783450]
10. Nijagal MA, Patel D, Lyles C, et al. Using human centered design to identify opportunities for reducing inequities in perinatal care. BMC Health Serv Res 2021 Jul 20;21(1):714. [doi: 10.1186/s12913-021-06609-8] [Medline: 34284758]
11. Fischer M, Safaeinili N, Haverfield MC, Brown-Johnson CG, Zionts D, Zulman DM. Approach to Human-Centered, Evidence-Driven Adaptive Design (AHEAD) for health care interventions: a proposed framework. J Gen Intern Med 2021 Apr;36(4):1041-1048. [doi: 10.1007/s11606-020-06451-4] [Medline: 33537952]
12. Kalsy M, Burant R, Ball S, Pohnert A, Dolansky MA. A human centered design approach to define and measure documentation quality using an EHR virtual simulation. PLoS One 2024;19(8):e0308992. [doi: 10.1371/journal.pone.0308992] [Medline: 39159187]
13. Mukherjee TI, Zerbe A, Falcao J, et al. Human-centered design for public health innovation: codesigning a multicomponent intervention to support youth across the HIV care continuum in Mozambique. Glob Health Sci Pract 2022 Apr;10(2):e2100664. [doi: 10.9745/GHSP-D-21-00664] [Medline: 35487546]
14. Harte R, Glynn L, Rodríguez-Molinero A, et al. A human-centered design methodology to enhance the usability, human factors, and user experience of connected health systems: a three-phase methodology. JMIR Hum Factors 2017 Mar 16;4(1):e8. [doi: 10.2196/humanfactors.5443] [Medline: 28302594]
15. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544. [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
16. Nelsestuen K, Smith J. Empathy interviews. Learn Prof 2020;41(5) [FREE Full text]
17. Thomas DR. A general inductive approach for analyzing qualitative evaluation data. Am J Eval 2006 Jun;27(2):237-246. [doi: 10.1177/1098214005283748]
18. Dedoose. 2024. URL: http://www.dedoose.com [accessed 2026-08-26]
19. Guest G, Namey E, Chen M. A simple method to assess and report thematic saturation in qualitative research. PLoS One 2020;15(5):e0232076. [doi: 10.1371/journal.pone.0232076] [Medline: 32369511]
20. Lee JJ, Patel D, Gadgil M, Langness S, von Hippel CD, Sammann A. Understanding barriers to home safety assessment adoption in older adults: qualitative human-centered design study. JMIR Hum Factors 2025 Jun 24;12:e66854. [doi: 10.2196/66854] [Medline: 40554799]

XSL•FO
RenderX

21. Bride BE. Prevalence of secondary traumatic stress among social workers. Soc Work 2007 Jan;52(1):63-70. [doi: 10.1093/sw/52.1.63] [Medline: 17388084]
22. Dean W, Talbot S, Dean A. Reframing clinician distress: moral injury not burnout. Fed Pract 2019 Sep;36(9):400-402. [Medline: 31571807]
23. Shanafelt TD, Noseworthy JH. Executive leadership and physician well-being: nine organizational strategies to promote engagement and reduce burnout. Mayo Clin Proc 2017 Jan;92(1):129-146. [doi: 10.1016/j.mayocp.2016.10.004] [Medline: 27871627]
24. Cobianchi L, Dal Mas F, Massaro M, et al. Team dynamics in emergency surgery teams: results from a first international survey. World J Emerg Surg 2021 Sep 16;16(1):47. [doi: 10.1186/s13017-021-00389-6] [Medline: 34530891]

Abbreviations

FQHC: Federally Qualified Health Center
HCD: human-centered design
SAMHSA: Substance Abuse and Mental Health Services Administration
TIC: trauma-informed care

Edited by I Steenstra, N Rohatgi; submitted 21.Jul.2025; peer-reviewed by L Alpay, S Schacher; revised version received 23.Jun.2026; accepted 07.Aug.2026; published 09.Sep.2026.

Please cite as:
McDaniel AT, Frye W, Rashidi N, von Hippel CD, Salvatori MG, Greer A, Diaz Tsuzuki M, Sammann A
Exploring Provider Perspectives on Gaps in Trauma-Informed Care Delivery Within the Acute Care Surgery Timeline at a Federally Qualified Health Center: Human-Centered Design Study
JMIR Perioper Med 2026;9:e81004
URL: https://periop.jmir.org/2026/1/e81004
doi:10.2196/81004

© Angelica T McDaniel, Willow Frye, Neema Rashidi, Christiana D von Hippel, Marianna G Salvatori, Adrienne Greer, Manami Diaz Tsuzuki, Amanda Sammann. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 9.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patient Blood Management Adoption and Transfusion-Threshold Concordance Among Romanian Anesthesiologists: National Survey Analysis

Mihai Gabriel Stefan[1,2*], MD; Daniela Filipescu[1,2*], MD, PhD, DEAA; Dan Corneci[1,3], MD, PhD; Cornelia Predoi[1,2], MD; Silvius Negoita[1,4], MD, PhD; Dorel Sandesc[5,6], MD, PhD; Liana Valeanu[1,7], MD, PhD; Serban Ion Bubenek Turconi[1,2], MD, PhD; Stefan Andrei[8,9], MD, PhD; Dana Tomescu[1,10], MD, PhD

[1]Discipline of Anaesthesiology and Intensive Care Medicine, Carol Davila University of Medicine and Pharmacy, Bucharest, Bucure ti, Romania

[10]Department of Anaesthesiology and Intensive Care, Institutul Clinic Fundeni, Bucharest, Bucure ti, Romania

[2]Department of Anaesthesiology and Intensive Care Medicine, Emergency Institute for Cardiovascular Diseases "Prof. C.C. Iliescu", Sos Fundeni, 258, Bucharest, Bucure ti, Romania

[3]Department of Anaesthesiology and Intensive Care, Spitalul Universitar de Urgenţă Militar Central Dr. Carol Davila, Bucharest, Bucure ti, Romania

[4]Department of Anaesthesiology and Intensive Care, Spitalul Universitar de Urgen ă Elias, Bucharest, Bucure ti, Romania

[5]Department of Anaesthesiology and Intensive Care, Victor Babe University of Medicine and Pharmacy Timi oara, Timisoara, Timi County, Romania

[6]Department of Anaesthesiology and Intensive Care, Spitalul Clinic Judeţean de Urgenţă, Timisoara, Timi County, Romania

[7]Department of Anesthesia and Perioperative Care, University of California, San Francisco, San Francisco, United States

[8]Department of Anaesthesiology and Intensive Care, Hôpital Bichat-Claude-Bernard, Paris, Île-de-France, France

[9]Group of Data Modeling, Computational Biology, and Predictive Medicine, Applied Mathematics, Institut de Biologie de l'École Normale Supérieure, Paris, Île-de-France, France

[*]these authors contributed equally

Corresponding Author:
Mihai Gabriel Stefan, MD
Discipline of Anaesthesiology and Intensive Care Medicine, Carol Davila University of Medicine and Pharmacy, Bucharest, Bucure ti, Romania

Abstract

Background: Romania issued national patient blood management (PBM) guidelines in 2018. A 2023 national survey described the overall uptake of PBM among anesthesiologists but did not examine which clinicians and institutions adopted PBM, nor whether stated transfusion thresholds aligned with a study-defined restrictive reference.

Objective: This study aimed to identify determinants of PBM adoption and to quantify transfusion-trigger concordance, using a planned, exploratory, cross-sectional secondary analysis of the 2023 survey.

Methods: We analyzed 511 responses to the 38-item Romanian Society of Anaesthesia and Intensive Care (SRATI) PBM survey. Three adoption outcomes (applying PBM, a formal hospital PBM group, and routine preoperative anemia treatment) were modeled by multivariable logistic regression against physician grade, hospital category, and local PBM educational events. Eight standardized clinical vignettes (age×cardiac disease×bleeding state) were scored against a restrictive hemoglobin reference (7 g/dL general; 8 g/dL with cardiac disease). The reported timing of the preanesthetic consultation (4 frequency items) was additionally examined as a determinant of preoperative anemia treatment. This was a planned, exploratory secondary analysis; the robustness of the main findings was examined in sensitivity analyses that accounted for survey attrition and for alternative transfusion-threshold scoring rules.

Results: Overall, 73.7% (373/506) reported applying PBM, 33.1% (158/477) reported a formal PBM group, and 33.5% (117/349) reported routine preoperative anemia treatment. Local PBM educational events were the strongest adjusted correlate of reported adoption (applying PBM: adjusted odds ratio [aOR] 9.78, 95% CI 4.33 - 22.07; *P*<.001; formal group: aOR 7.72, 95% CI 4.30 - 13.88; *P*<.001). Earlier preanesthetic consultation was associated with routine preoperative anemia treatment (aOR 1.79 per frequency level for consultation more than 7 days before surgery, 95% CI 1.36 - 2.35; *P*<.001). Transfusion-trigger concordance was lower for older than younger patients in otherwise identical scenarios (eg, active bleeding without cardiac disease: 195/345, 56.5% at ≥65 years vs 288/347, 83% at <65 years), and 56.5% (196/347) endorsed a liberal trigger (hemoglobin ≥8 g/dL) in at least one noncardiac scenario. Self-reported PBM application was not associated with a restrictive transfusion pattern (aOR 1.17, 95% CI 0.69 - 1.98; *P*=.56). Findings were unchanged across scoring rules, and reported PBM application was similar among respondents who did and did not complete the demographic block (246/329, 74.8% vs 127/177, 71.8%; *P*=.53).

XSL•FO
RenderX

Conclusions: In this exploratory cross-sectional analysis, reported PBM adoption was associated more strongly with local educational activity than with hospital category or physician seniority. Self-reported PBM application was not associated with restrictive stated thresholds in hypothetical scenarios, and respondents selected more liberal thresholds for older patients. Standardized multidisciplinary pathways, earlier anemia screening shared with the surgical team, and education focused on transfusion decisions may help address these gaps. Prospective studies using observed process measures, transfusion practice, and clinical outcomes are needed.

(JMIR Perioper Med 2026;9:e106731) doi:10.2196/106731

KEYWORDS

patient blood management; transfusion; anemia; hemoglobin threshold; anesthesiology; survey; implementation; guideline adherence; Romania

Introduction

Patient blood management (PBM) is a patient-centered, evidence-based approach to improve outcomes and the safety of patient care. It rests on 3 pillars: detecting and treating anemia, minimizing blood loss and bleeding, and harnessing physiological tolerance of anemia through judicious, restrictive transfusion [1,2].

Multimodal PBM programs are associated with reduced transfusion and improved outcomes [3,4]. Moreover, large randomized trials and meta-analyses have shown that restrictive red blood cell transfusion strategies are, for most patient groups, as safe as liberal strategies while reducing exposure to allogeneic blood [5-10]. International bodies, including the World Health Organization, have called for the urgent and systematic implementation of PBM and issued a comprehensive guidance document on the subject [11]. Romania issued national PBM recommendations, endorsed by the Romanian Society of Anaesthesia and Intensive Care (SRATI) and the Ministry of Health in 2018, addressing preoperative anemia, perioperative bleeding, and appropriate transfusion [12].

In 2023, 5 years after those recommendations were issued, SRATI conducted a national survey among anesthesiologists to assess PBM implementation. The primary report of that survey described overall uptake, showing, for example, that roughly three-quarters of respondents applied some PBM measures but only a third worked in a hospital with a formal PBM group or routinely treated preoperative anemia [13]. That analysis was, by design, descriptive: it reported the marginal frequency of each response without examining who adopted PBM or whether the transfusion thresholds clinicians endorsed were actually consistent with the restrictive strategy the guidelines recommend.

Two questions therefore remain unanswered and are clinically and organizationally important. First, which clinician- and hospital-level factors are associated with PBM adoption—information needed to target implementation efforts. Second, how closely do stated transfusion triggers align with a study-defined restrictive reference, and does reported PBM application correspond to restrictive stated thresholds? In this planned secondary analysis of the 2023 survey, we addressed both questions, modeling determinants of 3 adoption outcomes and scoring concordance across 8 standardized clinical vignettes.

Methods

Study Design and Data Source

This was a planned, exploratory secondary analysis of the 2023 SRATI national PBM survey [13]. The instrument comprised 38 items (37 closed-response items and 1 open-ended item) covering PBM awareness and application, institutional structures, anemia diagnosis and treatment, bleeding management, transfusion practice, perceived barriers, and respondent and hospital characteristics. The SRATI PBM working group developed the survey and tested its usability and technical functionality before fielding. It was administered using SurveyMonkey, with one question displayed per screen. Invitations were distributed by email to approximately 2000 members through the SRATI database, with additional announcements at society events and in the society newsletter. Respondents formed a self-selected convenience sample of anesthesiologists practicing in Romania. Data were collected between March 9 and September 9, 2023; 512 respondents opened the survey, and the analytical dataset contained 511 records. No new data were collected for this analysis. Reporting followed CHERRIES (Checklist for Reporting Results of Internet E-Surveys) [14]; the completed checklist is provided as a separate supplementary file for editorial review. Throughout, we distinguish reported PBM application (individual self-report), formal institutional PBM structures (reported presence of a hospital PBM group), and transfusion-threshold concordance (agreement of stated triggers with the study-defined restrictive reference).

Ethical Considerations

The parent survey was approved by the SRATI research ethics board (approval 5/3.03.2023). The invitation and introductory page explained the study purpose, investigator, expected completion time, and the planned use and storage of data. Participation was voluntary, questionnaire completion constituted informed consent, and no incentives or compensation were provided. No names or email addresses were collected with survey responses. Although the survey platform recorded IP addresses, these were not used in the analysis, and the analytical dataset contained no direct identifiers. The secondary analysis used the same deidentified dataset covered by the parent approval; no additional consent was required.

XSL•FO
RenderX

Outcomes and Variables

Three binary adoption outcomes were defined: (1) applying PBM in daily practice (yes vs no), (2) presence of a formal hospital PBM group (yes vs no or uncertain), and (3) routine preoperative anemia treatment in the respondent's hospital (usually or always vs sometimes, rarely, or never; the item asked whether preoperative anemia treatment other than transfusion was practiced in the respondent's hospital). Adoption of PBM was self-reported with the single item "Aplica i PBM în practica dumneavoastră curentă?" ("Do you apply PBM in your current practice?"; yes or no), which did not define the individual PBM pillars; all items used in this analysis are reproduced in Romanian and English in Multimedia Appendix 1. Candidate predictors were physician grade (senior consultant, specialist, and resident), hospital category (university or institute, county, municipal or town, and other) and whether PBM educational events had been held locally in the previous 3 years. In an exploratory extension, we also examined the reported timing of the preanesthetic consultation. Four items asked how often the consultation was performed more than 7 days before surgery, more than 1 day but less than 7 days before surgery, on the day before surgery, and on the morning of surgery. Each item used a 5-point frequency scale (never, rarely, sometimes, usually, and always), which we scored 0 to 4 and entered as an ordinal predictor.

Transfusion-Trigger Concordance

Eight survey items presented a balanced 2×2×2 set of clinical vignettes. The factors were patient age (<65 vs ≥65 years), cardiac disease (absent vs present), and active intraoperative bleeding (present vs absent). The older-patient bleeding vignettes specified patients aged 65 years or older, and the older-patient nonbleeding vignettes specified patients older than 65 years; both were analyzed as the ≥65-year stratum. For each vignette, respondents selected the hemoglobin threshold at which they would routinely transfuse (6, 7, 8, or ≥9 g/dL). The bleeding vignettes specified active intraoperative bleeding; the other vignettes described an anemic patient without active bleeding. Cardiac disease was stated explicitly in every scenario. Hemodynamic status, transfusion-related symptoms, frailty, and other comorbidities were not specified. Full wording is provided in Multimedia Appendix 1. Responses were scored against a study-defined restrictive reference consistent with the 2018 Romanian, European, and international recommendations [2,12,15]: a hemoglobin threshold of 7 g/dL for patients without cardiac disease and 8 g/dL for patients with cardiac disease. Each vignette response was scored with 1 of 3 per-vignette rules: the primary rule (concordant if at or below the reference), a strict rule (concordant only if equal to the reference), and a lenient rule (concordant if within 1 g/dL of the reference). The open-ended ≥9 g/dL option was coded as 9 g/dL; it was therefore discordant under the primary and strict rules and concordant under the lenient rule only when the reference is 8 g/dL. Throughout, "concordance" denotes agreement with this study-defined numerical reference and is not a judgment of the clinical appropriateness of an individual transfusion decision. This applies particularly to the active bleeding vignettes, in which hemodynamic status, bleeding rate, symptoms, and anticipated blood loss also govern transfusion in practice but were not specified. At the respondent level, a restrictive pattern was defined as concordance in at least 6 of the 8 vignettes under the rule in use; the primary rule was used for the main analyses and the strict and lenient rules in sensitivity analyses.

Statistical Analysis

Categorical variables are summarized as counts and percentages using item-specific denominators. Each adoption outcome was modeled by multivariable logistic regression using a fixed covariate set: physician grade, hospital category, and local educational events. Odds ratios (ORs) were reported with 95% CI, and model discrimination was reported as the area under the receiver operating characteristic curve (AUC). The covariate set was fixed in advance rather than selected from the data. Physician grade and hospital category were included because individual seniority and institutional setting are established determinants of guideline uptake, and local educational activity was included because it is the most consistently reported modifiable driver of PBM implementation. No stepwise or significance-based variable selection was performed, and the same covariates were used for all 3 outcomes; models extended with institutional infrastructure variables are reported in Multimedia Appendix 2. Hospital sector (public or private) was not entered as a covariate because it was strongly collinear with hospital category—private practice occurred almost exclusively within the "other" hospital category—which produced unstable estimates. Within-respondent contrasts between paired vignettes were tested with the McNemar test, and predictors of a restrictive transfusion pattern (concordance on at least 6 of the 8 vignettes) were examined by logistic regression. Because item completion declined across the questionnaire, the proportion applying PBM was compared between respondents who did and did not complete the demographic block to assess informative attrition. Two-sided $P<.05$ was considered significant; for the family of 8 vignette comparisons, the Benjamini-Hochberg procedure was applied. Analyses were performed in Python 3 (Python Software Foundation; NumPy, SciPy, and pandas); logistic regression coefficients were estimated by maximum likelihood with Wald 95% CIs, and single-predictor estimates were verified against the corresponding 2×2 OR. In this exploratory extension, each ordinal consultation timing score was added separately to the fixed covariate set, and its adjusted odds ratio per 1-level increase in frequency was estimated for each adoption outcome. Ordinal frequency scores were entered as linear terms on the log-odds scale; this assumption was verified by refitting the models with the score as a categorical variable and comparing fits by likelihood-ratio test. Models used complete cases for the covariate set, and the number of respondents and outcome events in each model was reported with each estimate.

Results

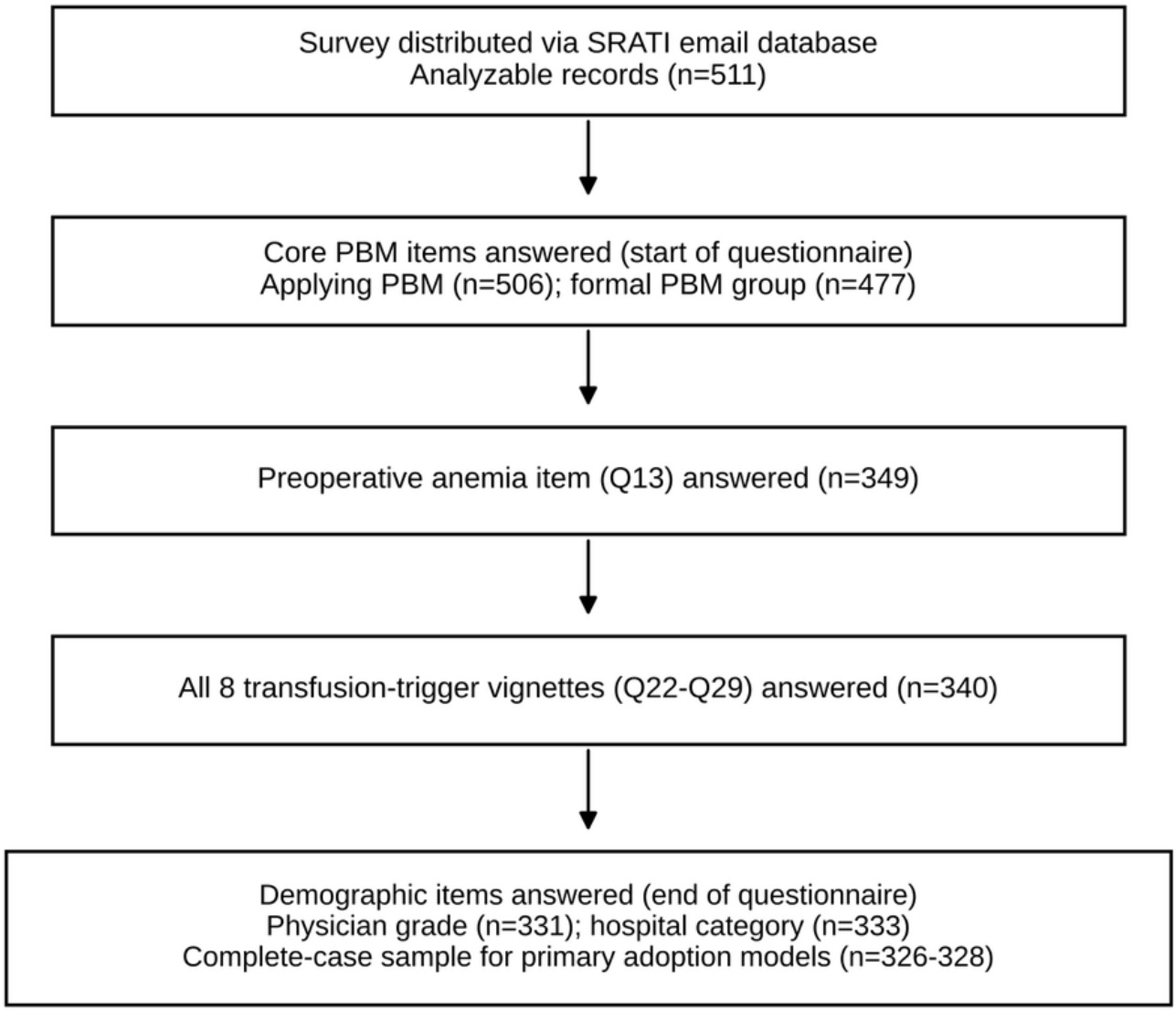
Respondents and Attrition

Of 511 analyzable records, 477 to 506 (93.3-99%) respondents answered the core PBM items near the beginning of the questionnaire. Relative to the approximately 2000 members invited, this corresponds to an approximate invitation-to-response proportion of 26%; the exact number of

XSL•FO
RenderX

active email addresses and the number who opened the invitation were unavailable. Fewer respondents completed the demographic block at the end of the questionnaire: 331 reported physician grade, and 333 reported hospital category (Figure 1). Earlier items, including educational exposure and availability of anemia diagnostic tests, had higher completion (n=348 and n=350, respectively; Table 1). Completion of the demographic block was defined by provision of physician grade, the least-completed item (n=331). Reported PBM application was similar among respondents who did and did not complete this block (246/329, 74.8% vs 127/177, 71.8%; *P*=.53), and the complete-case sample for the adjusted models was 326 to 328. Respondents represented 41 of Romania's 42 county-level administrative units. Among 324 respondents reporting location, 99 (30.6%) practiced in Bucharest.

Figure 1. Respondent flow and item completion through the questionnaire. SRATI: Romanian Society of Anaesthesia and Intensive Care.



XSL•FO
RenderX

Table . Respondent and hospital characteristics[a].

Characteristics	Respondents
Physician grade (n=331), n (%)[b]	
Senior consultant	175 (52.9)
Specialist	112 (33.8)
Resident	44 (13.3)
Hospital category (n=333), n (%)[b]	
University or institute	130 (39)
County	91 (27.3)
Municipal or town	67 (20.1)
Other	45 (13.5)
Sector (n=333), n (%)	
Public	298 (89.5)
Private	17 (5.1)
Public and private	18 (5.4)
Local patient blood management educational events in the previous 3 years (n=348), n (%)	
Yes	136 (39.1)
No	170 (48.9)
Unsure	42 (12.1)
Anemia diagnostic tests available (of 7; n=350), median (IQR)	5 (4 - 5)

[a]Descriptive data were previously reported in the primary survey publication [13] and are reproduced here to characterize the analytical sample.

[b]Physician grade and hospital category were end-of-survey items and had more dropout than the earlier educational and diagnostic items.

PBM Adoption

Overall, 73.7% (373/506) of respondents reported applying PBM, 33.1% (158/477) worked in a hospital with a formal PBM group, and 33.5% (117/349) reported routine preoperative anemia treatment in their hospital, reproducing the headline figures of the primary report. Unadjusted variation by hospital category, grade, and educational exposure is shown in Table 2.

Table . Adoption outcomes by hospital category and physician grade (unadjusted proportions)[a].

Stratum	Applies PBM[b], n/N (%)	Formal PBM group, n/N (%)	Routine preoperative anemia, n/N (%)
Hospital category			
University or institute	102/130 (78.5)	48/130 (36.9)	46/130 (35.4)
County	64/90 (71.1)	21/91 (23.1)	19/91 (20.9)
Municipal or town	44/67 (65.7)	18/67 (26.9)	20/66 (30.3)
Other	38/44 (86.4)	21/45 (46.7)	29/45 (64.4)
Physician grade			
Senior consultant	141/173 (81.5)	58/175 (33.1)	61/174 (35.1)
Specialist	78/112 (69.6)	30/112 (26.8)	37/112 (33)
Resident	27/44 (61.4)	18/44 (40.9)	15/44 (34.1)

[a]Descriptive data were previously reported in the primary survey publication [13] and are reproduced here to characterize the analytical sample. Univariable associations with local educational exposure are reported in Multimedia Appendix 2.

[b]PBM: patient blood management.

In adjusted models (Table 3 and Figure 2), local PBM educational events were the strongest and only consistently significant modifiable correlate across all 3 outcomes. Hospital category was largely not associated with the outcomes after adjustment, and residents were less likely than senior consultants to report applying PBM. Model discrimination was acceptable (AUC 0.77, 0.77, and 0.73). The models included 326 respondents (243 outcome events) for reported PBM application,

XSL•FO
RenderX

328 respondents (106 events) for a formal PBM group, and 327 respondents (113 events) for routine preoperative anemia treatment. The "other hospital" category was associated with routine preoperative anemia treatment but not with reported PBM application or a formal PBM group. This category was small and heterogeneous (n=45); 20 respondents (44.4%) worked wholly or partly in the private sector. Compared with respondents in university hospitals, those in the "other" category more often reported outpatient preanesthetic assessment (23/44, 52.3% vs 52/130, 40%) and consultation more than 7 days before surgery (29/45, 64.4% vs 54/130, 41.5%). These descriptive differences suggest possible pathway and case-mix explanations but do not establish why the association occurred.

Table . Adjusted odds ratios (aORs) from the primary multivariable logistic regression models[a].

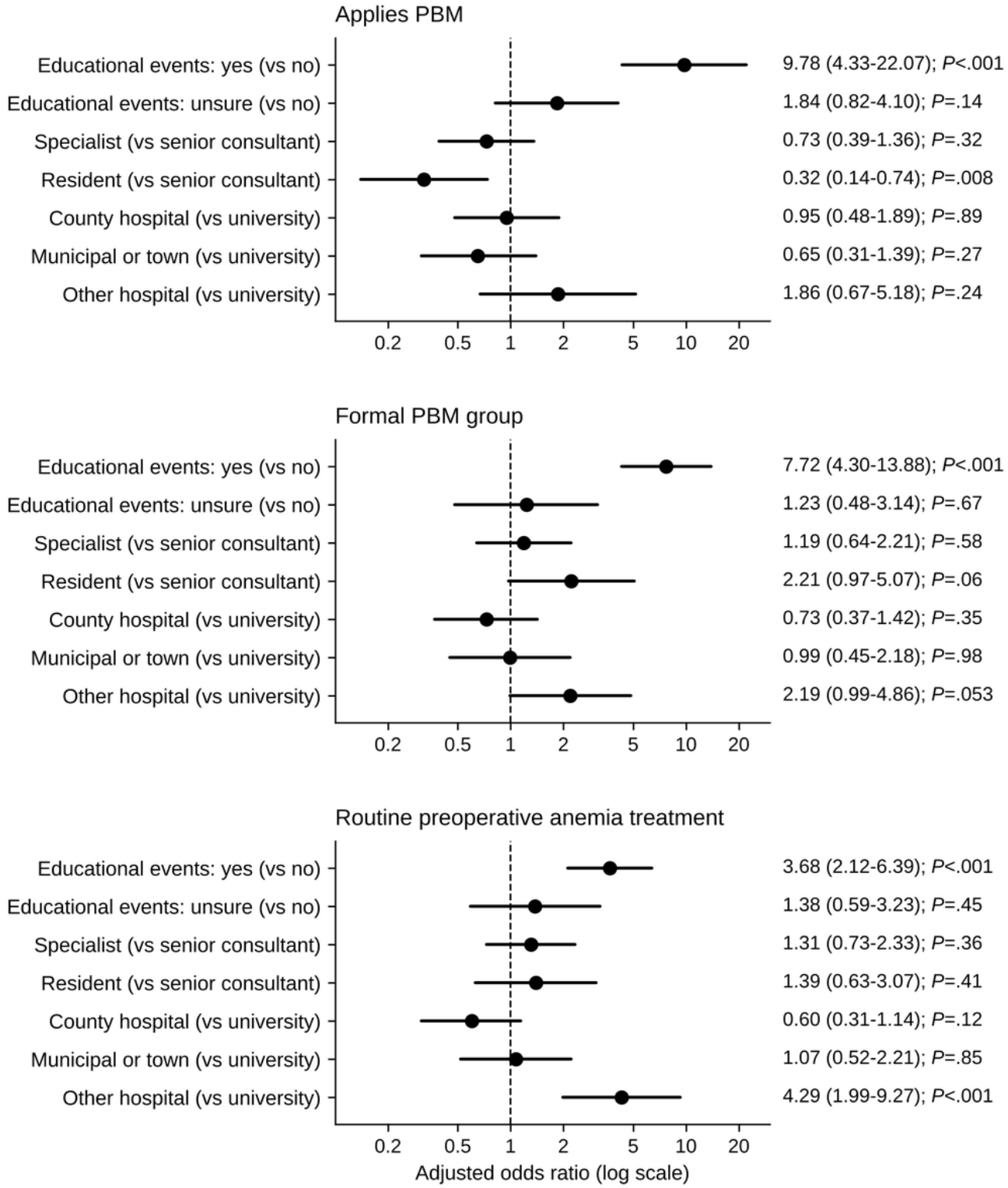
Predictor	Applies PBM[b], aOR (95% CI)	*P* value	Formal PBM group, aOR (95% CI)	*P* value	Routine preoperative anemia, aOR (95% CI)	*P* value
Educational events						
Yes (reference: no)	9.78 (4.33 - 22.07)	<.001	7.72 (4.30 - 13.88)	<.001	3.68 (2.12 - 6.39)	<.001
Unsure (reference: no)	1.84 (0.82 - 4.10)	.14	1.23 (0.48 - 3.14)	.67	1.38 (0.59 - 3.23)	.45
Physician grade						
Specialist (reference: senior consultant)	0.73 (0.39 - 1.36)	.32	1.19 (0.64 - 2.21)	.58	1.31 (0.73 - 2.33)	.36
Resident (reference: senior consultant)	0.32 (0.14 - 0.74)	.008	2.21 (0.97 - 5.07)	.06	1.39 (0.63 - 3.07)	.41
Hospital category						
County (reference: university)	0.95 (0.48 - 1.89)	.89	0.73 (0.37 - 1.42)	.35	0.60 (0.31 - 1.14)	.12
Municipal or town (reference: university)	0.65 (0.31 - 1.39)	.27	0.99 (0.45 - 2.18)	.98	1.07 (0.52 - 2.21)	.85
Other (reference: university)	1.86 (0.67 - 5.18)	.24	2.19 (0.99 - 4.86)	.05	4.29 (1.99 - 9.27)	<.001

[a]*P* values are from 2-sided Wald tests. The model for reported patient blood management (PBM) application included 326 respondents and 243 events (area under the receiver operating characteristic curve [AUC] 0.77); the model for a formal PBM group included 328 respondents and 106 events (AUC 0.77); and the model for routine preoperative anemia treatment included 327 respondents and 113 events (AUC 0.73).

[b]PBM: patient blood management.

XSL•FO
RenderX

Figure 2. Adjusted predictors of the 3 adoption outcomes. Points are adjusted odds ratios, horizontal lines are 95% CIs, and corresponding 2-sided Wald *P* values are shown. PBM: patient blood management.



Timing of the Preanesthetic Consultation

The reported timing of the preanesthetic consultation was itself associated with PBM adoption (n=326 - 328). Respondents who more often assessed patients well before surgery were more likely to report routine preoperative anemia treatment in their hospital. Each 1-level increase in the frequency of consulting more than 7 days before surgery carried an adjusted odds ratio (aOR) of 1.79 (95% CI 1.36 - 2.35) for routine preoperative anemia treatment and 1.87 (95% CI 1.33 - 2.65) for applying PBM. Both models adjusted for physician grade, hospital category, and local educational events. Crudely, routine preoperative anemia treatment was reported by 54.1% (66/122) of respondents who performed the more-than-7-day consultation at least sometimes, versus 22.1% (50/226) of those who did so rarely or never. The gradient reversed for late assessment: more frequent consultation on the morning of surgery was associated with lower odds of routine preoperative anemia treatment (aOR

XSL•FO
RenderX

0.75 per level, 95% CI 0.59 - 0.95). Consultation timing was not associated with the presence of a formal PBM group (aOR 1.13 per level for the more-than-7-day item, 95% CI 0.86 - 1.48). The per-level associations had *P*<.001 for routine preoperative anemia treatment, *P*<.001 for applying PBM, *P*=.39 for the formal PBM group, and *P*=.02 for morning-of-surgery consultation. Treating the timing score as categorical rather than linear did not improve model fit (likelihood-ratio *P*=.82 for preoperative anemia treatment and *P*=.86 for applying PBM). Category-specific ORs also rose monotonically: for preoperative anemia treatment, they were 1.48, 3.54, and 4.95 for "rarely," "sometimes," and "usually" versus "never." Both findings support the linear coding (Multimedia Appendix 2).

Transfusion-Trigger Concordance

Concordance with the study-defined restrictive reference varied across scenarios (Table 4 and Figure 3). Two patterns were consistent. First, concordance was lower for older than younger patients in otherwise identical scenarios (active bleeding without cardiac disease: 195/345, 56.5% at ≥65 years vs 288/347, 83% at <65 years). In the scenario involving an older patient with cardiac disease and active bleeding, 42.7% (148/347) selected hemoglobin ≥9 g/dL. Second, among older patients without cardiac disease, respondents selected more liberal thresholds during active bleeding than in the corresponding scenario without active bleeding (McNemar *P*<.001). Because fixed thresholds are least directly applicable during active bleeding, this contrast describes a difference in stated numerical triggers rather than a measure of decision quality. Overall, 56.5% (196/347) endorsed a liberal trigger (hemoglobin ≥8 g/dL) in at least 1 noncardiac scenario.

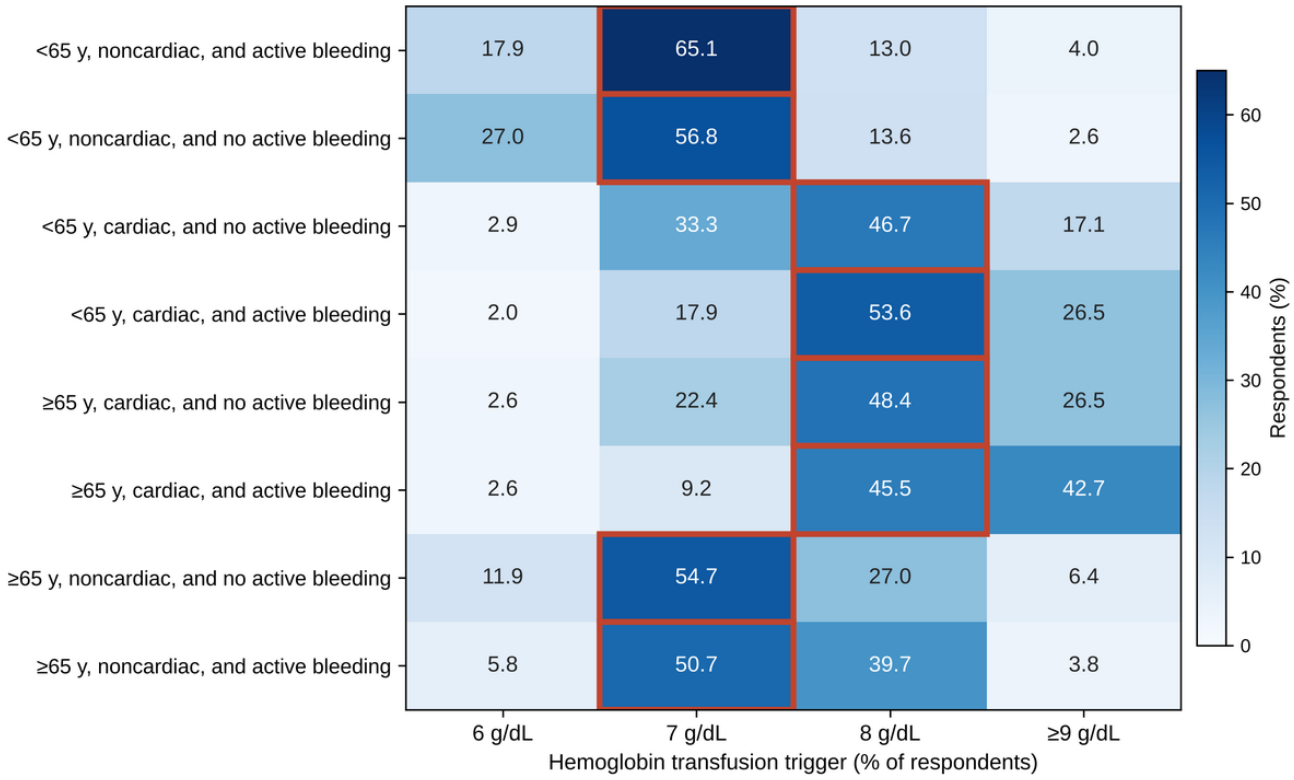
A restrictive pattern was present in 203 (59.7%) of the 340 respondents who answered all 8 vignettes. In the complete-case adjusted model (n=323), self-reported PBM application was not associated with a restrictive pattern (aOR 1.17, 95% CI 0.69 - 1.98; *P*=.56). Thus, reported PBM application did not correspond to more restrictive stated thresholds in these hypothetical scenarios.

Table . Per-scenario concordance, defined as agreement of the stated trigger with the study-defined restrictive hemoglobin reference.

Scenario	Concordant, n (%)
≥65 y, noncardiac disease, and active bleeding (n=345)	195 (56.5)
≥65 y, cardiac disease, and active bleeding (n=347)	199 (57.3)
<65 y, noncardiac disease, and active bleeding (n=347)	288 (83.0)
<65 y, cardiac disease, and active bleeding (n=347)	255 (73.5)
≥65 y, noncardiac disease, and no active bleeding (n=344)	229 (66.6)
≥65 y, cardiac disease, and no active bleeding (n=343)	252 (73.5)
<65 y, noncardiac disease, and no active bleeding (n=345)	289 (83.8)
<65 y, cardiac disease, and no active bleeding (n=345)	286 (82.9)

XSL•FO
RenderX

Figure 3. Distribution of stated hemoglobin transfusion triggers across the 8 scenarios. The outlined cell in each row marks the study-defined restrictive reference; numbers are percentages of respondents. "No active bleeding" reproduces the questionnaire wording and does not imply that hemodynamic stability or symptoms were specified.



Sensitivity Analyses

The direction of all findings was unchanged under the alternative concordance-scoring rules (Table 5); only the absolute level of measured concordance changed, as expected. Under the strict rule, 27.6% (94/340) met the restrictive pattern definition, versus 59.7% (203/340) under the primary rule and 97.9% (333/340) under the lenient rule. The age and bleeding state gradients persisted under all 3 rules.

For the formal PBM group outcome, in which "don't know" responses (91/477, 19.1%) were counted as absence of a group in the primary analysis, a sensitivity model excluding those responses (386 eligible; complete-case regression n=272) gave a materially similar educational events effect (aOR 8.44, 95% CI 4.55 - 15.65, vs 7.72, 95% CI 4.30 - 13.88 in the primary model). The only appreciable change was a strengthening of the resident association (aOR 3.56, 95% CI 1.31 - 9.64, vs 2.21, 95% CI 0.97 - 5.07). This is consistent with residents being more likely to answer "don't know," which itself indicates the limited visibility of PBM structures to trainees.

Table . Concordance under the 3 per-vignette scoring rules among the 340 respondents who answered all 8 vignettes[a].

Scoring rule	Per-scenario concordance, mean (%; SD)	Concordant scenarios (of 8), median (IQR)	Restrictive pattern (≥6/8), n (%)
Primary (less than or equal to the reference)	72.1 (11.1)	6 (4-8)	203 (59.7)
Strict (equal to the reference)	52.9 (6.2)	4 (3-6)	94 (27.6)
Lenient (±1 g/dL)	96.8 (1.4)	8 (8-8)	333 (97.9)

[a]Mean per-scenario concordance is the average of the 8 scenario-level proportions; the restrictive pattern is the respondent-level classification (≥6 of 8 vignettes concordant under the rule in use). The ≥9 g/dL option was coded as 9 g/dL.

Discussion

Principal Findings

This secondary analysis of a national survey yields 3 findings that the descriptive primary report did not reveal. First, reported adoption of PBM in Romania was associated far more strongly with local educational activity than with hospital size or seniority. Hospitals that had held PBM educational events showed roughly 10-fold higher odds of applying PBM and 8-fold higher odds of having a formal PBM group, independent of hospital category and grade. Second, stated transfusion triggers were systematically more liberal for older patients than for younger patients in otherwise identical situations, and most respondents endorsed a liberal threshold in at least 1 noncardiac scenario. Third, and most importantly for implementation, self-reported application of PBM was not associated with a

XSL•FO
RenderX

restrictive pattern of stated thresholds—reported adoption and stated bedside thresholds were disconnected.

Comparison With Prior Work

The primary survey report [13] and international surveys [16] documented incomplete PBM uptake, while a Romanian study identified organizational barriers to implementation [17]. The present analysis extends that work by examining correlates of reported adoption and by quantifying threshold concordance. The association with educational exposure is consistent with evidence supporting locally embedded education reinforced by audit and feedback [18,19]. Preoperative anemia treatment is likewise supported by evidence linking anemia to adverse surgical outcomes and by consensus and trial data on correction strategies [20-22].

Respondents selected more liberal thresholds for older patients, although major trials and meta-analyses have generally not shown harm from restrictive strategies in older or cardiovascular populations [5,10,23]. In the prespecified older subgroup of the Transfusion Requirements in Cardiac Surgery III trial, a restrictive strategy was associated with a lower risk of the composite clinical outcome than a liberal strategy (patients aged 75 years or older: OR 0.70, 95% CI 0.54 - 0.89). Respondents may nevertheless have implicitly attributed frailty or unreported comorbidity to older patients. Cardiac disease was held constant in the factorial design, but such implicit assumptions cannot be excluded in a stated-preference survey.

PBM also encompasses perioperative bleeding management, including tranexamic acid and cell salvage [24,25]. Variation in stated red blood cell thresholds is consistent with evidence that transfusion practice remains heterogeneous across institutions and specialties [26]. Registry data further suggest that transfusion need is concentrated in a minority of patients, supporting risk-stratified implementation [27]. Together, these findings indicate that closing the gap between reported adoption and stated practice requires interventions directed at transfusion decisions as well as institutional structures.

The consultation timing findings offer a plausible mechanism for the low rate of preoperative anemia treatment. Correcting anemia before surgery, whether with iron or by rescheduling, requires lead time. Respondents who routinely assessed patients more than 7 days before surgery were substantially more likely to report routine preoperative anemia treatment in their hospital, whereas those who typically saw patients only on the morning of surgery were less likely to do so. Because this first PBM pillar is time-dependent, organizational pathways that create lead time—not clinician knowledge alone—may be a prerequisite for delivering it. Bringing the preanesthetic consultation forward is only part of the solution: anesthesiologists usually meet the patient after the surgical decision has been made, so responsibility for anemia screening is best shared with the surgical team and initiated at the surgical consultation, when the indication for surgery is established. In some centers, surgeons start the diagnostic workup and, within an agreed multidisciplinary PBM protocol, prescribe iron and selective erythropoietin; 2 recent observational studies in elective arthroplasty illustrate the feasibility of such a surgeon-initiated, shared-care pathway [28,29]. These reports should be read as examples of coordinated pathway design rather than as support for routine erythropoietin in unselected surgical patients. Such a pathway is a modifiable target that complements the educational interventions our other findings support.

The higher odds of routine preoperative anemia treatment in the “other hospital” category require cautious interpretation because university centers might be expected to lead PBM practice. The association was specific to this outcome and was not observed for reported PBM application or a formal PBM group. University hospitals had the highest reported exposure to local educational events (65/129, 50.4%), and the prevalence of formal PBM groups in university hospitals (48/130, 36.9%) exceeded that in county and municipal hospitals, although it was lower than in the “other” category (21/45, 46.7%). The “other” category was small (n=45), heterogeneous, and contained most private or mixed-sector practice in the sample (20/45, 44.4%). Respondents in this category more often reported outpatient assessment and consultation more than 7 days before surgery, which may reflect an elective case mix with more time for anemia correction. Because the study was cross-sectional and case-mix information was unavailable, these observations cannot explain the association. The association should be considered hypothesis-generating rather than evidence that nonuniversity hospitals deliver better PBM overall.

Strengths and Limitations

Strengths include a large national sample, a balanced factorial vignette design enabling within-respondent contrasts, and sensitivity analyses addressing alternative scoring rules and attrition. Several limitations apply. The self-selected convenience sample likely overrepresented clinicians interested in PBM, potentially biasing reported adoption and concordance upward. The exact number of active email addresses and the number of invitation recipients who opened the survey were unavailable; therefore, only an approximate invitation-to-response proportion could be reported. Although respondents covered 41 of 42 county-level administrative units and included different hospital categories and physician grades, population benchmarks were unavailable, so representativeness by geography, institution, and professional seniority could not be established. Reported PBM application was measured with a single item that did not define the 3 PBM pillars and may overestimate implementation of the complete bundle. Similarly, the survey measured the reported presence of formal structures, not their activity. Vignette responses reflect stated intent rather than observed transfusion practice, and concordance with a study-defined hemoglobin reference is not equivalent to the appropriateness of an individual transfusion decision, which also depends on hemodynamic status, bleeding rate, symptoms, anticipated blood loss, and comorbidity; fixed hemoglobin thresholds are least directly applicable during active bleeding, so discordance in the bleeding vignettes should not be equated with inappropriate practice. Cross-sectional associations may be bidirectional: hospitals already committed to PBM may be more likely to organize education, so the education-adoption association should not be interpreted causally. PBM is multidisciplinary, but only anesthesiologists were surveyed; conclusions about hospital-level implementation must therefore be limited to their perspective. The survey did not measure

XSL•FO
RenderX

transfusion rates or clinical outcomes. Finally, the analysis was not prospectively registered and should be considered exploratory and hypothesis-generating.

Practical Implications

These findings suggest 2 complementary implementation priorities. First, locally delivered multidisciplinary education may support reported PBM adoption, but its effectiveness should be evaluated prospectively because the observed association may reflect reverse causation. Second, implementation should address the transfusion decision through approaches such as decision support, single-unit ordering, and audit and feedback, particularly in scenarios involving older patients. Earlier preanesthetic assessment may also create the lead time required for anemia treatment, and anemia screening initiated at the surgical consultation under a shared anesthetic-surgical PBM protocol may help secure that lead time [28,29]. A coordinated national program could involve anesthesiology, surgery, hematology, transfusion medicine, hospital management, residency training, and institutional PBM champions, potentially supported by dedicated funding, for which the economic case of preoperative anemia management in the Romanian setting has been modeled [30]. Linking survey responses to audited process indicators, transfusion use, and clinical outcomes, building on earlier work characterizing perioperative hemoglobin trajectories and predictors of transfusion [31], would allow the relationship between stated and observed practice to be tested directly.

Conclusions

Five years after publication of national recommendations, most responding Romanian anesthesiologists reported applying PBM, while only approximately one-third reported a formal institutional PBM group or routine preoperative anemia treatment. Local educational activity and earlier preanesthetic assessment were associated with reported implementation outcomes, but the cross-sectional design precludes causal interpretation. Self-reported PBM application was not associated with concordance with the study-defined restrictive reference in hypothetical scenarios, particularly those involving older patients. Sharing responsibility for early anemia screening with the surgical team is a practical route to the lead time that anemia correction requires. Prospective multidisciplinary studies using hospital-level process indicators, observed transfusion practice, and clinical outcomes are required.

Acknowledgments

The authors thank the members of the Romanian Society of Anaesthesia and Intensive Care who completed the survey.

During revision, the authors used Claude (Anthropic, Claude Opus 4.8) and Codex (OpenAI, Codex 5.3) to assist with language editing, organization of point-by-point reviewer responses, and consistency checking. All AI-assisted suggestions were critically reviewed and verified by the authors, who take full responsibility for the final content.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The deidentified dataset supporting this secondary analysis is available from the corresponding author on reasonable request, subject to Romanian Society of Anaesthesia and Intensive Care approval.

Authors' Contributions

MGS and DF contributed equally to the conception and design of the secondary analysis, interpretation of results, and drafting of the manuscript. MGS performed the statistical analysis and prepared the figures. All authors contributed to the interpretation of the results, critically revised the manuscript, approved the final version, and accept accountability for the work.

Conflicts of Interest

DF declares honoraria for consulting or lecturing from - CSL Vifor, Werfen Instrumentation Lab, Paloma Tours (organising GPATI Timisoara, Romania), Eurasia Heart Foundation & Krems Donau University University of Continuing Education - Department for Health Sciences, Medicine and Research, Austria, and American Society of Anesthesiology Institutional research grant support from - Vifor Pharma Romania, CSL Vifor, Octapharma, Clinical Trials Unit Strategic Partnerships in Health Excellence, Research, and Engagement (SPHERE), Unity Health Toronto Non financial support - Foundation for Health, Patient Safety and PBM, WFSA, SRATI. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Patient blood management survey Romania—raw questions.
[PDF File, 82 KB - periop_v9i1e106731_app1.pdf]

Multimedia Appendix 2

XSL•FO
RenderX

Full model outputs supporting the analyses reported in the main text.
[PDF File, 115 KB - periop_v9i1e106731_app2.pdf]

References

1. Shander A, Hardy JF, Ozawa S, et al. A global definition of Patient Blood Management. Anesth Analg 2022 Sep 1;135(3):476-488. [doi: 10.1213/ANE.0000000000005873] [Medline: 35147598]
2. Mueller MM, Van Remoortel H, Meybohm P, et al. Patient Blood Management: recommendations from the 2018 Frankfurt Consensus Conference. JAMA 2019 Mar 12;321(10):983-997. [doi: 10.1001/jama.2019.0554] [Medline: 30860564]
3. Althoff FC, Neb H, Herrmann E, et al. Multimodal Patient Blood Management program based on a three-pillar strategy: a systematic review and meta-analysis. Ann Surg 2019 May;269(5):794-804. [doi: 10.1097/SLA.0000000000003095] [Medline: 30418206]
4. tefan M, Tomescu D, Predoi C, et al. Less (transfusion) is more-enhancing recovery through implementation of Patient Blood Management in cardiac surgery: a retrospective, single-centre study of 1174 patients. J Cardiovasc Dev Dis 2023 Jun 22;10(7):266. [doi: 10.3390/jcdd10070266] [Medline: 37504522]
5. Carson JL, Stanworth SJ, Dennis JA, et al. Transfusion thresholds and other strategies for guiding red blood cell transfusion. Cochrane Database Syst Rev 2025 Oct 20;10(10):CD002042. [doi: 10.1002/14651858.CD002042.pub6] [Medline: 41114449]
6. Hébert PC, Wells G, Blajchman MA, et al. A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. Transfusion requirements in critical care investigators, Canadian Critical Care Trials Group. N Engl J Med 1999 Feb 11;340(6):409-417. [doi: 10.1056/NEJM199902113400601] [Medline: 9971864]
7. Carson JL, Brooks MM, Hébert PC, et al. Restrictive or liberal transfusion strategy in myocardial infarction and anemia. N Engl J Med 2023 Dec 28;389(26):2446-2456. [doi: 10.1056/NEJMoa2307983] [Medline: 37952133]
8. Villanueva C, Colomo A, Bosch A, et al. Transfusion strategies for acute upper gastrointestinal bleeding. N Engl J Med 2013 Jan 3;368(1):11-21. [doi: 10.1056/NEJMoa1211801] [Medline: 23281973]
9. Holst LB, Haase N, Wetterslev J, et al. Lower versus higher hemoglobin threshold for transfusion in septic shock. N Engl J Med 2014 Oct 9;371(15):1381-1391. [doi: 10.1056/NEJMoa1406617] [Medline: 25270275]
10. Mazer CD, Whitlock RP, Fergusson DA, et al. Restrictive or liberal red-cell transfusion for cardiac surgery. N Engl J Med 2017 Nov 30;377(22):2133-2144. [doi: 10.1056/NEJMoa1711818] [Medline: 29130845]
11. Guidance on implementing Patient Blood Management to improve global blood health status. World Health Organization. 2024. URL: https://iris.who.int/handle/10665/380784 [accessed 2026-10-05]
12. Filipescu D, Bănăţeanu R, Beuran M, et al. Perioperative Patient Blood Management programme. Multidisciplinary recommendations from the Patient Blood Management Initiative Group. Rom J Anaesth Intensive Care 2017 Oct;24(2):139-157. [doi: 10.21454/rjaic.7518.242.fil] [Medline: 29090267]
13. Filipescu D, tefan MG, Turconi , et al. Patient Blood Management in 2023: a nationwide survey of anesthesiologists in Romania following the 2018 guidelines. Blood Transfus 2025;23(3):232-241. [doi: 10.2450/BloodTransfus.776] [Medline: 39447187]
14. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
15. Kietaibl S, Ahmed A, Afshari A, et al. Management of severe peri-operative bleeding: guidelines from the European Society of Anaesthesiology and Intensive Care: second update 2022. Eur J Anaesthesiol 2023 Apr 1;40(4):226-304. [doi: 10.1097/EJA.0000000000001803] [Medline: 36855941]
16. Gammon R, Jindal A, Dubey R, et al. An international survey of Patient Blood Management practices. Transfusion 2025 Mar;65(3):505-513. [doi: 10.1111/trf.18136] [Medline: 39865438]
17. Kazamer A, Ilinca R, Stanescu-Spinu II, et al. Perceptions of the conditions and barriers in implementing the Patient Blood Management standard by anesthesiologists and surgeons. Health Care (Don Mills) 2024;12(7):760. [doi: 10.3390/healthcare12070760]
18. Meybohm P, Richards T, Isbister J, et al. Patient Blood Management bundles to facilitate implementation. Transfus Med Rev 2017 Jan;31(1):62-71. [doi: 10.1016/j.tmrv.2016.05.012] [Medline: 27317382]
19. Ivers N, Yogasingam S, Lacroix M, et al. Audit and feedback: effects on professional practice. Cochrane Database Syst Rev 2025 Mar 25;3(3):CD000259. [doi: 10.1002/14651858.CD000259.pub4] [Medline: 40130784]
20. Fowler AJ, Ahmad T, Phull MK, Allard S, Gillies MA, Pearse RM. Meta-analysis of the association between preoperative anaemia and mortality after surgery. Br J Surg 2015 Oct;102(11):1314-1324. [doi: 10.1002/bjs.9861] [Medline: 26349842]
21. Muñoz M, Acheson AG, Auerbach M, et al. International consensus statement on the peri-operative management of anaemia and iron deficiency. Anaesthesia 2017 Feb;72(2):233-247. [doi: 10.1111/anae.13773] [Medline: 27996086]
22. Richards T, Baikady RR, Clevenger B, et al. Preoperative intravenous iron to treat anaemia before major abdominal surgery (PREVENTT): a randomised, double-blind, controlled trial. Lancet 2020 Oct 24;396(10259):1353-1361. [doi: 10.1016/S0140-6736(20)31539-7] [Medline: 32896294]
23. Simon GI, Craswell A, Thom O, Fung YL. Outcomes of restrictive versus liberal transfusion strategies in older adults from nine randomised controlled trials: a systematic review and meta-analysis. Lancet Haematol 2017 Oct;4(10):e465-e474. [doi: 10.1016/S2352-3026(17)30141-2] [Medline: 28919087]

XSL•FO
RenderX

24. Myles PS, Smith JA, Forbes A, et al. Tranexamic acid in patients undergoing coronary-artery surgery. N Engl J Med 2017 Jan 12;376(2):136-148. [doi: 10.1056/NEJMoa1606424] [Medline: 27774838]
25. Lloyd TD, Geneen LJ, Bernhardt K, et al. Cell salvage for minimising perioperative allogeneic blood transfusion in adults undergoing elective surgery. Cochrane Database Syst Rev 2023 Sep 8;9(9):CD001888. [doi: 10.1002/14651858.CD001888.pub5] [Medline: 37681564]
26. Gabarin N, Nguyen PU, Li N, et al. Plasma transfusion practice: a five-year audit of plasma transfusion at 23 hospitals. Transfusion 2025 Oct;65(10):1839-1850. [doi: 10.1111/trf.18400] [Medline: 40914907]
27. Rumpf F, Choorapoikayil S, Hof L, et al. Who needs most? Multicenter subanalysis of blood transfusion profiles in the German Patient Blood Management Network Registry. J Clin Med 2026 Feb 26;15(5):1759. [doi: 10.3390/jcm15051759] [Medline: 41827175]
28. Hourlier H, Fricault G, Fennema P. Blood management protocol for baseline anemic patients undergoing hip arthroplasty. Arch Orthop Trauma Surg 2025 Jan 24;145(1):143. [doi: 10.1007/s00402-025-05764-5] [Medline: 39853461]
29. Hourlier H, Fricault G, Fennema P. Low-dose erythropoietin in anemic arthroplasty patients within contemporary Patient Blood Management pathways. J Orthop Surg Res 2026 Apr 18;21(1):340. [doi: 10.1186/s13018-026-06857-6] [Medline: 42001126]
30. Bába LI, Farkas-Ráduly S, Bradacs IA, et al. Budget impact of preoperative anemia management, the first pillar of Patient Blood Management, on the Romanian healthcare system. Clin Drug Investig 2026 Jan;46(1):91-100. [doi: 10.1007/s40261-025-01497-w] [Medline: 41205029]
31. tefan M, Lupu AR, Andrei , et al. Perioperative trajectory of haemoglobin, predictors of blood transfusion and tailoring targets for Patient Blood Management interventions: a single-centre, retrospective study of non-emergent cardiac surgery patients. Rom J Anaesth Intensive Care 2021 Dec;28(2):47-56. [doi: 10.2478/rjaic-2021-0008] [Medline: 36844120]

Abbreviations

aOR: adjusted odds ratio
AUC: area under the receiver operating characteristic curve
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
OR: odds ratio
PBM: patient blood management
SRATI: Romanian Society of Anaesthesia and Intensive Care

Edited by N Rohatgi; submitted 10.Jul.2026; peer-reviewed by H Hourlier, M Aoki; revised version received 08.Sep.2026; accepted 25.Sep.2026; published 16.Oct.2026.

Please cite as:
Stefan MG, Filipescu D, Corneci D, Predoi C, Negoita S, Sandesc D, Valeanu L, Bubenek Turconi SI, Andrei S, Tomescu D
Patient Blood Management Adoption and Transfusion-Threshold Concordance Among Romanian Anesthesiologists: National Survey Analysis
JMIR Perioper Med 2026;9:e106731
URL: https://periop.jmir.org/2026/1/e106731
doi:10.2196/106731

© Mihai Gabriel Stefan, Daniela Filipescu, Dan Corneci, Cornelia Predoi, Silvius Negoita, Dorel Sandesc, Liana Valeanu, Serban Ion Bubenek Turconi, Stefan Andrei, Dana Tomescu. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 16.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing the User Experience of a Perioperative Digital Health Tool for Information Exchange Using a Human-Centered Design Thinking Approach: Qualitative Observational Study

Charlé Steyl[1], MBChB, DCH, DA, MMed; Carljohan Orre[2], MSc, PhD; Greg Foster[3], MSc, PhD; Hanel Duvenage[4], MSc; Michelle S Chew[5,6], MBBS, PhD; Hyla Louise Kluyts[1], MBChB, MMed, DMed

[1]Department of Anaesthesiology, School of Medicine, Sefako Makgatho Health Sciences University, Pretoria, South Africa

[2]Department of Computer Science and Media Technology, Malmö University, Nordenskiöldsgatan 1, Malmö, Sweden

[3]Department of Information Systems, Rhodes University, Makhanda, South Africa

[4]Safe Surgery South Africa, Johannesburg, South Africa

[5]Department of Anaesthesiology and Intensive Care, Linköping University, Linkoping, Sweden

[6]Department of Perioperative Medicine and Intensive Care, Karolinska University Hospital, Stockholm, Sweden

Corresponding Author:
Carljohan Orre, MSc, PhD
Department of Computer Science and Media Technology, Malmö University, Nordenskiöldsgatan 1, Malmö, Sweden

Abstract

Background: Perioperative patient-reported outcomes (PROs) allow patients to share their experiences of surgical procedures with their health care teams using standardized measures. Despite increasing recognition of their value, PROs are not routinely used in clinical practice, partly due to limited evidence of their impact on traditional clinical outcomes and uncertainty among clinicians about their use. Digital health tools offer a promising way to integrate PROs into clinical workflows and enhance patient-clinician interaction, but their success depends on person-centered design to ensure usability and relevance. Safe Surgery South Africa, a nonprofit organization, developed the Perioperative Shared Health Record (PSHR), a secure web-based tool that enables patients to share personal health information and PROs with their anesthetist and surgeon before and after surgery. Initial implementation revealed significant user experience challenges, which contributed to poor uptake.

Objective: This study aimed to explore factors influencing the PSHR user experience in a low- and middle-income country (LMIC) using human-centered design principles.

Methods: This observational qualitative user experience study followed the 5 design thinking stages: empathize, define, ideate, prototype, and test. Semistructured interviews were conducted with postoperative patients from both the public and private health care sectors, including those with and with no prior experience using the PSHR. Thematic analysis followed the 6-phase framework described by Braun and Clarke and was structured using Karagianni's Optimized Honeycomb user experience model. A problem statement was developed, followed by ideation to explore solutions. Paper prototypes were created, refined, and tested through observation, interviews, and validated usability questionnaires.

Results: In the *empathize* stage, 22 interviews were conducted in the private and public health care sectors in South Africa; 7 participants had previous experience using the PSHR. In the *define* stage, participants emphasized the need for connection, feedback, information, and support through their surgical journey. Contrary to expectations, patients were not discouraged by the length of questionnaires if they perceived them as purposeful. In the *ideate* stage, the team considered user expectations and PSHR integration into care processes. In the *prototype* stage, low-fidelity mock-ups were created and refined into paper prototypes. In the *test* stage, testing with 5 participants highlighted the importance of trust, communication, and user-friendly interfaces. Feedback loops and clinician engagement were identified as key motivators for sustained use. The mean usability questionnaire scores indicated excellent usability and high levels of user satisfaction across most domains.

Conclusions: This study is one of the first to apply human-centered design principles to a perioperative digital health tool in an LMIC setting, addressing usability challenges and patient engagement. Key user experience factors influencing patient engagement included communication, feedback, and access to information throughout the surgical journey. Digital health tools such as the PSHR can strengthen communication and support person-centered perioperative care by integrating PROs into clinical workflows and care processes.

(JMIR Perioper Med 2026;9:e79349) doi:10.2196/79349

XSL•FO
RenderX

KEYWORDS

patient-reported outcome measures; person-centered care; patient participation; digital health; perioperative care; user-centered design; user experience research; human-centered design; universal design; design thinking

Introduction

Patients presenting for surgical procedures often feel vulnerable and may become overwhelmed by information that lies outside their usual frame of reference. Many also experience significant physical and emotional symptoms before and after their operation [1,2]. Time constraints and brief interactions during busy ward rounds can limit opportunities for patients to voice concerns or seek clarification [1,3]. In this context, there is a risk that patients feel depersonalized: reduced to passive participants in a system rather than active participants in their own treatment [1]. Furthermore, perioperative clinicians such as surgeons and anesthesiologists often prioritize traditional problem-focused postoperative outcomes such as morbidity and mortality rates, which do not necessarily reflect the outcomes that matter most to individual patients [4-6].

Person-centered care addresses these challenges by recognizing the individual behind the patient: a human being with values, emotions, and goals, and by fostering a partnership that supports patient autonomy and active participation in care decisions [7,8]. Evidence suggests that patients who are empowered and engaged in their own health care may have better outcomes [3,9-11]. Perioperative patient-reported outcomes (PROs) provide one means of achieving this by allowing patients to communicate their experiences of surgical procedures using standardized questionnaires, known as patient-reported outcome and experience measures (PROMs and PREMs) [2,4,12-14]. Various PROs have been defined in the perioperative sphere, including patient satisfaction, quality of recovery (a short-term outcome after surgery), and quality of life (a longer-term outcome after surgery) [14]. Perioperative PROMs and PREMs give individual patients a means to communicate how they are recovering after surgery in such a way that it can be compared between patient groups and procedure types [5,14]. The data can be used to track quality of care over time [6].

Consensus guidelines recommend the use of PROMs and PREMs in clinical research, and their implementation in perioperative care is increasingly studied [2,6,15-18]. In daily practice, however, the routine use of PROs is hampered by their time-intensive nature, limited evidence linking them to traditional outcomes such as complications or mortality, and uncertainty among clinicians and patients regarding their value [2,6,12,16,17]. These challenges may reflect insufficient person-centeredness in the application of PROs and a lack of responsiveness of health care teams to the information provided by patients [2,19].

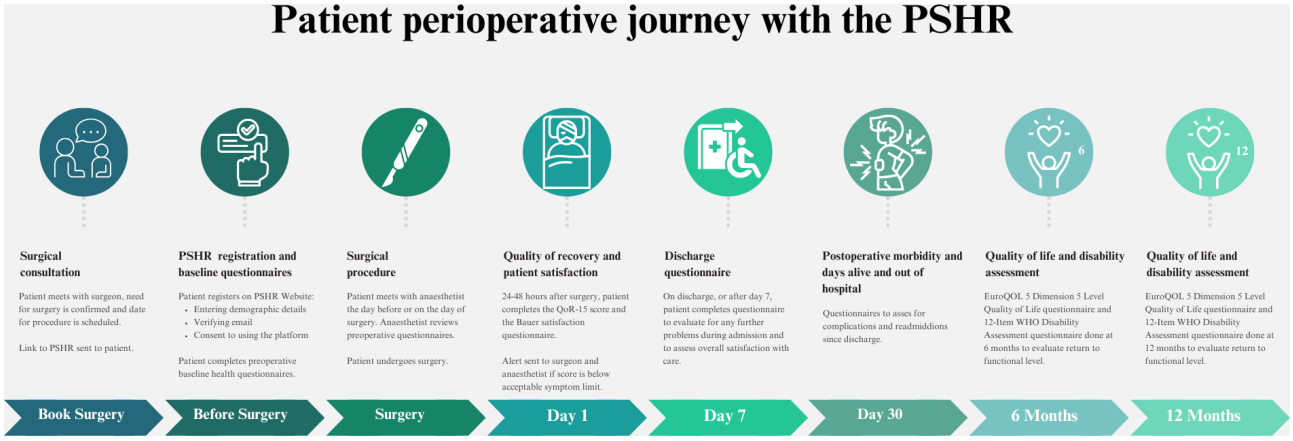
Digital health tools offer a promising way to integrate PROs into clinical workflows and to enhance communication between patients and clinicians [17,18,20]. Achieving this, however, requires a human-centered design approach to ensure that digital tools are responsive to diverse user needs and care contexts [21]. Human-centered design forms part of the broader design thinking framework: an empathetic, iterative process that involves end users throughout development to create tools that are understandable, useful, and enjoyable to use [22-25]. Applying these principles through user experience research, which uses interviews, surveys, and usability testing to explore how people interact with digital systems, helps developers create tools that are more intuitive, engaging, and relevant to real-world care [21,26]. Learning from established digital health platforms and implementation of electronic medical record systems can help create digital health tools that support care across the patient journey [27-31].

Perioperative digital health tools have shown promise in high-income countries (HICs) [17,18], but their implementation and use remain underexplored in low- and middle-income countries (LMICs). In South Africa, barriers to large-scale adoption of digital health solutions include limited digital literacy and unequal access to technology and internet across social, economic, and geographic groups [32,33]. These disparities reflect broader inequalities within the health system, where a tax-funded public sector provides care for most of the population but is underresourced, while a private sector funded through medical schemes and out-of-pocket payments serves a minority yet absorbs a large share of resources [34-36]. The public sector continues to rely largely on paper-based documentation, with uneven implementation of systems to capture routine health information and limited electronic record keeping compared with the private sector [37-41]. The private sector is data-rich, with more electronic data systems, but its datasets are typically siloed and not routinely accessible to public governance systems, clinicians, or patients [39,42]. Neither sector currently supports routine or large-scale capture of PROs, limiting opportunities to measure and improve perioperative care from the patient perspective. Addressing these challenges requires context-specific digital solutions that can be designed to strengthen perioperative care in South Africa.

In response to these challenges, Safe Surgery South Africa [43], a research-driven nonprofit organization, developed the Perioperative Shared Health Record (PSHR) [44], a web-based digital tool enabling patients to share baseline preoperative data and postoperative PROs with their surgeon and their anesthetist for up to a year after surgery. Preoperative data can be used in risk stratification and shared decision-making, whereas postoperative PRO data can improve patient care. Data are stored on a secure server but are accessible to both patient and clinician. The system was designed to function across the public and private health care sectors to promote broader accessibility. Figure 1 describes the use of the PSHR in capturing the perioperative journey of a surgical patient.

XSL•FO
RenderX

Figure 1. Patient journey when using the PSHR. EuroQOL: European Quality of Life questionnaire; PSHR: Perioperative Shared Health Record; QoR-15: 15-Item Quality-of-Recovery questionnaire; WHO: World Health Organization.



Initial use of the PSHR in the private health care sector, during the South African Collaborative Surgical Outcomes Study (SACSOS; ClinicalTrials.gov NCT05052021), identified numerous user experience challenges that led to low patient and clinician engagement, which reduced the effectiveness of the PSHR. The registration process was cumbersome, requiring active support for patient users to complete it. In addition, some questionnaires were perceived as lengthy and burdensome, potentially discouraging patients from completing subsequent assessments. As many of the questionnaires are standardized tools designed for specific purposes, content modifications were not always feasible.

The aim of this observational study was to determine the factors that influence the patient user experience of the PSHR as a tool to support perioperative care. The primary objective was to evaluate the user experience of patients in South Africa who had previously used the PSHR during SACSOS. Further objectives were to explore the user needs of patients who have no prior experience with the PSHR and to gain deeper insights into the future design requirements of the PSHR.

To achieve this aim, this study used a human-centered design thinking approach, which required a multidisciplinary team that could combine clinical insight, technical expertise, and practical experience. In this study, anesthesiologists contributed their understanding of perioperative workflows and patient-clinician communication. The information systems, computer science, and media technology researchers applied user experience and design thinking principles to translate patient needs into feasible design solutions. One of the anesthesiologists (CS), with research expertise in PROs, and one of the information systems researchers (CJO), with personal experience as both a patient and a hospital representative, brought perspectives that ensured that the patient remained central throughout the course of the project. The researchers brought together expertise from South Africa and Sweden, combining experience with emerging and advanced digital health systems and perspectives from LMIC and HIC settings. These efforts aim to improve the PSHR's usability and provide insights for more person-centered digital health design.

Methods

Ethical Considerations

This observational qualitative user experience study was approved by the Sefako Makgatho Health Sciences University Research Ethics Committee on November 16, 2023 (SMUREC/M/513/2023:IR), and registered on the National Health Research Database on January 28, 2025 (NHRD: GP_202501_070). Written consent was obtained from all participants and patient privacy and confidentiality was respected by deidentifying patient data and removing any identifiable features from images used in publication. Participation was voluntary with no compensation paid to participants.

Setting

The study took place in South Africa, with postoperative patients and carers recruited via purposeful sampling in both the private (insurance-funded) and public (tax-funded) health care sectors. These 2 sectors are vastly different in South Africa, with a different patient demographic and a significant difference in availability of resources.

Research Team

The research team consisted of 4 female researchers and 2 male researchers. Three of the female researchers (CS, HK, and MC) are practicing anesthesiologists. HD has a research and development background. HK is the founder of Safe Surgery South Africa and focuses on data-driven solutions to improve perioperative risk stratification and surgical outcomes. CS has a clinical and research interest in PROs and previously recruited patients for the PSHR as part of the SACSOS study, maintaining professional relationships with these participants. One of the male researchers, CJO, comes from a digital health and media background and is a patient representative on a hospital management board after surviving cancer. CJO is also involved in an online cancer rehabilitation program spearheading the use of PROMs and PREMs to improve care processes. The other male researcher, GF, comes from a user experience and design science research background.

XSL•FO
RenderX

Methodology

This study was informed by human-centered design principles. The design thinking process was used to structure the study around 5 phases: empathize, define, ideate, prototype, and test [23,25]. In keeping with a person-centered approach, participants informed the project from the outset by sharing their perioperative experiences and needs. These insights guided ideation and design decisions by the investigators, and participants were reengaged during prototype testing to evaluate and improve solutions based on earlier input.

Phase 1: Empathize

The first phase of the project focused on empathizing with PSHR users by interviewing 2 distinct groups. The first group, recruited from the private health care sector, had prior experience using the PSHR before and after surgery through SACSOS (group 1: PSHR experience). The second group included individuals from both the public and private sectors who had undergone surgery but who had no prior exposure to the PSHR (group 2: no PSHR experience). Participants were invited to take part via email, telephone, or by in-person invitation.

For ease of reference when presenting participant quotations, each participant is assigned a letter prefix. “P” denotes participants in group 1, who had used the PSHR before, “U” denotes participants in group 2 from the public sector, and “I” denotes participants in group 2 from the private sector.

Semistructured interviews were conducted between November 2023 and January 2024. Through storytelling, empathy maps and patient journey maps were created. Demographic data were recorded in REDCap (Research Electronic Data Capture) [45] and exported to MS Excel (version 2411; Microsoft Corp). The interviews began by exploring all patients’ perioperative experiences. Group 1 participants were then asked about their experiences using the PSHR, while group 2 participants received a brief demonstration of the PSHR before discussing their expectations of a digital tool for perioperative information exchange. The interview guide is included in Multimedia Appendix 1.

Interviews were recorded and transcribed using transcription software (Transcribe—Speech to Text, version 4.20.5; DENIVIP Group LLC) on an iPad dedicated to the project. Transcriptions and audiovisual files were stored in a password-protected online folder. Transcriptions were checked for accuracy by CS and HD and reviewed by all investigators before data analysis. Transcriptions of the initial interviews were systematically coded and analyzed thematically using Nvivo (release 14.23.4(49); Lumiere) by CS and HD. The thematic analysis followed the 6-phase framework described by Braun and Clarke [46]. To explore the main themes related to participants’ experiences and expectations of the PSHR, responses were systematically coded and categorized using Karagianni’s Optimized Honeycomb model [47]. This model, commonly used in user experience research, structures the analysis of how users interact with a product by breaking down their experiences into 3 primary dimensions: Use, Feel, and Think [47-51]. By applying this framework, we identified patterns in the data and gained deeper insight into the factors influencing user experience of the PSHR.

Phase 2: Define

The information obtained during phase 1 was used to create a problem statement and summary of findings.

Phase 3: Ideate

Insights from the initial interviews and the defining phase informed an ideation phase, during which various solutions were brainstormed by CS, HD, HK, CJO, and GF to enhance future implementation of the PSHR and also taking into account the interoperability with electronic health records.

Phase 4: Prototype

Paper prototypes for the PSHR were created in Balsamiq Wireframes for Desktop (version 4.8.1; Balsamiq Studios LLC). Paper prototypes were refined based on research team group discussions and during user testing.

Phase 5: Test

User testing with the paper prototypes took place in December 2024 with patients and carers who were recruited via email, telephonically, and in person, with the aim to recruit both patients and carers who had used the PSHR before (“expert users”) to determine whether insights learned from them during initial interviews had improved their user experience. and patients and carers who had no prior experience of the PSHR (“novice users”), to determine their first time user experience with the system.

As all the expert users would be from the private health care sector, novice users were recruited from the public health care sector. To recruit the expert users, attempts were made to contact all 7 participants from group 1; 3 could not be reached at all, 1 initially agreed but later withdrew, and 3 consented and participated. For the novice users, we intentionally sought individuals with no prior exposure to the platform, including through earlier interview phases, to ensure unbiased, first-time user perspectives. This necessitated recruitment of new participants. Eligibility for user testing included being conversant in English, having basic familiarity with mobile phone and computer use and with the use of the internet. Testing was undertaken by 4 investigators (CJO, CS, GF, and HD), 1 acting as the “computer” to change paper “screens” based on user actions, 1 facilitating the scenario, and 2 observing the interaction; sessions were also audio-recorded for later analysis. For ease of reference, “T” denotes participant responses in the user testing phase.

Participants were asked to complete four tasks during the prototype testing: (1) registering and consenting, (2) completing preoperative baseline questionnaires, (3) finding additional information on the PSHR, and (4) completing postoperative quality of recovery and patient satisfaction questionnaires (Multimedia Appendix 2).

The System Usability Scale (SUS) was used to assess usability after the user prototype testing, as this is a well-established tool that has been found to have good reliability to evaluate the usability of digital systems. The SUS is a 10-statement scale

XSL•FO
RenderX

for usability of electronic health applications with good reliability (Cronbach α=0.911) and good face validity [52]. The SUS score ranges from 0 to 100 where higher scores indicate greater usability [53]. A mean SUS score of 68 (SD 12.5) represents the average benchmark for digital health apps [54].

User experience and usability aspects were assessed after prototype testing with the User Experience Questionnaire (UEQ), a well-established tool that evaluates 6 aspects including attractiveness, effectiveness, perspicuity, dependability, stimulation, and novelty with 26 pairs of terms that are scored from 1 to 7 [55]. The questionnaire has good construct validity and good reliability (Cronbach α for 5 of the 6 aspects is above 0.7) [55]. Scoring is done with a downloadable tool, with values ranging between −3 (horribly bad) and +3 (extremely good) [55]. Scores should be evaluated against current benchmarks, freely available for download [56,57].

Prototype user testing was analyzed by 4 investigators (CJO, CS, GF, and HD) who took part in the process using interviews and observation. Thematic analysis of the user tests was done based on research team discussions following the user tests. The SUS and the UEQ were scored in MS Excel according to the guidelines in their reference papers [53-55,57].

Results

Phase 1: Empathize

A total of 22 initial semistructured interviews were conducted as part of empathizing with users. Participant demographics are summarized in (Table 1). All the participants had access to a mobile phone; all but 2 participants in group 2 in the public sector had access to the internet on their mobile phone. Most participants (16/22, 73%) usually used their mobile phones for accessing the internet, whereas 4 out of 22 (18%) participants preferred to use a computer for internet access, and 2 out of 22 (9%) participants did not use the internet at all.

XSL•FO
RenderX

Table . Demographic data for semistructured interviews conducted for phase 1: empathize.

ID[a]	Age (years)	Sex	Language	Race	Prior PSHR[b] experience	Education	Surgery
P1	33	Female	English	Black	Yes	After school qualification	Major abdominal
P2	42	Female	English	White	Yes	After school qualification	Major abdominal
P3	62	Female	Afrikaans	White	Yes	After school qualification	Major abdominal
P4	22	Female	English	Black	Yes	After school qualification	Major abdominal
P5	80	Male	Afrikaans	White	Yes	After school qualification	Major abdominal
P6	72	Male	Afrikaans	White	Yes	After school qualification	Major abdominal
P7	46	Male	Afrikaans	White	Yes	After school qualification	N/A[c]—assisted family member
U1	34	Female	Setswana	Black	No	Secondary school not completed	Major abdominal
U2	22	Male	Northern Sotho	Black	No	Secondary school not completed	Major abdominal
U3	43	Female	Tsonga	Black	No	Secondary school completed	Vascular surgery
U4	65	Male	Setswana	Black	No	Secondary school not completed	Vascular surgery
U5	48	Female	Afrikaans	White	No	Secondary school not completed	Bariatric surgery
U6	29	Female	Northern Sotho	Black	No	After school qualification	Bariatric surgery
U7	42	Female	Afrikaans	Colored[d]	No	After school qualification	Bariatric surgery
U8	39	Female	English	Black	No	After school qualification	Bariatric surgery
U9	42	Male	French	Black	No	Secondary school completed	Vascular surgery
U10	45	Female	Setswana	Black	No	After school qualification	Vascular surgery
I1	64	Female	Afrikaans	White	No	After school qualification	Orthopedic
I2	41	Female	Afrikaans	White	No	After school qualification	Bariatric surgery
I3	54	Female	Afrikaans	White	No	After school qualification	Breast surgery
I4	76	Male	Afrikaans	White	No	After school qualification	Orthopedic
I5	42	Male	Setswana	Black	No	After school qualification	Orthopedic

[a]User ID explanation: “P” denotes participants who had used the PSHR before, “U“ denotes participants from the public sector with no prior experience

XSL•FO
RenderX

of the PSHR, and "I" denotes participants from the private sector with no prior experience of the PSHR.

[b]PSHR: Perioperative Shared Health Record.

[c]N/A: not applicable.

[d]In South Africa, the term "Colored" refers to a distinct cultural and ethnic group with mixed ancestry, recognized as a separate demographic category.

Thematic analysis of the interviews identified 3 main themes: Patient Journey (both groups), PSHR Experience (group 1), and PSHR Expectations (group 2), each with subthemes related to patient engagement and user experience. Detailed findings are provided in Multimedia Appendix 3.

Patient Journey

In understanding the patient perioperative journey (main theme), the following subthemes were identified in both user groups: information-seeking behaviors, emotional response, postoperative difficulties, interaction with health care providers, and advice to other patients. Some participants actively sought more information, either by doing online searches or talking to family members or patients who have been through a similar situation.

> *I was checking [online] how long it's going to be the operation. Okay. Yeah. And how was going to be the pain? How I was cut, a lot of it.* [U1]
>
> *I go and check like, like the food I have to eat. And then the thing I didn't Google about it is the pills the most. But you know that if you want to go look, you can go, you can go find.* [U3]
>
> *So I had my sister-in-law, who is a general practitioner, check for the results and then she was the one [that told me].* [P2]
>
> *And I was also following [on social media], uh, people that will talk about their experience, you know.* [P1]
>
> *I must say the information from other patients helped a lot, knowing what someone else went through, their experiences, how they felt, what the cost implications were, how they paid it, all of that helped a lot.* [I2]

In both the private and public sectors, there were participants who indicated that they avoided looking for any additional information:

> *I don't really want to Google stuff because you always, there's always stuff. Too much information.* [P3]
>
> *I think that would've scared me off a little bit more if I knew truly what was to come.* [P4]
>
> *...because you know when you Google things you don't always get the right information. And it can be very scary.* [I3]
>
> *So, I give up to an extent that I did not even want to stress myself about the Google information. Because others there are just making some speculations.* [U10]

All the participants experienced some form of emotional turmoil in the time after their diagnosis and before they had surgery, with some describing being in denial, feeling helpless, and isolated:

> *I started like shaking and getting worried. Yes. Like now it's getting worse. And like I took it easy, like okay, fine. I went back to work instead of going to the doctor.* [I5]
>
> *But now it all became too much. It just felt like you take one step forward and like five steps back… It felt like I was in constant pain and I also felt very helpless.* [P1]
>
> *It was very... because nobody can come in with you and then you're there alone and then they don't communicate well, doctors all the time, some of them.* [P2]

One participant said an uplifting conversation with her surgeon gave her hope before her surgery and this helped her carry on with her treatment:

> *That answer, that one sentence, and with such conviction, uh, brought back my, um, my hope.* [P3]

Participants in both groups described some physical difficulties in the postoperative period:

> *I think the first two weeks were the hardest really. And the vomiting was much worse at home. Yeah. Uh, the pain also from eating was really terrible.* [P4]
>
> *But that was also the worst thing that I had the operation, because it was very painful! I didn't expect it would be so painful!* [U7]

Two participants commented that being informed and being able to contact their surgeon made the perioperative journey easier.

> *So being informed. Yeah. Makes you feel more reassured.* [P3]
>
> *The interesting thing about this surgeon's practice, that I have not come across before, is that he gives you a 24hr whatsapp number that you can use any time of the day if you have problems or questions. There is always someone that responds—that is not something that everyone would do.* [I4]

Participants in both the public and private sectors described their interaction with their health care providers in positive terms, and they valued in-person communication:

> *It made such a difference that [the anaesthetist] were there and [she] could explain to us what was going to happen, it made us feel a lot more secure and calm.* [P6]
>
> *...what I felt was more these people are taking care of me. I was positive. I could see these other people (points to other patients in ward), they are getting more healthy.* [U2]

Participants offered advice to others that reflected both practical and emotional preparation for surgery. Some emphasized the need to prepare physically by doing breathing exercises and

XSL•FO
RenderX

maintaining mobility, while others recognized the emotional impact of surgery on the patient and their families.

> *It's the emotional side of things that takes quite a toll. But not only on [the patient], but on [the family] too. So I think the emotional strain on both was tough.* [P7]

Several advised future patients to listen to their doctors, follow instructions carefully, and trust the care team. Others highlighted the importance of patience and realistic expectations, especially regarding the time needed for healing:

> *I would tell them that it is very important to listen to what the doctor tells you. To stick to the rules...And I would tell them that they shouldn't be scared to go through with it.* [I1]

> *I would explain my journey the way it is, then they can come here and get that help because it's a better help than any other.* [U3]

> *As a patient, I will say first thing first you need to be patient. You need like...healing is a mercy. It won't just happen overnight.* [I5]

PSHR Experience

Group 1 included 6 patients and 1 family member who had used the PSHR. Their user experience, analyzed according to the Optimized Honeycomb model, is summarized in (Table 2), with supporting quotations in Multimedia Appendix 3.

XSL•FO
RenderX

Table . Codes related to experience or expectations of the Perioperative Shared Health Record[a].

	Subthemes[b,c]						
	Use			Feel		Think	
	Findable	Accessible	Usable	Desirable	Credible[d]	Useful	Valuable
PSHR[e] experience: Group 1: PSHR exposed; private sector; 7 interviews conducted with 4 women and 3 men (2 Black and 5 White, aged 18 - 80 years)	• WhatsApp link preferred (9) • Email not used frequently (5) • Desktop used initially (1)	• Mobile phone access preferred (6) • Device limitation (2) • Postoperation no access to glasses (1) • Not comfortable with technology (1)	• Easy to use (4) • Technical problems/bugs (4) • Login process difficult (2) • Importance of feedback (2) • Font size too small (1) • Medical language tricky (1) • Loss of interest over time (1)	• Can complete at home (2)	• No codes	• Personal connection to doctor (10) • Means to feedback from doctor (8) • Patients' ability to express their needs (5) • Benchmarking (4) • Ability to give family access to platform (1)	• Improved care (5) • Patient involvement (1) • Altruism (1)
PSHR expectations: group 2: PSHR unexposed; public sector; 10 Interviews conducted with 7 women and 3 men (7 Black, 2 White, 1 Colored[f], aged 22 - 65 years)	• No codes	• WhatsApp link preferred (12) • Device limitation (4) • Mobile phone access preferred (3) • Email not used frequently (3) • No access to phone postoperatively (2)	• No codes	• No codes	• No codes	• Communication channel (9) • Means to feedback (5) • Benchmarking (4) • Efficiency (3)	• Improved care (2) • Personal connection (2)
PSHR expectations: group 2: PSHR unexposed; private sector; 5 interviews conducted with 3 women and 2 men (1 Black and 4 White, aged 41 - 76 years)	• No codes	• WhatsApp or SMS (2) • Email not easily accessible (2) • Email link useful (1) • No access to phone postoperatively (1)	• No codes	• No codes	• Negative feedback may impact care (1)	• Efficiency (3) • Source of information (3) • Curated list of information (1) • Information about doctor (1) • Communication channel (1) • Link to other patients/support groups (1)	• Altruism (6)

[a]Main theme: PSHR Experience (group 1) and PSHR Expectations (group 2).

[b]Codes from the text were sorted according to the 7 aspects of user experience from the Honeycomb Model (Findable, Accessible, Usable, Desirable,

XSL•FO
RenderX

Credible, Useful, and Valuable) [47]. Code names are included under each aspect, with the code count in parentheses.

[c]Subthemes: Use, Feel, and Think according to Karagiani's Optimized Honeycomb Model [47].

[d]Note that the aspect Credible falls under both Feel and Think.

[e]PSHR: Perioperative Shared Health Record.

[f]In South Africa, the term "Colored" refers to a distinct cultural and ethnic group with mixed ancestry, recognized as a separate demographic category.

Most participants preferred to follow WhatsApp links to find their way to the PSHR on their smartphones:

> *I do prefer that it was easy to use on my phone. So if it can be improved, it must still just be improved. Mainly for, for like a smartphone.* [P1]
>
> *Yes, because on personal email you're not visiting that often, so it gets lost with all the other stuff. So, I believe your preferred communication is WhatsApp.* [P7,]

One participant commented on the potential difficulty of using a smartphone on the first day postoperatively:

> *It was a bit difficult. Different. If you wear glasses and you don't have glasses on, and you're on morphine. But it wasn't, it wasn't impossible.* [P2]

One of the older participants indicated that they were not comfortable with technology, which is a potential barrier to using a digital tool such as the PSHR:

> *No, I'm not so comfortable with my phone. The internet on there is not something that I usually use.* [P5]

Contrary to expectations, the length of the questionnaires was not perceived in a negative light:

> *It did just go on and on and on. Um, I, I think in my head it was just all part of, just part of the process, preparation and the process you had to do, you know, and making sure that everything is fine.* [P1]
>
> *It was easy to answer. It doesn't take too long.* [P4.]

However, there was some concern about medical jargon and font size:

> *There's some, um, uh, of the wordings and stuff that I really didn't understand.* [P3]
>
> *I think the only complaint if I need to complain about improvements will be the size of the font perhaps.* [P7]

Participants valued the PSHR for enhancing their engagement and improving the quality of care they received:

> *...[the doctor] was able to quickly know and come back and improve my care, you know?* [P1]
>
> *You feel that you were more involved in the planning of your care.* [P2]

Feedback from the surgeon or anesthetist emerged as an important motivator for continued use:

> *[The anaesthetist] had read what was going on. She came and she asked what was going on, and I explained that and she worked around it and talked to the staff.* [P1]
>
> *If I didn't get feedback, I wouldn't have filled in anymore. I would've done the first one and left it at that.* [P2]

Participants appreciated being able to reflect on their recovery:

> *All of this is quite relevant because it lets you think about your own wellbeing and progress.* [P6]

Interestingly, some participants were also motivated by altruism, expressing a desire to help others:

> *...if my information can help somebody else get through a very difficult situation...then I feel it's worth it.* [P3]

Suggestions for improving the PSHR centered around information sharing and being able to contact patients who had been through a similar procedure:

> *I suppose especially for, for large operations, it might help people to know who the anaesthetist is and have like a name and a, a maybe a photograph of your doctor on the system...And maybe info about postoperative care. Because I mean all these ops have different things and I didn't know I was gonna go to need dietary requirements after the first liver operation...So having that as a portal to kind of find information may be useful.* [P2]
>
> *I think having someone else who knows, you know, what you've been through would be nice. Yeah. They can give you kind of, like a perspective on what to expect.* [P4]
>
> *I would prefer to see a video, just a more informal video and then follow up with a verbal conversation just before the operation.* [P7]

PSHR Expectations

Following a brief demonstration of the PSHR, 10 public and 5 private sector postoperative patients with no prior experience of the PSHR (group 2) were interviewed about their expectations of a digital information-sharing tool. Their expectations are summarized in the second part of Table 2, with illustrative quotations provided in Multimedia Appendix 3.

Most participants indicated a preference for accessing the PSHR via a WhatsApp link on their smartphones:

> *I think overall on one's phone is just better, it is more accessible.* [I2]
>
> *We have emails, but we don't use it so much.* [U2]
>
> *Anything that is easy for you is easy for me. But really Whatsapp is easiest.* [U10]

Participants indicated that they would value features such as curated information, feedback from their surgeon or anesthetist, and the ability to track their recovery progress.

XSL•FO
RenderX

> *I don't want to get the information by doing a google search. I want information that comes from the doctor themselves, so that I know it is correct.* [I1]
>
> *You know, if I think back to my work again, clients want to be heard… and now in my setting I am not upset about anything, but it may still be nice to be acknowledged, if I fill something in, it would be nice to get a message or a call to confirm that my responses were seen.* [I3]
>
> *But when you check, keep on checking on your patient, it's good because if I feel something on me, I have to let you know. Then you'll ask me maybe then to come back at hospital. Then you can check that and sort it out.* [U3]
>
> *I mean, if they know what your baseline is, what my baseline is, how my life, my, my health is, you know, then they'll know how to proceed. With any procedure for that matter.* [U5]

Participants also indicated that they would be motivated to use a tool such as the PSHR by knowing that they would contribute data that could help others:

> *I would actually do it more for the greater good to contribute to ongoing medical knowledge and learning.* [I1]
>
> *I think I would still contribute my data if I knew it went for a good cause and if my doctor asked for it.* [I2]
>
> *If it'll help someone with the same problem that I have, it's important to share it.* [I5]

Potential barriers to using the PSHR are high costs of data and low digital confidence:

> *When I'm at home, I don't see that airtime. Because it's a cost of money. It's very expensive.* [U3]
>
> *...all the fancy phones, the internet, all that stuff, that's not for me.* [U4]

Phase 2: Define

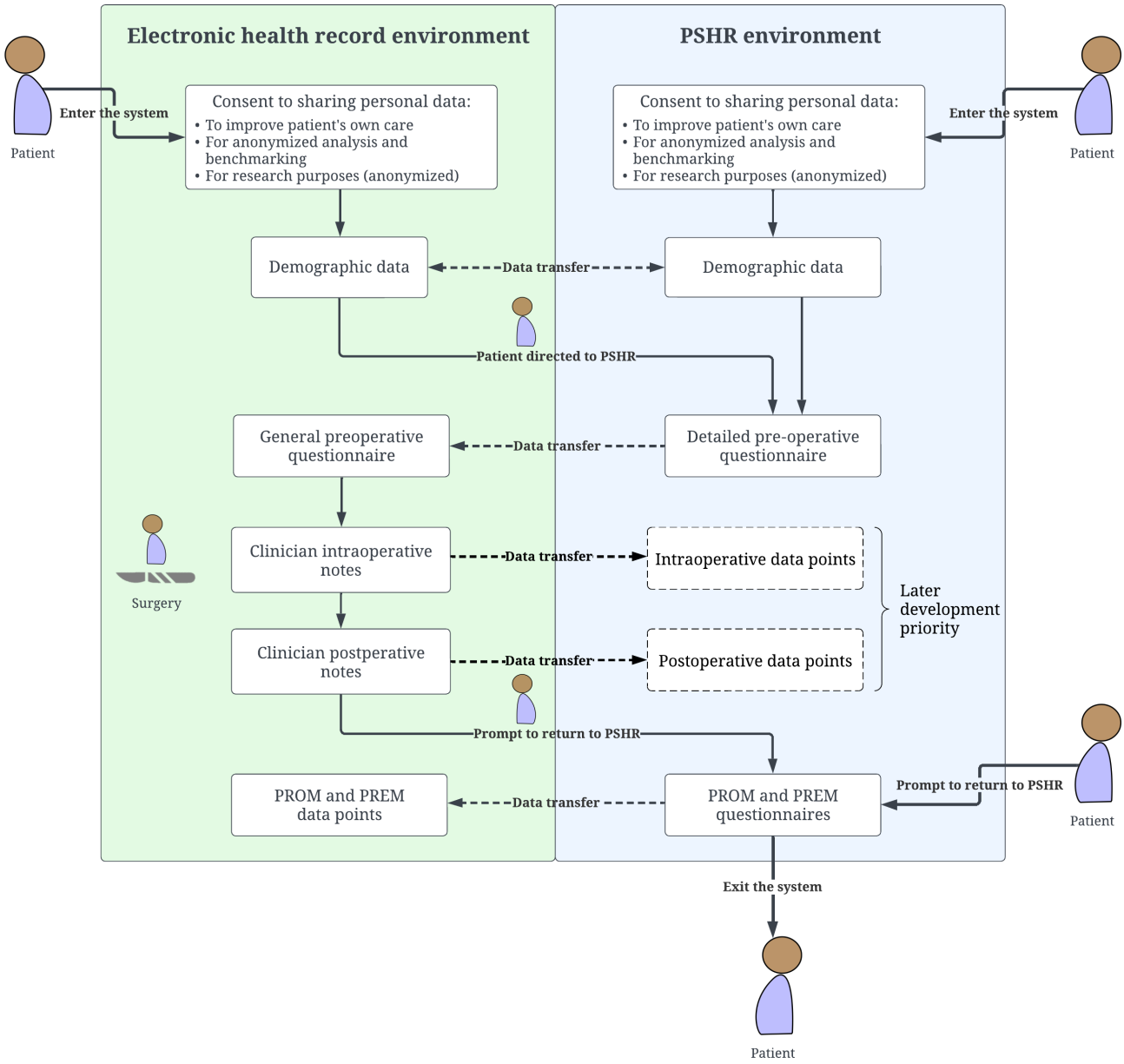
Insights from the initial interviews were that individuals presenting for surgical procedures have a need for connection with and feedback from their health care providers and a willingness to engage in actions necessary to navigate a challenging phase in their lives. Contrary to the investigators' expectations, participants who had used the PSHR were not discouraged by the length of the questionnaires, provided they perceived a clear and meaningful purpose to their completion. In addition, patients expressed a need for information related to their surgical procedures, highlighting the importance of incorporating targeted educational content to support informed decision-making throughout the perioperative journey.

Phase 3: Ideate

The research team brainstormed suggestions and expectations from patient groups, and how the PSHR can be integrated into usual care processes. While standardized questionnaires remained unchanged, their sequence was reorganized to group similar questions, particularly in the PSHR preoperative questionnaire, where multiple risk assessments and surveys are consolidated into a single comprehensive questionnaire. Feedback messages were developed to provide patients with information tailored to their questionnaire responses. Various approaches were explored to support patient-clinician communication through the PSHR. The research team also considered the potential interaction of the PSHR with electronic health records (Figure 2).

XSL•FO
RenderX

Figure 2. Proposed PSHR interaction and interoperability with electronic health records. PREM: patient-reported experience measure; PROM: patient-reported outcomes measure; PSHR: Perioperative Shared Health Record.



Phase 4: Prototype

A series of low-fidelity wireframes were created based on the potential solutions obtained during phase 3. These wireframes were refined into a low-fidelity paper prototype of the PSHR.

Phase 5: User Test

Five individuals consented to participate in prototype testing. Three participants had previously used the PSHR ("expert users"), and 2 participants had no prior experience of the PSHR ("novice users"). Five users have been reported as sufficient for undertaking user testing [58,59]. Demographics are summarized in Table 3. All participants reported having their own mobile phone and usually accessing the internet and their email on their mobile phone and not on a computer. Each testing session took approximately 60 minutes to complete, with the most time spent on the second task. Figure 3 shows paper prototype testing in action; Figure 3A shows a participant discussing task 2 (completing the preoperative questionnaire), and Figure 3B shows a participant responding to a pop-up notification during task 4 (completing postoperative questionnaires).

XSL•FO
RenderX

Table . Demographic data for participants taking part in phase 5: prototype testing.

ID[a]	Age (years)	Sex	Language	Race	Prior PSHR[b] experience	Education	Surgery
T1	42	Female	English	White	Yes	After school qualification	Major abdominal
T2	63	Female	Afrikaans	White	Yes	After school qualification	Major abdominal
T3	46	Male	Afrikaans	White	Yes	After school qualification	N/A[c]—assisted family member
T4	21	Male	Northern Sotho	Black	No	After school qualification	Head and neck
T5	32	Male	Setswana	Black	No	Secondary school completed	Head and neck

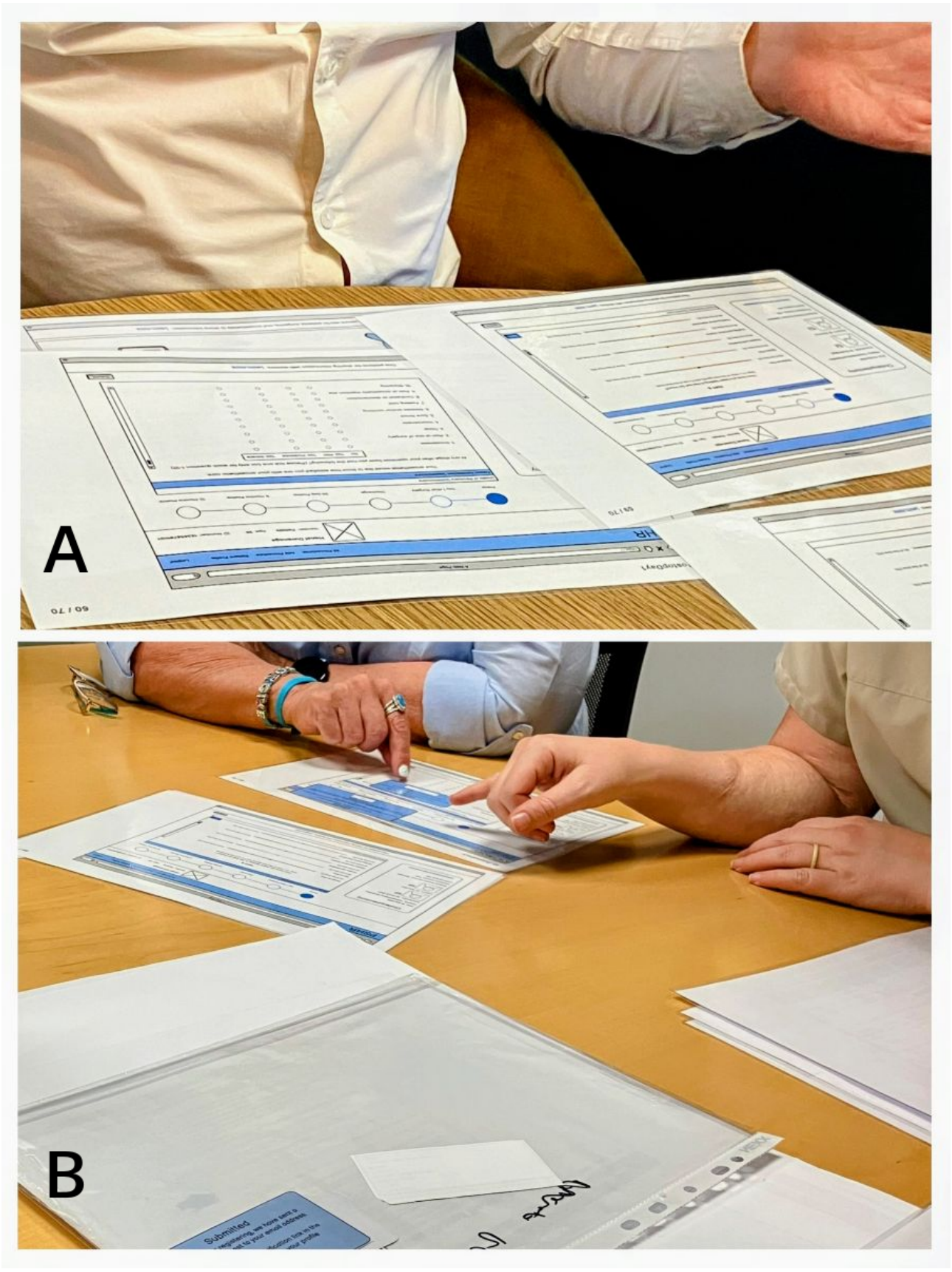
[a]User ID explanation: “T” denotes participant responses in the user testing phase.

[b]PSHR: Perioperative Shared Health Record.

[c]N/A: not applicable.

XSL•FO
RenderX

Figure 3. Paper prototype testing in action. (A) A participant discussing task 2 (completing the preoperative questionnaire). (B) A participant responding to a pop-up notification during task 4 (completing postoperative questionnaires).



For the first task (registering and consenting), trust was an important factor for 2 of the participants. Prior notification by their doctor to expect a registration email would help improve trust when receiving a link to an unknown website:

XSL•FO
RenderX

> *Yeah, knowing that this is safe, yes, because I spoke to you and I know that you will give me something like this. I think a personal call, direct, to say I'm sending you something now. I trust it more, because I know that I'm protected.* [T4]

Two of the participants read the consent form in detail, 2 participants scrolled through with minimal reading, and 1 participant indicated that he would abort the process when confronted with a lengthy consent process:

> *Unless if I'm buying a car, I don't read the details.* [T1]

> *As soon as I get this information, when I get to this one, I will say, yoh aha, this is too much. Pause, pause, pause.* [T4]

The 2 participants who read the consent form doubted that most users would engage deeply with the consent process. One participant suggested that using illustrations or icons could clarify abstract concepts. During the second task (preoperative questionnaire), various data input methods were tested, including radio buttons, colored numerical sliders, and free text blocks. Users seemed to appreciate color coding to interpret the numerical sliders. One participant suggested modifications to the order of questions.

Participants were able to navigate to the information portal (task 3); all participants found general information links useful but preferred procedure-specific content, with 3 favoring video links over text, and 1 mentioning that they would refer to written information only if the desired content was not available in the video links.

By the fourth task, participants were familiar with the layout of the home page and the questionnaires. Their understanding of the timeline had improved, but several suggestions were made to enhance its visual clarity. Feedback messages following questionnaire completion elicited mixed responses; some participants expressed concern that alerts about poor recovery outcomes could cause anxiety:

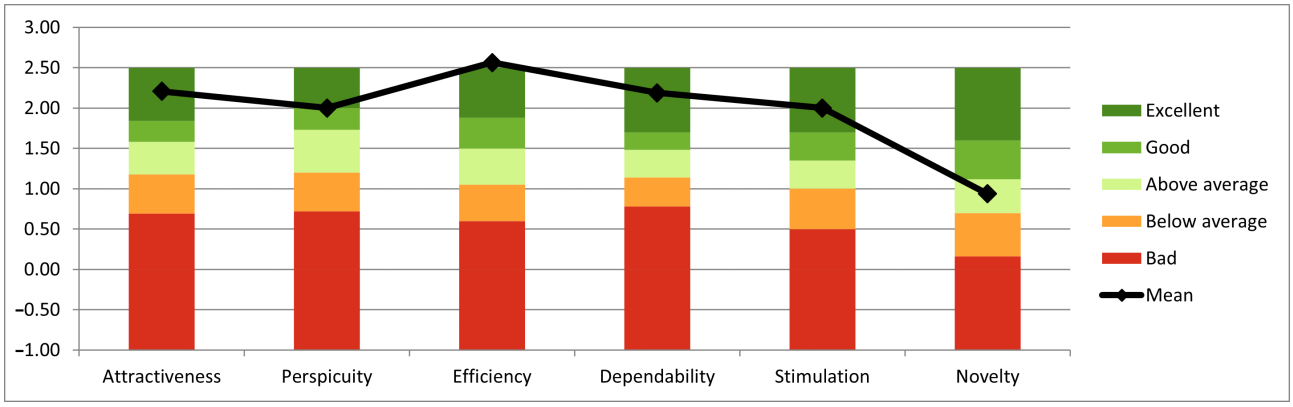
> *It makes me feel worried...I will go back to what I completed [to check] that I completed it correctly. Okay. Because there might be something that I said that might alarm the system.* [T5]

All participants indicated that they would value automatic feedback from their surgeon or anesthetist if they recorded poor scores on their postoperative questionnaires, with expectations for response times ranging from 30 to 60 minutes to up to 48 hours. Participants also noted that a lack of clinician feedback would reduce their motivation to continue using the platform. The potential for the PSHR to enhance patient-doctor relationships is illustrated in this quote:

> *So you still need to get to the buy-in from this. Where's my buy-in coming from? It's coming from the aftercare service, from the doctor, building that relationship. Because [of feedback through the PSHR] I've got a relationship again with the doctor, the surgeon or the rooms. I would be their patient for life! I think that's for me, that's how you get the buy-in to carry on the rest of the process.* [T2]

Four participants (3 experts and 1 novice) completed the SUS and UEQ questionnaires related to the paper prototype testing. The overall mean SUS score was 91.3 (SD 5.7), which indicates very good usability. The mean SUS scores per usability aspect were learnability 91.3 (SD 10.2), efficiency 93.4 (SD 6.9), and satisfaction 89.1 (SD 6.8). The UEQ mean scores for the attractiveness scale are 2.21 (SD 0.6), perspicuity 2.0 (SD 0.9), efficiency 2.6 (SD 0.4), dependability 2.2 (SD 0.6), stimulation 2.0 (SD 0.8), and novelty 0.94 (SD 1.4). The UEQ scores ranged from good to excellent, except for novelty, which scored above average (Figure 4).

Figure 4. Comparing the Perioperative Shared Health Record (PSHR) paper prototype User Experience Questionnaire scores to benchmark data. The measured scale means from our study are compared in relation to existing values from a benchmark dataset, which allows conclusions about the relative quality of the PSHR paper prototypes according to user testing compared with other products.



Discussion

Principal Results

This study applied a human-centered design approach to evaluate and improve patient user experience of a digital health tool developed to capture perioperative patient-reported outcomes. While identifying key usability challenges, it also showed how digital health tools such as the PSHR can help enhance connection between patients and clinicians through information sharing and timely feedback. The findings contribute new evidence from an LMIC setting, where practical integration of PROs into perioperative care is limited. By drawing on patient experiences as a resource for design, the study demonstrates how patient involvement can inform iterative

XSL•FO
RenderX

improvements to digital health tools and strengthen person-centered perioperative care [7,21,22,60,61].

The findings of our study align with established user experience principles in digital health design, emphasizing the importance of empathy, communication, accessibility, regulatory compliance, and data privacy and security [21].

Mapping the patient journey revealed the emotional strain of the perioperative period and the value of designing digital health tools that provide empathetic support during this vulnerable time [21,60]. This aligns with previous research highlighting that emotional engagement and good information provision are central to person-centered perioperative care and patient satisfaction [1,62,63]. Experiences reported in HIC show that access to targeted digital health tools can improve patient well-being and empowerment as well as improve postoperative outcomes [18,61,63,64]. Providing patients with a digital resource that offers clear, accessible information may therefore strengthen patient engagement by improving understanding of recovery and fostering a sense of partnership in care [1,18,65].

A key finding was the importance of communication and feedback from clinicians to create trust and to maintain motivation to continue using the system. This aligns with studies on implementation of electronic health records and patient perspectives on digital health tools [61,66]. It was interesting to note that patients were not deterred by lengthy questionnaires if they perceived them as purposeful. Participants valued knowing that their submitted data would inform their care, similar to evidence that perceived purpose and clinician responsiveness may increase adherence to digital platforms [61,64,67]. However, clinicians may not always see the value of using digital health tools to strengthen relationships with patients, especially if these tools are perceived as adding to their workload [28,66,68]. It is important to note that for patients who used the PSHR, a lack of clinician feedback following the completion of the postoperative questionnaires reduced their motivation to continue engaging with the tool. Therefore, it is important that patient needs are balanced with clinical feasibility. One potential solution would be to use automated alerts from PROM data that notify clinicians when patient responses are below a predefined threshold, prompting timely feedback to those patients who need them. This could enhance the perceived usefulness and reliability of the PSHR and strengthen the patient-clinician relationship, without overburdening the clinician [1].

Using the PSHR to track and benchmark recovery progress can offer reassurance or prompt patients to seek help when needed. Such features can promote self-management by helping patients to understand their recovery and to feel more in control of their health [61,64,67]. From our user testing, it emerged that when automated feedback messages to patients flag potential concerns, these messages should balance the communication of information with reassurance and clear guidance on next steps.

Despite the benefits of using a digital tool such as the PSHR, barriers such as limited digital literacy, the high cost of data, and inconsistent access to internet connectivity are significant obstacles to digital health implementation in South Africa [33,39] Patient preference for accessing the PSHR via WhatsApp links highlights the importance of incorporating widely used, low-barrier communication channels in LMIC settings. This aligns with priorities outlined in the South African Digital Health Strategy and the WHO Global Strategy on Digital Health, which highlights the need for equitable access, user-centered design, and interoperability of digital health tools [20,69].

Regulatory constraints, particularly around the consent process, present a design challenge for the PSHR. While simplifying consent forms with icons and condensed text may improve accessibility, maintaining careful attention to detail and to legal and ethical standards is important to ensure the integrity of the consent process. Providing reliable and up-to-date medical information is resource-intensive, whether creating original content or vetting existing material. One potential solution is to involve patients in content development and curation, fostering a collaborative platform. However, ensuring the accuracy of medical content would still require professional oversight and quality control.

An additional consideration in the design of the PSHR is safeguarding data privacy and security, especially as it collects personal and health information subject to the South African Protection of Personal Information Act, comparable with the Health Insurance Portability and Accountability Act in the United States and the General Data Protection Regulation in Europe. In the South African private health care sector, access to patient data can be restricted to the patient and their designated surgeon and anesthetist, which enhances data security. However, in the public health care sector, where care is provided by teams rather than individuals, maintaining data privacy may be more difficult. Concerns about cybersecurity and a lack of trust in an unknown system were seen by some patients as barriers to engaging with the PSHR.

Strengths and Limitations

A key strength of this study lies in its adherence to design thinking and human-centered design principles [21,22]. The use of Karagianni's Optimized Honeycomb model provided a structured lens for analyzing user experience and expectations, capturing functional cognitive and emotional factors influencing user experience [47]. Involving individual patients in the design process enabled the research team to draw on their lived experience as a form of expertise. By prioritizing the needs of actual patients rather than relying on personas, we aim to advance our mission of developing an intuitive, efficient, user-friendly, and person-centered tool. Furthermore, the inclusion of a diverse sample of patients from both the public and private health care sectors in South Africa enhances the study's relevance, particularly as the country progresses toward universal health coverage. Including in the research team a clinician focused on patient-reported outcomes and a hospital patient representative kept the group focused on a patient-centered approach.

The main limitation of this project is the inclusion of a relatively small patient sample across the various phases, limiting the generalizability of findings. However, from a user experience research perspective, idea saturation in initial interviews suggests sufficient theme coverage, and prototype testing revealed consistent usability issues, aligning with Nielsen and

XSL•FO
RenderX

Landauer's 5-user rule [58,59]. The original study protocol aimed to include patients from Sweden to allow comparisons with a high-resource setting, but logistical and resource constraints prevented this. Future research will focus on expanding data collection and user testing, with a comparative analysis between HICs with well-established digital health platforms and LMICs where digital health systems are still being developed. An additional limitation was that user testing was conducted only with low-fidelity paper prototypes. However, this is a recognized approach within design thinking methodology, enabling iterative development without significant cost investment during the early stages when design elements remain subject to change [23,25,70].

Considering that one of the researchers has a professional interest in PROs and had established rapport with several participants, this may have introduced a subtle positive bias in how participants perceived and articulated the value of the PSHR. Furthermore, as the majority of patients had undergone intermediate or major surgical procedures, their emphasis on the need for information and emotional support may not be generalizable to patients presenting for minor operations.

Another limitation is that the study does not capture the user experience of health care providers. The original project plan included workshops with surgeons and anesthetists; however, time constraints necessitated postponing these activities. Ongoing work by the South African authors includes intentional network weaving to promote data-driven surgery which will include engagement with perioperative clinicians.

Although initially unfamiliar to the anesthesiologist investigators, the qualitative and user experience methodologies provided valuable learning. This collaboration highlighted the importance of such approaches in helping clinicians understand patient needs and to develop intuitive digital health tools.

Future Research

Findings from this study will inform further design iterations of the PSHR, both to optimize its use in individual patient care and to generate future research outputs. Next steps include testing a high-fidelity prototype and evaluating the final product in real care settings, with particular attention to patient experiences over time: for example, how patients respond to repeated questionnaires that may appear similar at different intervals. Data from the PSHR will in time become a resource for organizational development and quality improvement. Ongoing development will require the active involvement of both clinicians and patients to ensure that the tool remains relevant, feasible, and responsive to real-world clinical processes and workflows.

Conclusions

This study is one of the first to apply human-centered design principles to a perioperative digital health tool in an LMIC setting, addressing usability challenges and patient engagement. Key user experience factors influencing patient engagement included communication, feedback, and access to information throughout the surgical journey. Digital health tools such as the PSHR can strengthen communication and support person-centered perioperative care by integrating PROs into clinical workflows and care processes. As health care systems worldwide move toward digital integration, our findings provide valuable insights into the factors to consider when digital health tools are introduced in diverse health care contexts. Future research should focus on integrating digital health tools into clinical workflows and assessing their impact on person-centered outcomes and care delivery, with particular emphasis on involving all relevant stakeholders, both clinicians and patients, to ensure that the tools are contextually appropriate and aligned with real-world processes and workflow needs.

Funding

This work is based on the research supported in part by the National Research Foundation of South Africa (reference no. SASUF23041894320) through a seed grant for collaborative research with the South African Swedish University Forum (SASUF-NRF). The research reported in this manuscript is partially supported by the South African Medical Research Council (SAMRC) through its Division of Research Capacity Development under the Research Capacity Development Initiative from funding received from the South African National Treasury. The content and findings reported/illustrated are the sole deduction, view, and responsibility of the researcher, and do not reflect the official position and sentiments of the SAMRC/SASUF-NRF.

Authors' Contributions

CS participated in study concept and design, data collection, data analysis, data interpretation, writing of the manuscript, and critical revision of manuscript. HD contributed to data collection, data analysis, data interpretation, and critical revision of manuscript. CJO and GF participated in study concept and design, data collection, data analysis, data interpretation, and critical revision of manuscript. MC and HK participated in study concept and design and critical revision of manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Patient semistructured interviews guide.
[DOCX File, 17 KB - periop_v9i1e79349_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Key tasks during paper prototype testing.
[DOCX File, 21 KB - periop_v9i1e79349_app2.docx]

Multimedia Appendix 3
Detailed thematic analysis.
[DOCX File, 57 KB - periop_v9i1e79349_app3.docx]

References

1. Bavin E, Tobiano G, Gillespie BM. Patient experience of elective general surgery: an integrative review. Collegian 2023 Oct;30(5):676-685. [doi: 10.1016/j.colegn.2023.08.002]
2. Wang XS, Gottumukkala V. Patient-reported outcomes: Is this the missing link in patient-centered perioperative care? Best Pract Res Clin Anaesthesiol 2021 Dec;35(4):565-573. [doi: 10.1016/j.bpa.2020.10.006] [Medline: 34801218]
3. Steenkamp L, Silva SA, Collie S. 8-year review of patient satisfaction surveys: Medical scheme members who score initial hospital stay poorly, more likely to have unplanned readmission. 2024. URL: https://www.discovery.co.za/corporate/health-insights-review-patient-surveys-hospitals [accessed 2024-05-02]
4. Roberts GP, Levy N, Lobo DN. Patient-centric goal-oriented perioperative care. Br J Anaesth 2021 Mar;126(3):559-564. [doi: 10.1016/j.bja.2020.12.004] [Medline: 33419527]
5. Myles PS, Boney O, Botti M, et al. Systematic review and consensus definitions for the Standardised Endpoints in Perioperative Medicine (StEP) initiative: patient comfort. Br J Anaesth 2018 Apr;120(4):705-711. [doi: 10.1016/j.bja.2017.12.037] [Medline: 29576111]
6. Billig JI, Sears ED, Travis BN, Waljee JF. Patient-reported outcomes: understanding surgical efficacy and quality from the patient's perspective. Ann Surg Oncol 2020 Jan;27(1):56-64. [doi: 10.1245/s10434-019-07748-3] [Medline: 31489556]
7. Ekman I, Swedberg K, Taft C, et al. Person-centered care--ready for prime time. Eur J Cardiovasc Nurs 2011 Dec;10(4):248-251. [doi: 10.1016/j.ejcnurse.2011.06.008] [Medline: 21764386]
8. Leplege A, Gzil F, Cammelli M, Lefeve C, Pachoud B, Ville I. Person-centredness: conceptual and historical perspectives. Disabil Rehabil 2007;29(20-21):1555-1565. [doi: 10.1080/09638280701618661] [Medline: 17922326]
9. Knutsson B, Kadum B, Eneqvist T, Mukka S, Sayed-Noor AS. Patient satisfaction with care is associated with better outcomes in function and pain 1 year after lumbar spine surgery. J Patient Cent Res Rev 2022;9(1):7-14. [doi: 10.17294/2330-0698.1883] [Medline: 35111878]
10. Çakmak C, Uğurluoğlu Ö. The effects of patient-centered communication on patient engagement, health-related quality of life, service quality perception and patient satisfaction in patients with cancer: a cross-sectional study in Türkiye. Cancer Control 2024;31:10732748241236327. [doi: 10.1177/10732748241236327] [Medline: 38411086]
11. Olsson LE, Karlsson J, Ekman I. The integrated care pathway reduced the number of hospital days by half: a prospective comparative study of patients with acute hip fracture. J Orthop Surg Res 2006 Sep 25;1(1):3. [doi: 10.1186/1749-799X-1-3] [Medline: 17150123]
12. Kingsley C, Patel S. Patient-reported outcome measures and patient-reported experience measures. BJA Educ 2017 Apr;17(4):137-144. [doi: 10.1093/bjaed/mkw060]
13. Weldring T, Smith SMS. Patient-reported outcomes (PROs) and patient-reported outcome measures (PROMs). Health Serv Insights 2013;6(6):61-68. [doi: 10.4137/HSI.S11093] [Medline: 25114561]
14. Moonesinghe SR, Jackson AIR, Boney O, et al. Systematic review and consensus definitions for the Standardised Endpoints in Perioperative Medicine initiative: patient-centred outcomes. Br J Anaesth 2019 Nov;123(5):664-670. [doi: 10.1016/j.bja.2019.07.020] [Medline: 31493848]
15. Myles PS, Grocott MPW, Boney O, Moonesinghe SR, COMPAC-StEP Group. Standardizing end points in perioperative trials: towards a core and extended outcome set. Br J Anaesth 2016 May;116(5):586-589. [doi: 10.1093/bja/aew066] [Medline: 27106961]
16. Sokas C, Hu F, Edelen M, Sisodia R, Pusic A, Cooper Z. A review of PROM implementation in surgical practice. Ann Surg 2022 Jan 1;275(1):85-90. [doi: 10.1097/SLA.0000000000005029] [Medline: 34183512]
17. Tsang C, Lee KS, Richards H, Blazeby JM, Avery KNL. Electronic collection of patient-reported outcomes following discharge after surgery: systematic review. BJS Open 2021 Mar 5;5(2):zraa072. [doi: 10.1093/bjsopen/zraa072] [Medline: 33782708]
18. van der Meij E, Anema JR, Otten RHJ, Huirne JAF, Schaafsma FG. The effect of perioperative E-health interventions on the postoperative course: a systematic review of randomised and non-randomised controlled trials. PLoS One 2016;11(7):e0158612. [doi: 10.1371/journal.pone.0158612] [Medline: 27383239]
19. Coulter A, Locock L, Ziebland S, Calabrese J. Collecting data on patient experience is not enough: they must be used to improve care. BMJ 2014 Mar 26;348:g2225. [doi: 10.1136/bmj.g2225] [Medline: 24671966]
20. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://iris.who.int/server/api/core/bitstreams/1f4d4a08-b20d-4c36-9148-a59429ac3477/content [accessed 2024-08-11]

XSL•FO
RenderX

21. Familoni BT, Odetunde Babatunde S. User experience (UX) design in medical products: theoretical foundations and development best practices. Eng Sci technol j 2024;5(3):1125-1148. [doi: 10.51594/estj.v5i3.975]
22. Melles M, Albayrak A, Goossens R. Innovating health care: key characteristics of human-centered design. Int J Qual Health Care 2021 Jan 12;33(Suppl 1):37-44. [doi: 10.1093/intqhc/mzaa127] [Medline: 33068104]
23. Krolikowski KA, Bi M, Baggott CM, Khorzad R, Holl JL, Kruser JM. Design thinking to improve healthcare delivery in the intensive care unit: promise, pitfalls, and lessons learned. J Crit Care 2022 Jun;69:153999. [doi: 10.1016/j.jcrc.2022.153999] [Medline: 35104692]
24. Doorley S, Holcomb S, Klebahn P, Segovia K, Utley J. Design thinking bootleg. In: Elbogen E, editor. Standford Design School 2018. URL: https://dschool.sfo3.digitaloceanspaces.com/documents/dschool_bootleg_deck_2018_final_sm2-6.pdf [accessed 2025-12-12]
25. An introduction to design thinking: process guide. Institute of Design at Stanford. URL: https://web.stanford.edu/~mshanks/MichaelShanks/files/509554.pdf [accessed 2025-03-02]
26. Gilbert RM. Reimagining digital healthcare with a patient-centric approach: the role of user experience (UX) research. Front Digit Health 2022;4(August):899976. [doi: 10.3389/fdgth.2022.899976] [Medline: 36016600]
27. Essén A, Scandurra I, Gerrits R, et al. Patient access to electronic health records: differences across ten countries. Health Policy Technol 2018 Mar;7(1):44-56. [doi: 10.1016/j.hlpt.2017.11.003]
28. Cijvat CD, Cornet R, Hägglund M. Factors influencing development and implementation of patients' access to electronic health records-a comparative study of Sweden and the Netherlands. Front Public Health 2021;9:621210. [doi: 10.3389/fpubh.2021.621210] [Medline: 34169054]
29. Wiljer D, Urowitz S, Apatu E, et al. Patient accessible electronic health records: exploring recommendations for successful implementation strategies. J Med Internet Res 2008 Oct 31;10(4):e34. [doi: 10.2196/jmir.1061] [Medline: 18974036]
30. Hägglund M, Scandurra I. Patients' online access to electronic health records: current status and experiences from the implementation in Sweden. Stud Health Technol Inform 2017;245:723-727. [Medline: 29295193]
31. Emilsson L, Lindahl B, Koster M, Lambe M, Ludvigsson JF. Review of 103 Swedish healthcare quality registers. Int J Epidemiol 2015 Oct;44(suppl_1):i174-i174. [doi: 10.1093/ije/dyv096.249]
32. Bridgwood A. General household survey. J Public Health Med 1999 Dec;21(4):482. [Medline: 11469378]
33. Khumalo MH, Moloi TS. Barriers to digital transformation in Gauteng's municipal health clinics. J Local Gov Res Innov 2025 Aug 22;6:1-12. [doi: 10.4102/jolgri.v6i0.282]
34. Reddy CL, Makasa EM, Biccard B, et al. Surgery as a component of universal healthcare: where is South Africa? S Afr Med J 2019 Aug 28;109(9):624-625. [doi: 10.7196/SAMJ.2019.v109i9.14233] [Medline: 31635583]
35. Coovadia H, Jewkes R, Barron P, Sanders D, McIntyre D. The health and health system of South Africa: historical roots of current public health challenges. Lancet 2009 Sep 5;374(9692):817-834. [doi: 10.1016/S0140-6736(09)60951-X] [Medline: 19709728]
36. Day C, Gray A, Cois A, Ndlovu N, Massyn N, Boerma T. Is South Africa closing the health gaps between districts? Monitoring progress towards universal health service coverage with routine facility data. BMC Health Serv Res 2021 Sep 13;21(Suppl 1):194. [doi: 10.1186/s12913-021-06171-3] [Medline: 34511085]
37. Mayosi BM, Lawn JE, van Niekerk A, et al. Health in South Africa: changes and challenges since 2009. Lancet 2012 Dec 8;380(9858):2029-2043. [doi: 10.1016/S0140-6736(12)61814-5] [Medline: 23201214]
38. Thabethe SS, Mathews VE. Quality of routine health data in DHIS2 in South Africa: Eastern Cape province from 2017–2020. S Afr J Inf Manag 2024;26(1):1903. [doi: 10.4102/sajim.v26i1.1903]
39. Zharima C, Griffiths F, Goudge J. Exploring the barriers and facilitators to implementing electronic health records in a middle-income country: a qualitative study from South Africa. Front Digit Health 2023;5:1207602. [doi: 10.3389/fdgth.2023.1207602] [Medline: 37600481]
40. Kluyts HL. Strengthening surgical systems in LMICs: data-driven approaches. Lancet Global Health 2024 Nov;12(11):e1744-e1745. [doi: 10.1016/S2214-109X(24)00375-9] [Medline: 39245052]
41. Nicol E, Hanmer LA, Mukumbang FC, Basera W, Zitho A, Bradshaw D. Is the routine health information system ready to support the planned national health insurance scheme in South Africa? Health Policy Plan 2021 Jun 1;36(5):639-650. [doi: 10.1093/heapol/czab008] [Medline: 33822055]
42. Matshidze P, Hanmer L. Health information systems in the private health sector. In: South African Health Review: Health Systems Trust; 2007. URL: https://www.hst.org.za/publications/South%20African%20Health%20Reviews/11_SAHR_2006-2007_Section%206.pdf [accessed 2025-12-12]
43. Transforming surgical care in Africa through data-driven innovation. Safe Surgery SA. URL: https://safesurgery.co.za [accessed 2020-10-15]
44. Perioperative shared health record. Safe Surgery SA. URL: https://safesurgery.co.za/pshr/ [accessed 2023-04-08]
45. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data Capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
46. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]

XSL•FO
RenderX

47. Karagianni K. Optimizing the UX honeycomb. Medium. 2018. URL: https://uxdesign.cc/optimizing-the-ux-honeycomb-1d10cfb38097 [accessed 2024-10-13]
48. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
49. Im D, Pyo J, Lee H, Jung H, Ock M. Qualitative research in healthcare: data analysis. J Prev Med Public Health 2023 Mar;56(2):100-110. [doi: 10.3961/jpmph.22.471] [Medline: 37055353]
50. Mansson L, Wiklund M, Öhberg F, Danielsson K, Sandlund M. Co-creation with older adults to improve user-experience of a smartphone self-test application to assess balance function. Int J Environ Res Public Health 2020 May 26;17(11):3768. [doi: 10.3390/ijerph17113768] [Medline: 32466484]
51. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004 Feb;24(2):105-112. [doi: 10.1016/j.nedt.2003.10.001] [Medline: 14769454]
52. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the System Usability Scale. Int J Hum Comput Interact 2008 Jul 29;24(6):574-594. [doi: 10.1080/10447310802205776]
53. Brooke J. SUS: A “quick and dirty” usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation in Industry, 1st edition: Taylor & Francis; 1996.
54. Hyzy M, Bond R, Mulvenna M, et al. System Usability Scale benchmarking for digital health apps: meta-analysis. JMIR Mhealth Uhealth 2022 Aug 18;10(8):e37290. [doi: 10.2196/37290] [Medline: 35980732]
55. Laugwitz B, Held T, Schrepp M. Construction and Evaluation of a User Experience Questionnaire: Springer; 2008, Vol. 5298 LNCS:63-76. [doi: 10.1007/978-3-540-89350-9_6]
56. A fast and reliable questionnaire to measure the User Experience of interactive products. User Experience Questionnaire (UEQ). URL: https://www.ueq-online.org/ [accessed 2025-01-19]
57. Schrepp M, Hinderks A, Thomaschewski J. Construction of a benchmark for the User Experience Questionnaire (UEQ). Int J Interact Multim Artif Intell 2017;4(4):40. [doi: 10.9781/ijimai.2017.445]
58. Nielsen J, Landauer TK. A mathematical model of the finding of usability problems. 1993 Presented at: Conference on Human Factors in Computing Systems—Proceedings Published by ACM; Apr 24-29, 1993 URL: http://portal.acm.org/citation.cfm?doid=169059 [accessed 2025-12-19] [doi: 10.1145/169059.169166]
59. Nielsen J. Why you only need to test with 5 users. Nielsen Norman Group. 2000. URL: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/ [accessed 2025-01-19]
60. Molina-Recio G, Molina-Luque R, Jiménez-García AM, Ventura-Puertos PE, Hernández-Reyes A, Romero-Saldaña M. Proposal for the user-centered design approach for health apps based on successful experiences: integrative review. JMIR Mhealth Uhealth 2020 Apr 22;8(4):e14376. [doi: 10.2196/14376] [Medline: 32319965]
61. Madanian S, Nakarada-Kordic I, Reay S, Chetty T. Patients’ perspectives on digital health tools. PEC Innov 2023;2:100171. [doi: 10.1016/j.pecinn.2023.100171]
62. Jones CH, O’Neill S, McLean KA, Wigmore SJ, Harrison EM. Patient experience and overall satisfaction after emergency abdominal surgery. BMC Surg 2017;17(1):76. [doi: 10.1186/s12893-017-0271-5]
63. Sansalone E, White J, Zucca A, et al. Acceptability of the RecoverEsupport digital health intervention among patients undergoing breast cancer surgery: qualitative study. J Med Internet Res 2025 Oct 17;27:e77567. [doi: 10.2196/77567] [Medline: 41105868]
64. ElKefi S, Asan O. How technology impacts communication between cancer patients and their health care providers: a systematic literature review. Int J Med Inform 2021 May;149:104430. [doi: 10.1016/j.ijmedinf.2021.104430] [Medline: 33684711]
65. Hladkowicz E, Yachnin D, Boland L, et al. Evaluation of a preoperative personalized risk communication tool: a prospective before-and-after study. Can J Anaesth 2020 Dec;67(12):1749-1760. [doi: 10.1007/s12630-020-01809-y] [Medline: 32929659]
66. Hägglund M, Scandurra I. Usability of the Swedish accessible electronic health record: qualitative survey study. JMIR Hum Factors 2022 Jun 23;9(2):e37192. [doi: 10.2196/37192] [Medline: 35737444]
67. Esper SA, Holder-Murray J, Meister KA, et al. A novel digital health platform with health coaches to optimize surgical patients: feasibility study at a large academic health system. JMIR Perioper Med 2024 Apr 4;7(1):e52125. [doi: 10.2196/52125] [Medline: 38573737]
68. Ahmed A, Ho CW, Grant Y, Archer S, Carrington EV. Acceptability of digital health interventions in perioperative care: a systematic review and narrative synthesis of clinician perspectives. BMJ Open 2025 Mar 15;15(3):e086412. [doi: 10.1136/bmjopen-2024-086412] [Medline: 40090692]
69. National digital health strategy for South Africa 2019-2024. National Department of Health. 2019 May. URL: https://knowledgehub.health.gov.za/system/files/elibdownloads/2023-04/national-digital-strategy-for-south-africa-2019-2024-b.pdf [accessed 2025-10-20]
70. Altman M, Huang TTK, Breland JY. Design thinking in health care. Prev Chronic Dis 2018 Sep 27;15:E117. [doi: 10.5888/pcd15.180128] [Medline: 30264690]

XSL•FO
RenderX

Abbreviations

HIC: high-income country
LMIC: low- and middle-income country
PREM: patient-reported experience measure
PRO: patient-reported outcome
PROM: patient-reported outcome measure
PSHR: Perioperative Shared Health Record
REDCap: Research Electronic Data Capture
SACSOS: South African Collaborative Surgical Outcomes Study
SUS: System Usability Scale
UEQ: User Experience Questionnaire

Edited by J Shiffermiller; submitted 19.Jun.2025; peer-reviewed by J Ohde, S Nilsson; revised version received 28.Oct.2025; accepted 18.Nov.2025; published 12.Jan.2026.

Please cite as:
Steyl C, Orre C, Foster G, Duvenage H, Chew MS, Kluyts HL
Enhancing the User Experience of a Perioperative Digital Health Tool for Information Exchange Using a Human-Centered Design Thinking Approach: Qualitative Observational Study
JMIR Perioper Med 2026;9:e79349
URL: https://periop.jmir.org/2026/1/e79349
doi:10.2196/79349

© Charlé Steyl, Carljohan Orre, Greg Foster, Hanel Duvenage, Michelle S Chew, Hyla Louise Kluyts. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 12.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Immersive Virtual Reality for Pain and Relaxation in Older Adults Following Elective Inpatient Abdominal Surgery: Single-Arm Study Examining Feasibility and Acceptability

Christina Keny[1], PhD, MS, RN; Ujala Shafiq[1], MS; Karl Lorenz[2], MD, MSHS; Marcia Russell[3], MD; Heather Leutwyler[1], PhD, NP; Laura M Wagner[1], PhD, NP; Victoria Tang[4], MD, MAS; Linda G Park[1], PhD, NP

[1]Department of Community Health Systems, School of Nursing, University of California, San Francisco, Box 0608, 490 Illinois Street, #93P, San Francisco, CA, United States

[2]Department of Primary Care and Population Health (PCPH), Stanford University School of Medicine, Palo Alto, CA, United States

[3]Department of Surgery, David Geffen School of Medicine at UCLA, Los Angeles, CA, United States

[4]Department of Medicine, Division of Geriatrics, University of California, San Francisco, San Francisco, CA, United States

Corresponding Author:
Christina Keny, PhD, MS, RN
Department of Community Health Systems, School of Nursing, University of California, San Francisco, Box 0608, 490 Illinois Street, #93P, San Francisco, CA, United States

Abstract

Background: There is mounting evidence to suggest that immersive virtual reality (IVR) can improve pain in older adults in community settings, yet the use of IVR postoperatively in the acute postoperative period following major elective abdominal surgery remains largely underexplored.

Objective: This single-arm pilot study aimed to assess the feasibility, acceptability, and preliminary impact of IVR on self-reported postoperative pain and relaxation levels in older adults following elective major abdominal surgery.

Methods: We recruited individuals aged 55 years and older undergoing elective abdominal surgery at an academic medical center from October 2023 to February 2024. We evaluated feasibility through accrual rate, intervention completion, and questionnaire compliance; acceptability via the System Usability Scale (SUS) and a user experience survey; and tolerability by monitoring self-reported side effects. The preliminary impact of IVR on self-reported pain intensity and relaxation levels was assessed through pre- and postintervention comparisons.

Results: A total of 29 participants, with a median age of 73 (IQR 55 - 81) years, were enrolled and completed at least 1 IVR session, with 19 also completing a second session. Perceived usability and overall acceptance of IVR were high, with minimal side effects reported. In terms of the preliminary impact of IVR, statistically significant improvements were observed in both pain and relaxation levels from pre- to post-IVR on day 1 and day 2.

Conclusions: This study suggests the feasibility and acceptability of IVR as a potential future intervention for postoperative pain management and enhancing relaxation among older adults following elective inpatient abdominal surgery. The preliminary findings suggest the need for large-scale studies across additional complex inpatient abdominal surgeries to confirm the acceptance and efficacy of IVR as a postoperative pain management intervention across a wide range of diverse older demographics. Future research is critical to evaluating the therapeutic potential of IVR in a variety of surgical and patient-specific contexts.

Trial Registration: ClinicalTrials.gov NCT06095661; https://clinicaltrials.gov/study/Nct06095661

(JMIR Perioper Med 2026;9:e81791) doi:10.2196/81791

KEYWORDS

geriatrics; surgery; pain; virtual reality; postoperative pain; geriatric surgery

Introduction

Nearly 4 million operations are performed annually on individuals aged 65 years or older in the United States, a number that is expected to rise significantly as the population continues to age [1,2]. Optimal pain management following major surgery is crucial for older adults who face unique risks associated with uncontrolled postoperative pain, such as delirium, functional decline, and reduced psychosocial well-being [3-6]. Effectively managing pain in older surgical adults is often complicated by challenges stemming from age-related physiological changes, the presence of multiple comorbidities, and the intricacies of polypharmacy [7,8]. Moreover, older adults have an increased risk of developing opioid-related adverse events, addiction,

XSL•FO
RenderX

and/or chronic pain following surgery when compared to younger age groups [9-12]. These factors may contribute to the heightened risks associated with traditional pharmacologic pain management in the older adult [13-15].

Given the complexities and inherent risks in managing acute postoperative pain in this demographic, there is a growing interest in exploring innovative nonpharmacological methods for pain [16-18]. Among these emerging solutions, immersive virtual reality (IVR) has garnered significant interest from both the clinical and research communities [19]. Defined by the use of a head-mounted display (HMD) with motion tracking capabilities, IVR effectively provides users with a believable sense of reality while engaged in a virtual environment [20]. This profound sense of "presence" within the virtual environment often provides positive distraction away from pain stimuli [21-23].

Research on IVR has shown significant potential in reducing acute pain across various clinical settings, such as burn wound care [24,25]. Recent studies have also demonstrated that IVR is effective in reducing postoperative pain in pediatric and young to middle-aged adults following various surgical procedures [25,26]. Contrary to the common belief that older adults are hesitant to embrace new technologies, there is mounting evidence to suggest that using IVR for pain management in the older adult demographic is promising [27-30]. Research focused on IVR use for chronic pain management in community-dwelling older adults has demonstrated improved pain tolerance with a high degree of acceptance as part of pain management [31,32]. Finally, studies have also indicated the acceptability and efficacy of IVR use for postoperative pain among older adults, specifically in elective total knee arthroplasty operations [33,34].

It is key to recognize that the user experience of IVR for postoperative pain management may differ across various older adult subgroups and types of surgical procedures [30,35]. Appreciating these distinctions is vital for tailoring IVR applications to effectively manage pain in a wide range of surgical scenarios and across different older age demographics. Thus, IVR could serve as a promising alternative or adjunct to traditional pain management techniques. However, the feasibility, acceptability, and tolerability of IVR for postoperative pain among older adults across a spectrum of major surgical procedures, including complex abdominal operations, remains largely underexplored. Therefore, the aim of this single-arm study was to investigate the initial feasibility and acceptability (primary outcome), and the preliminary impact (secondary outcome) of IVR on postoperative pain and relaxation levels in older adults during the initial days following inpatient elective abdominal surgery.

Methods

Study Design

This study used a prospective pretest-posttest single-arm study design to ascertain the feasibility, acceptability, and preliminary impact of IVR on pain and relaxation outcomes among older adults following elective major abdominal surgery. The participant enrollment sample size was set for practical reasons and not driven by power analysis for this initial study [36]. This study followed the CONSORT (Consolidated Standards of Reporting Trials) extension to pilot and feasibility studies statement, which is recommended for adaptation in nonrandomized feasibility studies [37]. This study design also aligns with the recommendations and methodological framework proposed by the Virtual Reality Clinical Outcomes Research Experts (VR-CORE) on best practices for the development and testing of IVR treatments in clinical care [38].

Ethical Considerations

The study received approval from the University of California, San Francisco Institutional Review Board (IRB #19 - 28391) and is registered under clinicaltrials.gov (NCT06095661). All participants were provided with written and verbal informed consent before taking part in the study. Data were stored on secure, password-protected servers, and all analyses used deidentified data in accordance with the approved protocol. All responses to study questionnaires were deidentified and entered directly into an electronic tablet that was password-protected. A one-time US $25 gift card was provided to all participants for their participation in at least 1 IVR session. A subgroup of participants was additionally offered the option to complete a single user experience survey prior to hospital discharge. Those who opted to participate in the survey received an additional US $25 gift card.

Participants

We used purposeful sampling to recruit adults aged ≥55 years undergoing elective inpatient abdominal surgery at the University of California, San Francisco, Colorectal and General Surgery clinics. Those who were potentially eligible for the study and who were interested in participation were screened via telephone. Inclusion criteria were individuals anticipated to have an elective abdominal operation requiring hospitalization for at least 48 hours after surgery and those who were able to speak and write in English. Exclusion criteria encompassed individuals with a reported history of self-reported motion sickness, severe cognitive impairment, epilepsy, eye, neck, or face injuries, blindness or severe visual impairment, severe hearing loss, or acute illness hindering postsurgery IVR use. Participants with immediate preintervention nausea, vomiting, or dizziness were also excluded. All participants continued to receive their usual surgical care as per the recommendations of clinical providers and were not asked to decline or change any adjunct strategies for pain management, as the intent of the study was to understand the initial feasibility and acceptability of IVR during the first days after surgery on an inpatient hospital unit.

Intervention

This study used the REAL System i-Series IVR HMD from Penumbra, Inc, featuring built-in audio and gaze-controlled navigation [39]. This IVR system, preloaded with various 360-degree immersive environments for positive distraction and relaxation, includes experiences such as mindful meditation, travel, nature scenes, and games. The IVR system provides motion tracking through sensors embedded in the headset,

XSL•FO
RenderX

capturing all possible participant movements, thus facilitating an extensive immersive experience. The decision to leverage this specific IVR device was 2-fold: (1) the HMD unit was preloaded with a built-in library of experiences, allowing the user access to a wide variety of IVR-positive distraction environments, and (2) the device offered gaze-controlled navigation, potentially allowing ease of use in the immediate postoperative phase of care following major abdominal surgery.

Procedure

During the preoperative surgery visit, clinical staff provided possible participants with an informational flyer and an email introducing the study. Participant information was gathered from the electronic medical record, and potential participants were then contacted by phone and screened for eligibility by the research team. Screening included assessment of cognitive function using the Short Portable Mental Status Questionnaire (SPMSQ), self-reported history of motion sickness, epilepsy, blindness, severe hearing deficits, and any current eye, face, or neck injuries. If deemed eligible and interested in study participation following the initial screening, participants electronically received the informed consent form. Once the consent form was signed, participants were asked to electronically complete an online questionnaire for sociodemographic and clinical data prior to their date of surgery. Prior to surgery, participants received an instructional video link demonstrating the use of the IVR headset, along with a catalog of available immersive content options. Immediately before each intervention session, the study team again reviewed headset use and content selection with the participant. A trained study team member with expertise in the IVR protocol remained present throughout the session to assist with setup, navigation, and troubleshooting as needed.

All participants enrolled in the study were provided with the opportunity to engage in at least 1 IVR session within their hospital room. These sessions were made available starting from the next day following surgery and could extend up to the second day after surgery, ensuring a maximum offering of 2 IVR sessions in total. The IVR intervention was administered to the patient in a seated or lying position by a member of the research team, who was present during and up to 15 minutes after each IVR session. Participants then choose their desired experience within the IVR content library. IVR program preference selection and length of the session were determined by the participant, up to a maximum of 30 minutes per session.

Immediately before and after the IVR intervention, participants reported their pain intensity level and state of relaxation on an 11-point Numeric Rating Scale (NRS) ranging from “0” representing “no pain” or “not relaxed at all” to “10” representing “pain as bad as you can imagine” or “as relaxed as possible.” Adverse outcomes were assessed up through the first 15 minutes after each IVR session using an adapted 4-item Simulator Sickness Questionnaire (SSQ) [40]. All responses were deidentified and entered directly into an electronic tablet that was password-protected.

Outcome Measures

Acceptability

In the context of this study, acceptability refers to participants’ willingness to use IVR in the initial 2 days of their hospitalization and their ability to tolerate IVR use, with minimal side effects reported. Acceptability was assessed by the System Usability Scale (SUS) in all participants (N=29), with a subgroup of 21 participants also completing a user experience survey. After at least 1 session of IVR, all participants were asked to complete the SUS once. SUS is considered a valid and reliable instrument for measuring perceived usability of a technology system and consists of a 10-item questionnaire with 5 response options, with “1” representing “strongly disagree” and “5” representing “strongly agree” [41]. A higher score indicates greater self-reported usability, reflecting a positive attitude toward using the system [42]. An 8-item user experience survey created by the research team using a 5-point Likert scale, ranging from 1 (“totally disagree”) to 5 (“totally agree”), was also administered to quantify participant satisfaction with IVR. The scores from the user experience survey for each sentiment level (strongly disagree-1, disagree-2, neutral-3, agree-4, and strongly agree-5) are presented as a mean and SD. Tolerability refers to the evaluation of adverse events that occurred as a result of IVR use, related to either the hardware or software components [38]. Adverse outcomes, which include symptoms such as nausea, headache, blurred vision, and dizziness, were assessed using a 4-item adapted questionnaire of the SSQ [40]. This questionnaire was administered immediately after each IVR session, allowing participants to indicate the presence or absence of these symptoms with a “Yes” or “No” response. Participants also had the option to free-text any side effects they felt occurred during or immediately after IVR use.

Feasibility

In this study, feasibility is defined as the extent to which potentially eligible participants consented to join the study during the recruitment phase and the degree to which enrolled participants successfully completed the IVR intervention and all questionnaires. To evaluate feasibility, we measured the rate of participant accrual, reasons for nonparticipation, the successful completion of the intervention on the first and second days after surgery, and the mean duration time spent using IVR during each session. We also captured reasons for not completing a second IVR session.

We also evaluated the feasibility by the rate at which participants completed baseline questionnaires. The baseline characteristics captured through questionnaires were self-reported perceived health status, anxiety, depression, and pain catastrophizing prior to the operation. Perceived health status was measured using the EQ-5D-5L questionnaires’ Visual Analog Scale (VAS), with scores ranging from 0 to 100, with a higher score indicating higher perceived health [43,44]. Anxiety was measured using the Generalized Anxiety Disorder-7 (GAD-7) scale, with higher scores indicating higher anxiety levels (total score for GAD is 0 - 21) [45]. Depression was assessed using the Patient Health Questionnaire-8 (PHQ-8), which ranges from 0 to 24, with a higher score indicating higher levels of depression [46]. Finally,

XSL•FO
RenderX

we leveraged the Pain Catastrophizing Scale (PCS), where higher scores signify more intense negative thoughts and feelings toward pain [47].

Preliminary Clinical Impact on Pain and Relaxation

Preliminary clinical impact of IVR on pain intensity levels and state of relaxation was measured through pre- and postintervention mean differences using independent paired sample *t* tests if the data were deemed normal. Nonparametric continuous data were evaluated using the Wilcoxon signed-rank test. Pain intensity level was measured on an 11-point NRS and ranged from "0" representing "no pain" to "10" representing extreme pain. State of relaxation is also measured on an 11-point NRS with "0" representing "not relaxed at all" to "10" representing "as relaxed as one could imagine." Both pain and relaxation were assessed immediately prior to and after each IVR session.

Data Analysis

Acceptability, feasibility, and tolerability are reported as descriptive statistics. User experience surveys for each sentiment level were presented as a mean and SD. Prior to analyzing the preliminary effects of the IVR intervention, we assessed the normality of the distribution of pain and relaxation levels using the Shapiro-Wilk test. Normal data were compared as pre-post mean differences with independent paired 2-sample *t* tests. Nonparametric data were reported as a median and compared using the Wilcoxon signed-rank test. The significance level was set at $P<.05$. All statistical analyses were performed using the STATA statistical software, version 18 SE (StataCorp LLC) [48].

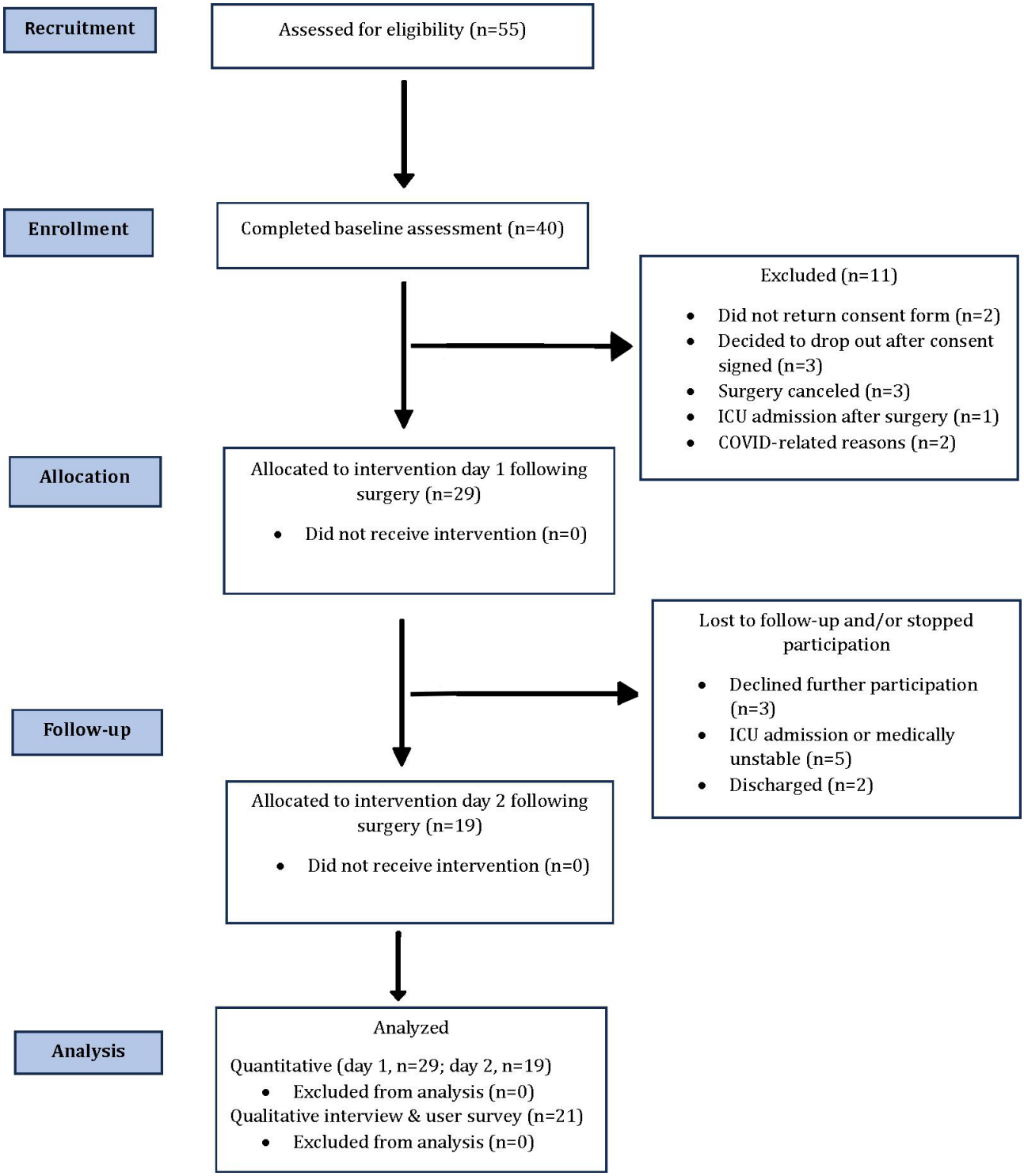
Results

Characteristics of Participants

Fifty-five participants scheduled for elective inpatient abdominal surgery were assessed for study eligibility between October 2023 and February 2024 (Figure 1). Among possible participants, 15 in total were excluded due to not meeting the inclusion criteria. A total of 40 participants completed the baseline questionnaires prior to surgery. Of the 40 participants enrolled, 11 withdrew for several reasons, including failure to complete the written informed consent prior to surgery (n=2). Postoperative consent was not pursued in these cases because participants had undergone general anesthesia. Additional reasons for withdrawal included surgery cancellation (n=3), postoperative Intensive Care Unit admission (n=1), and testing positive for COVID-19 following surgery (n=2). Three additional participants withdrew after providing consent for unknown reasons. A total of 29 enrolled participants were allocated to and completed the first IVR intervention the next day following their surgery, with 19 (65.5%) additionally completing a second intervention the next day. Most participants reported no prior experience with IVR (27/29, 93.1%).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram.



Among 29 participants, the median age was 73 (IQR 55-81) years. A total of 79.3% (23/29) identified as White, 62.10% (18/29) were female, and all but 2 participants reported at least some level of college education (27/29, 93.1%; Table 1). Nearly half of the participants (14/29, 48.3%) underwent a low anterior resection abdominal operation, with cancer as the most common indication for surgery (19/29, 65.5%). More than half of participants reported pain in the 2 weeks prior to surgery (16/29, 55.2%), with a few individuals taking opioid medications for pain (n=3). In our evaluation of the baseline characteristics of the study sample, we found that participants typically described their perceived health status as generally good. Additionally, they reported experiencing mild levels of anxiety and depression, alongside minimal tendencies toward pain catastrophizing, before undergoing surgery and the IVR intervention (Table 1).

XSL•FO
RenderX

Table . Sociodemographic, baseline characteristics, and clinical descriptives (N=29).

Characteristics	Values
Age (years), median (range)	73 (55 - 81)
Sex, n (%)	
Male	11 (37.9)
Female	18 (62.1)
Race or ethnicity, n (%)	
White or Caucasian	23 (79.3)
Asian	4 (13.8)
Hispanic	2 (6.9)
Black or African American	0 (0)
Relationship status, n (%)	
Currently married	14 (48.3)
Divorced	1 (3.4)
Single	13 (44.8)
Widowed	1 (3.4)
Level of education, n (%)	
High school diploma	2 (6.9)
Some college, no degree	5 (17.2)
Any college, graduate, or professional degree	22 (75.9)
Primary indication for surgery, n (%)	
Cancer	19 (65.5)
Primary types of abdominal procedures, n (%)	
Laparoscopic low anterior resection	11 (37.9)
Robotic-assisted low anterior resection	3 (10.3)
Laparoscopic colectomy	5 (17.2)
Open colectomy	2 (6.8)
Open colostomy revision or takedown	3 (10.3)
Ileostomy takedown	3 (10.3)
Open abdominal perineal resection	1 (3.4)
Robotic-assisted rectopexy	1 (3.4)
Prior pain and virtual reality use, n (%)	
Pain in the past week prior to surgery	16 (55.2)
Current opioid use for pain prior to surgery	3 (10.3)
Prior virtual reality use	2 (6.9)
Baseline scores, mean (SD)	
Pain Catastrophizing Scale (PCS)[a]	14.56 (14.21)
Health State Questionnaire – EQ-5D-5L[b]	71.10 (22.3)
Generalized Anxiety Disorder 7-item (GAD-7)[c]	5.56 (5.9)
Patient Health Questionnaire-8 (PHQ-8)[d]	5.03 (5.5)

[a]Pain Catastrophizing Scale consists of 13 items, with each item on a scale from 0 to 4 based on their thoughts when experiencing pain. The total score can range from 0 to 52, with higher scores indicating greater levels of pain catastrophizing.

[b]The EuroQol 5-Dimension 5-Level questionnaire uses a Visual Analog Scale (VAS), where the end points are labeled as the “best imaginable health state” and the “worst imaginable health state.” The VAS score ranges from 0 to 100, with the higher score indicating higher perceived health.

XSL•FO
RenderX

[c]Generalized anxiety disorder 7-item scale is a self-reported questionnaire used to assess the severity of anxiety symptoms with each item scored from 0 (not at all) to 3 (nearly every day). The total score ranges from 0 to 21, with higher scores indicating higher anxiety levels.

[d]The Patient Health Questionnaire-8 measures the severity of depressive symptoms. Each item is scored on a scale from 0 to 3. The sum of all items is a range of 0 to 24, with a higher score indicating greater levels of depressive symptoms.

Acceptability

The mean duration of IVR use in the initial session for 29 participants was 19.14 (SD 7.67) minutes. Moreover, 19 participants additionally completed a second session, during which the mean usage time was 16.78 (SD 6.13) minutes (Table 2). The most common IVR experiences chosen by study participants were guided travel, followed by mindfulness and meditation, with nearly half of the participants choosing more than 1 IVR experience during a single session. The results indicated high perceived usability of IVR in this sample as demonstrated by a high mean SUS score of 88.10 (SD 6.15). SUS scores above 68 are considered above average and are an indicator of good usability [41]. Individual adjusted raw mean scores for each SUS item were generally >3, indicating positive usability for each SUS item statement (adjusted items ranged from 0 to 4, with 4 as more desirable per each item). Higher scores after adjustment indicate better usability (Table 3).

Table . Mean time spent using immersive virtual reality and content selection.

Usage time and content selection	1 day after surgery (N=29)	2 days after surgery (n=19)
Time spent in IVR[a] (min), mean (SD; range)	19.14 (7.67; 6-30)	16.78 (6.13; 3-30)
Participant IVR content selection[b], n (%)		
Guided travel	20 (68.9)	10 (52.6)
Mindfulness and meditation	7 (24.1)	5 (26.3)
Arctic cold and/or underwater	4 (13.7)	3 (15.8)
Forests and/or wildlife	2 (6.8)	1 (5.3)
Games	1 (3.4)	1 (5.3)

[a]IVR: immersive virtual reality.

[b]Participants had the option to choose as many experiences as desired, up to a maximum of 30 minutes of use per session, in any of the content categories offered within the preloaded software library.

Table . System Usability Scale (SUS) ratings.

SUS[a] adjusted raw score per item (N=29)	Score, mean (SD)	Range
I think that I would like to use this system frequently.	3.5 (0.58)	2 - 4
I found the system unnecessarily complex.	3.8 (0.77)	0 - 4
I thought the system was easy to use.	3.7 (0.47)	3 - 4
I think that I would need the support of a technical person to be able to use this system.	3.2 (0.85)	1 - 4
I found the various functions in this system were well integrated.	2.48 (0.87)	1 - 4
I thought there was too much inconsistency in this system.	3.7 (0.66)	2 - 4
I would imagine that most people would learn to use this system very quickly.	3.6 (0.50)	3 - 4
I found the system very cumbersome to use.	3.9 (0.37)	2 - 4
I felt very confident using the system.	3.7 (0.47)	3 - 4
I needed to learn a lot of things before I could get going with this system.	3.8 (0.41)	3 - 4

[a]SUS: System Usability Scale.

The original responses are given on a Likert scale from 1 (strongly disagree) to 5 (strongly agree) for each of the 10 items. After adjusting the scale of negatively worded questions (items 2, 4, 6, 8, and 10) by subtracting their scores from 5 and for positively worded questions (items 1, 3, 5, 7, and 9) obtained by subtracting 1, adjusted scores will range from 0 to 4. After adjustment, “0” represents a negative usability experience as related to the item statement and “4” represents a positive usability experience for each item. Higher scores after

XSL•FO
RenderX

adjustment indicate better usability (adapted from the study by Brooke [42]).

Study participants reported an overall positive experience with using IVR as indicated in the survey questionnaire (Table 4). All study participants marked responses of "agree-4" or "strongly agree-5" to the statement "I liked the virtual reality experience." Most of the study participants (27/29, 93.1%) also marked "agree-4" or "strongly agree-5" that IVR improved their postoperative pain. Furthermore, most participants agreed that they would use IVR again for pain (26/29, 89.6%) or anxiety (25/29, 86.2%). Finally, all participants (N=29) marked "strongly agree" to the statement, "I would recommend virtual reality to other older surgical patients."

Table . User experience survey responses (N=29).[a]

User experience item	Scores, mean (SD)
Would recommend VR[b] to others	5.0 (0.0)
Would use VR again for pain	4.8 (0.4)
I liked the VR experience	4.8 (0.7)
Would use VR again for anxiety	4.7 (0.7)
The audio sound was pleasant	4.6 (0.6)
VR improved my pain	4.5 (0.8)
The image quality was pleasant	3.8 (0.8)
The headset was comfortable	3.1 (1.4)

[a]Participants rated each statement on a 5-point Likert scale (1=totally disagree, 2=disagree, 3=neutral, 4=agree, 5=totally agree). Mean scores and SDs are reported for each item and are ordered from highest to lowest mean rating.

[b]VR: virtual reality.

Nearly all participants (28/29, 96.6%) completed one or more IVR sessions without self-reported side effects (eg, dizziness, headache, eye strain, and nausea). One participant reported mild face and chest skin redness that occurred nearly 24 hours after the first IVR session. Upon further investigation, it was unclear as to the exact cause of the skin irritation, whether due to the IVR headset foam padding or related to recent medications as part of surgical care. Although unlikely related to the use of IVR equipment, the possible adverse event was reported out of an abundance of caution.

Feasibility

To evaluate overall feasibility, we measured the rate of accrual and reasons for nonparticipation, as well as successful completion of IVR on the first and second day after inpatient abdominal surgery (Figure 1). Of the 55 potential participants assessed for eligibility, 7 individuals did not meet inclusion requirements during telephone screening, 7 declined to participate after learning more about the study, stating extreme anxiety over surgery (n=5) or fear of new technology (n=2) as the main reasons for nonparticipation, and 1 individual refused to state a reason for declining participation (Figure 1). The remaining participants (n=40) were deemed eligible and agreed to participate. Of the 40 participants, a total of 11 either dropped out or were excluded before the intervention was administered due to the following reasons: did not return the consent forms (n=2), decided to drop out of the study prior to the date of surgery for unspecified reasons (n=3), surgery was canceled (n=3), Intensive Care Unit admission directly from the operating room (n=1), and COVID-19–related reasons (n=2).

A total of 29 enrolled participants allocated to the intervention on the day following surgery all completed the first IVR session, with 19 additionally completing the second IVR session the next day. The most common reason for not completing a second IVR session the next day was typically due to a change in complexity of care or severe nausea or vomiting not related to IVR (n=5). All participants allocated to the first and second IVR sessions completed all baseline questionnaires, as well as all pre- and postintervention questionnaires, with no missing or unanswered items.

Preliminary Clinical Impact of IVR on Pain and Relaxation

Significant improvements were observed in both pain and relaxation levels from pre- to post-IVR on both day 1 and day 2 following surgery (Table 5). The preliminary impact of IVR on pain levels was analyzed using a paired 2-sample *t* test. On day 1 following surgery, post-IVR mean pain levels showed a significant reduction as compared to pre-IVR pain levels, with a mean improvement of 2.65 (SD 2.0), representing approximately a 50% reduction in pain scores (95% CI 1.89-3.41; $P<.001$). Similarly, on day 2 following surgery, results indicated a significant decrease in pain levels from pre- to post-IVR, with a mean pain level decrease of 2.05 (SD 1.5; 95% CI 1.33-2.78; $P<.001$).

XSL•FO
RenderX

Table . Pre-to-post immersive virtual reality changes in self-reported pain and relaxation.

Pre-post IVR[a] pain and relaxation levels	Mean (SD) or median (IQR)	95% CI	Test statistic	*P* value
Pain level day 1, N=29, mean (SD)			N/A[b]	.001
Pre-IVR	5.17 (2.1)	4.37 - 5.97		
Post-IVR	2.51 (1.5)	1.91 - 3.16		
Change	2.65 (2.0)	1.89 - 3.41		
Pain level day 2, N=19, mean (SD)			N/A	.001
Pre-IVR	4.84 (1.6)	4.08 - 5.60		
Post-IVR	2.79 (1.5)	2.06 - 3.51		
Change	2.05 (1.5)	1.33 - 2.78		
Relaxation level day 1, N=29, median (IQR)			Z=−4.6	<.01
Pre-IVR	3 (3-6)			
Post-IVR	8 (7-9)			
Change	__[c]			
Relaxation level day 2, n=19, median (IQR)			Z=−3.85	<.01
Pre-IVR	4 (4-5)			
Post-IVR	8 (7-9)			
Change	—			

[a]IVR: immersive virtual reality.

[b]N/A: not applicable.

[c]Not available.

Finally, participants reported a significantly higher level of relaxation immediately following IVR as compared to pre-IVR relaxation levels on both day 1 and day 2 following surgery. The Wilcoxon signed-rank test, applied due to the nonnormal distribution of relaxation scores, revealed significant findings. One day following surgery, the median relaxation score prior to using IVR was 3 (IQR 3-6), which increased to a median score of 8 (IQR 7-9) immediately following IVR usage, indicating a statistically significant improvement in relaxation levels from pre- to postintervention (Z=−4.6; *P*<.01). On the second day after surgery, the median score before intervention was 4 (IQR 4-5), which increased to a median relaxation score of 8 (IQR 7-9) following the intervention, demonstrating a significant improvement in relaxation scores from pre- to post-IVR use (Z=–3.85; *P*<.01).

Discussion

Principal Findings

Our study indicates that IVR is a feasible, acceptable, and well-tolerated intervention for postoperative pain management and relaxation in older adults in the initial days following elective major inpatient abdominal surgery. Participants reported high perceived usability and acceptance, with minimal side effects. IVR use resulted in statistically significant reductions in self-reported pain and improvements in relaxation from pre- to postintervention on both day 1 and day 2, with most participants indicating a willingness to use IVR again and recommend it to other older surgical patients. While prior studies have established the feasibility of IVR for preoperative anxiety and postoperative exercise in older adults undergoing colorectal surgery [49,50], our study is among the few to extend these findings to postoperative pain and relaxation across a wider range of elective abdominal procedures. These results support the potential of IVR as a nonpharmacologic option for pain management and relaxation in this population.

Acceptability, Feasibility, and Tolerability

Our research both corroborates and diverges from existing literature on the application of IVR for managing pain and facilitating relaxation. Recent meta-analyses have suggested that IVR is efficacious in reducing acute pain following a variety of surgical and medical procedures [25,51]. However, except for total knee arthroplasty operations, there is a notable gap in research regarding the acceptability, feasibility, and tolerability of IVR among older surgical adults in the initial days following various types of major elective inpatient surgeries [52]. Some studies suggest that hospitalized older adults older than the age of 60 are more likely to decline participation in IVR studies, related to a lack of understanding or perceived usefulness of IVR [53,54]. Other research indicates a possibly lower acceptance of IVR among older adults in acute care settings due to negative attitudes and anxiety toward using technology [52,55]. Furthermore, except for those who developed postoperative clinical complications precluding the use of IVR, the dropout rate among our study participants was notably low at 12.5% (n=5). In both community and residential contexts, as well as within the scope of colorectal cancer care, existing

XSL•FO
RenderX

research has indicated that the use of IVR as a positive distraction is both feasible and highly accepted among older adults for alleviating chronic pain in nonhospital settings [31,56-58]. Moreover, studies on the adoption of emerging technologies such as IVR suggest that older adults are generally more receptive to using new technology when they are offered a broad range of choices and the autonomy to select content and information according to their unique preferences [59-63].

Regarding the tolerability of IVR use in older adults, there have been minor reports of motion sickness and occasional discomfort with the IVR headset noted in prior research [52,64,65]. In our study, there were no reports of motion sickness during or after use. This finding may have been due to limiting each IVR session to a maximum of 30 minutes. A recent review of IVR use in older adults across a wide spectrum of nonsurgical settings found cybersickness (eg, dizziness, nausea, eye strain, etc) to be minimal across 39 studies [66,67]. Also related to the tolerability of IVR, in our study, 1 participant reported skin irritation, an adverse effect for which there is little to no existing research quantifying its incidence rate [68]. Finally, our user experience survey results suggest that while the overall satisfaction with the IVR system was positive, a significant number of participants expressed discomfort with the headset. This reported discomfort is consistent with prior research indicating that while older adults in community settings have found the headsets to be bulky and uncomfortable, their overall experience with IVR was enjoyable [30].

Clinical Implications

The integration of IVR into the care of older adults following major abdominal surgery presents a promising avenue for enhancing pain management and relaxation strategies, with the potential to improve patient-reported outcomes, patient satisfaction, and the overall recovery experience. First, despite existing concerns about older adults' willingness to engage with new technologies, our study found a high level of acceptability and tolerability of IVR for pain and relaxation. This finding suggests that with proper introduction and support, as we provided in our study, older adults might be open to using innovative technology such as IVR in the acute care setting as part of their care [28,69]. Second, our results hold particular significance in the surgical care of older adults, where IVR could act as a cost-effective and safe alternative or adjunct to conventional pain management approaches [9,70,71]. Some studies indicate that IVR retains the possibility to lessen opioid medication usage and its associated side effects [72,73]. In the last decade, government agencies and clinical professional organizations, including the American College of Surgeons' Geriatric Surgery Verification program, have increased emphasis on the importance of incorporating opioid-sparing methods into pain treatment strategies, including nonpharmacological interventions, especially in vulnerable groups such as older adults [16,74]. The prominence of this push can also be exemplified through a recent ruling by the Centers for Medicare & Medicaid Services, indicating that one type of an integrated software or hardware IVR device may be eligible for Medicare insurance coverage [75]. In the near future, it is expected that improved insurance coverage will greatly increase the accessibility of such technologies for older adults as part of their pain management.

Furthermore, introducing IVR into the care of older adults may initially be met with uncertainty, by both clinicians and older adults, due to prevailing beliefs about technological proficiency or appropriateness in this age group. This view may originate from widespread societal perceptions on aging and technology, which commonly depict older adults as less skilled or less inclined to engage with new technology [76,77]. Such perspectives can unintentionally result in exclusion from IVR clinical trials or use in the clinical setting [78]. Thus, we may underestimate the older adults' capacity to accept or learn new technology such as IVR. Personalized education, tailored training, and sufficient support are essential actions to enhance the adoption of IVR among older adults [60].

Finally, aligning IVR content with older adults' personal interests, such as cultural and spiritual practices, as well as hobbies and previous life experiences, is vital for its wider acceptance in this demographic [30]. Future research is needed to investigate whether the level of pain or relaxation correlates with specific types of IVR content or exhibits a dose-response relationship based on duration and frequency of use. Finally, the insights gained from this study could pave the way for more extensive randomized controlled trials aimed at assessing the acceptance and overall efficacy of IVR in older adults undergoing a variety of major elective surgeries.

Limitations

The findings of this study should be considered in view of certain limitations. The study design and limited sample size restrict the generalizability of the findings; therefore, final conclusions about the efficacy of IVR to improve pain and relaxation cannot be made. Additionally, this study was not designed or powered to reliably assess the impact of predictors such as age, gender, specific surgeries, or other baseline characteristics on study outcomes. Postoperative pain and relaxation outcomes are influenced by multiple perioperative factors that were not controlled in this initial feasibility study, including baseline pain levels prior to surgery, preoperative opioid consumption, extent of colorectal surgery, use of regional anesthesia or local blocks, and concurrent use of multimodal opioid-sparing medications. Moreover, the reliance on self-report introduces potential for bias and subjectivity in our findings. Although targeted recruitment efforts were undertaken to increase diversity, the study sample reflected the demographic composition of elective colorectal surgery patients treated within a single academic health care system during the study period. As a result, these findings may not fully capture the experiences of individuals from racially or ethnically diverse backgrounds, non-English speakers, those with lower educational attainment, or patients with frailty or cognitive impairment. This limitation highlights the importance of future multisite studies to more robustly evaluate the feasibility, accessibility, and broader applicability of IVR interventions in colorectal surgical populations. Also, because pain and relaxation outcomes were assessed only immediately before and after each IVR session, we were unable to determine the duration of the intervention's effects beyond the short-term period. Given the short average

XSL•FO
RenderX

postoperative length of stay in this cohort (just over 3 days), the intervention and assessments were necessarily limited to the immediate inpatient recovery window. It is key to note that although questionnaires were brief, repeated pre-post assessments across multiple IVR sessions may have contributed to respondent burden and fatigue in the immediate postoperative setting, although not mentioned by participants. Survey methodology literature suggests that repeated measures can affect response quality. Importantly, the assessment schedule remained feasible in this pilot study; however, larger and more longitudinal studies should explicitly evaluate participant burden related to repeated questionnaires and assessment frequency [79]. Finally, the IVR intervention was not standardized, allowing variation in both content selection and duration of use. Given that some participants selected briefer immersive content experiences while others completed longer sessions, it remains unclear whether shorter exposure reflected content preference, postoperative fatigue, discomfort, or other barriers to sustained use. Future studies should consider controlling content type, duration, and frequency of use to better assess the impact of IVR on acceptability, feasibility, and clinical outcomes.

Conclusion

This study supports the feasibility, acceptability, and tolerability of IVR as a potential intervention for postoperative pain among older adults following elective inpatient abdominal surgery. Our findings add to the growing body of evidence supporting nonpharmacological approaches for postoperative pain management. While our preliminary findings are promising, larger-scale studies are needed to confirm the acceptance and efficacy of IVR as a postoperative pain management intervention across more diverse populations of older adults, including those from underrepresented minority groups and individuals with physical or cognitive limitations. Importantly, this study highlights the potential of IVR to enhance patient-reported outcomes and improve the perioperative care experience for a demographic that is often considered vulnerable and is frequently underrepresented in technology-based research.

Acknowledgments

We declare that artificial intelligence was not used in any aspect of this research study or manuscript development.

Funding

This work was supported by the National Institute of Nursing Research of the National Institutes of Health (T32NR01692), the University of California, San Francisco (UCSF), School of Nursing Century Research award, the UCSF Graduate Mentorship Fellowship award, the UCSF Phillip R. Lee Health Policy Fellowship award, the UCSF Office of the Executive Vice Chancellor and Provost (EVCP) Social and Population Science Fellowship Research Award, and by the Department of Veterans Affairs, Veterans Health Administration, Office of Academic Affiliations VA Quality Scholars Advanced Fellowship Program (OAA 3Q072019C).

Conflicts of Interest

None declared.

References

1. Schwarze ML, Barnato AE, Rathouz PJ, et al. Development of a list of high-risk operations for patients 65 years and older. JAMA Surg 2015 Apr;150(4):325-331. [doi: 10.1001/jamasurg.2014.1819] [Medline: 25692282]
2. Cheung C, Meissner MA, Garg T. Incorporating outcomes that matter to older adults into surgical research. J Am Geriatr Soc 2021 Mar;69(3):618-620. [doi: 10.1111/jgs.17028] [Medline: 33462830]
3. Aubrun F, Marmion F. The elderly patient and postoperative pain treatment. Best Pract Res Clin Anaesthesiol 2007 Mar;21(1):109-127. [doi: 10.1016/j.bpa.2006.12.005] [Medline: 17489223]
4. Hwang SS, Chang VT, Kasimis B. Dynamic cancer pain management outcomes: the relationship between pain severity, pain relief, functional interference, satisfaction and global quality of life over time. J Pain Symptom Manage 2002 Mar;23(3):190-200. [doi: 10.1016/s0885-3924(01)00418-3] [Medline: 11888717]
5. Apfelbaum JL, Chen C, Mehta SS, Gan TJ. Postoperative pain experience: results from a national survey suggest postoperative pain continues to be undermanaged. Anesth Analg 2003 Aug;97(2):534-540. [doi: 10.1213/01.ANE.0000068822.10113.9E] [Medline: 12873949]
6. Borys M, Zyzak K, Hanych A, et al. Survey of postoperative pain control in different types of hospitals: a multicenter observational study. BMC Anesthesiol 2018 Jul 18;18(1):83. [doi: 10.1186/s12871-018-0551-3] [Medline: 30021520]
7. Jin F, Chung F. Minimizing perioperative adverse events in the elderly. Br J Anaesth 2001 Oct;87(4):608-624. [doi: 10.1093/bja/87.4.608] [Medline: 11878732]
8. Falzone E, Hoffmann C, Keita H. Postoperative analgesia in elderly patients. Drugs Aging 2013 Feb;30(2):81-90. [doi: 10.1007/s40266-012-0047-7] [Medline: 23288604]
9. Kessler ER, Shah M, Gruschkus SK, Raju A. Cost and quality implications of opioid-based postsurgical pain control using administrative claims data from a large health system: opioid-related adverse events and their impact on clinical and economic outcomes. Pharmacotherapy 2013 Apr;33(4):383-391. [doi: 10.1002/phar.1223] [Medline: 23553809]

XSL•FO
RenderX

10. Baker J, Brovman EY, Rao N, Beutler SS, Urman RD. Potential opioid-related adverse drug events are associated with decreased revenue in hip replacement surgery in the older population. Geriatr Orthop Surg Rehabil 2020;11:2151459320915328. [doi: 10.1177/2151459320915328] [Medline: 32231864]
11. Huhn AS, Strain EC, Tompkins DA, Dunn KE. A hidden aspect of the U.S. opioid crisis: rise in first-time treatment admissions for older adults with opioid use disorder. Drug Alcohol Depend 2018 Dec 1;193:142-147. [doi: 10.1016/j.drugalcdep.2018.10.002] [Medline: 30384321]
12. Hamilton HJ, Gallagher PF, O'Mahony D. Inappropriate prescribing and adverse drug events in older people. BMC Geriatr 2009 Jan 28;9:5. [doi: 10.1186/1471-2318-9-5] [Medline: 19175914]
13. Mangoni AA, Jackson SHD. Age-related changes in pharmacokinetics and pharmacodynamics: basic principles and practical applications. Br J Clin Pharmacol 2004 Jan;57(1):6-14. [doi: 10.1046/j.1365-2125.2003.02007.x] [Medline: 14678335]
14. McKeown JL. Pain management issues for the geriatric surgical patient. Anesthesiol Clin 2015 Sep;33(3):563-576. [doi: 10.1016/j.anclin.2015.05.010] [Medline: 26315638]
15. Weiss AJ, Heslin KC, Barrett ML, Izar R, Bierman AS. Opioid-Related Inpatient Stays and Emergency Department Visits Among Patients Aged 65 Years and Older, 2010 and 2015: Agency for Healthcare Research and Quality (US); 2018. URL: https://www.ncbi.nlm.nih.gov/books/NBK534106 [accessed 2026-03-17]
16. Non-pharmacologic and non-opioid solutions for pain management. : The Joint Commission; 2018 URL: https://digitalassets.jointcommission.org/api/public/content/b79588e914734c62ba3d7a4427ee3a5a?v=84bb0b4d [accessed 2026-04-09]
17. Martinez V, Beloeil H, Marret E, Fletcher D, Ravaud P, Trinquart L. Non-opioid analgesics in adults after major surgery: systematic review with network meta-analysis of randomized trials. Br J Anaesth 2017 Jan;118(1):22-31. [doi: 10.1093/bja/aew391] [Medline: 28039239]
18. Opioids in medicare part D: concerns about extreme use and questionable prescribing. : U.S. Department of Health & Human Services; 2017 URL: https://oig.hhs.gov/OEI/REPORTS/OEI-02-17-00250.pdf [accessed 2026-03-17]
19. Baker NA, Polhemus AH, Haan Ospina E, et al. The state of science in the use of virtual reality in the treatment of acute and chronic pain: a systematic scoping review. Clin J Pain 2022 Jun 1;38(6):424-441. [doi: 10.1097/AJP.0000000000001029] [Medline: 35537072]
20. Furht B, editor. Immersive virtual reality. In: Encyclopedia of Multimedia: Springer; 2008:345-346. [doi: 10.1007/978-0-387-78414-4_85]
21. Gupta A, Scott K, Dukewich M. Innovative technology using virtual reality in the treatment of pain: does It reduce pain via distraction, or is there more to it? Pain Med 2018 Jan 1;19(1):151-159. [doi: 10.1093/pm/pnx109] [Medline: 29025113]
22. Malloy KM, Milling LS. The effectiveness of virtual reality distraction for pain reduction: a systematic review. Clin Psychol Rev 2010 Dec;30(8):1011-1018. [doi: 10.1016/j.cpr.2010.07.001] [Medline: 20691523]
23. Dan A, Reiner M. EEG-based cognitive load of processing events in 3D virtual worlds is lower than processing events in 2D displays. Int J Psychophysiol 2017 Dec;122:75-84. [doi: 10.1016/j.ijpsycho.2016.08.013] [Medline: 27592084]
24. Chan E, Foster S, Sambell R, Leong P. Clinical efficacy of virtual reality for acute procedural pain management: a systematic review and meta-analysis. PLoS One 2018;13(7):e0200987. [doi: 10.1371/journal.pone.0200987] [Medline: 30052655]
25. Ding L, Hua H, Zhu H, et al. Effects of virtual reality on relieving postoperative pain in surgical patients: a systematic review and meta-analysis. Int J Surg 2020 Oct;82:87-94. [doi: 10.1016/j.ijsu.2020.08.033] [Medline: 32882400]
26. Iannicelli AM, Vito D, Dodaro CA, et al. Does virtual reality reduce pain in pediatric patients? A systematic review. Ital J Pediatr 2019 Dec 30;45(1):171. [doi: 10.1186/s13052-019-0757-0] [Medline: 31888710]
27. Syed-Abdul S, Malwade S, Nursetyo AA, et al. Virtual reality among the elderly: a usefulness and acceptance study from Taiwan. BMC Geriatr 2019 Aug 19;19(1):223. [doi: 10.1186/s12877-019-1218-8] [Medline: 31426766]
28. Seifert A, Schlomann A. The use of virtual and augmented reality by older adults: potentials and challenges. Front Virtual Real 2021;2. [doi: 10.3389/frvir.2021.639718]
29. Healy D, Carr E, Conlan O, Browne AC, Walsh JC. Exploring the content of the STAND-VR intervention: a qualitative interview study. PLoS Digit Health 2023 Mar;2(3):e0000210. [doi: 10.1371/journal.pdig.0000210] [Medline: 36913343]
30. Healy D, Flynn A, Conlan O, McSharry J, Walsh J. Older adults' experiences and perceptions of immersive virtual reality: systematic review and thematic synthesis. JMIR Serious Games 2022 Dec 6;10(4):e35802. [doi: 10.2196/35802] [Medline: 36472894]
31. Benham S, Kang M, Grampurohit N. Immersive virtual reality for the management of pain in community-dwelling older adults. OTJR (Thorofare N J) 2019 Apr;39(2):90-96. [doi: 10.1177/1539449218817291] [Medline: 30595096]
32. Goudman L, Jansen J, Billot M, et al. Virtual reality applications in chronic pain management: systematic review and meta-analysis. JMIR Serious Games 2022 May 10;10(2):e34402. [doi: 10.2196/34402] [Medline: 35536641]
33. Peng L, Zeng Y, Wu Y, Si H, Shen B. Virtual reality-based rehabilitation in patients following total knee arthroplasty: a systematic review and meta-analysis of randomized controlled trials. Chin Med J (Engl) 2021 Dec 13;135(2):153-163. [doi: 10.1097/CM9.0000000000001847] [Medline: 34908004]
34. Su W, Zhou Y, Qiu H, Wu H. The effects of preoperative rehabilitation on pain and functional outcome after total knee arthroplasty: a meta-analysis of randomized controlled trials. J Orthop Surg Res 2022 Mar 21;17(1):175. [doi: 10.1186/s13018-022-03066-9] [Medline: 35313897]

XSL•FO
RenderX

35. Huang Q, Lin J, Han R, Peng C, Huang A. Using virtual reality exposure therapy in pain management: a systematic review and meta-analysis of randomized controlled trials. Value Health 2022 Feb;25(2):288-301. [doi: 10.1016/j.jval.2021.04.1285] [Medline: 35094802]
36. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. Pilot Feasibility Stud 2016;2(1):64. [doi: 10.1186/s40814-016-0105-8] [Medline: 27965879]
37. Lancaster GA, Thabane L. Guidelines for reporting non-randomised pilot and feasibility studies. Pilot Feasibility Stud 2019;5(1):114. [doi: 10.1186/s40814-019-0499-1] [Medline: 31608150]
38. Birckhead B, Khalil C, Liu X, et al. Recommendations for methodology of virtual reality clinical trials in health care by an International Working Group: iterative study. JMIR Ment Health 2019 Jan 31;6(1):e11973. [doi: 10.2196/11973] [Medline: 30702436]
39. Immersive healthcare. Penumbra. URL: https://www.penumbrainc.com/immersive-healthcare/ [accessed 2026-04-09]
40. Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG. Simulator Sickness Questionnaire: an enhanced method for quantifying simulator sickness. Int J Aviat Psychol 1993 Jul;3(3):203-220. [doi: 10.1207/s15327108ijap0303_3]
41. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009;4(3):114-123 [FREE Full text]
42. Brooke J. SUS: a “quick and dirty” usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. Usability Evaluation in Industry: Taylor & Francis; 1996:207-212. [doi: 10.1201/9781498710411-35]
43. Brooks R. EuroQol: the current state of play. Health Policy 1996 Jul;37(1):53-72. [doi: 10.1016/0168-8510(96)00822-6] [Medline: 10158943]
44. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011 Dec;20(10):1727-1736. [doi: 10.1007/s11136-011-9903-x] [Medline: 21479777]
45. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing Generalized Anxiety Disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
46. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
47. Sullivan MJL, Bishop SR, Pivik J. The Pain Catastrophizing Scale: development and validation. Psychol Assess 1995;7(4):524-532. [doi: 10.1037/1040-3590.7.4.524]
48. Sullivan MJL. The complete statistical software for data science. Stata. URL: https://www.stata.com/ [accessed 2021-03-10]
49. Schrempf MC, Zanker J, Arndt TT, et al. Immersive virtual reality fitness games to improve recovery after colorectal surgery: a randomized single blind controlled pilot trial. Games Health J 2023 Dec;12(6):450-458. [doi: 10.1089/g4h.2023.0004] [Medline: 37428543]
50. Ugras GA, Kanat C, Yaman Z, Yilmaz M, Turkmenoglu MO. The effects of virtual reality on preoperative anxiety in patients undergoing colorectal and abdominal wall surgery: a randomized controlled trial. J Perianesth Nurs 2023 Apr;38(2):277-283. [doi: 10.1016/j.jopan.2022.07.005] [Medline: 36319521]
51. Smith RC. Making the biopsychosocial model more scientific-its general and specific models. Soc Sci Med 2021 Mar;272:113568. [doi: 10.1016/j.socscimed.2020.113568] [Medline: 33423810]
52. Wang S, Lim SH, Aloweni F. Virtual reality interventions and the outcome measures of adult patients in acute care settings undergoing surgical procedures: an integrative review. J Adv Nurs 2022 Mar;78(3):645-665. [doi: 10.1111/jan.15065] [Medline: 34633112]
53. Tashjian VC, Mosadeghi S, Howard AR, et al. Virtual reality for management of pain in hospitalized patients: results of a controlled trial. JMIR Ment Health 2017 Mar 29;4(1):e9. [doi: 10.2196/mental.7387] [Medline: 28356241]
54. Mosadeghi S, Reid MW, Martinez B, Rosen BT, Spiegel BMR. Feasibility of an immersive virtual reality intervention for hospitalized patients: an observational cohort study. JMIR Ment Health 2016 Jun 27;3(2):e28. [doi: 10.2196/mental.5801] [Medline: 27349654]
55. Hauk N, Hüffmeier J, Krumm S. Ready to be a silver surfer? A meta-analysis on the relationship between chronological age and technology acceptance. Comput Human Behav 2018 Jul;84:304-319. [doi: 10.1016/j.chb.2018.01.020]
56. Dermody G, Whitehead L, Wilson G, Glass C. The role of virtual reality in improving health outcomes for community-dwelling older adults: systematic review. J Med Internet Res 2020 Jun 1;22(6):e17331. [doi: 10.2196/17331] [Medline: 32478662]
57. Moore RC, Hancock JT, Bailenson JN. From 65 to 103, older adults experience virtual reality differently depending on their age: evidence from a large-scale field study in nursing homes and assisted living facilities. Cyberpsychol Behav Soc Netw 2023 Dec;26(12):886-895. [doi: 10.1089/cyber.2023.0188] [Medline: 38011717]
58. Kelleher SA, Fisher HM, Winger JG, et al. Virtual reality for improving pain and pain-related symptoms in patients with advanced stage colorectal cancer: a pilot trial to test feasibility and acceptability. Palliat Support Care 2022 Aug;20(4):471-481. [doi: 10.1017/S1478951521002017] [Medline: 35078545]
59. Hawley-Hague H, Boulton E, Hall A, Pfeiffer K, Todd C. Older adults’ perceptions of technologies aimed at falls prevention, detection or monitoring: a systematic review. Int J Med Inform 2014 Jun;83(6):416-426. [doi: 10.1016/j.ijmedinf.2014.03.002] [Medline: 24798946]

XSL•FO
RenderX

60. Roberts AR, De Schutter B, Franks K, Radina ME. Older adults' experiences with audiovisual virtual reality: perceived usefulness and other factors influencing technology acceptance. Clin Gerontol 2019;42(1):27-33. [doi: 10.1080/07317115.2018.1442380] [Medline: 29505343]
61. Mitzner TL, Boron JB, Fausset CB, et al. Older adults talk technology: technology usage and attitudes. Comput Human Behav 2010 Nov 1;26(6):1710-1721. [doi: 10.1016/j.chb.2010.06.020] [Medline: 20967133]
62. Lee LN, Kim MJ, Hwang WJ. Potential of augmented reality and virtual reality technologies to promote wellbeing in older adults. Appl Sci (Basel) 2019;9(17):3556. [doi: 10.3390/app9173556]
63. Rawlins CR, Veigulis Z, Hebert C, Curtin C, Osborne TF. Effect of immersive virtual reality on pain and anxiety at a Veterans Affairs health care facility. Front Virtual Real 2021;2. [doi: 10.3389/frvir.2021.719681]
64. Bekelis K, Calnan D, Simmons N, MacKenzie TA, Kakoulides G. Effect of an immersive preoperative virtual reality experience on patient reported outcomes: a randomized controlled trial. Ann Surg 2017 Jun;265(6):1068-1073. [doi: 10.1097/SLA.0000000000002094] [Medline: 27906757]
65. Huang MY, Scharf S, Chan PY. Effects of immersive virtual reality therapy on intravenous patient-controlled sedation during orthopaedic surgery under regional anesthesia: a randomized controlled trial. PLoS One 2020;15(2):e0229320. [doi: 10.1371/journal.pone.0229320] [Medline: 32092098]
66. Tian N, Lopes P, Boulic R. A review of cybersickness in head-mounted displays: raising attention to individual susceptibility. Virtual Real 2022 Dec;26(4):1409-1441. [doi: 10.1007/s10055-022-00638-2]
67. Drazich BF, McPherson R, Gorman EF, et al. In too deep? A systematic literature review of fully-immersive virtual reality and cybersickness among older adults. J Am Geriatr Soc 2023 Dec;71(12):3906-3915. [doi: 10.1111/jgs.18553] [Medline: 37560978]
68. Facebook technologies recalls removable foam facial interfaces for oculus quest 2 virtual reality headsets due to skin irritation hazard (recall alert). US Consumer Product Safety Commission. URL: https://www.cpsc.gov/Recalls/2021/Facebook-Technologies-Recalls-Removable-Foam-Facial-Interfaces-for-Oculus-Quest-2-Virtual-Reality-Headsets-Due-to-Skin-Irritation-Hazard-Recall-Alert [accessed 2024-04-01]
69. Høeg ER, Andersen NB, Malmkjær N, Vaaben AH, Uth J. Hospitalized older adults' experiences of virtual reality-based group exercise therapy with cycle ergometers: an early feasibility study. Comput Hum Behav Rep 2023 Aug;11:100301. [doi: 10.1016/j.chbr.2023.100301]
70. Delshad SD, Almario CV, Fuller G, Luong D, Spiegel BMR. Economic analysis of implementing virtual reality therapy for pain among hospitalized patients. NPJ Digital Med 2018;1(1):1-8. [doi: 10.1038/s41746-018-0026-4]
71. Florence CS, Zhou C, Luo F, Xu L. The economic burden of prescription opioid overdose, abuse, and dependence in the United States, 2013. Med Care 2016 Oct;54(10):901-906. [doi: 10.1097/MLR.0000000000000625] [Medline: 27623005]
72. Pandrangi VC, Shah SN, Bruening JD, et al. Effect of virtual reality on pain management and opioid use among hospitalized patients after head and neck surgery: a randomized clinical trial. JAMA Otolaryngol Head Neck Surg 2022 Aug 1;148(8):724-730. [doi: 10.1001/jamaoto.2022.1121] [Medline: 35679057]
73. Slomski A. Virtual reality lessens postoperative pain. JAMA 2022 Aug 16;328(7):610. [doi: 10.1001/jama.2022.13772] [Medline: 35972497]
74. Geriatric surgery verification. American College of Surgeons. URL: https://www.facs.org/quality-programs/accreditation-and-verification/geriatric-surgery-verification/ [accessed 2026-04-09]
75. CMS issues first HCPCS code and Medicare DME benefit category determination for therapeutic Virtual Reality Device. DLA Piper. URL: https://www.dlapiper.com/en-ca/insights/publications/2023/04/cms-issues-first-hcpcs-code-and-medicare-dme-benefit-category-determination [accessed 2024-02-24]
76. Heinz M, Martin P, Margrett JA, et al. Perceptions of technology among older adults. J Gerontol Nurs 2013 Jan;39(1):42-51. [doi: 10.3928/00989134-20121204-04] [Medline: 23244061]
77. Vaportzis E, Clausen MG, Gow AJ. Older adults perceptions of technology and barriers to interacting with tablet computers: a focus group study. Front Psychol 2017 Oct 4;8:1687. [doi: 10.3389/fpsyg.2017.01687] [Medline: 29071004]
78. Mace RA, Mattos MK, Vranceanu AM. Older adults can use technology: why healthcare professionals must overcome ageism in digital health. Transl Behav Med 2022 Dec 30;12(12):1102-1105. [doi: 10.1093/tbm/ibac070] [Medline: 36073770]
79. Rolstad S, Adler J, Rydén A. Response burden and questionnaire length: is shorter better? A review and meta-analysis. Value Health 2011 Dec;14(8):1101-1108. [doi: 10.1016/j.jval.2011.06.003] [Medline: 22152180]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
GAD-7: Generalized Anxiety Disorder 7-item scale
HMD: head-mounted display
IVR: immersive virtual reality
NRS: Numeric Rating Scale
PCS: Pain Catastrophizing Scale
PHQ-8: Patient Health Questionnaire-8

XSL•FO
RenderX

SPMSQ: Short Portable Mental Status Questionnaire
SSQ: Simulator Sickness Questionnaire
SUS: System Usability Scale
VAS: Visual Analog Scale
VR-CORE: Virtual Reality Clinical Outcomes Research Experts

Edited by J Shiffermiller; submitted 03.Aug.2025; peer-reviewed by S Langenfeld, Sunny, CL Au; revised version received 28.Feb.2026; accepted 03.Mar.2026; published 17.Apr.2026.

Please cite as:
Keny C, Shafiq U, Lorenz K, Russell M, Leutwyler H, Wagner LM, Tang V, Park LG
Immersive Virtual Reality for Pain and Relaxation in Older Adults Following Elective Inpatient Abdominal Surgery: Single-Arm Study Examining Feasibility and Acceptability
JMIR Perioper Med 2026;9:e81791
URL: https://periop.jmir.org/2026/1/e81791
doi:10.2196/81791

© Christina Keny, Ujala Shafiq, Karl Lorenz, Marcia Russell, Heather Leutwyler, Laura M Wagner, Victoria Tang, Linda G Park. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 17.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Clinical Feasibility and Outcomes of Surgeon-Performed Laparoscopic-Guided Subcostal Transversus Abdominis Plane Block in Laparoscopic Cholecystectomy: Prospective Observational Study

Sarun Mahasupachai, MD; Thawatchai Tullavardhana, MD

Department of Surgery, Faculty of Medicine, Srinakharinwirot University, Ransit - Nakhon nayok Road, Ongkharak, Nakhon nayok, Thailand

Corresponding Author:
Thawatchai Tullavardhana, MD
Department of Surgery, Faculty of Medicine, Srinakharinwirot University, Ransit - Nakhon nayok Road, Ongkharak, Nakhon nayok, Thailand

Abstract

Background: Laparoscopic-guided subcostal transversus abdominis plane (TAP) block has been introduced as a surgeon-performed approach to postoperative analgesia in laparoscopic cholecystectomy (LC), allowing direct visual confirmation of local anesthetic delivery without ultrasound guidance. However, evidence regarding its clinical outcomes, particularly in patients with complicated gallstone disease, remains limited.

Objective: This study aimed to evaluate postoperative analgesic outcomes and identify factors associated with opioid requirement following laparoscopic-guided subcostal TAP block.

Methods: A prospective observational study was conducted between November 2023 and October 2024 at Srinakharinwirot University Hospital, Thailand. Patients (aged 18 - 80 years) undergoing LC for uncomplicated or complicated gallstone disease received a laparoscopic-guided subcostal TAP block with 0.25% bupivacaine. Postoperative pain was assessed using the Visual Analog Scale at 2, 4, 6, 8, 12, and 24 hours. Morphine administration within the first 24 hours was recorded. Associations between perioperative variables and opioid requirement were analyzed using univariate and exploratory multivariable logistic regression.

Results: A total of 42 patients were included in the analysis. Of these, 21 (50%) did not require postoperative opioids, while the remaining patients (n=21, 50%) received a mean cumulative morphine dose of 3.86 (SD 1.39) mg within 24 hours. Pain scores were lower during the early postoperative period (2, 4, and 12 h) in patients who did not require opioids. Higher American Society of Anesthesiologists classification was independently associated with postoperative morphine requirement (odds ratio 6.51, 95% CI 1.37 - 30.96; *P*=.01). No major complications or local anesthetic toxicity were observed.

Conclusions: In this prospective observational cohort, laparoscopic-guided subcostal TAP block may be associated with favorable early postoperative analgesic profiles and relatively low opioid requirements after LC, including in patients with gallstone-related complications. Higher American Society of Anesthesiologists classification may be associated with increased opioid demand, highlighting the importance of individualized, risk-adapted analgesic strategies. Although limited by the absence of a control group and modest sample size, these findings support the clinical feasibility of surgeon-performed TAP block for consideration within multimodal analgesia approaches in enhanced recovery after surgery–oriented perioperative care.

Trial Registration: Thai Clinical Trials Registry TCTR20250314002; https://www.thaiclinicaltrials.org/show/TCTR20250314002

(JMIR Perioper Med 2026;9:e87622) doi:10.2196/87622

KEYWORDS

laparoscopic cholecystectomy; minimally invasive surgery; postoperative pain; transversus abdominis plane block; enhanced recovery after surgery

Introduction

Laparoscopic cholecystectomy (LC) is the standard surgical approach for gallstone disease, offering distinct advantages over open cholecystectomy, including shorter recovery times and earlier return to normal activities [1]. Despite these benefits, LC is associated with moderate to severe postoperative pain, particularly within the first 24 hours, often necessitating opioid analgesia [2]. High-dose opioid use, however, is frequently complicated by nausea, vomiting, dizziness, abdominal distension, and urinary retention, which may delay recovery and prolong hospitalization [3,4].

Enhanced recovery after surgery (ERAS) protocols have been widely implemented to optimize perioperative care and expedite

XSL•FO
RenderX

recovery. Multimodal analgesia represents a cornerstone of these protocols, aiming to minimize opioid use while maintaining effective pain control [5,6]. Within this framework, the subcostal transversus abdominis plane (TAP) block has emerged as a valuable component of multimodal analgesia, providing targeted pain relief following LC. When performed under ultrasound guidance using 0.25% bupivacaine, this block reliably anesthetizes thoracic (T7-T12) and lumbar (L1) nerves, thereby improving pain control and facilitating earlier mobilization [7,8]. Nevertheless, its dependence on anesthesiologist expertise and specialized equipment limits feasibility in certain clinical environments.

To address these limitations, the laparoscopic-guided subcostal TAP block has been developed as a surgeon-performed technique seamlessly incorporated into the operative workflow. Under direct laparoscopic visualization, local anesthetic can be precisely delivered into the TAP, providing consistent parietal analgesia while obviating the need for ultrasound equipment or additional personnel. This method has been demonstrated to be safe, efficient, and time-effective, offering a practical alternative for postoperative pain control following LC [9].

Nevertheless, most previous studies evaluating TAP block for LC have primarily focused on patients with uncomplicated gallstone disease and were conducted in controlled trial settings. A recent systematic review and meta-analysis demonstrated that TAP block is effective in reducing postoperative pain and opioid consumption after LC, with most included studies using ultrasound-guided techniques [10]. In contrast, this study evaluates a surgeon-performed, laparoscopic-guided subcostal TAP block integrated into routine operative workflow.

This study aimed to assess the clinical feasibility and outcomes of surgeon-performed, laparoscopic-guided subcostal TAP block for postoperative pain management in patients undergoing LC for both uncomplicated and complicated gallstone disease, including acute cholecystitis and biliary tract obstruction. Additionally, perioperative predictors of postoperative opioid requirement were explored.

Methods

Overview

A single-center observational study was conducted at the Department of Surgery, Faculty of Medicine, Srinakharinwirot University, Thailand, between November 2023 and October 2024. Eligible patients were aged 18 to 80 years and diagnosed with symptomatic cholelithiasis or gallstone-related complications. Uncomplicated gallstone disease was defined as symptomatic cholelithiasis or chronic cholecystitis without evidence of systemic inflammation or biliary complications. Complicated gallstone disease was defined as gallstone-related conditions associated with acute inflammation or biliary obstruction, including acute cholecystitis, acute cholangitis, or biliary obstruction requiring endoscopic retrograde cholangiopancreatography. Patients who were converted to open surgery or had a known allergy to bupivacaine were excluded. The target sample size was 40 patients; however, to enhance reliability, a total of 50 patients were enrolled.

Study Design Declaration

The initial ethics-approved protocol was designed as a randomized comparison between subcostal TAP block and port-site local infiltration. However, due to limited patient recruitment, randomization could not be executed. This report therefore represents an observational analysis of patients who received the surgeon-performed, laparoscopic-guided subcostal TAP block in accordance with the originally approved protocol. No additional procedures, interventions, or deviations from the ethics approval were undertaken.

Ethical Considerations

This study was approved by the Institutional Ethics Committee of Srinakharinwirot University (ethics code: SWUEC-004/2566F). Written informed consent was obtained from all participants prior to enrollment, and all patient data were collected and managed in accordance with institutional and international standards for data confidentiality and ethical research practice.

The study was retrospectively registered with the Thai Clinical Trials Registry (TCTR20250314002) following a change in study execution from the originally approved randomized protocol to a prospective observational design. Importantly, no protocol deviations occurred beyond the scope approved by the institutional ethics committee.

Laparoscopic-Guided Subcostal TAP Block Technique

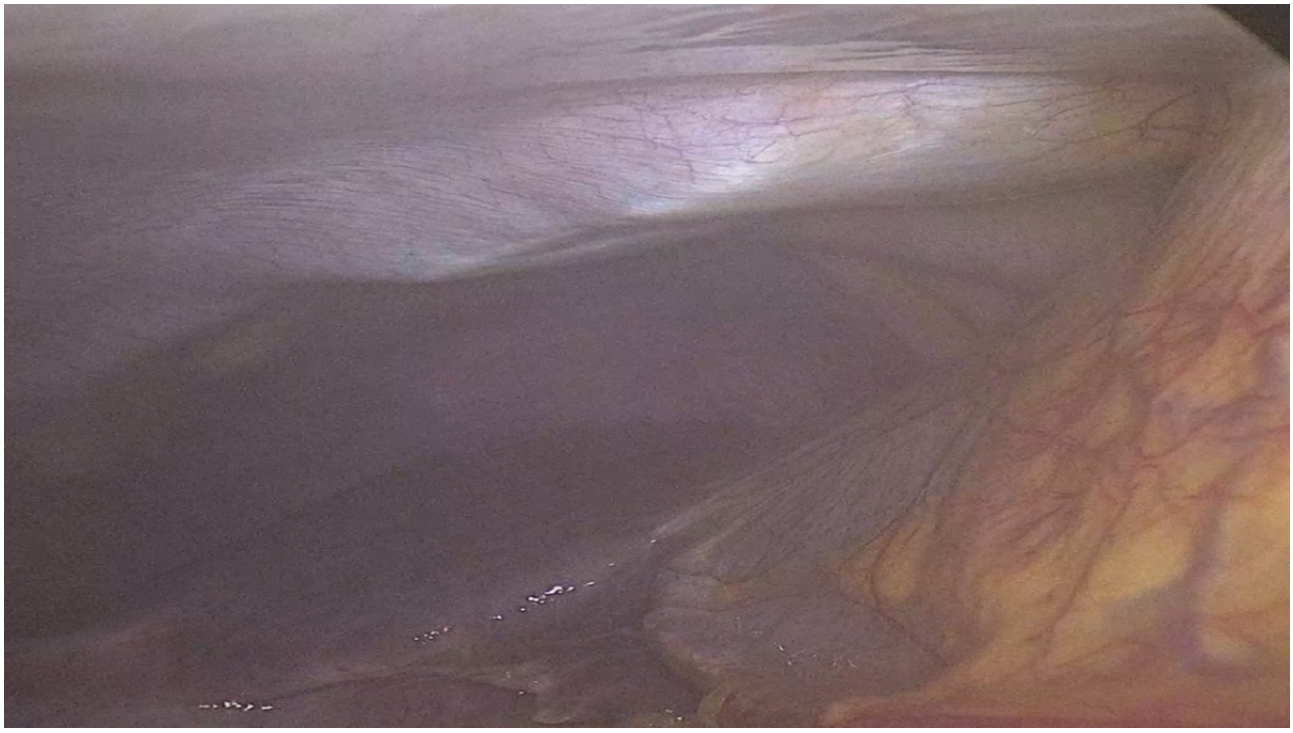
LC was performed under general anesthesia. Intraoperative analgesia was standardized according to the institutional protocol and was not analyzed separately. Prior to skin incision, 5 mL of 0.25% bupivacaine was infiltrated at the umbilical site for local analgesia. A 12-mm umbilical port was inserted for the laparoscopic camera, and pneumoperitoneum was established at an intra-abdominal pressure of 8 to 12 mmHg.

After establishment of pneumoperitoneum, the right upper quadrant was inspected under laparoscopic visualization. The injection point was identified at the right subcostal region, approximately 2 to 3 cm inferior to the costal margin and lateral to the midline, corresponding to the junction between the posterior rectus sheath and the transversus abdominis muscle.

Under direct laparoscopic visualization, a long spinal needle was inserted percutaneously toward the TAP, with initial advancement at an angle of approximately 60 to 80 degrees relative to the abdominal wall. Upon reaching the target fascial plane, the needle angle was adjusted to approximately 30 degrees to facilitate controlled anesthetic delivery. After negative aspiration, 20 mL of 0.25% bupivacaine was injected incrementally. Correct local anesthetic deposition was inferred from laparoscopic visualization of the Doyle bulge, indicating separation of the posterior rectus sheath and transversus abdominis muscle as a surrogate marker of appropriate fascial plane injection (Figure 1).

XSL•FO
RenderX

Figure 1. Illustration of Doyle bulge observed during laparoscopic-guided subcostal transversus abdominis plane block.



In our contemporary LC practice, routine use of a 12-mm epigastric working port has been replaced by three 5-mm working ports positioned along the right subcostal margin, in addition to the 12-mm umbilical optical port. This port configuration reflects an evolving minimally invasive approach aimed at reducing port-site trauma while maintaining adequate operative ergonomics and visualization. After cholecystectomy, the gallbladder was retrieved in a sterile bag through the 12-mm umbilical port. A supplementary video demonstrating the laparoscopic-guided subcostal TAP block technique is provided in Multimedia Appendix 1.

Postoperative Pain Management and Monitoring

All patients received standardized postoperative analgesia. Oral acetaminophen (500 mg every 6 hours) and naproxen (250 mg twice daily) were prescribed as first-line agents; tramadol (50 mg twice daily) was substituted for patients with nonsteroidal anti-inflammatory drug intolerance. Persistent pain with a Visual Analog Scale (VAS) score >5 despite oral medication was managed with intravenous morphine (0.1 mg/kg every 4 hours as needed). Ondansetron (4 mg every 8 hours as required) was given for nausea and vomiting prophylaxis.

Pain intensity was evaluated using the VAS (0=no pain and 10=worst pain) at 2, 4, 6, 8, 12, and 24 hours postoperatively and during mobilization. Total morphine consumption and opioid-related adverse events (nausea, vomiting, and urinary retention) were recorded within the first 24 hours. Patients were observed for possible bupivacaine toxicity during this period.

Clinical Variables

Data collection included patient demographics, comorbidities, laboratory parameters, operative details, postoperative outcomes, and histopathological findings. Laboratory parameters analyzed in this study were preoperative values obtained within 24 hours before surgery. Operative variables comprised surgical duration, estimated blood loss, and intraoperative complications.

Postoperative variables included pain scores, total morphine consumption, and postoperative complications. All data were prospectively recorded in a predesigned spreadsheet (Microsoft Excel) for statistical analysis. The dataset was stored in password-protected files on secure institutional computers, with access restricted to study investigators only. Patient identifiers were removed and replaced with coded study numbers to ensure confidentiality.

Statistical Analysis

Statistical analysis was conducted using SPSS Statistics (version 27.0; IBM Corp). Descriptive statistics (mean, SD, frequencies, and percentages) were used to summarize demographic and clinical data. Comparisons between subgroups were conducted using the chi-square or Fisher exact test for categorical variables and the independent 2-tailed *t* test (or 1-way ANOVA where applicable) for continuous variables.

Exploratory multivariable logistic regression was applied to assess potential associations between perioperative factors and postoperative morphine requirement, acknowledging the limited sample size. Variables included in the multivariable model were selected based on clinical relevance and prior literature, rather than solely on statistical significance in univariate analyses. Given the limited number of outcome events, the analysis was conducted with a restricted number of covariates to maintain an acceptable events per variable ratio and minimize the risk of overfitting. A 2-tailed *P* value <.05 was considered statistically significant.

XSL•FO
RenderX

Results

Overview

During the study period, 50 patients underwent LC. In total, 8 (16%) patients were excluded due to conversion to open surgery, leaving 42 (84%) patients for analysis. Of these, 21 (50%) patients required postoperative morphine within the first 24 hours. Among patients who received opioids (n=21, 50%), the mean cumulative morphine dose was 3.86 (SD 1.39) mg.

Patient Characteristics

Baseline demographic and clinical characteristics are summarized in Table 1. There were no significant differences between the morphine-required and morphine-free groups in age (*P*=.55), BMI (*P*=.55), or sex distribution (*P*=.22). Comorbid conditions were more frequent in the morphine-required group; however, this difference did not reach statistical significance (*P*=.10). The mean American Society of Anesthesiologists (ASA) classification was significantly higher among patients who required morphine (mean 2.14, SD 0.57 vs mean 1.67, SD 0.73; *P*=.024), reflecting a higher baseline perioperative risk profile.

Table . Baseline demographic and clinical characteristics of the study population.

Variable	Morphine required (n=21)	Morphine free (n=21)	*P* value
Age (years), mean (SD)	57.1 (14.1)	54.3 (15.8)	.55
Sex (female), n (%)	13 (61.9)	8 (38.1)	.22
BMI (kg/m²), mean (SD)	25.8 (5.3)	24.8 (4.9)	.55
ASA[a] classification, mean (SD)	2.14 (0.57)	1.67 (0.73)	.02
Any comorbidity, n (%)	17 (81.0)	10 (47.6)	.10
Diabetes mellitus, n (%)	10 (47.6)	6 (28.6)	.21
Hypertension, n (%)	13 (61.9)	10 (47.6)	.36
Cardiovascular disease, n (%)	6 (28.6)	4 (19.0)	.48
Indication for surgery, n (%)[b]			
Symptomatic gallstone	17 (81.0)	15 (71.4)	.48
Acute cholecystitis	2 (9.5)	4 (19.0)	.39
Interval LC[c] after conservative treatment of acute cholecystitis	4 (19.0)	4 (19.0)	>.99
Previous ERCP[d], n (%)	6 (28.6)	5 (23.8)	.73

[a]ASA: American Society of Anesthesiologists.

[b]Categories are not mutually exclusive.

[c]LC: laparoscopic cholecystectomy.

[d]ERCP: endoscopic retrograde cholangiopancreatography.

Preoperative Laboratory Parameters

Preoperative laboratory findings are presented in Table 2. All laboratory values represent measurements obtained within 24 hours prior to surgery. Patients in the morphine-required group had significantly lower baseline hemoglobin (*P*=.01) and hematocrit levels (*P*=.002). Additionally, no significant between-group differences were observed in white blood cell count, neutrophil to lymphocyte ratio, liver function tests, or serum albumin levels.

XSL•FO
RenderX

Table . Preoperative laboratory parameters.

Variable	Morphine required (n=21), mean (SD)	Morphine free (n=21), mean (SD)	*P* value
Hemoglobin (g/dL)	11.95 (1.78)	13.34 (1.60)	.01
Hematocrit (%)	35.9 (4.8)	40.2 (4.0)	.002
White blood cell count (×10³/μL)	11.6 (19.1)	10.7 (11.1)	.85
Neutrophil to lymphocyte ratio	6.92 (13.8)	4.54 (6.54)	.48
Aspartate aminotransferase (U/L)	41.2 (54.4)	45.2 (55.5)	.81
Alanine aminotransferase (U/L)	38.8 (63.7)	46.0 (69.6)	.73
Alkaline phosphatase (U/L)	83.8 (41.2)	95.1 (84.9)	.58
Total bilirubin (mg/dL)	0.70 (0.51)	0.99 (1.30)	.35
Serum albumin (g/dL)	4.33 (0.39)	4.21 (0.34)	.56

Operative Details and Postoperative Outcomes

Operative and postoperative outcomes are summarized in Table 3. Mean operative time (mean 63.5, SD 15.6 minutes vs mean 58.4, SD 18.0 minutes; *P*=.32), estimated blood loss (mean 15.7, SD 11.3 mL vs mean 13.3, SD 12.5 mL; *P*=.52), and length of hospital stay (mean 2.57, SD 0.98 days vs mean 2.33, SD 0.86 days; *P*=.40) did not differ significantly between groups. Furthermore, no significant differences were observed between groups regarding trocar placement, operative technique, operative time, or intraoperative complications, and background analgesic regimens were comparable. Although patients with gallstone-related complications were included as indications for surgery, final histopathological examination of gallbladder specimens was reported as acute or chronic cholecystitis.

Table . Operative details, analgesic regimen, and postoperative outcomes.

Variable	Morphine required (n=21)	Morphine free (n=21)	*P* value
Analgesic regimen			
Acetaminophen+naproxen, n (%)	7 (33.3)	12 (57.1)	.12
Acetaminophen+tramadol, n (%)	14 (66.7)	9 (42.9)	.22
Operative details			
Complete cholecystectomy, n (%)	20 (95.2)	19 (90.5)	.56
Partial cholecystectomy, n (%)	1 (4.7)	2 (9.5)	.56
Operative time (minutes), mean (SD)	63.5 (15.6)	58.4 (18.0)	.32
Estimated blood loss (mL), mean (SD)	15.7 (11.3)	13.3 (12.5)	.52
Postoperative outcomes			
Postoperative complication, n (%)	1 (4.7)	1 (4.7)	.44
Nausea and vomiting, n (%)	1 (4.7)	1 (4.7)	.44
Hospital stay (days), mean (SD)	2.57 (0.98)	2.33 (0.86)	.40
Pathology: chronic cholecystitis, n (%)	20 (95.3)	17 (81.0)	.16
Pathology: acute cholecystitis, n (%)	1 (4.7)	4 (19.0)	.30

Factors Associated With Postoperative Morphine Requirement

Factors associated with postoperative morphine requirement in exploratory multivariable analysis are summarized in Table 4. Higher ASA class was associated with increased odds of morphine use within 24 hours after surgery (*P*=.01; odds ratio 6.51, 95% CI 1.37 - 30.96). Lower hemoglobin level demonstrated a trend toward association with morphine requirement but did not reach statistical significance (*P*=.07; odds ratio 0.58, 95% CI 0.32 - 1.06). Other variables, including age, sex, gallstone-related complications, and history of endoscopic retrograde cholangiopancreatography, were not significantly associated with opioid use.

XSL•FO
RenderX

Table . Multivariable logistic regression analysis of factors associated with postoperative morphine requirement.

Variable	Odds ratio (95% CI)	*P* value
Female	1.81 (0.26-12.80)	.55
ASA[a] classification	6.51 (1.37-30.96)	.01
Hemoglobin (per g/dL)	0.58 (0.32-1.06)	.07
Age (years)	0.98 (0.92-1.04)	.46
Gallstone-related complication	0.91 (0.16-5.15)	.91
Previous ERCP[b]	0.75 (0.12-4.56)	.75

[a]ASA: American Society of Anesthesiologists.

[b]ERCP: endoscopic retrograde cholangiopancreatography.

Given the limited number of outcome events, regression analyses were conducted within an exploratory framework constrained by events per variable considerations.

Postoperative Pain Scores

Postoperative pain scores are summarized in Table 5. Patients who required morphine reported higher VAS scores during the early postoperative period, particularly at 2 hours (mean 3.29, SD 1.45 vs mean 1.93, SD 0.96; *P*=.009), 4 hours (mean 3.95, SD 1.39 vs mean 2.07, SD 1.00; *P*<.001), and 12 hours (mean 3.57, SD 1.44 vs mean 2.43, SD 1.06; *P*=.02). At 6 hours, 24 hours, and during mobilization, pain scores remained numerically higher in the morphine-required group but did not reach statistical significance.

Table . Comparison of postoperative pain scores between morphine-required and morphine-free groups.

Time point	Morphine required, mean (SD)	Morphine free, mean (SD)	*P* value
2 hours	3.29 (1.45)	1.93 (0.96)	.009
4 hours	3.95 (1.39)	2.07 (1.00)	<.001
6 hours	3.33 (1.43)	2.73 (1.42)	.26
12 hours	3.57 (1.44)	2.43 (1.06)	.02
24 hours	2.57 (1.21)	1.97 (0.81)	.07
During mobilization	4.29 (1.33)	3.97 (1.25)	.40

These findings descriptively reflect differences in pain experience between groups and are presented to contextualize postoperative opioid requirement rather than to infer comparative analgesic effectiveness.

Discussion

Principal Findings

This prospective observational study suggests a clinically relevant opioid-sparing association of laparoscopic-guided subcostal TAP block in patients undergoing LC. Approximately half of the patients (n=21, 50%) did not require postoperative opioids, while those who did (n=21, 50%) received only a modest cumulative dose (mean 3.86, SD 1.39 mg).

The observed analgesic pattern was most evident during the early postoperative period, particularly between 2 and 12 hours, which is consistent with the expected pharmacodynamic profile of 0.25% bupivacaine. Within the context of this observational cohort, these findings support the feasibility of surgeon-performed TAP block as a practical adjunct to multimodal analgesia strategies, with the potential to limit postoperative opioid exposure while maintaining adequate pain control [11-14].

The results of this study are directionally consistent with previous reports suggesting that thoracoabdominal and subcostal TAP blocks are associated with improved early postoperative pain control following LC [15-17]. Importantly, this study extends the existing literature by including patients with complicated gallstone disease, a population that has been relatively underrepresented in prior research.

From a mechanistic perspective, the observed analgesic association may be attributable to localized somatic blockade of the upper abdominal wall corresponding to trocar insertion sites. Such coverage is thought to attenuate incisional and parietal peritoneal pain, which may explain the more pronounced pain relief observed during the first 12 postoperative hours [10,18]. Clinically, these observations underscore the importance of integrating regional analgesic techniques with scheduled nonopioid coanalgesics and appropriately timed rescue analgesia to maintain adequate pain control within ERAS-oriented perioperative pathways [5,6,13,14].

Exploratory predictor analysis suggested that patient-related factors were more strongly associated with postoperative opioid requirement than intraoperative variables. Higher ASA classification was independently associated with postoperative morphine use, indicating that greater comorbidity burden may be linked to increased analgesic needs despite regional blockade.

XSL•FO
RenderX

These findings support a risk-adapted approach to perioperative pain management for patients at higher risk.

Although the TAP block primarily targets somatic abdominal wall pain, unmeasured factors such as visceral pain burden, neuropathic pain components, and subtle variations in block accuracy may have influenced postoperative pain perception and opioid requirement. These factors were not objectively assessed in this study and may contribute to residual variability beyond patient-level characteristics such as ASA classification.

Beyond its analgesic association, surgeon-performed subcostal TAP block offers several practical advantages, particularly in settings with limited anesthesiology support or restricted access to ultrasound equipment. Incorporation of this technique into the laparoscopic workflow allows a consistent and equipment-independent approach to regional analgesia, aligning with broader initiatives in opioid stewardship and sustainable perioperative care [9,15-17,19].

Several limitations should be acknowledged. The single-center, nonrandomized design limits causal inference, and the absence of a control group precludes direct comparison with standard port-site local anesthetic infiltration. The modest sample size restricts statistical precision; therefore, multivariable analyses were conducted within an exploratory framework with selective variable inclusion to reduce the risk of overfitting.

In addition, no formal postoperative sensory testing was performed to objectively verify block success, primarily due to practical considerations within the perioperative workflow and the study's focus on clinically relevant outcomes. Instead, correct local anesthetic deposition was inferred from laparoscopic visualization of the Doyle bulge, indicating separation of the posterior rectus sheath and transversus abdominis muscle as a surrogate marker of appropriate fascial plane injection [20,21]. However, this visual confirmation cannot fully substitute for objective sensory testing and does not confirm the extent or consistency of dermatomal blockade.

Substitution of tramadol for nonsteroidal anti-inflammatory drugs in a small number of patients may have introduced minor confounding, although perioperative analgesic protocols were otherwise standardized. In addition, psychosocial factors known to influence postoperative pain, including anxiety, pain catastrophizing, and prior opioid exposure, were not assessed [22,23]. These limitations should be considered when interpreting the findings.

Despite these constraints, this study has several notable strengths. The prospective data collection, use of standardized perioperative analgesic pathways, and inclusion of patients with both uncomplicated and complicated gallstone disease enhance the clinical relevance of the study. Overall, the findings indicate that surgeon-performed, laparoscopic-guided subcostal TAP block is a technically straightforward and reproducible adjunct within multimodal analgesia strategies. The observed early opioid-sparing association supports the feasibility of considering this technique in ERAS-oriented perioperative pathways and broader opioid reduction efforts.

The relatively longer hospital stay observed in this cohort reflects local institutional practice, in which LC is not routinely performed as a day-case procedure. Inclusion of patients with complicated gallstone disease required preoperative admission and postoperative observation for safety. In addition, routine preoperative admission at least 1 day prior to surgery, according to institutional protocol, contributed to the overall length of hospital stay.

Future research should include multicenter randomized controlled trials comparing laparoscopic-guided TAP block with standard port-site local anesthetic infiltration. On the basis of the observed effect estimates, a sample size of approximately 80 to 100 patients per arm may provide 80% power to detect a 1-point difference in mean VAS pain score at an α level of .05. Incorporation of cost-effectiveness analyses, patient-reported outcome measures, and longer-term follow-up would further clarify the clinical value, scalability, and role of this technique in minimally invasive surgery.

Conclusions

This study suggests that laparoscopic-guided subcostal TAP block may be associated with lower early postoperative pain scores and reduced opioid requirements in patients undergoing LC, including those with gallstone-related complications. Although limited by sample size, the findings support the feasibility of considering surgeon-performed subcostal TAP block as part of multimodal analgesia strategies within ERAS-oriented perioperative pathways.

Acknowledgments

The authors thank the surgical staff at the Department of Surgery, Faculty of Medicine, Srinakharinwirot University, for their assistance with data acquisition. During the preparation of this work, the authors used ChatGPT to improve grammar and language clarity. After using this tool, the authors carefully reviewed, revised, and approved the content and take full responsibility for the published article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' Contributions

TT contributed to the conception and design of the study, data acquisition, data analysis and interpretation, and manuscript drafting. SM contributed to study design, data collection and analysis, and manuscript preparation.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Laparoscopic-guided subcostal transversus abdominis plane (TAP block).
[MP4 File, 98727 KB - periop_v9i1e87622_app1.mp4]

References

1. Tullavardhana T. Critical view of safety: a safe method to prevent bile duct injury from laparoscopic cholecystectomy. J Med Health Sci 2015;22:49-57 [FREE Full text]
2. Keir A, Rhodes L, Kayal A, Khan OA. Does a transversus abdominis plane (TAP) local anaesthetic block improve pain control in patients undergoing laparoscopic cholecystectomy? A best evidence topic. Int J Surg 2013;11(9):792-794. [doi: 10.1016/j.ijsu.2013.05.039] [Medline: 23770342]
3. Law LS, Lo EA, Gan TJ, et al. Preoperative antiemetic and analgesic management. In: Gan TJ, Thacker JK, Miller TE, editors. Enhanced Recovery for Major Abdominopelvic Surgery: Professional Communications Inc; 2016:105-120.
4. Ljungqvist O, Scott M, Fearon KC. Enhanced recovery after surgery: a review. JAMA Surg 2017 Mar 1;152(3):292-298. [doi: 10.1001/jamasurg.2016.4952] [Medline: 28097305]
5. Kaye AD, Urman RD, Rappaport Y, et al. Multimodal analgesia as an essential part of enhanced recovery protocols in the ambulatory settings. J Anaesthesiol Clin Pharmacol 2019 Apr;35(Suppl 1):S40-S45. [doi: 10.4103/joacp.JOACP_51_18] [Medline: 31142958]
6. Hebbard P, Fujiwara Y, Shibata Y, Royse C. Ultrasound-guided transversus abdominis plane (TAP) block. Anaesth Intensive Care 2007 Aug;35(4):616-617. [doi: 10.1177/0310057X0703500417] [Medline: 18020088]
7. Wang W, Wang L, Gao Y. A meta-analysis of randomized controlled trials concerning the efficacy of transversus abdominis plane block for pain control after laparoscopic cholecystectomy. Front Surg 2021;8:700318. [doi: 10.3389/fsurg.2021.700318] [Medline: 34422893]
8. Elamin G, Waters PS, Hamid H, et al. Efficacy of a laparoscopically delivered transversus abdominis plane block technique during elective laparoscopic cholecystectomy: a prospective, double-blind randomized trial. J Am Coll Surg 2015 Aug;221(2):335-344. [doi: 10.1016/j.jamcollsurg.2015.03.030] [Medline: 25899736]
9. Venkatraman R, Saravanan R, Dhas M, Pushparani A. Comparison of laparoscopy-guided with ultrasound-guided subcostal transversus abdominis plane block in laparoscopic cholecystectomy - a prospective, randomised study. Indian J Anaesth 2020 Dec;64(12):1012-1017. [doi: 10.4103/ija.IJA_528_20] [Medline: 33542563]
10. Alsharari AF, Abuadas FH, Alnassrallah YS, Salihu D. Transversus abdominis plane block as a strategy for effective pain management in patients with pain during laparoscopic cholecystectomy: a systematic review. J Clin Med 2022 Nov 22;11(23):6896. [doi: 10.3390/jcm11236896] [Medline: 36498471]
11. Simpson JC, Bao X, Agarwala A. Pain Management in Enhanced Recovery after Surgery (ERAS) protocols. Clin Colon Rectal Surg 2019 Mar;32(2):121-128. [doi: 10.1055/s-0038-1676477] [Medline: 30833861]
12. Cao L, Yang T, Hou Y, Yong S, Zhou N. Efficacy and safety of different preemptive analgesia measures in pain management after laparoscopic cholecystectomy: a systematic review and network meta-analysis of randomized controlled trials. Pain Ther 2024 Dec;13(6):1471-1497. [doi: 10.1007/s40122-024-00647-w] [Medline: 39227523]
13. Barazanchi AWH, MacFater WS, Rahiri JL, et al. Evidence-based management of pain after laparoscopic cholecystectomy: a PROSPECT review update. Br J Anaesth 2018 Oct;121(4):787-803. [doi: 10.1016/j.bja.2018.06.023] [Medline: 30236241]
14. Cho HY, Hwang IE, Lee M, Kwon W, Kim WH, Lee HJ. Comparison of modified thoracoabdominal nerve block through perichondral approach and subcostal transversus abdominis plane block for pain management in laparoscopic cholecystectomy: a randomized-controlled trial. Korean J Pain 2023 Jul 1;36(3):382-391. [doi: 10.3344/kjp.23114] [Medline: 37336756]
15. Tolchard S, Davies R, Martindale S. Efficacy of the subcostal transversus abdominis plane block in laparoscopic cholecystectomy: Comparison with conventional port-site infiltration. J Anaesthesiol Clin Pharmacol 2012 Jul;28(3):339-343. [doi: 10.4103/0970-9185.98331] [Medline: 22869941]
16. Ozciftci S, Sahiner Y, Sahiner IT, Akkaya T. Is Right Unilateral Transversus Abdominis Plane (TAP) block successful in postoperative analgesia in laparoscopic cholecystectomy? Int J Clin Pract 2022;2022:2668215. [doi: 10.1155/2022/2668215] [Medline: 35685608]
17. Emile SH, Elfeki H, Elbahrawy K, Sakr A, Shalaby M. Ultrasound-guided versus laparoscopic-guided subcostal transversus abdominis plane (TAP) block versus No TAP block in laparoscopic cholecystectomy; a randomized double-blind controlled trial. Int J Surg 2022 May;101:106639. [doi: 10.1016/j.ijsu.2022.106639] [Medline: 35487422]
18. Hinrichs-Rocker A, Schulz K, Järvinen I, Lefering R, Simanski C, Neugebauer EAM. Psychosocial predictors and correlates for chronic post-surgical pain (CPSP) - a systematic review. Eur J Pain 2009 Aug;13(7):719-730. [doi: 10.1016/j.ejpain.2008.07.015] [Medline: 18952472]

XSL•FO
RenderX

19. Boonyapalanant C, Woranisarakul V, Jitpraphai S, et al. The efficacy of inside-out transversus abdominis plane block vs local infiltration before wound closure in pain management after kidney transplantation: a double-blind, randomized trial. Siriraj Med J 2022;74(4):233-238. [doi: 10.33192/Smj.2022.29]
20. Vindal A, Sarda H, Lal P. Laparoscopically guided transversus abdominis plane block offers better pain relief after laparoscopic cholecystectomy: results of a triple blind randomized controlled trial. Surg Endosc 2021 Apr;35(4):1713-1721. [doi: 10.1007/s00464-020-07558-9] [Medline: 32297054]
21. Suragul W, Tantawanit A, Rungsakulkij N, et al. Effect of local anaesthetic infiltration on postoperative pain after laparoscopic cholecystectomy: randomized clinical trial. BJS Open 2022 May 2;6(3):zrac066. [doi: 10.1093/bjsopen/zrac066] [Medline: 35639946]
22. Ip HYV, Abrishami A, Peng PWH, Wong J, Chung F. Predictors of postoperative pain and analgesic consumption: a qualitative systematic review. Anesthesiology 2009 Sep;111(3):657-677. [doi: 10.1097/ALN.0b013e3181aae87a] [Medline: 19672167]
23. Riddle DL, Wade JB, Jiranek WA, Kong X. Preoperative pain catastrophizing predicts pain outcome after knee arthroplasty. Clin Orthop Relat Res 2010 Mar;468(3):798-806. [doi: 10.1007/s11999-009-0963-y] [Medline: 19585177]

Abbreviations

ERAS: enhanced recovery after surgery
LC: laparoscopic cholecystectomy
TAP: transversus abdominis plane
VAS: Visual Analog Scale

Edited by N Rohatgi, S Pan; submitted 11.Nov.2025; peer-reviewed by A Jahangeer, ADK Hill, H Abdallah; revised version received 22.Feb.2026; accepted 25.Feb.2026; published 25.Mar.2026.

Please cite as:
Mahasupachai S, Tullavardhana T
Clinical Feasibility and Outcomes of Surgeon-Performed Laparoscopic-Guided Subcostal Transversus Abdominis Plane Block in Laparoscopic Cholecystectomy: Prospective Observational Study
JMIR Perioper Med 2026;9:e87622
URL: https://periop.jmir.org/2026/1/e87622
doi:10.2196/87622

© Sarun Mahasupachai, Thawatchai Tullavardhana. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 25.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Preoperative Mortality Risk Prediction and Equity Using Preoperative Score to Predict Postoperative Mortality in a Diverse US Surgical Population: Retrospective Cohort Study

Jackson Hall*, BS; Haley Warzecha*, BA; Chase Penhallurick, BA; Venkatsai Bellala, BS; Shyamal Asher, MD, MBA

Department of Anesthesiology, Brown University Health, Brown University, 593 Eddy Street, Davol 129, Providence, RI, United States

*these authors contributed equally

Corresponding Author:
Shyamal Asher, MD, MBA
Department of Anesthesiology, Brown University Health, Brown University, 593 Eddy Street, Davol 129, Providence, RI, United States

Abstract

Background: The Preoperative Score to Predict Postoperative Mortality (POSPOM), a point-based score using only preoperative variables, was developed and validated to predict in-hospital postoperative mortality in European surgical cohorts. Other widely used clinical risk tools perform differently across racial and socioeconomic groups. Although postoperative mortality has been studied in diverse US surgical populations, POSPOM performance across racial and ethnic subgroups remains insufficiently characterized.

Objective: This study aimed to evaluate POSPOM performance for 30-day postoperative mortality in a diverse US cohort and explore its performance for 1-year mortality among high-risk surgical patients and across racial and ethnic subgroups.

Methods: We performed a retrospective cohort study using the All of Us Research Program. Adults aged 18 to 91 years who underwent 1 of 21 POSPOM-eligible surgical procedures were included, and only each participant's most recent eligible procedure was analyzed. POSPOM scores were calculated from age, procedure type, and 15 weighted comorbidities. The primary outcome was 30-day postoperative mortality, as incomplete discharge data precluded reliable assessment of in-hospital mortality; 1-year mortality was examined as an exploratory secondary outcome. Logistic regression estimated odds ratios (ORs) per 1-point increase in the POSPOM score, and discrimination was evaluated using the area under the receiver operating characteristic curve (AUC). A high-risk subgroup (procedure points ≥14) was analyzed separately, and outcomes across racial and ethnic subgroups were explored descriptively. Given the small subgroup sizes and few events, subgroup analyses were interpreted cautiously.

Results: Among 17,262 patients, 12 (0.1%) died within 30 days and 45 (0.3%) within 1 year. For 30-day mortality, POSPOM yielded an AUC of 0.764, with higher odds of death per point (OR 1.112, 95% CI 1.040-1.190; *P*=.002). Discrimination was lower for 1-year mortality (AUC 0.638; OR 1.054, 95% CI 1.017-1.092; *P*=.004). In the high-risk subgroup (n=536; 30-day mortality 1.1%, 1-year mortality 2.1%), the 30-day AUC fell to 0.583, with no significant association (OR 0.942, 95% CI 0.812-1.092; *P*=.43). For 1-year mortality, the AUC was 0.692, with an inverse association (OR 0.882, 95% CI 0.781-0.998; *P*=.046). Descriptively, high-risk Black patients had higher 30-day mortality than White patients (2/61, 3.3% vs 4/348, 1.1%); no deaths were recorded among Asian, other-race, or Hispanic high-risk patients.

Conclusions: POSPOM demonstrated an AUC of 0.764 for 30-day mortality overall but provided limited risk stratification among high-risk surgical patients. Higher observed mortality among high-risk Black patients suggests the score may not fully capture baseline risk across racial groups, possibly reflecting structural and social factors outside its scope. As the 1-year and subgroup analyses were exploratory with few events, larger, adequately powered US studies are needed to evaluate POSPOM at its intended in-hospital end point, assess calibration across diverse populations, and inform recalibration before broader clinical use.

(JMIR Perioper Med 2026;9:e99623) doi:10.2196/99623

KEYWORDS

preoperative care; postoperative mortality; risk assessment; health care disparities; socioeconomic factors; comorbidity

XSL•FO
RenderX

Introduction

The Preoperative Score to Predict Postoperative Mortality (POSPOM) is a validated, point-based tool that estimates postoperative mortality using only preoperative variables, including age, comorbidities, and surgical procedure type [1]. Developed to support early risk stratification, POSPOM intentionally excludes subjective clinician judgment and intraoperative data, making it relatively simple to implement across different hospitals and practice settings [1]. POSPOM was originally developed and validated to predict in-hospital postoperative mortality, with "good calibration and excellent discrimination" reported in large European surgical cohorts [1].

Postoperative mortality has previously been examined in large, demographically diverse US surgical cohorts. Smith et al [2] evaluated postoperative mortality risk estimates and 1-year survival among Veterans Health Administration surgical patients, while Kaiser et al [3] examined 30-day mortality in a large multicenter cohort of noncardiac surgical patients. However, these studies did not specifically evaluate POSPOM performance across racial and ethnic subgroups. Prior studies have shown that widely used clinical prediction tools, including the Sequential Organ Failure Assessment (SOFA) score, the American College of Surgeons National Surgical Quality Improvement Program surgical risk calculator, and cardiovascular risk equations, can exhibit differential performance across racial and socioeconomic groups [1,4-8]. These findings have raised concerns that risk models may not perform equally across different patient populations, treatment decisions, and outcomes [1,4-8].

Although prior studies have examined postoperative mortality prediction in diverse US surgical populations, POSPOM performance across racial and ethnic subgroups has not been specifically characterized [2,3]. Evaluating POSPOM across these subgroups may help determine whether its predictive performance remains consistent across heterogeneous surgical populations.

Hospital discharge dates were not available for all participants in the All of Us dataset, preventing reliable assessment of in-hospital mortality across the full cohort [9]. We therefore selected 30-day mortality as the primary outcome because it provided a standardized postoperative interval that could be applied consistently across participants. We additionally examined 1-year mortality as an exploratory secondary outcome to assess the longer-term association between preoperative POSPOM scores and postoperative mortality.

The primary objective of this study was to evaluate POSPOM performance for 30-day postoperative mortality in a large, demographically diverse US cohort from the All of Us Research Program [9]. Secondary exploratory objectives were to evaluate performance for 1-year mortality, performance among high-risk surgical patients, and performance across racial and ethnic subgroups. We hypothesized that, similar to other clinical risk prediction tools, POSPOM may not perform consistently across demographic subgroups, particularly among patients undergoing high-risk surgical procedures [1,4-6].

Methods

Study Design and Data Source

This retrospective cohort study evaluated POSPOM performance for 30-day postoperative mortality in a large-scale, diverse, real-world US population, with secondary exploratory analyses of 1-year mortality and demographic subgroups. Data were derived from the National Institutes of Health (NIH) All of Us Research Program's Registered Tier Dataset version 8, which includes electronic health record (EHR) data from more than 1 million participants [9]. Clinical information was extracted from the Observational Medical Outcomes Partnership (OMOP) Common Data Model version 5.3.1, using tables for demographics, procedures, condition occurrences, surveys, and death records [10]. This retrospective study was not prospectively registered, and no formal study protocol was published.

Study Population

The initial cohort was identified from participants with recorded condition and procedure domains. To ensure independent observations among participants with multiple eligible surgeries, only the most recent procedure was included for the analysis. Participants were included in the final analytic cohort if they were aged 18 to 91 years at the time of data collection and underwent 1 of the 21 POSPOM-eligible surgical procedures.

Participants were excluded from the analysis if they fell outside the specified age range, had missing demographic or surgical data required for cohort eligibility or POSPOM score calculation, or underwent surgical procedures not covered by the POSPOM classification system.

Following these criteria, the final analytic cohort consisted of 17,262 patients.

Data Extraction and Processing

Demographic variables, including sex at birth, race, and ethnicity, were cleaned and standardized for analysis. Race was categorized as White, Black, Asian, or other, and ethnicity was coded as a binary variable (Hispanic vs non-Hispanic). Sex at birth was recoded as male or female.

POSPOM Score Calculation

Each participant's total POSPOM score was computed as the sum of 3 components—age, procedure type, and comorbidities—with total possible scores ranging from 0 to 54 points. The complete scoring criteria, including all variables and point allocations, are detailed in Table 1.

XSL•FO
RenderX

Table . Components and point allocations for the Preoperative Score to Predict Postoperative Mortality (POSPOM)[a].

Category and characteristic	POSPOM points
Patient age (y)	
<45	0
45-49	1
50-54	2
55-59	3
60-64	4
65-69	5
70-74	6
75-79	7
80-84	8
85-89	9
≥90	10
Comorbidities	
Ischemic heart disease	1
Cardiac arrhythmia or heart blocks	1
Chronic heart failure or cardiomyopathy	4
Peripheral vascular disease	1
Dementia	2
Cerebrovascular disease	1
Hemiplegia	4
Chronic obstructive pulmonary disease	1
Chronic respiratory failure	3
Chronic alcohol abuse	4
Cancer	4
Diabetes	1
Transplanted organs	2
Preoperative chronic dialysis	1
Chronic renal failure	2
Surgical category	
Endoscopic digestive	0
Ophthalmologic	0
Gynecologic	6
Other orthopedic	6
Interventional cardiorhythmology	8
Arthroplasty and spine	9
Ear, nose, and throat	9
Minor urologic	9
Plastic	9
Major urologic	12
Other surgery	12
Minor hepatic	12
Minor gastrointestinal	13

XSL•FO
RenderX

Category and characteristic	POSPOM points
Renal transplant	13
Minor vascular	13
Orthopedic trauma	14
Major hepatic	15
Thoracic	15
Neurosurgery	15
Major vascular	16
Major gastrointestinal	16
Interventional neuroradiology	17
Cardiac	17
Transplant	22
Multiple trauma related	22

[a]Adapted from Le Manach et al [1].

Points were assigned based on 5-year age brackets at the time of surgery, ranging from 0 points for patients younger than 45 years to 10 points for patients aged 90 years or older, as detailed in Table 1.

Procedures were mapped to POSPOM categories using Current Procedural Terminology and Systematized Nomenclature of Medicine Clinical Terms codes. Points associated with these procedures ranged from 0 for low-risk procedures (eg, colonoscopy and cataract surgery) to 22 for high-risk procedures (eg, liver transplantation).

Fifteen distinct comorbidities were identified using OMOP concept identifiers and mapped to diagnostic codes. To ensure comprehensive capture, structured query language queries used OMOP concept identifiers to include all descendant codes. Weights were applied to each condition, ranging from 1 point for conditions such as diabetes or chronic obstructive pulmonary disease to 4 points for chronic heart failure or malignant neoplastic disease.

Outcome Measures

The primary outcome was 30-day postoperative mortality, defined as death occurring within 30 days of the index surgical procedure. POSPOM was originally developed to predict in-hospital mortality; however, hospital discharge dates were not available for all participants, preventing reliable assessment of this outcome across the full cohort [1]. Thirty-day mortality was therefore selected as a standardized postoperative outcome that could be consistently assessed across participants. The secondary outcome was 1-year postoperative mortality, defined as death occurring within 365 days of the index surgical procedure. Mortality status and dates were derived from the OMOP death domain, and time from surgery to death was calculated. Participants without a recorded death date in the EHR were assumed to have survived through the relevant follow-up interval.

Statistical Analysis

A logistic regression model was used for the primary analysis, with the continuous POSPOM score as the predictor and 30-day mortality as the binary outcome. Model performance was evaluated through discrimination and calibration. Discrimination was measured by the concordance statistic (C-statistic or area under the receiver operating characteristic curve [AUC]). In accordance with recent consensus, qualitative labels for model discrimination were avoided, and continuous AUC values were reported directly [11]. Calibration was assessed using the Hosmer-Lemeshow goodness-of-fit test and calibration plots comparing observed vs predicted mortality rates. Effect estimation used logistic regression to calculate odds ratios (ORs) for mortality per 1-point increase in the POSPOM score, reported with 95% CIs.

To evaluate model fairness and calibration across different demographic strata, a stratified analysis was conducted specifically for high-risk procedures, defined as a POSPOM procedure score of 14 or greater [1]. Exploratory subgroup analyses were conducted by race and ethnicity.

Data Quality and Validation

Data quality checks included assessment of missingness for key variables, validation of date ranges (restricting surgery dates to 1980-2023), verification of age calculations (≥18 years), and evaluation of mortality data completeness. The low number of mortality events limited statistical precision, particularly in subgroup analyses; these findings were therefore interpreted cautiously.

Software and Tools

Data extraction and processing were performed using Python (version 3.10; Python Software Foundation) with the pandas and NumPy libraries. Statistical analyses were conducted using *statsmodels* and scikit-learn, while visualizations were generated with Matplotlib and Seaborn. All queries and data management were conducted through the BigQuery environment within the All of Us researcher workbench.

Ethical Considerations

All data were deidentified in accordance with All of Us privacy protocols. No direct participant contact or intervention occurred.

XSL•FO
RenderX

Analyses were performed under an approved data use agreement, and all findings are presented in aggregate form to protect participant confidentiality. The Brown University Institutional Review Board (IRB) formally reviewed this study and determined that it did not constitute human subjects research under federal regulations because it exclusively used deidentified registered tier data from the NIH All of Us Research Program. The overarching All of Us data collection protocol is independently reviewed and approved by the program's central IRB. Patients and members of the public were not directly involved in the design, conduct, analysis, or reporting of this study.

Results

Cohort Characteristics and Overall Mortality

The study analyzed an analytic cohort of 17,262 participants who underwent POSPOM-eligible surgical procedures. Within this overall population, the absolute mortality event rate was low: 12 (0.1%) participants died within 30 days of their index surgery and 45 (0.3%) participants died within 1 year. A subset of 536 (3.1%) participants were categorized as high-risk based on a POSPOM surgical procedure score of 14 or greater. Mortality rates were notably higher in this high-risk subset, with 6 (1.1%) deaths occurring at 30 days and 11 (2.1%) deaths occurring at 1 year. The detailed predictive performance of the POSPOM score across these groups is summarized in Table 2.

Table . Predictive performance of the Preoperative Score to Predict Postoperative Mortality (POSPOM) score across the overall cohort and high-risk subgroups.

Group	Outcome type	Participants, n	Deaths, n (%)	AUC[a]	OR[b] (95% CI)	*P* value
All patients	30-d	17,262	12 (0.1)	0.764	1.112 (1.040-1.190)	.002
All patients	1-y	17,262	45 (0.3)	0.638	1.054 (1.017-1.092)	.004
High-risk (POSPOM score ≥14)	30-d	536	6 (1.1)	0.583	0.942 (0.812-1.092)	.43
High-risk (POSPOM score ≥14)	1-y	536	11 (2.1)	0.692	0.882 (0.781-0.998)	.046
Asian (High-risk)	30-d	9	0 (0)	—[c]	—	—
Black (High-risk)	30-d	61	2 (3.3)	—	—	—
Other (High-risk)	30-d	118	0 (0)	—	—	—
White (High-risk)	30-d	348	4 (1.2)	—	—	—

[a]AUC: area under the receiver operating characteristic curve.

[b]OR: odds ratio.

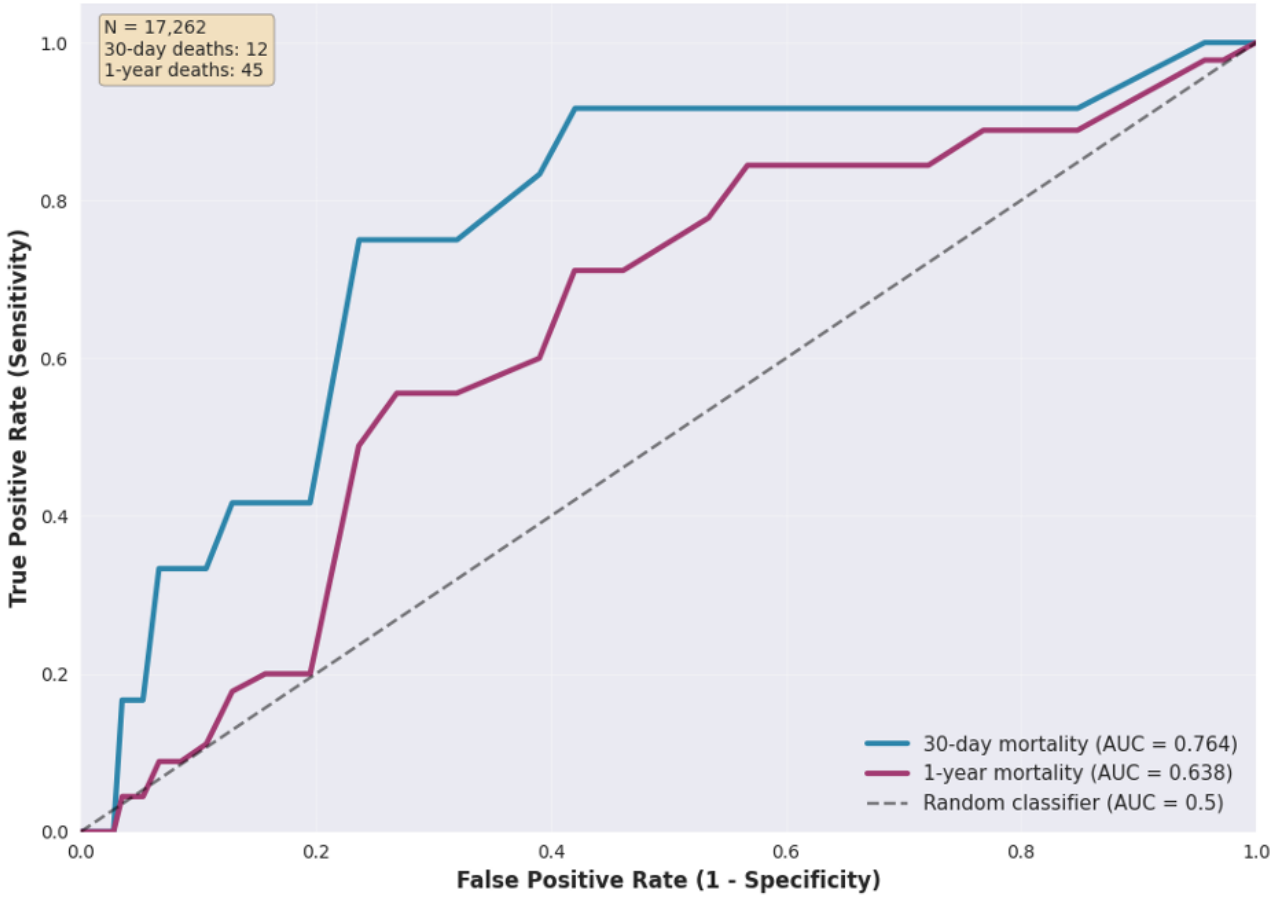
[c]Not applicable.

Overall POSPOM Score Performance

In the primary analysis of the overall cohort, the POSPOM model demonstrated an AUC of 0.764 for 30-day mortality (Figure 1). Logistic regression indicated that each 1-point increase in the POSPOM score was associated with a statistically significant increase in the odds of 30-day mortality (OR 1.112, 95% CI 1.040-1.190; *P*=.002). When evaluating longer-term outcomes, the model's discriminative ability decreased; for 1-year mortality, the model produced an AUC of 0.638 (Figure 1), and each 1-point score increase was associated with a smaller increase in mortality odds (OR 1.054, 95% CI 1.017-1.092; *P*=.004).

XSL•FO
RenderX

Figure 1. Receiver operating characteristic curves showing Preoperative Score to Predict Postoperative Mortality score performance for 30-day vs 1-year mortality in the overall cohort. AUC: area under the receiver operating characteristic curve.

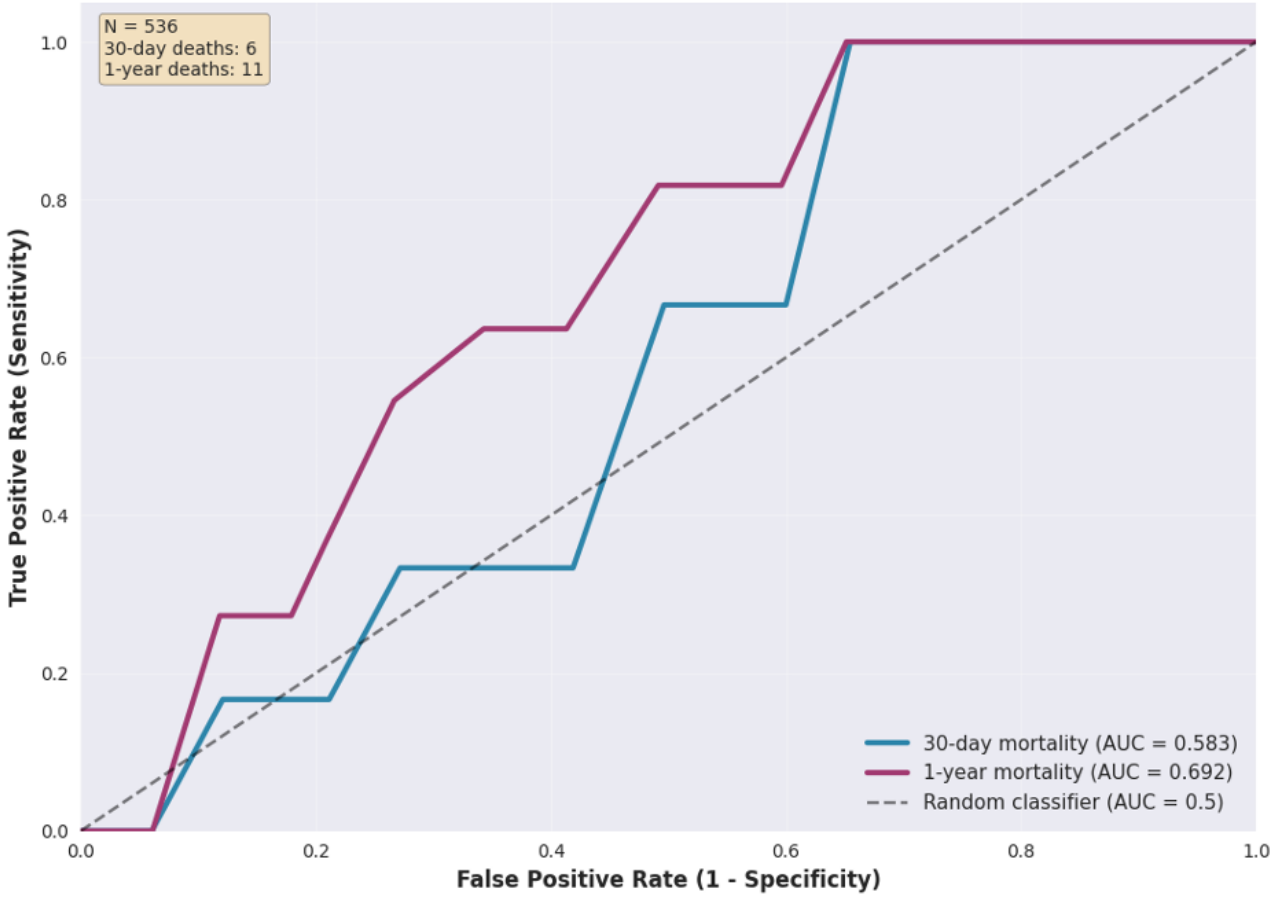


Performance in High-Risk and Demographic Subgroups

When the analysis was restricted to the high-risk surgical cohort (n=536), model discrimination dropped considerably. For 30-day mortality in this subgroup, the POSPOM score yielded an AUC of 0.583 (Figure 2), with the score no longer showing a positive association with mortality (OR 0.942, 95% CI 0.812-1.092; *P*=.43). For 1-year mortality in the high-risk cohort, the AUC was 0.692 (Figure 2), and a 1-point increase in the POSPOM score was inversely associated with mortality odds (OR 0.882, 95% CI 0.781-0.998; *P*=.046).

XSL•FO
RenderX

Figure 2. Receiver operating characteristic curves showing Preoperative Score to Predict Postoperative Mortality score performance for 30-day vs 1-year mortality in the high-risk surgical cohort. AUC: area under the receiver operating characteristic curve.



Race-stratified analyses within the high-risk cohort were evaluated descriptively due to small sample sizes. Among Black patients (n=61), there were 2 (3.3%) recorded 30-day deaths. Among White patients (n=348), there were 4 (1.2%) recorded 30-day deaths. No mortality events were recorded among Asian (n=9) or other (n=118) racial groups within the high-risk cohort. In ethnicity-stratified analyses, the Hispanic subgroup (n=72) recorded no mortality events, while the non-Hispanic group (n=464) recorded 11 deaths at 1 year.

Discussion

Principal Findings

This retrospective cohort study evaluated POSPOM performance in a diverse US surgical population, with an exploratory assessment of outcomes across racial and ethnic subgroups. For 30-day mortality in the overall surgical cohort, POSPOM demonstrated an AUC of 0.764, but its performance decreased among high-risk patients (AUC 0.583) and for the 1-year outcome (AUC 0.638). In exploratory subgroup analyses, discrimination appeared to be higher among Black participants; however, event counts were small and should be interpreted cautiously. Although a higher absolute mortality rate was observed among Black patients compared with White patients within this high-risk subset, our data are insufficient to conclude whether the model differentially accounts for underlying baseline risk. These findings underscore the need for larger, adequately powered US validation studies to assess model calibration across minoritized populations.

General Performance and High-Risk Limitations

The initial validation of POSPOM in large European cohorts suggested that it was a robust and objective alternative to more subjective risk assessments [1,4]. The AUC of 0.764 observed for 30-day mortality supports the potential utility of POSPOM for standardized short-term postoperative risk assessment in diverse US populations. However, the decline in discrimination among high-risk surgical patients (AUC 0.583 for 30-day mortality in those with surgical points ≥14) indicates limited predictive performance in this subgroup. In these patients, the score did not provide meaningful risk stratification. For the 1-year outcome, an inverse association was observed within the high-risk subgroup. Given the small number of events and the fact that POSPOM was not developed for long-term mortality prediction, this finding should be interpreted cautiously and warrants evaluation in a larger cohort.

The outcomes evaluated in this study differ from the in-hospital mortality end point for which POSPOM was originally developed [1]. Because hospital discharge dates were unavailable for some participants, in-hospital mortality could not be reliably assessed across the full cohort. We therefore used 30-day mortality as a standardized primary outcome and examined 1-year mortality as an exploratory secondary outcome. Future studies with complete admission and discharge data

XSL•FO
RenderX

should evaluate POSPOM performance at its intended in-hospital end point in diverse US populations.

Given the low number of events, all of the estimates, especially those from subgroup analyses, are subject to statistical uncertainty.

Equity and Racial Disparities

A central objective of this study was to assess whether POSPOM exhibits performance disparities across demographic lines, similar to other clinical tools [4,6,8]. Due to small sample sizes in the high-risk subgroup, outcomes across racial groups were evaluated descriptively. Notably, despite similar preoperative risk categorization, Black patients had a 30-day mortality rate 3 times higher than that of White patients (2/61, 3.3% vs 4/348, 1.2%) and a 1-year mortality rate more than triple that of their White counterparts (6.6% vs 2%).

These descriptive findings suggest that the point-based POSPOM system may underrepresent the absolute risk faced by Black patients. This disparity likely reflects broader structural inequities in health care access, the impact of social determinants of health, or differences in hospital quality that the POSPOM score is not designed to capture [1]. Although we did not analyze insurance status, hospital volume, or nonsurgical participants, prior literature indicates these systemic factors may heavily influence US surgical outcomes [12-14]. As a purely clinical tool, POSPOM does not account for these external health care system variances [1,4].

Comparison With Existing Risk Models

The performance variations and racial mortality gaps observed in this study align with emerging literature regarding algorithmic bias in clinical risk calculators [5,8,14-17]. For instance, studies evaluating the SOFA score have demonstrated that this model may overpredict mortality for certain racial groups while underpredicting mortality for others, potentially due to the inclusion of variables that do not account for baseline physiological differences or chronic exposure to systemic stressors [8,14,17]. Similarly, investigations into the National Surgical Quality Improvement Program surgical risk calculator have found that while the model provides strong aggregate discrimination, it often underestimates complications and mortality for minority patients and those from lower socioeconomic backgrounds [5,15,16]. This study suggests that POSPOM exhibits a similar pattern where the model fails to calibrate for the higher baseline incidence of adverse outcomes in minoritized cohorts. This consistent trend across various clinical tools suggests that preoperative risk models may require the inclusion of social and structural variables to achieve true equity in predictions.

Limitations in Subgroup Analyses

The absence of mortality events among several racial and ethnic subgroups highlights the small sample size challenge associated with these exploratory analyses [18]. Similarly, the absence of events in the Hispanic subgroup limited the ability to assess POSPOM performance in this population. Larger sample sizes are needed to evaluate fairness and calibration across ethnic groups. In addition, the extremely low number of mortality events in the high-risk surgical subgroup limits the statistical power of the model's discrimination in this specific cohort; therefore, our high-risk analyses should be interpreted as exploratory and signal-generating rather than definitive. Furthermore, while the All of Us database provided a more diverse sample than many traditional datasets, the overall low event rate for 30-day mortality (0.1%) limits the precision of our estimates [9]. These exploratory findings primarily underscore the need for larger, fully powered datasets to accurately evaluate subgroup performance.

Clinical and Policy Implications

These results exemplify the potential danger of assuming that a clinically validated tool is neutral simply because it avoids subjective physician input. If POSPOM is used to guide surgical decision-making or resource allocation without recalibration for US populations, it may inadvertently exacerbate existing health disparities [6,19,20]. For example, if a high-risk Black patient receives a risk score that discriminates well but overlooks a systematically higher baseline mortality, they may be given less aggressive preoperative optimization or postoperative surveillance than their true risk warrants. Ultimately, our findings suggest that the clinical utility of POSPOM in diverse US health care settings would be significantly enhanced by regional and demographic recalibration. Prioritized next steps should include using larger, multicenter US registries to reweight the tool's clinical variables and explicitly investigating whether the integration of specific social determinants of health can optimize baseline calibration across heterogeneous patient populations.

Generative AI Use

The authors used Cursor (an AI-powered editor) to help draft, debug, and refactor Python code for data processing and statistical analysis. All queries, analytic decisions, and outputs were reviewed and validated by the authors, who retain full responsibility for the manuscript.

Funding

No funding was received for this study.

Data Availability

The analytical code used for data processing and statistical analysis is not publicly available.

XSL•FO
RenderX

The data used in this study are available to registered, approved researchers through the National Institutes of Health All of Us Research Program's cloud-based platform. Detailed information on how to apply for data access can be found on the All of Us Research Program website [21].

Authors' Contributions

Conceptualization: JH (lead), HW (equal)
Formal analysis: HW (lead)
Methodology: HW, JH, CP
Project administration: JH (lead), HW (equal)
Supervision: SA
Visualization: HW, CP
Writing—original draft: JH (lead), CP (supporting), VB (supporting)
Writing—review and editing: JH, HW, CP, VB, SA

Conflicts of Interest

None declared.

References

1. Le Manach Y, Collins G, Rodseth R, et al. Preoperative Score to Predict Postoperative Mortality (POSPOM): derivation and validation. Anesthesiology 2016 Mar;124(3):570-579. [doi: 10.1097/ALN.0000000000000972] [Medline: 26655494]
2. Smith T, Li X, Nylander W, Gunnar W. Thirty-day postoperative mortality risk estimates and 1-year survival in Veterans Health Administration surgery patients. JAMA Surg 2016 May 1;151(5):417-422. [doi: 10.1001/jamasurg.2015.4882] [Medline: 26747331]
3. Kaiser HA, Saied NN, Kokoefer AS, Saffour L, Zoller JK, Helwani MA. Incidence and prediction of intraoperative and postoperative cardiac arrest requiring cardiopulmonary resuscitation and 30-day mortality in non-cardiac surgical patients. PLoS One 2020 Jan 22;15(1):e0225939. [doi: 10.1371/journal.pone.0225939] [Medline: 31967987]
4. Layer YC, Menzenbach J, Layer YL, et al. Validation of the Preoperative Score to Predict Postoperative Mortality (POSPOM) in Germany. PLoS One 2021 Jan 27;16(1):e0245841. [doi: 10.1371/journal.pone.0245841] [Medline: 33503043]
5. Vigna C, Ore AS, Fabrizio A, Messaris E. Evaluation of racial/ethnic disparities in surgical outcomes after rectal cancer resection: an ACS-NSQIP analysis. Surg Pract Sci 2024 Apr 17;17:100248. [doi: 10.1016/j.sipas.2024.100248] [Medline: 39845641]
6. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
7. Brindle P, May M, Gill P, et al. Primary prevention of cardiovascular disease: a web-based risk score for seven British Black and minority ethnic groups. Heart 2006 Nov;92(11):1595-1602. [doi: 10.1136/hrt.2006.092346] [Medline: 16762981]
8. Ashana DC, Anesi GL, Liu VX, et al. Equitably allocating resources during crises: racial differences in mortality prediction models. Am J Respir Crit Care Med 2021 Mar 30;204(2):178-186. [doi: 10.1164/rccm.202012-4383OC] [Medline: 33751910]
9. All of Us Research Program Investigators, Denny JC, Rutter JL, et al. The "All of Us" Research Program. N Engl J Med 2019 Aug 15;381(7):668-676. [doi: 10.1056/NEJMsr1809937] [Medline: 31412182]
10. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(1):54-60. [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]
11. de Hond AA, Steyerberg EW, van Calster B. Interpreting area under the receiver operating characteristic curve. Lancet Digit Health 2022 Dec;4(12):e853-e855. [doi: 10.1016/S2589-7500(22)00188-1] [Medline: 36270955]
12. Reames BN, Birkmeyer NJ, Dimick JB, Ghaferi AA. Socioeconomic disparities in mortality after cancer surgery: failure to rescue. JAMA Surg 2014 May;149(5):475-481. [doi: 10.1001/jamasurg.2013.5076] [Medline: 24623106]
13. Wakeam E, Hevelone ND, Maine R, et al. Failure to rescue in safety-net hospitals: availability of hospital resources and differences in performance. JAMA Surg 2014 Mar;149(3):229-235. [doi: 10.1001/jamasurg.2013.3566] [Medline: 24430015]
14. Tolchin B, Oladele C, Galusha D, et al. Racial disparities in the SOFA score among patients hospitalized with COVID-19. PLoS One 2021 Sep 17;16(9):e0257608. [doi: 10.1371/journal.pone.0257608] [Medline: 34535009]
15. Matar DY, Knoedler S, Matar AY, et al. Surgical outcomes and sociodemographic disparities across all races: an ACS-NSQIP and NHIS multi-institutional analysis of over 7.5 million patients. Ann Surg Open 2024 Jul 16;5(3):e467. [doi: 10.1097/AS9.0000000000000467] [Medline: 39310358]
16. Betancourt-Garcia MM, Vatcheva K, Gupta PK, et al. The effect of Hispanic ethnicity on surgical outcomes: an analysis of the NSQIP database. Am J Surg 2019 Apr;217(4):618-633. [doi: 10.1016/j.amjsurg.2018.10.004] [Medline: 30466953]
17. Sarkar R, Martin C, Mattie H, Gichoya JW, Stone DJ, Celi LA. Performance of intensive care unit severity scoring systems across different ethnicities in the USA: a retrospective observational study. Lancet Digit Health 2021 Apr;3(4):e241-e249. [doi: 10.1016/S2589-7500(21)00022-4] [Medline: 33766288]

XSL•FO
RenderX

18. Walker R, Behman R, Bubis L, Karanicolas P. Prospective and retrospective cohort studies. In: Thoma A, Sprague S, Voineskos SH, Serrano PE, Goldsmith CH, editors. Evidence-Based Surgery: A Guide to Understanding and Interpreting the Surgical Literature: Springer; 2025:191-204. [doi: 10.1007/978-3-031-87083-5_17]
19. Chin MH, Afsar-Manesh N, Bierman AS, et al. Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. JAMA Netw Open 2023 Dec 1;6(12):e2345050. [doi: 10.1001/jamanetworkopen.2023.45050] [Medline: 38100101]
20. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight - reconsidering the use of race correction in clinical algorithms. N Engl J Med 2020 Aug 27;383(9):874-882. [doi: 10.1056/NEJMms2004740] [Medline: 32853499]
21. All of Us Research Program. URL: https://www.nih.gov/allofus [accessed 2026-11-13]

Abbreviations

AUC: area under the receiver operating characteristic curve
EHR: electronic health record
IRB: institutional review board
NIH: National Institutes of Health
OMOP: Observational Medical Outcomes Partnership
OR: odds ratio
POSPOM: Preoperative Score to Predict Postoperative Mortality
SOFA: Sequential Organ Failure Assessment

Edited by I Steenstra, K Cummings; submitted 28.Apr.2026; peer-reviewed by D Riddle, KF Hollis; revised version received 27.Oct.2026; accepted 27.Oct.2026; published 18.Nov.2026.

Please cite as:
Hall J, Warzecha H, Penhallurick C, Bellala V, Asher S
Preoperative Mortality Risk Prediction and Equity Using Preoperative Score to Predict Postoperative Mortality in a Diverse US Surgical Population: Retrospective Cohort Study
JMIR Perioper Med 2026;9:e99623
URL: https://periop.jmir.org/2026/1/e99623
doi:10.2196/99623

© Jackson Hall, Haley Warzecha, Chase Penhallurick, Venkatsai Bellala, Shyamal Asher. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 18.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Association Between Complications and Death Within 30 Days After Orthopedic Surgery: Vascular Events in Noncardiac Surgery Patients Cohort Evaluation (VISION) Substudy

Lily J Park[1,2], MD, MSc; P J Devereaux[2,3,4], MD, PhD; Ameen Patel[2,5], MD; Vikas Tandon[2,5], MD; Diane Heels-Ansdell[6], MSc; Lehana Thabane[6], PhD; Pablo E Serrano[1], MD, PhD; Matthew T V Chan[7], MBBS, PhD; Wojciech Szczeklik[8], MD, PhD; Sadeesh Srinathan[9], MD, MSc; Ignacio Garutti[10], MD, PhD; Gerard Urrutia[11], MS, MD, PhD; Ernesto Guerra-Farfan[12], MD; Hassaan Abdel Khalik[13], MD; Emmanuelle Duceppe[14], MD, PhD; Sandra Ofori[2,3], MSc, MD, MBBS, PhD; Maura Marcucci[2], MD, MSc; David Conen[2,3], MD, MPH; Michael K Wang[2,3], MD; Jessica Spence[2,15], MD, PhD; Daniel Tushinski[13], MD, MSc; Kamal Bali[13], MBBS, MSc; Anthony Adili[13], MD; Vickas Khanna[13], MD, MHA; Ana Claudia Tonelli[16], MD, PhD; Francesca Mulazzani[17], MD; Wenjun Jiang[18]; Olufemi R Ayeni[13], MD, PhD; Gerard Slobegean[19], MD, MPH; Theodore Miclau[20], MD; Mohit Bhandari[6,13], MSc, MD, PhD; Flavia K Borges[2,5,6], MD, PhD

[1]Department of Surgery, Division of General Surgery, McMaster University, Hamilton, ON, Canada
[10]Department of Medicine, Universidad Complutense de Madrid, Madrid, Spain
[11]Institut de Recerca Sant Pau (IR SANT PAU) – CIBERESP, Barcelona, Spain
[12]Department of Surgery, Division of Orthopedic surgery, Vall d'Hebron Hospital Universitari, Barcelona, Catalonia, Spain
[13]Department of Surgery, Division of Orthopedic Surgery, McMaster University, Hamilton, ON, Canada
[14]Department of Medicine, Université de Montréal, Montreal, QC, Canada
[15]Department of Anesthesia, McMaster University, Hamilton, ON, Canada
[16]Internal Medicine Service, Hospital de Clínicas de Porto Alegre, Porto Alegre, Rio Grande do Sul, Brazil
[17]Department of Medicine, University of Milano-Bicocca, Milan, Lombardy, Italy
[18]Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada
[19]Department of Orthopedics, School of Medicine, University of California, Irvine, CA, United States
[2]Population Health Research Institute, 237 Barton St E, C1-109, Hamilton, ON, Canada
[20]Department of Orthopaedic Surgery, University of California, San Francisco, San Francisco, CA, United States
[3]Division of Cardiology, Department of Medicine, McMaster University, Hamilton, ON, Canada
[4]World Health Research, Hamilton, ON, Canada
[5]Department of Medicine, McMaster University, Hamilton, ON, Canada
[6]Department of Health Research Methods, Evidence & Impact, McMaster University, Hamilton, ON, Canada
[7]Chinese University of Hong Kong, Hong Kong, China (Hong Kong)
[8]Centre for Intensive Care and Perioperative Medicine, Jagiellonian College, Krakow, Kujawsko-Pomorskie, Poland
[9]Department of Surgery, University of Manitoba, Winnipeg, MB, Canada

Corresponding Author:
Flavia K Borges, MD, PhD
Population Health Research Institute, 237 Barton St E, C1-109, Hamilton, ON, Canada

Abstract

Background: The contemporary causes of postoperative mortality in orthopedic surgery are not well characterized.

Objective: This study aimed to describe the epidemiology of postoperative complications among adult patients who underwent orthopedic surgery and inform their relationships with 30-day mortality.

Methods: Vascular Events in Noncardiac Surgery Patients Cohort Evaluation (VISION) was a prospective cohort study involving 40,004 adult patients who underwent noncardiac surgery across 28 centers in 14 countries. For the subset of patients who underwent orthopedic surgery, a Cox proportional hazards model was used to determine time-dependent associations between various surgical complications and 30-day postoperative mortality. Analyses were adjusted for preoperative and surgical variables.

Results: Among 8385 patients who underwent an orthopedic surgery in VISION, 1.6% (n=132) died within 30 days of surgery. Of the 132 deaths, 63.6% (n=84) occurred in hospital during the index hospitalization, while 36.4% (n=48) occurred after discharge.

XSL•FO
RenderX

The incidence of death across the subcategories of orthopedic surgery was above-knee amputation (30/221, 13.6%), internal fixation of femur (29/750, 3.9%), lower leg amputation (9/252, 3.6%), major hip or pelvic surgery (49/2898, 1.7%), major spine surgery (8/1405, 0.6%), and knee arthroplasty (7/2876, 0.2%). A total of 6 postoperative complications (myocardial injury after noncardiac surgery [MINS], major bleeding, infection without sepsis, sepsis, stroke, and atrial fibrillation) were associated with death in adjusted analyses. The greatest attributable fractions of postoperative mortality (ie, proportion of deaths in the cohort that can be attributed to each complication, if causality were established) were from MINS (1454/8385, 17.3%; hazard ratio [HR] 2.08, 95% CI 1.38 - 3.14; *P*<.001; attributable fraction=20.6%), major bleeding (2422/8385, 28.9%; HR 1.95, 95% CI 1.34 - 2.85; *P*<.001; attributable fraction=16.5%), and sepsis (318/8385, 3.8%; HR 6.24, 95% CI 3.85 - 10.12; *P*<.001; attributable fraction=9.7%).

Conclusions: The complications most attributable to 30-day mortality following orthopedic surgery were MINS, major bleeding, and sepsis. These findings highlight areas for further study to mitigate perioperative mortality in orthopedic surgery. MINS demonstrated the highest attributable fraction for mortality (20.6%), emphasizing the importance of appropriate MINS screening, diagnosis, and management.

(JMIR Perioper Med 2026;9:e90823) doi:10.2196/90823

KEYWORDS

perioperative medicine; orthopedic surgery; morbidity; postoperative complications; mortality

Introduction

Globally, the number of orthopedic surgeries performed annually is estimated to have approached 31.4 million procedures in 2024 [1]. Among these patients, the mortality rate is broadly estimated to be between 0.6% to 8.7%. Even at the lower end of this estimated range, mortality after orthopedic surgery represents a substantial health issue [2-5]. Much of the existing mortality data are derived from patients with hip fracture and orthopedic trauma, which represents only a small fraction of the entire orthopedic surgery population [2-5]. This highlights the need to better elucidate mortality rates in a broad sample of general contemporary patients who underwent orthopedic surgery.

The largest study to date in this area used the National Hospital Discharge Survey to identify risk factors for mortality in orthopedic surgery across a nationwide sample of hospitals in the United States [5]. However, this study is limited by the retrospective nature of the study, reliance on administrative data, and older data. Furthermore, the National Hospital Discharge Survey is considered less reliable in reflecting morbidity and complications [5-10]. Considering this, an updated and accurate understanding of modifiable risk factors for death in orthopedic surgery is needed.

Vascular Events in Noncardiac Surgery Patients Cohort Evaluation (VISION) was a large prospective cohort study that included patients who underwent noncardiac surgery and were systematically followed to document postoperative complications, including mortality [11]. We previously reported the incidence of perioperative complications and associated mortality for the entire VISION study population. Patients undergoing orthopedic surgery represent a unique population that increasingly includes older adults with rising prevalence of cardiovascular disease [12-14]. This necessitates specialty-specific epidemiologic data.

The objective of this prospective cohort substudy was to describe the epidemiology of postoperative complications and death, evaluate the associations between these postoperative complications and death, and report the attributable fractions of each complication for death among a contemporary orthopedic surgery cohort.

Methods

Ethical Considerations

The research ethics board at each participating site approved the protocol before patient recruitment. This study was approved by the Hamilton Integrated Research Ethics Board under the project number 07-220. All participants provided informed consent and we followed the ethical standard principles of the Declaration of Helsinki, including maintaining privacy and confidentiality of research participants' data. No compensation was provided to participants.

Study Design

The design and methods of the VISION study have been previously described [11,15]. In summary, VISION was an international prospective cohort study, which enrolled 40,004 patients who underwent noncardiac surgery across 28 centers in 14 countries from August 2007 to November 2013. Patients aged ≥45 years, who underwent noncardiac surgery, receiving general or regional anesthesia, and requiring at least 1 overnight hospital stay after surgery were eligible for inclusion. This inclusion criteria was determined a priori with the intention to efficiently identify those who are at higher risk of mortality and capture a greater number of mortality outcomes and perioperative complications. Each participating hospital obtained approval from their research ethics board prior to the start of patient enrollment. Refer to Multimedia Appendix 1 for funding sources. The Strengthening the Reporting of Observational Studies in Epidemiology Statement was followed (Checklist 1).

Analysis Population

Patients who were enrolled in the VISION study and underwent an orthopedic procedure (including spine surgery) were included in this substudy. According to the VISION definitions, this would include surgeries categorized as major hip or pelvic surgery, internal fixation of femur, knee arthroplasty,

XSL•FO
RenderX

above-knee amputation, lower leg amputation, and major spine surgery.

Follow-Up

Patients were followed for 30 days following their surgery and were censored at the time of their last assessment if the 30-day follow-up was not complete.

Complications Variables

Definitions for all variables included in the analyses are provided in (Multimedia Appendix 2). The primary outcome was time to all-cause mortality. The following postoperative complications during the first 30 days after surgery were investigated: myocardial injury after noncardiac surgery (MINS), venous thromboembolism (VTE), stroke, major bleeding, acute kidney injury (AKI) resulting in dialysis, sepsis, infection without sepsis, new clinically important atrial fibrillation, and congestive heart failure [11,12,16-19]. These variables were selected for clinical relevance while being parsimonious to preserve model stability. Of note, major bleeding was defined using a previously validated definition in the perioperative setting demonstrating independent association with 30-day mortality [18,19].

Statistical Analysis

A statistical analysis plan was written before undertaking analyses for this study, which was finalized in June 2023. Descriptive statistics were used to report patient characteristics, as well as the incidence of death and postoperative complications. As AKI resulting in dialysis occurred at the rate of 0.2%, this variable was omitted from the model to preserve model stability.

To determine the relationship between complications and 30-day mortality, we used a time-dependent Cox proportional hazards model and adjusted for baseline characteristics that were known to be independently associated with 30-day mortality, according to previous VISION analyses [11]. Complications were modeled as time-dependent covariates, allowing for multiple events per patient and accounting for their timing relative to death. This approach enabled assessment of the association between each complication with mortality rather than assignment of a single causal event. This model was adjusted for age category (65 - 75 vs 45 - 65 years, and age ≥75 vs 45 - 65 years), cancer at the time of surgery, history of chronic obstructive pulmonary disease, surgery urgency (emergent: <24 hours from admission for an acute surgical condition, urgent: 24 - 72 hours from admission for an acute surgical condition, nonemergent: all other surgeries), and history of peripheral arterial disease. All variables included in the model were selected a priori. We reported the adjusted hazard ratios (HRs) and corresponding 95% CIs. To avoid optimistic estimation of the c-statistic from possible overfitting, model performance was assessed using the c-statistic corrected for optimism using 1000 bootstrapped samples [20]. This involved the generation of 1000 simulated datasets through random sampling and replacement from the original dataset to estimate model performance while accounting for overfitting. Attributable fractions for death were calculated using an established method [21]. The attributable fraction represents the proportion of deaths that would not have occurred in the VISION orthopedic cohort if the complication had not occurred, if we assumed a causal relationship between the complication and death. Attributable fraction methods have been increasingly used in perioperative and epidemiologic research. Prior applications include large cohort analyses in noncardiac surgery and population health studies evaluating modifiable contributors to mortality [11,17,21-23]. There were 10 events per variable included in the model, which supports model stability [24].

For all tests, we used α <.05 as the level of significance. Analyses were performed using R statistical software (version 4.2.2; R Foundation for Statistical Computing) using the "survival," "epiR," and "ggplot2" packages.

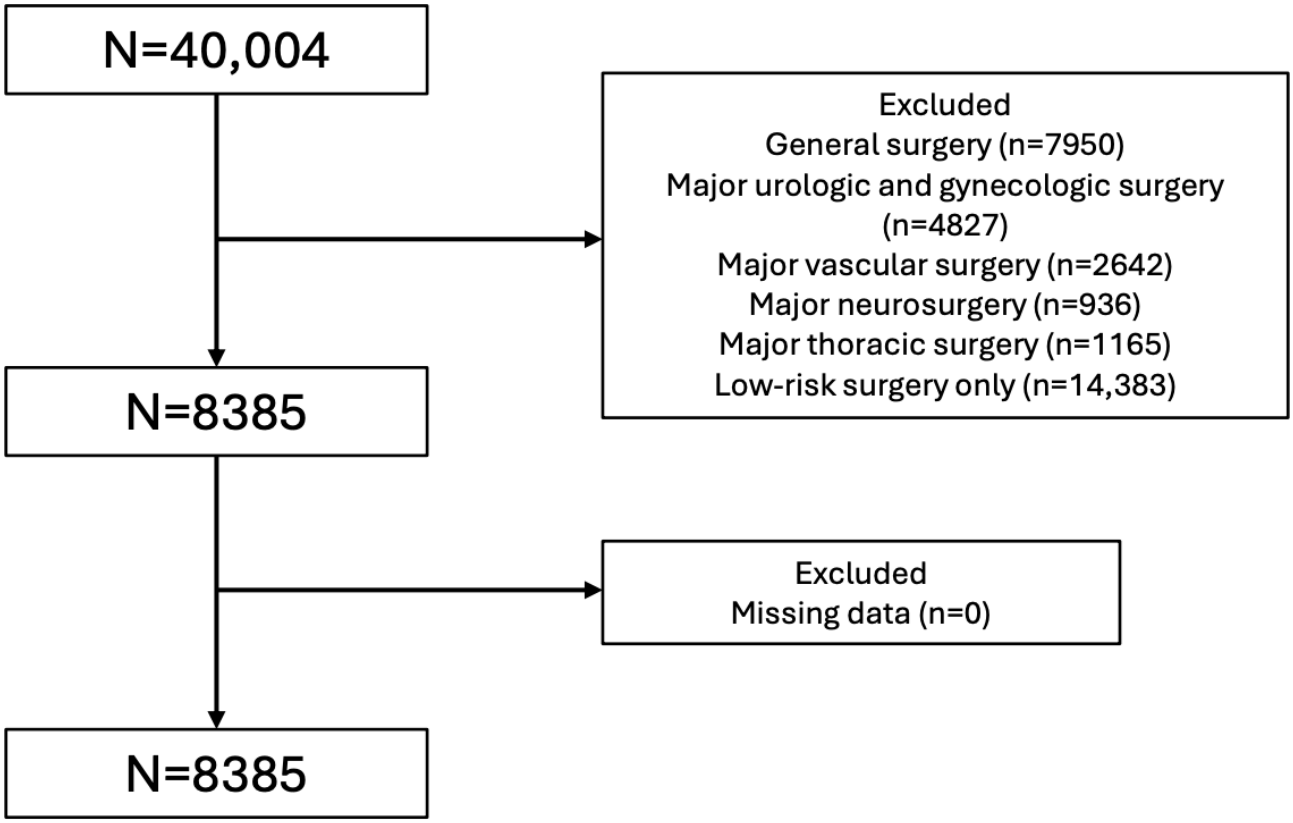
Results

Patient Characteristics

Within the 40,004 patients enrolled in the overall VISION study, 21% (8385/40,004) underwent an orthopedic surgery and were included in these analyses. Figure 1 demonstrates the flow of patient inclusion. Table 1 presents the baseline characteristics of the participants. More than half of the patients were aged >64 years, and 57.3% (4802/8385) were female. There was no loss to follow-up in the orthopedic surgery population.

XSL•FO
RenderX

Figure 1. Flow of patient inclusion in the Vascular Events in Noncardiac Surgery Patients Cohort Evaluation (VISION) orthopedic surgery substudy. Note: in the overall VISION study, 284 patients had 2 or 3 surgical categories.



XSL•FO
RenderX

Table . Baseline characteristics of the orthopedic surgery cohort.

Characteristics	Participants, n (%)	Number of deaths within 30 days, n (%)
Age (years)		
45 - 64	3411 (40.7)	31 (0.9)
65 - 74	2418 (28.8)	20 (0.8)
≥75	2556 (30.5)	81 (3.2)
Sex: female	4802 (57.3)	76 (1.6)
History of		
Hypertension	4948 (59)	89 (1.8)
Diabetes	1823 (21.8)	44 (2.4)
Coronary artery disease	1253 (15)	41 (3.3)
Peripheral arterial disease	519 (6.2)	39 (7.5)
Chronic obstructive pulmonary disease	640 (7.6)	33 (5.2)
Coronary revascularization	470 (5.6)	6 (1.3)
Stroke	398 (4.7)	20 (5)
Congestive heart failure	401 (4.8)	20 (5)
Active cancer	320 (3.8)	13 (4.1)
Atrial fibrillation	323 (3.9)	11 (3.4)
Preoperative estimated glomerular filtration rate (mL/min/1.73 m^2)		
<30	294 (3.7)	26 (8.8)
30 - 44	503 (6.2)	22 (4.4)
45 - 59	988 (12.3)	12 (1.2)
≥60	6264 (77.8)	67 (1.1)
Types of orthopedic surgery		
Major hip or pelvic surgery	2898 (34.6)	49 (1.7)
Internal fixation of femur	750 (8.9)	29 (3.9)
Knee arthroplasty	2876 (34.3)	7 (0.2)
Above-knee amputation	221 (2.6)	30 (13.6)
Lower leg amputation	252 (3)	9 (3.6)
Major spine surgery	1405 (16.8)	8 (0.6)
Subcategory of surgery		
Nonemergent	7369 (87.9)	81 (1.1)
Urgent or emergent	1016 (12.1)	51 (5)
Type of anesthesia		
General only	2436 (29.1)	36 (1.5)
Neuraxial (spinal or epidural) only	4182 (49.9)	72 (1.7)
General with nitrous oxide only	424 (5.1)	10 (2.4)
General and thoracic epidural only	10 (0.1)	0 (0)
General and nerve block only	424 (5.1)	2 (0.5)
Other	903 (10.8)	12 (1.3)

Among 8385 patients, 34.6% (n=2898) underwent a major hip or pelvic surgery, 34.3% (n=2876) underwent knee arthroplasty, 16.8% (n=1405) underwent a major spine surgery, 8.9% (n=750) underwent internal fixation of femur, 3% (n=252) underwent lower leg amputation, and 2.6% (n=221) underwent above-knee amputation. The majority of patients (n=7370, 87.9%) underwent nonemergent surgery.

XSL•FO
RenderX

Incidence of Death Within 30 Days After Surgery

There were 132 (1.6%) deaths in the orthopedic surgery cohort within 30 days after surgery (Table 2). Of the 132 deaths, 63.6% (84/132) occurred in hospital during the index hospitalization, while the remaining 36.4% (48/132) occurred after discharge. In 1.5% (2/132) of the deaths, patients died in the operating room. The median time to death was 13.5 (IQR 6 - 21) days. Refer to Table 2 for incidence of death according to subcategories of orthopedic surgery. Death rates by nonemergent, urgent, and emergent surgeries were 1.1% (81/7288), 5.7% (46/812), and 3.3% (5/153), respectively. Risk of 30-day mortality was increased in emergent and urgent surgeries compared to nonemergent surgeries in adjusted analyses (HR 2.30, 95% CI 1.61 - 3.42; *P*<.001).

Table . Thirty-day perioperative complications, overall and by subtypes of orthopedic surgery[a].

Outcome	Patients, n (%)						
	All orthopedic surgery (N=8385)	Subtypes of orthopedic surgery					
		Major hip or pelvic surgery (n=2898)	Internal fixation of femur (n=750)	Knee arthroplasty (n=2876)	Above-knee amputation (n=221)	Lower leg amputation (n=252)	Major spine surgery (n=1405)
Major bleeding	2422 (28.9)	965 (33.3)	287 (38.3)	725 (25.2)	103 (46.6)	96 (38.1)	258 (18.4)
MINS[b]	1454 (17.3)	563 (19.4)	165 (22)	347 (12.1)	95 (43)	95 (37.7)	197 (14)
Sepsis	318 (3.8)	124 (4.3)	33 (4.4)	53 (1.8)	30 (13.6)	20 (7.9)	60 (4.3)
Infection without sepsis	562 (6.7)	251 (8.7)	63 (8.4)	153 (5.3)	14 (6.3)	29 (11.5)	54 (3.8)
Acute kidney injury with dialysis	17 (0.2)	4 (0.1)	3 (0.4)	2 (0.1)	2 (0.9)	3 (1.2)	3 (0.2)
Stroke	27 (0.3)	9 (0.3)	6 (0.8)	6 (0.2)	1 (0.5)	2 (0.8)	3 (0.2)
Venous thromboembolism	123 (1.5)	29 (1)	16 (2.1)	69 (2.4)	1 (0.5)	0 (0)	8 (0.6)
Congestive heart failure	124 (1.5)	50 (1.7)	29 (3.9)	23 (0.8)	13 (5.9)	6 (2.4)	4 (0.3)
New, clinically important atrial fibrillation	93 (1.1)	46 (1.6)	13 (1.7)	28 (1)	2 (0.9)	1 (0.4)	4 (0.3)
Death	132 (1.6)	49 (1.7)	29 (3.9)	7 (0.2)	30 (13.6)	9 (3.6)	8 (0.6)

[a]Subtypes of orthopedic surgery were not mutually exclusive (ie, 1 patient could have been categorized to have undergone more than 1 subtype of orthopedic surgery).

[b]MINS: myocardial injury after noncardiac surgery.

Postoperative Complications and Relationship with 30-Day Mortality

The most common complication was major bleeding (2422/8385, 28.9% patients), followed by MINS (1454/8385, 17.3% patients), infection without sepsis (562/8385, 6.7% patients), sepsis (318/8385, 3.8% patients), then VTE (123/8385, 1.5% patients), congestive heart failure (124/8385, 1.5% patients), new atrial fibrillation (93/8385, 1.1% patients), stroke (27/8385, 0.3% patients), and AKI resulting in dialysis (17/8385, 0.2% patients). Refer to Table 2 for details. The median time from surgery to major bleeding was 2 (IQR 1 - 3) days, MINS 2 (IQR 1 - 3) days, infection without sepsis 10 (IQR 5 - 17) days, and sepsis 8 (IQR 4 - 13) days. Tables 3 and 4 demonstrate the proportion of patients that underwent surgery for an acute fracture as well as the mortality rates according to the timing of surgery performed within each subcategory of orthopedic surgery.

XSL•FO
RenderX

Table . Surgery for acute fracture according to subcategories of orthopedic surgery[a].

Subcategory of orthopedic surgery	Surgery for acute fracture[b], n (%)	
	No (n=6767)	Yes (n=1616)
Major hip or pelvic surgery[c]	1982 (29.3)	916 (56.7)
Internal femoral fixation	119 (1.8)	622 (38.5)
Knee arthroplasty	2854 (42.2)	21 (1.3)
Above-knee amputation	216 (3.2)	3 (0.2)
Lower leg amputation	243 (3.6)	6 (0.4)
Major spine surgery	1353 (20)	48 (3)

[a]Surgery for acute fracture was considered nonelective surgery.

[b]2 patients with missing subcategory of orthopedic surgery.

[c]Surgery for hip and pelvis were not mutually exclusive.

Table . Mortality rates according to timing of surgery (emergent or urgent vs nonemergent surgery) within each subcategory of orthopedic surgery.

Subcategory and timing of surgery	Proportions[a], n (%)[b]	Mortality, n	Mortality rate[c], % (95% CI)
Major hip or pelvic surgery (N=2898)			
Emergent or urgent	525 (18.1)	22	4.2 (2.7 - 6.4)
Nonemergent	2373 (81.9)	27	1.13 (0.8 - 1.7)
Internal femoral fixation (N=741)			
Emergent or urgent	313 (42.2)	14	4.5 (2.6 - 7.6)
Nonemergent	428 (57.8)	15	3.5 (2.1 - 5.8)
Knee arthroplasty (N=2875)			
Emergent or urgent	35 (1.2)	0	0
Nonemergent	2840 (98.8)	7	0.24 (0.1 - 0.5)
Above-knee amputation (N=219)			
Emergent or urgent	48 (21.9)	11	23 (12.5 - 37.7)
Nonemergent	171 (78.1)	19	11.2 (7.0 - 17.0)
Lower leg amputation (N=249)			
Emergent or urgent	49 (19.7)	3	6 (1.6 - 17.9)
Nonemergent	200 (80.3)	6	3.0 (1.2 - 6.7)
Major spine surgery (N=1403)			
Emergent or urgent	46 (3.3)	1	2 (0.1 - 13.0)
Nonemergent	1357 (96.7)	7	0.51 (0.2 - 1.1)
All (N=8385)			
Emergent or urgent	1016 (12.1)	51	5.01 (3.8 - 6.6)
Nonemergent	7369 (87.9)	81	1.09 (0.9 - 1.4)

[a]2 patients with missing subcategory of orthopedic surgery

[b]The proportion was calculated using the total number of patients within each surgical subcategory as the denominator.

[c]The proportion is out of total patients receiving urgent or emergent surgery versus nonemergent, within each subcategory of surgery.

Postoperative complications associated with 30-day mortality included major bleeding (adjusted hazard ratio [aHR] 1.95, 95% CI 1.34 - 2.85), MINS (aHR 2.08, 95% CI 1.38 - 3.14), sepsis (aHR 6.24, 95% CI 3.85 - 10.12), infection without sepsis (aHR 2.74, 95% CI 1.54 - 4.85), stroke (aHR 6.01, 95% CI 2.19 - 16.56), and new clinically important atrial fibrillation (aHR 2.65, 95% CI 1.25 - 5.65). Refer to Table 5 for details. The c-statistic for model performance before and after correction for optimism was 0.87 and 0.85, respectively. Figure 2 is a cumulative hazard curve of the postoperative complications associated with 30-day mortality.

XSL•FO
RenderX

Table . Relation between perioperative complications and 30-day mortality in orthopedic surgery.

Outcome	Patients who died, % (n) (95% CI)	Adjusted HR[a,b] (95% CI)	Attributable fraction (%)
Major bleeding (n=2422)	3.1 (75) (2.44 - 3.87)	1.95 (1.34 - 2.85)	16.5
No major bleeding (n=5960)	0.9 (57) (0.73 - 1.24)	Reference	N/A[c]
MINS[d] (n=1454)	4.3 (63) (3.35 - 5.51)	2.08 (1.38 - 3.14)	20.6
No MINS (n=6931)	0.9 (69) (0.78 - 1.26)	Reference	N/A
Sepsis (n=318)	9.4 (30) (6.46 - 13.19)	6.24 (3.85 - 10.12)	9.7
Infection without sepsis (n=562)	3 (17) (1.77 - 4.80)	2.74 (1.54 - 4.85)	3.8
No sepsis or infection (n=7503)	1.1 (85) (0.91 - 1.40)	Reference	N/A
Acute kidney injury with dialysis (n=17)	41.2 (7) (18.44 - 67.07)	N/A	N/A
No acute kidney injury with dialysis (n=8366)	1.5 (125) (1.25 - 1.78)	N/A	N/A
Stroke (n=27)	14.8 (4) (4.19 - 33.73)	6.01 (2.19 - 16.56)	1.5
No stroke (n= 8356)	1.5 (128) (1.28 - 1.82)	Reference	N/A
Venous thromboembolism (n=123)	2.4 (3) (0.51 - 6.96)	2.24 (0.70 - 7.13)	N/A
No venous thromboembolism (n=8261)	1.6 (129) (1.31 - 1.85)	Reference	N/A
Congestive heart failure (n=124)	15.3 (19) (9.48 - 22.89)	1.54 (0.81 - 2.94)	N/A
No congestive heart failure (n=8259)	1.4 (13) (1.13 - 1.64)	Reference	N/A
New, clinically important atrial fibrillation (n=93)	10.8 (10) (5.28 - 18.89)	2.65 (1.25 - 5.65)	2.2
No new, clinically important atrial fibrillation (n=8290)	1.5 (122) (1.22 - 1.75)	Reference	N/A

[a]HR: hazard ratio.

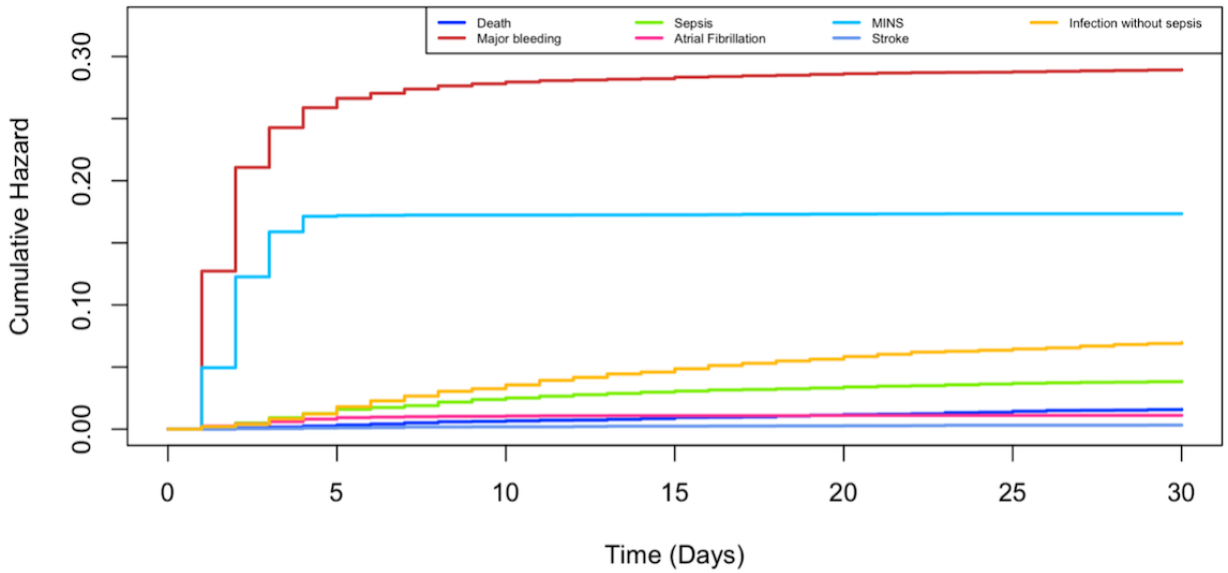
[b]Adjusted variables were as follows: age, history of peripheral vascular disease, history of chronic obstructive pulmonary disease, surgery urgency, and active cancer.

[c]N/A: not applicable.

[d]MINS: myocardial injury after noncardiac surgery.

XSL•FO
RenderX

Figure 2. Cumulative hazard curve of postoperative complications associated with 30-day mortality in orthopedic surgery. MINS: myocardial injury after noncardiac surgery.



Among the postoperative complications significantly associated with mortality, the greatest attributable fraction was for MINS (20.6%), followed by major bleeding (16.5%), then sepsis (9.7%), infection without sepsis (3.8%), new atrial fibrillation (2.2%), and stroke (1.5%).

Post hoc subgroup analyses performed according to surgery urgency demonstrated similar findings. Among those who underwent nonemergent surgeries, there were 81 deaths. Following adjustment, MINS (HR 2.02, 95% CI 1.22 - 3.32), atrial fibrillation (HR 2.64, 95% CI 1.04 - 6.89), major bleeding (HR 2.30, 95% CI 1.41 - 3.76), sepsis (HR 6.27, 95% CI 3.43 - 11.46), and stroke (HR 10.49, 95% CI 3.18 - 34.63) were significantly associated with 30-day mortality. Among those who underwent urgent and emergent surgeries, there were 51 deaths. In this subgroup, with further reduced number of mortality events, infection without sepsis (HR 3.49, 95% CI 1.85 - 6.24), MINS (HR 3.40, 95% CI 1.85 - 6.24), and sepsis (HR 5.65, 95% CI 2.51 - 12.70) were significantly associated with 30-day mortality.

Discussion

Principal Findings

This large international prospective cohort study involving 8385 patients who underwent orthopedic surgery demonstrated a 1.6% (132/8385) mortality rate. Major bleeding, MINS, infection without sepsis, sepsis, stroke, and new clinically important atrial fibrillation were significantly associated with 30-day mortality. Among these, the greatest attributable fraction for death in our cohort was from MINS (1454/8385, 21%), major bleeding (2422/8385, 17%), sepsis (318/8385, 10%), infection without sepsis (562/8385, 4%), stroke (27/8385, 2%), and new atrial fibrillation (93/8385, 2%).

Consistent with prior studies that reported 30-day mortality rates ranging from 0.6% to 8.7%, this work demonstrated a 1.6% (132/8385) mortality rate [2-4]. Our study also demonstrated substantial variations in mortality across subcategories of orthopedic surgery and surgery urgency. Traditional preoperative risk scoring systems do not consider the nuances between surgical procedures that contribute to notable variations in mortality risk [25]. We found substantial mortality differences across surgical procedures within orthopedic surgery, which urges greater consideration of procedure type for preoperative risk stratification and consideration of higher-level postoperative monitoring (eg, admission to intensive cardiac care unit and telemetry).

Among the few studies that investigate complications associated with mortality in orthopedic surgery, pneumonia, acute renal failure, stroke, myocardial infarction, and MINS have been reported [5,26]. Our results are unique in identifying major bleeding to be a risk factor for death in the orthopedic surgery cohort. Previous studies did not select postoperative bleeding as a potential variable to explore, despite 80% of 739 orthopedic surgeons and 50 anesthesiologists in an international survey responding that they were either concerned or very concerned for bleeding in orthopedic surgery populations [27]. Our results demonstrate that major bleeding in orthopedic surgery is common and is associated with mortality, necessitating further investigations in this area to improve patient outcomes.

Although a major bleeding event was the most common complication in our cohort, MINS demonstrated the greatest attributable fraction for death. In other words, although an overall smaller number of patients experienced MINS, a greater number of patients with MINS died within 30 days of surgery compared to the number of patients who experienced major bleeding. This is unique to the orthopedic surgery cohort compared to the overall VISION mortality analyses and general surgery mortality substudy, where major bleeding had the highest attributable fraction for death [11,17]. Certainly, the varying risk profiles, propensity for bleeding, and cardiovascular stress by surgical subspecialties contribute to the differences in

XSL•FO
RenderX

these results. This emphasizes the need for specialty-specific data to inform surgery-specific practice.

Another potential explanation is the unique practice of bleeding prophylaxis that is commonplace in orthopedic surgery [28-31]. Consistent demonstration of perioperative bleeding reduction with tranexamic acid use in orthopedic surgery–specific research has led to the routine use of prophylactic tranexamic acid in orthopedic surgery for many years, particularly in trauma, joint, and spine surgery [28-33]. This is not commonplace in other noncardiac surgical specialties, despite recent evidence to support its safety and efficacy in these contexts [28-30,34-37]. The long-standing practice of routine prophylaxis for perioperative bleeding in orthopedic surgery may have contributed to reduced bleeding severity and lower attributable fraction for death in this cohort compared to the overall and general surgery cohorts. If this were to be true, it is also worth noting that the rate of VTE in this orthopedic cohort was only around 1% (123/8385), which is similar to the incidence found in the noncardiac and general surgery cohorts [11].

Of the postoperative complications investigated, MINS was the second most common complication occurring in 17.3% (1454/8385) of the orthopedic surgery cohort. We also demonstrated a 12% (347/2876) incidence of MINS in the knee arthroplasty subcategory. Previously reported rates of myocardial infarction (MI) after knee arthroplasty have ranged between 0.3% and 2.2% [38,39]. Considering MINS encompasses a broader spectrum of myocardial injury than MI, the larger incidence of MINS was expected, especially as high-sensitivity troponin assays were used for monitoring. Among the few existing studies investigating the incidence of MINS, the results have varied significantly, likely due to limitations in very small sample sizes (ie, 1/82, 1.2% incidence vs 68/160, 42% incidence) [40,41]. Our study included 2876 patients who underwent knee arthroplasty, representing the largest cohort in which the epidemiology of MINS has been studied. As MINS is largely asymptomatic, these findings highlight the importance of identifying MINS for postoperative risk stratification and downstream management, as emerging evidence suggests targeted treatment may reduce future cardiovascular events [12,16,25,42].

Finally, subgroup analyses by surgery urgency demonstrated minor differences in postoperative complications that were found to be associated with mortality between the emergent and nonemergent groups. Specifically, major bleeding was not associated with mortality in the urgent-emergent subgroup. Clinical interpretation of these findings is limited as these are post hoc analyses with reduced events per variable and therefore, model instability. However, these findings suggest differences in mortality determinants highlighting tailored strategies to address perioperative risks according to surgery urgency.

Limitations

There are a few limitations to consider. Only patients aged ≥45 years were included; thus, these findings may not apply to younger patients. Baseline variables (ie, age and estimated glomerular filtration rate) were modeled as a categorical variable to align with prior VISION analyses, although this approach may not fully capture potential nonlinear relationships. Additionally, as our inclusion criteria were targeted to include a higher risk surgical population, the present results would not apply for patients undergoing low-risk, same-day surgeries. The granularity of the data are limited to the subcategories of orthopedic surgery that were predefined before the start of the study [43-45]. Although orthopedic procedures are heterogeneous, combining them enabled sufficient power to evaluate associations between relatively infrequent complications and mortality. We report subtype-specific complication rates (Table 2), demonstrating that major bleeding, MINS, and sepsis were consistently among the most common complications across procedures. While procedure-specific attributable fractions were not feasible due to limited events, our findings identify broadly relevant targets for perioperative risk reduction. The stability of a Cox regression model is dependent on the events per variable. The base model meets the lower acceptable cutoff at 10 events per variable, suggesting adequate power for the primary analysis; there remains potential risk for overfitting [24,46]. However, the negligible change in the c-statistic after correcting for optimism suggests overfitting may not be a significant issue [20]. Variation in VTE prophylaxis practices, which were not captured in detail in this study, may contribute to differences in bleeding risk and represents a potential unmeasured confounder. Finally, complications such as infection and sepsis may occur later in the postoperative course, potentially reducing their relative contribution in time-dependent analyses within a 30-day follow-up period, particularly given their lower event frequency compared to earlier complications such as bleeding or MINS.

To our knowledge, this is the first large prospective study exploring the association of various postoperative complications and death in a global prospective orthopedic surgery study. The inclusion of diverse participants across 14 countries increases the generalizability of our results. Furthermore, with no loss to follow-up and the collection of time-dependent data, we were able to increase precision and reduce bias in fulfilling our objectives by creating a realistic model that accounts for the changes in risk factors for survival that occur over time.

Conclusions

This large international prospective cohort study of patients undergoing orthopedic surgery (7369/8385, 88% nonemergent) demonstrated a 30-day mortality rate of 1.6% (132/8385). Adjusted analyses demonstrated major bleeding, MINS, sepsis, infection without sepsis, stroke, and atrial fibrillation to be associated with mortality. The highest attributable fraction of death in our cohort was contributed by MINS, major bleeding, sepsis, and infection without sepsis, which highlights areas for further study to reduce mortality among patients who underwent orthopedic surgery.

XSL•FO
RenderX

Funding

Please see Multimedia Appendix 1 for funding sources.

Data Availability

The datasets generated or analyzed during this study are not publicly available due but are available from the corresponding author on reasonable request.

Conflicts of Interest

On the basis of study questions PJD has originated and grants he has written, he has received grants from Abbott Diagnostics, AOP Pharma, AstraZeneca, Bayer, Boehringer Ingelheim, Bristol-Myers-Squibb, CloudDX, Covidien, Octapharma, Philips Healthcare, Roche Diagnostics, Siemens, and Stryker. PJD has participated in advisory board meetings for GlaxoSmithKline, Boehringer Ingelheim, Bayer, and Quidel Canada. He attended expert panel meetings with AstraZeneca and Boehringer Ingelheim; served as a consultant for a call with Roche Pharma; and conducted consultancy work for Abbott, AstraZeneca, Renibus Therapeutics, Roche Canada, and Trimedic. He has also been invited as a speaker with Bayer Inc, Novartis Pharma Canada, and Abbott Diagnostics. FKB holds a PHRI Career Award and a McMaster University Department of HEI Mid-Career Research Award. She received investigator-initiated research grants from Roche Diagnostics and SIEMENS, and attended advisory board meetings from Abbott. ED received investigator-initiated research grants from Roche Laboratories and Abbott Laboratories, outside of the submitted work. She receives a clinician-researcher salary award from the Fonds de recherche du Québec-Santé. DC received consulting fees from Roche Diagnostics and Trimedics. PES has received consulting fees from Incyte Biosciences Canada and Hoffman La Roche. ORA reports a relationship with Stryker Canada that includes speaking and lecture fees. He is the tier 2 Canada Research Chair in Joint Preservation and President/Owner of Notch Academy. MKW is supported by the CANTRAIN Clinical Trials Training Platforms – Doctoral Studentship Award. DT has received funding from Stryker Canada.

Multimedia Appendix 1
VISION funding sources.
[DOCX File, 17 KB - periop_v9i1e90823_app1.docx]

Multimedia Appendix 2
VISION postoperative complications and baseline variable definitions.
[DOCX File, 20 KB - periop_v9i1e90823_app2.docx]

Checklist 1
STROBE checklist.
[DOCX File, 22 KB - periop_v9i1e90823_app3.docx]

References

1. Kamel I, Ahmed MF, Sethi A. Regional anesthesia for orthopedic procedures: what orthopedic surgeons need to know. World J Orthop 2022 Jan 18;13(1):11-35. [doi: 10.5312/wjo.v13.i1.11] [Medline: 35096534]
2. Banerjee S, Suresh G, Kale AB, Sathe AH. An audit of admissions and mortality of orthopedic indoor patients in a tertiary care hospital of India. J Clin Orthop Trauma 2020 Jul;11(Suppl 4):S518-S521. [doi: 10.1016/j.jcot.2020.04.012] [Medline: 32774021]
3. Sheikh HQ, Hossain FS, Aqil A, Akinbamijo B, Mushtaq V, Kapoor H. A comprehensive analysis of the causes and predictors of 30-day mortality following hip fracture surgery. Clin Orthop Surg 2017 Mar;9(1):10-18. [doi: 10.4055/cios.2017.9.1.10] [Medline: 28261422]
4. Tan HB, MacDonald DA, Matthews SJ, Giannoudis PV. Incidence and causes of mortality following acute orthopaedic and trauma admissions. Ann R Coll Surg Engl 2004 May;86(3):156-160. [doi: 10.1308/003588404323043256] [Medline: 15140297]
5. Bhattacharyya T, Iorio R, Healy WL. Rate of and risk factors for acute inpatient mortality after orthopaedic surgery. J Bone Joint Surg Am 2002 Apr;84(4):562-572. [doi: 10.2106/00004623-200204000-00009] [Medline: 11940616]
6. Ramirez LF, Thisted R. Using a national health care data base to determine surgical complications in community hospitals: lumbar discectomy as an example. Neurosurgery 1989 Aug;25(2):218-225. [doi: 10.1097/00006123-198908000-00011] [Medline: 2770986]
7. Ciol MA, Deyo RA, Howell E, Kreif S. An assessment of surgery for spinal stenosis: time trends, geographic variations, complications, and reoperations. J Am Geriatr Soc 1996 Mar;44(3):285-290. [doi: 10.1111/j.1532-5415.1996.tb00915.x] [Medline: 8600197]
8. Derkay CS. Pediatric otolaryngology procedures in the United States: 1977-1987. Int J Pediatr Otorhinolaryngol 1993 Jan;25(1-3):1-12. [doi: 10.1016/0165-5876(93)90004-m] [Medline: 8436452]

XSL•FO
RenderX

9. Gillum RF. Peripheral arterial occlusive disease of the extremities in the United States: hospitalization and mortality. Am Heart J 1990 Dec;120(6 Pt 1):1414-1418. [doi: 10.1016/0002-8703(90)90257-x] [Medline: 2248187]
10. Owings MF, Kozak LJ. Ambulatory and inpatient procedures in the United States, 1996. Vital Health Stat 13 1998 Nov;13(139):1-119 [FREE Full text] [Medline: 9866429]
11. Vascular Events in Noncardiac Surgery Patients Cohort Evaluation (VISION) Study Investigators, Spence J, LeManach Y, et al. Association between complications and death within 30 days after noncardiac surgery. CMAJ 2019 Jul 29;191(30):E830-E837. [doi: 10.1503/cmaj.190221] [Medline: 31358597]
12. Thomas S, Borges F, Bhandari M, et al. Association between myocardial injury and cardiovascular outcomes of orthopaedic surgery: a vascular events in noncardiac surgery patients cohort evaluation (VISION) substudy. J Bone Joint Surg Am 2020 May 20;102(10):880-888. [doi: 10.2106/JBJS.18.01305] [Medline: 32118652]
13. Salerno SM, Carlson DW, Soh EK, Lettieri CJ. Impact of perioperative cardiac assessment guidelines on management of orthopedic surgery patients. Am J Med 2007 Feb;120(2):185. [doi: 10.1016/j.amjmed.2005.11.009] [Medline: 17275461]
14. Devereaux PJ, Goldman L, Cook DJ, Gilbert K, Leslie K, Guyatt GH. Perioperative cardiac events in patients undergoing noncardiac surgery: a review of the magnitude of the problem, the pathophysiology of the events and methods to estimate and communicate risk. CMAJ 2005 Sep 13;173(6):627-634. [doi: 10.1503/cmaj.050011] [Medline: 16157727]
15. Vascular Events In Noncardiac Surgery Patients Cohort Evaluation (VISION) Study Investigators, Devereaux PJ, Chan MT, et al. Association between postoperative troponin levels and 30-day mortality among patients undergoing noncardiac surgery. JAMA 2012 Jun 6;307(21):2295-2304. [doi: 10.1001/jama.2012.5502] [Medline: 22706835]
16. Writing Committee for the VISION Study Investigators, Devereaux PJ, Biccard BM, et al. Association of postoperative high-sensitivity troponin levels with myocardial injury and 30-day mortality among patients undergoing noncardiac surgery. JAMA 2017 Apr 25;317(16):1642-1651. [doi: 10.1001/jama.2017.4360] [Medline: 28444280]
17. Park LJ, Borges FK, Ofori S, et al. Association between complications and death within 30 days after general surgery: a vascular event in noncardiac surgery patients cohort evaluation (VISION) substudy. Ann Surg 2025 Dec 1;282(6):1007-1013. [doi: 10.1097/SLA.0000000000006372] [Medline: 38860367]
18. Roshanov PS, Eikelboom JW, Crowther M, et al. Bleeding impacting mortality after noncardiac surgery: a protocol to establish diagnostic criteria, estimate prognostic importance, and develop and validate a prediction guide in an international prospective cohort study. CMAJ Open 2017 Aug 4;5(3):E594-E603. [doi: 10.9778/cmajo.20160106] [Medline: 28943515]
19. Roshanov PS, Eikelboom JW, Sessler DI, et al. Bleeding Independently associated with Mortality after noncardiac Surgery (BIMS): an international prospective cohort study establishing diagnostic criteria and prognostic importance. Br J Anaesth 2021 Jan;126(1):163-171. [doi: 10.1016/j.bja.2020.06.051] [Medline: 32768179]
20. Smith GC, Seaman SR, Wood AM, Royston P, White IR. Correcting for optimistic prediction in small data sets. Am J Epidemiol 2014 Aug 1;180(3):318-324. [doi: 10.1093/aje/kwu140] [Medline: 24966219]
21. Rückinger S, von Kries R, Toschke AM. An illustration of and programs estimating attributable fractions in large scale surveys considering multiple risk factors. BMC Med Res Methodol 2009 Jan 23;9(1):7. [doi: 10.1186/1471-2288-9-7] [Medline: 19166593]
22. von Cube M, Timsit JF, Schumacher M, Motschall E, Schumacher M. Quantification and interpretation of attributable mortality in core clinical infectious disease journals. Lancet Infect Dis 2020 Dec;20(12):e299-e306. [doi: 10.1016/S1473-3099(20)30485-0] [Medline: 32916101]
23. Mansournia MA, Altman DG. Population attributable fraction. BMJ 2018 Feb 22;360:k757. [doi: 10.1136/bmj.k757] [Medline: 29472187]
24. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol 2007 Mar 15;165(6):710-718. [doi: 10.1093/aje/kwk052] [Medline: 17182981]
25. Duceppe E, Parlow J, MacDonald P, et al. Canadian Cardiovascular Society guidelines on perioperative cardiac risk assessment and management for patients who undergo noncardiac surgery. Can J Cardiol 2017 Jan;33(1):17-32. [doi: 10.1016/j.cjca.2016.09.008] [Medline: 27865641]
26. Hu W, Chen Y, Zhao K, et al. Association of perioperative myocardial injury with 30-day and long-term mortality in older adult patients undergoing orthopedic surgery in China. Med Sci Monit 2021 Dec 25;27:e932036. [doi: 10.12659/MSM.932036] [Medline: 34952895]
27. Ginzburg E, Dujardin F. Physicians' perceptions of the definition of major bleeding in major orthopedic surgery: results of an international survey. J Thromb Thrombolysis 2011 Feb;31(2):188-195. [doi: 10.1007/s11239-010-0498-9] [Medline: 20640486]
28. Wong J, George RB, Hanley CM, Saliba C, Yee DA, Jerath A. Tranexamic acid: current use in obstetrics, major orthopedic, and trauma surgery. Can J Anesth 2021 Jun;68(6):894-917. [doi: 10.1007/s12630-021-01967-7] [Medline: 33993459]
29. Jennings JD, Solarz MK, Haydel C. Application of tranexamic acid in trauma and orthopedic surgery. Orthop Clin North Am 2016 Jan;47(1):137-143. [doi: 10.1016/j.ocl.2015.08.014] [Medline: 26614928]
30. Lin ZX, Woolf SK. Safety, efficacy, and cost-effectiveness of tranexamic acid in orthopedic surgery. Orthopedics 2016;39(2):119-130. [doi: 10.3928/01477447-20160301-05] [Medline: 26942474]
31. Colomina MJ, Contreras L, Guilabert P, Koo M, M Ndez E, Sabate A. Clinical use of tranexamic acid: evidences and controversies. Braz J Anesthesiol 2022;72(6):795-812. [doi: 10.1016/j.bjane.2021.08.022] [Medline: 34626756]

XSL•FO
RenderX

32. Gandhi R, Evans HM, Mahomed SR, Mahomed NN. Tranexamic acid and the reduction of blood loss in total knee and hip arthroplasty: a meta-analysis. BMC Res Notes 2013 May 7;6:184. [doi: 10.1186/1756-0500-6-184] [Medline: 23651507]
33. Goldstein K, Jones C, Kay J, Shin J, de Sa D. Tranexamic acid administration in arthroscopic surgery is a safe adjunct to decrease postoperative pain and swelling: a systematic review and meta-analysis. Arthroscopy 2022 Apr;38(4):1366-1377. [doi: 10.1016/j.arthro.2021.10.001] [Medline: 34655767]
34. Montroy J, Hutton B, Fergusson DA, et al. Lysine analogue use during cancer surgery: a survey from a Canadian tertiary care centre. Curr Oncol 2020 Dec;27(6):e560-e568. [doi: 10.3747/co.27.6613] [Medline: 33380871]
35. Devereaux PJ, Marcucci M, Painter TW, et al. Tranexamic acid in patients undergoing noncardiac surgery. N Engl J Med 2022 May 26;386(21):1986-1997. [doi: 10.1056/NEJMoa2201171] [Medline: 35363452]
36. Tsan SE, Viknaswaran NL, Cheong CC, et al. Prophylactic intravenous tranexamic acid and thromboembolism in non-cardiac surgery: a systematic review, meta-analysis and trial sequential analysis. Anaesthesia 2023 Sep;78(9):1153-1161. [doi: 10.1111/anae.16058] [Medline: 37314744]
37. Houston BL, Fergusson DA, Falk J, et al. Prophylactic tranexamic acid use in non-cardiac surgeries at high risk for transfusion. Transfus Med 2021 Aug;31(4):236-242. [doi: 10.1111/tme.12780] [Medline: 33938051]
38. Lu N, Misra D, Neogi T, Choi HK, Zhang Y. Total joint arthroplasty and the risk of myocardial infarction: a general population, propensity score–matched cohort study. Arthritis Rheumatol 2015 Oct;67(10):2771-2779. [doi: 10.1002/art.39246] [Medline: 26331443]
39. Petersen PB, Kehlet H, Jørgensen CC, Lundbeck Foundation Center for Fast-track Hip and Knee Replacement Collaborative Group. Myocardial infarction following fast-track total hip and knee arthroplasty-incidence, time course, and risk factors: a prospective cohort study of 24,862 procedures. Acta Orthop 2018 Dec;89(6):603-609. [doi: 10.1080/17453674.2018.1517487] [Medline: 30326755]
40. Van Zyl RD, Burger MC, Jordaan JD. Prevalence of a postoperative troponin leak in patients with cardiac risk factors undergoing knee and hip arthroplasty in a South African population. S Afr Med J 2020 Mar 30;110(4):320-326. [doi: 10.7196/SAMJ.2020.v110i4.14133] [Medline: 32657745]
41. Lee S, Allen AJ, Morley E, et al. Perioperative myocardial injury risk after elective knee and hip arthroplasty in patients with a high risk of obstructive sleep apnea. Sleep Breath 2021 Mar;25(1):513-515. [doi: 10.1007/s11325-020-02068-3] [Medline: 32253610]
42. Devereaux PJ, Duceppe E, Guyatt G, et al. Dabigatran in patients with myocardial injury after non-cardiac surgery (MANAGE): an international, randomised, placebo-controlled trial. Lancet 2018 Jun 9;391(10137):2325-2334. [doi: 10.1016/S0140-6736(18)30832-8] [Medline: 29900874]
43. Thakker A, Briggs N, Maeda A, Byrne J, Davey JR, Jackson TD. Reducing the rate of post-surgical urinary tract infections in orthopedic patients. BMJ Open Qual 2018;7(2):e000177. [doi: 10.1136/bmjoq-2017-000177] [Medline: 29719874]
44. Al-Mulhim FA, Baragbah MA, Sadat-Ali M, Alomran AS, Azam MQ. Prevalence of surgical site infection in orthopedic surgery: a 5-year analysis. Int Surg 2014;99(3):264-268. [doi: 10.9738/INTSURG-D-13-00251.1] [Medline: 24833150]
45. Seidelman JL, Mantyh CR, Anderson DJ. Surgical site infection prevention: a review. JAMA 2023 Jan 17;329(3):244-252. [doi: 10.1001/jama.2022.24075] [Medline: 36648463]
46. Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. J Clin Epidemiol 2016 Aug;76:175-182. [doi: 10.1016/j.jclinepi.2016.02.031] [Medline: 26964707]

Abbreviations

aHR: adjusted hazard ratio
AKI: acute kidney injury
HR: hazard ratio
MI: myocardial infarction
MINS: myocardial injury after noncardiac surgery
VISION: Vascular Events in Noncardiac Surgery Patients Cohort Evaluation
VTE: venous thromboembolism

XSL•FO
RenderX

Edited by J Shiffermiller; submitted 04.Jan.2026; peer-reviewed by M Pappas, S Cohn; revised version received 28.Apr.2026; accepted 01.May.2026; published 09.Jun.2026.

Please cite as:
Park LJ, Devereaux PJ, Patel A, Tandon V, Heels-Ansdell D, Thabane L, Serrano PE, Chan MTV, Szczeklik W, Srinathan S, Garutti I, Urrutia G, Guerra-Farfan E, Abdel Khalik H, Duceppe E, Ofori S, Marcucci M, Conen D, Wang MK, Spence J, Tushinski D, Bali K, Adili A, Khanna V, Tonelli AC, Mulazzani F, Jiang W, Ayeni OR, Slobegean G, Miclau T, Bhandari M, Borges FK
Association Between Complications and Death Within 30 Days After Orthopedic Surgery: Vascular Events in Noncardiac Surgery Patients Cohort Evaluation (VISION) Substudy
JMIR Perioper Med 2026;9:e90823
URL: https://periop.jmir.org/2026/1/e90823
doi:10.2196/90823

© Lily J Park, P J Devereaux, Ameen Patel, Vikas Tandon, Diane Heels-Ansdell, Lehana Thabane, Pablo E Serrano, Matthew T V Chan, Wojciech Szczeklik, Sadeesh Srinathan, Ignacio Garutti, Gerard Urrutia, Ernesto Guerra-Farfan, Hassaan Abdel Khalik, Emmanuelle Duceppe, Sandra Ofori, Maura Marcucci, David Conen, Michael K Wang, Jessica Spence, Daniel Tushinski, Kamal Bali, Anthony Adili, Vickas Khanna, Ana Claudia Tonelli, Francesca Mulazzani, Wenjun Jiang, Olufemi R Ayeni, Gerard Slobegean, Theodore Miclau, Mohit Bhandari, Flavia K Borges. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 9.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing the Effects of eHealth Literacy and the Area Deprivation Index on Barriers to Electronic Patient Portal Use for Orthopedic Surgery: Cross-Sectional Observational Study

Audrey Lynn Litvak[1*], BA; Nicholas Lin[2*], MD; Kelly Hynes[3*], MD; Jason Strelzow[3*], MD; Jeffrey G Stepan[2*], MD, MSc

[1]Pritzker School of Medicine, University of Chicago, 924 E 57th Street #104, Chicago, IL, United States

[2]Department of Orthopaedic Surgery and Rehabilitation Services, University of Chicago, Chicago, IL, United States

[3]Department of Orthopaedic Surgery, Washington University in St Louis, St Louis, MO, United States

[*]all authors contributed equally

Corresponding Author:
Audrey Lynn Litvak, BA
Pritzker School of Medicine, University of Chicago, 924 E 57th Street #104, Chicago, IL, United States

Abstract

Background: As electronic patient portals (EPPs) continue to gain popularity and systems transition to online tools for scheduling, communication, and telehealth, patients without access or skills to use these tools may be overlooked.

Objective: This study analyzed patient and neighborhood-level factors, including eHealth literacy level and the Area Deprivation Index (ADI), that may limit EPP access for orthopedic surgery.

Methods: A cross-sectional, survey-based study was performed at a single urban tertiary academic medical center in the United States across foot and ankle, hand and upper extremity, and orthopedic trauma subspecialty clinics from June 21, 2022, to August 12, 2022. Survey responses (N=287) provided information on sociodemographic characteristics; barriers to EPP use and frequency of EPP use; the eHealth Literacy Scale; and the ADI, which is an address-generated national census measure of neighborhood-level disadvantage. Barriers to EPP use were inductively coded into barrier types, classified as physical access, technology discomfort, or preference. The primary outcome measure was patient-reported barriers to EPP use, which was treated as a binary outcome (1=barrier; 0=no barrier). Bivariate analyses and multivariable binary logistic regressions were performed.

Results: The percentage of patients who self-reported barriers to EPP access was 43.2% (124/287), which related to physical access (13/124, 10.4%), technology discomfort (55/124, 44.3%), and preference (78/124, 63.0%). In the adjusted regressions, only low eHealth literacy and older age predicted barriers to EPP use (low eHealth literacy, adjusted odds ratio [AOR] 1.32, 95% CI 1.13-1.54; *P*<.001; older age, AOR 1.007, 95% CI 1.003-1.009; *P*<.001), including barriers of technology discomfort (low eHealth literacy, AOR 1.25, 95% CI 1.11-1.40; *P*<.001; older age, AOR 1.004, 95% CI 1.002-1.007; *P*<.001) and preference (low eHealth literacy, AOR 1.33, 95% CI 1.17-1.51; *P*<.001; older age, AOR 1.004, 95% CI 1.00-1.01; *P*<.01). Patients with physical access–related barriers as opposed to technology discomfort or preference barriers had the lowest median eHealth literacy scores (17.0, IQR 12.0-14.0 vs 27.0, IQR 16.0-32.0 vs 27.0, IQR 20.0-32.0, respectively) and roughly a quartile higher median ADI (73.0, IQR 41.0-92.0 vs 53.5, IQR 31.2-76.0 vs 58.0, IQR 38.8-83.8, respectively).

Conclusions: Low eHealth literacy was the most significant determinant of overall barriers to EPP use for orthopedic surgery, followed by older age. Neighborhood-level disadvantage as measured through the ADI had no mediating effect on patient-reported barriers to EPP use when adjusting for eHealth literacy level. While patients with physical access barriers had higher ADIs, overall, few patients reported physical access barriers compared to barriers related to technology discomfort or preference. Patient preference for EPP versus non-EPP communications should be documented. Point-of-care screening using the eHealth Literacy Scale may also identify patients who require follow-up outside of the EPP during critical perioperative periods.

(JMIR Perioper Med 2026;9:e72035) doi:10.2196/72035

KEYWORDS

electronic health records; eHealth literacy; online systems; health equity; social determinants of health; SDOH; orthopedic surgery

XSL•FO
RenderX

Introduction

Improving digital health information transparency and transmission via electronic patient portals (EPPs) has been a central focus of health IT policy in the United States during the last decade [1]. EPPs facilitate patients' communication with their treating care teams and direct access to their electronic personal health record and tools to request prescription refills, participate in e-visits, and complete patient-reported outcome questionnaires, among other functions. These interactions empower patients to take an active role in their health care [1,2]. However, few studies have examined patient portal use in the surgical setting or patient factors that may limit EPP use in this context [3-8].

Benefits of EPP enrollment among orthopedic patients include improved patient outcomes [4,7], medication adherence [3], higher patient satisfaction and psychosocial health [3,7], and fewer missed appointments [3]. Patient engagement via the EPP may additionally facilitate more effective screening for commonly avoidable complications that delay patients' return to function, such as soaking of splints or patient-prolonged immobilization due to unanticipated postoperative pain. Moreover, as routine messaging and completion of patient-reported outcomes via the EPP becomes standard, it is likely that patient engagement via the EPP beyond enrollment may become another critical quality metric tied to physician reimbursement.

Despite advantages of and health care provider interest in adoption of EPP tools, prior studies in orthopedic surgery have shown that patient factors, including older age and lower educational level, may limit EPP enrollment [3,4,6,8], which is analogous to observations in the internal medicine setting [9-12]. The 2020 Health Information National Trends Survey (HINTS) found that the most commonly cited reason for patient nonuse within a large US sample was desire to speak directly with a health care provider (ie, physician, nurse practitioner)—a sentiment shared by 69% of patients [11]. Furthermore, roughly 30% of patients expressed discomfort with the technology [11]. Traditional health literacy refers to the capacity to find, understand, and use health information to inform health-related decisions and actions, whereas eHealth literacy specifically refers to the capacity and skills to seek, assess, and make use of health information via electronic media. In the hospital medicine setting, low eHealth literacy in particular is associated with less awareness, use, and perceived usefulness of EPPs [13].

Lack of examination of granular patient-level use data beyond EPP activation status, explicit barriers to EPP use, and associated patient factors such as health literacy are described as significant limitations and directions for future work in orthopedic surgery [3,4,8]. Additionally, no study across any prior setting has assessed the effect of structural or neighborhood-level determinants on barriers to EPP use, nor have they assessed barriers among patients who are actively enrolled in EPPs. Individual-level determinants may refer to patient demographics or skill sets such as health literacy, whereas neighborhood-level determinants refer to unmeasured social factors conferred by the geographic environment in which a patient lives, often described via census variables related to percentage of unemployment, percentage of individuals with a high school education, and food and housing quality, among others.

For digital health uptake in particular, distinguishing among types and levels of determinants is critical to realizing equity-informed intervention and policy [14,15]. For example, digital literacy is an individual-level factor for which a policy-level solution such as improving broadband connectivity or personal device accessibility may be ineffective in the absence of community-responsive interventions to provide individuals with digital skill training [15]. In particular, it is important to analyze whether neighborhood disadvantage may amplify the impact of low eHealth literacy on barriers to patient portal use, which single-level analyses of eHealth literacy cannot capture.

This study aimed to contribute to the existing body of literature on barriers to EPP use in orthopedic surgery by analyzing how both individual-level and neighborhood-level social determinants, including eHealth literacy and the Area Deprivation Index (ADI), may relate to patient-reported barriers to EPP access and use across foot and ankle, hand and upper extremity, and orthopedic trauma surgery.

Methods

Study Design and Setting

This was a cross-sectional survey–based study conducted via an anonymized paper survey administered at a single urban tertiary academic medical center in the United States between June 21, 2022, and August 12, 2022. The survey was administered in the clinic following each patient visit and consisted of sociodemographic questions, the eHealth Literacy Scale (eHEALS), and 2 questions regarding EPP access and use detailed below. This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) design and reporting guidelines for cross-sectional studies.

Participants

All English-speaking patients aged >18 years presenting for orthopedic surgery evaluation at foot and ankle, trauma, and hand and upper extremity clinics were included and approached consecutively during the aforementioned period. This study was limited to foot and ankle, trauma, and hand and upper extremity surgery clinics where faculty involvement at our institution was feasible. Approached patients were excluded if they had not received a tablet to complete their in-office patient-reported outcome questionnaire per routine standard of care due to external technological or capacity constraints (ie, tablet out of battery or too few working tablets in the clinic on a particular day) unless that patient had already completed the questionnaire via their patient portal. While this practical limitation resulted in potential selection bias, it randomly affected only a small portion of patients (<10) and was remedied to avoid recurrence of the problem for continued recruitment efforts. Patients who could not read or write were included and read aloud the study survey.

XSL•FO
RenderX

Sample Size Calculation

The HINTS 2020 reported that approximately 40% of adults in the United States accessed their patient portal in the previous year, whereas 59% were nonusers [11]. Using a baseline nonuse or potential barrier rate of roughly 40% to 60% and setting an α value of .05, we estimated that a sample of 270 to 290 patients would provide at least 80% power to detect large effect sizes (odds ratio 2.0 - 3.0) in a binary regression model constrained by 10 events per covariate included.

Ethical Considerations

Biological Sciences Division/University of Chicago Institutional Review Board approval (IRB22-0230) was obtained with waivers for written consent and HIPAA (Health Insurance Portability and Accountability Act) authorization. Anonymized survey data were transcribed into a REDCap (Research Electronic Data Capture; Vanderbilt University) database for secure storage [16]. No compensation was provided for participation.

Primary and Secondary Outcomes

The primary outcome measure was patient-reported barriers to EPP access. Barriers to access were derived from the HINTS and secondarily classified as barriers of physical access, discomfort with technology, or patient preferences for nonelectronic provider communication (Multimedia Appendix 1) [11]. These categories were inductively coded by the research team after data collection. This classification is not validated, which we discuss as a limitation. Patients were instructed to mark preference for nonelectronic provider communication as a selection only if they perceived their preferences as barriers to using their portals. Importantly, "I do not have a patient portal account" was listed as an option on the original survey but was analyzed separately. The secondary outcome measure was patient-reported level of EPP use classified into 2 categories: routine use and nonroutine use. Per Maroney et al [17], level of use was characterized as routine if at least monthly use was indicated and as nonroutine if use a few times a year or less frequently was indicated, including those who did not have an EPP (Multimedia Appendix 1). Importantly, the level of EPP use included use for any clinic, not limited only to their orthopedic surgery care.

Variables and Demographics

eHealth literacy was measured via patient responses to the eHEALS tool to determine its association with barriers to EPP use (Multimedia Appendix 2) [18]. This tool has been validated in the orthopedic outpatient setting, among others [18-20]. As in prior literature, a cumulative score of 25 or less indicated low eHealth literacy, and a score of 26 or greater indicated high eHealth literacy [13,21,22]. Neighborhood-level disadvantage was assessed using the ADI, which is calculated via publicly available census data in the domains of income, educational level, employment, and housing quality to assign numeric scores of societal disadvantage to particular geographical regions [23]. Higher scores indicate higher levels of societal disadvantage. Self-reported demographic data were additionally collected.

Statistical Analysis

Survey data were analyzed using the Python statistical program (version 3.10.6; Python Software Foundation) [24-26]. Missing values were excluded pointwise across the relevant analyses given the low frequency. Numerical data presented as medians were reported with the IQR. The level of significance was set at P=.05.

Bivariate analyses were performed to examine the association between both patient-reported barriers and level of EPP use and demographic variables, the ADI, and eHealth literacy level. Three multivariable logistic regressions were conducted wherein barriers were treated as a binary outcome (1=barrier; 0=no barrier). Categorical variables were converted to binary dummy variables for regression. The main regression considered all barriers, whereas the 2 subsequent regressions examined only barriers of technology discomfort or preference-related barriers. Regression was not performed for physical access–related barriers due to outcome size of 13, which is discussed as a limitation. Variable selection for each model was determined via outcome size (at least 10 outcomes per covariate to avoid overfitting) and one-at-a-time sensitivity analyses (via examination of the McFadden pseudo-R^2) in a backward regression approach. To avoid collinearity, only covariates with a variance inflation factor of <5 were included together. Model fit was assessed using McFadden pseudo-R^2 values, with 0.2 considered excellent if not overfit (corroborated via df).

Demographic categories and regression reference levels were selected based on breakdowns and historical controls used in prior related literature [9,27,28]. Income was treated categorically, and per the analogous literature, we selected 3 levels representative of low, medium, and high income based on the median household income cutoffs for the zip code tied to the authors' institution.

Results

Participant Characteristics

A convenience sample of 339 eligible patients was approached, of whom 52 (15.3%) declined participation, leaving 287 (84.7%) for analysis. The median age of the study participants was 48.5 (IQR 35.0-64.2) years; 58.2% (167/287) of the study participants self-identified as non-Hispanic Black individuals, and 26.1% (75/287) identified as non-Hispanic White individuals. The median cumulative eHEALS score was 32 (IQR 27-35), with 21.3% (61/287) of the study participants having low eHealth literacy (eHEALS score of 25 or less). The median ADI was 53.0 (IQR 32.0-74.8).

Of the study participants, 63.1% (181/287) were routine users, and only 9.8% (28/287) did not have EPPs. One or more barriers to accessing their EPPs were reported by 43.2% (124/287) of all patients. Moreover, among the 90.2% (259/287) of patients who were enrolled in the EPP, 42.5% (110/259) still reported barriers to access or use of their portal. The remaining patient characteristics were compared by self-reported barriers to EPP access and self-reported EPP use (Table 1). Patients reporting one or more barriers had higher median age than patients who did not report barriers (57.0, IQR 42.2-71.0 years vs 43.0, IQR

XSL•FO
RenderX

29.2-57.0 years; *P*<.001; Table 1). A higher percentage of non-Hispanic Black patients (*P*=.04), retirees (*P*=.001), and patients who fell into the income bracket of US $30,000 or less (*P*=.03) reported barriers to access (Table 1). Patients who did not routinely use the EPP had a lower educational level than routine users (*P*=.005; Table 1).

Table . Descriptive characteristics and exploratory comparison of self-reported barriers to electronic patient portal (EPP) access and self-reported EPP use by patient characteristics.[a]

Characteristic	Overall	No barriers	One or more barriers	*P* value	Routine use	Nonroutine use	*P* value
Subspecialty clinic, n/N (%)				.07			.77
Hand	187/287 (65.2)	111/187 (59.4)	76/187 (40.6)		116/187 (62.0)	71/187 (38.0)	
Foot and ankle	68/287 (23.7)	40/68 (58.8)	28/68 (41.2)		43/68 (63.2)	25/68 (36.8)	
Trauma	32/287 (11.1)	12/32 (37.5)	20/32 (62.5)		22/32 (68.8)	10/32 (31.3)	
Age (years), median (IQR)	48.5 (35.0-64.2)	43.0 (29.2-57.0)	57.0 (42.2-71.0)	<.001	47.0 (35.0-62.0)	52.0 (37.0-67.0)	.28
Race or ethnicity, n/N (%)				.04			.09
Hispanic or Latino	24/287 (8.4)	16/24 (66.7)	8/24 (33.3)		12/24 (50.0)	12/24 (50.0)	
Non-Hispanic Black	167/287 (58.2)	83/167 (49.7)	84/167 (50.3)		103/167 (61.7)	64/167 (38.3)	
Non-Hispanic White	75/287 (26.1)	51/75 (68.0)	24/75 (32.0)		48/75 (64.0)	27/75 (36.0)	
Other identity or preferred not to answer	21/287 (7.3)	13/21 (61.9)	8/21 (38.0)		18/21 (85.7)	3/21 (14.3)	
Highest educational level attained, n/N (%)				.07			.005
High school or lower	82/284 (28.9)	39/82 (47.6)	43/82 (52.4)		41/82 (50.0)	41/82 (50.0)	
Some college	73/284 (25.7)	40/73 (54.8)	33/73 (45.2)		45/73 (61.6)	28/73 (38.4)	
College or higher	129/284 (45.4)	82/129 (63.6)	47/129 (36.4)		93/129 (72.1)	36/129 (27.9)	
Annual income bracket (US $), n/N (%)				.03			.09
≤30,000	89/260 (34.2)	42/89 (47.2)	47/89 (52.8)		52/89 (58.4)	37/89 (41.6)	
30,001-50,000	48/260 (18.5)	26/48 (54.2)	22/48 (45.8)		30/48 (62.5)	18/48 (37.5)	
>50,000	123/260 (47.3)	80/123 (65.0)	43/123 (35.0)		89/123 (72.4)	34/123 (27.6)	
Current employment status, n/N (%)				.001			.19
Employed	160/285 (56.1)	105/160 (65.6)	55/160 (34.4)		109/160 (68.1)	51/160 (31.9)	
Unemployed or on disability	66/285 (23.2)	35/66 (53.0)	31/66 (47.0)		38/66 (57.6)	28/66 (42.4)	
Retired	59/285 (20.7)	22/59 (37.3)	37/59 (62.7)		34/59 (57.6)	25/59 (42.4)	

[a]Percentages may not add up to 100 due to rounding.

Exploratory Analysis of eHealth and ADI as Potential Barriers to EPP Access and Use

In the analysis of eHealth literacy, patients who reported barriers had a lower median eHEALS score than patients who did not report barriers (29.0, IQR 22.0-32.2 vs 32.0, IQR 30.0-38.0; *P*<.001; Table 2). Conversely, a higher percentage of patients with low eHealth literacy compared to high eHealth literacy reported barriers to using their EPPs (45/61, 74% vs 79/224, 35%, respectively; *P*<.001; Table 2). Patients who reported barriers also had higher median national ADI than patients who did not report barriers (55.5, IQR 37.5-78.2 vs 51.0, IQR 32.0-70.0; *P*=.06), and a higher percentage of patients from the most deprived ADI quartile also reported barriers compared to patients from the least deprived ADI quartiles (most deprived ADI quartile [76-100], 37/70, 53% vs second most deprived ADI quartile [51-75], 35/87, 40% vs second least deprived ADI quartile [26-50], 29/70, 41% vs least deprived ADI quartile [1-25], 19/55, 35%; *P*=.19); however, these results did not reach statistical significance (Table 2).

XSL•FO
RenderX

Table . Comparison of self-reported barriers to electronic patient portal access by eHealth literacy level and the Area Deprivation Index (ADI).[a]

	Overall	No barriers	Any barriers	*P values*
eHEALS[b] score, median (IQR)	32.0 (27.0-35.0)	32.0 (30.0-38.0)	29.0 (22.0-32.2)	<.001
eHealth literacy level (eHEALS score), n/N (%)				<.001
High eHealth literacy	224/285 (78.6)	145/224 (64.7)	79/224 (35.3)	
Low eHealth literacy	61/285 (21.4)	16/61 (26.2)	45/61 (73.8)	
National ADI, median (IQR)	53.0 (32.0-74.8)	51.0 (32.0-70.0)	55.5 (37.5-78.2)	.06
National ADI quartile, n/N (%)				.19
Least deprived ADI quartile (1-25)	55/282 (19.5)	36/55 (65.4)	19/55 (34.5)	
Second least deprived ADI quartile (26-50)	70/282 (24.8)	41/70 (58.6)	29/70 (41.4)	
Second most deprived ADI quartile (51-75)	87/282 (30.9)	52/87 (59.8)	35/87 (40.2)	
Most deprived ADI quartile (76-100)	70/282 (24.8)	33/70 (47.1)	37/70 (52.9)	

[a]Percentages may not add up to 100 due to rounding and missing values.

[b]eHEALS: eHealth Literacy Scale.

Most of the patient-reported barriers were related to technology discomfort (55/124, 44.4%) or preference (78/124, 62.9%) rather than physical access (13/124, 10.5%). However, the group reporting physical access barriers had the lowest levels of eHealth literacy (median eHEALS score 17.0 vs 27.0 vs 27.0, respectively) and highest ADI (median 73.0 vs 53.5 vs 58.0, respectively) compared to groups reporting barriers related to technology discomfort or preference (Table 3). Patients from the most deprived ADI quartile had higher percentages of barriers (Table 3). Bivariate analysis was not performed as barrier type was nonexclusive.

Table . Exploratory analysis of barrier type by eHealth literacy level and the Area Deprivation Index (ADI).[a]

	No barriers	Physical access barriers	Technology discomfort barriers	Preference barriers
eHEALS[b] score, median (IQR)	32.0 (30.0-38.0)	17.0 (12.0-24.0)	27.0 (16.0-32.0)	27.0 (20.0-32.0)
eHealth literacy level (eHEALS), n/N (%)				
High eHealth literacy	145/224 (64.7)	2/224 (0.9)	30/224 (13.4)	44/224 (19.6)
Low eHealth literacy	16/61 (26.2)	11/61 (18.0)	25/61 (41.0)	34/61 (55.7)
National ADI, median (IQR)	51.0 (32.0-70.0)	73.0 (41.0-92.0)	53.5 (31.2-76.0)	58.0 (38.8-83.8)
National ADI quartile, n/N (%)				
Least deprived ADI quartile (1-25)	36/55 (65.5)	0/55 (0.0)	10/55 (18.2)	10/55 (18.2)
Second least deprived ADI quartile (26-50)	41/70 (58.6)	4/70 (5.7)	13/70 (18.6)	19/70 (27.1)
Second most deprived ADI quartile (51-75)	52/87 (59.8)	3/87 (3.4)	17/87 (19.5)	15/87 (17.2)
Most deprived ADI quartile (76-100)	33/70 (47.1)	6/70 (8.6)	14/70 (20.0)	30/70 (42.9)

[a]Percentages may not add up to 100 due to rounding and missing values.

[b]eHEALS: eHealth Literacy Scale.

Regression Analysis of Barriers to EPP Use

In the overall regression including demographic variables, the ADI, and eHealth literacy level, only low eHealth literacy level (adjusted odds ratio [AOR] 1.32, 95% CI 1.13-1.54; *P*<.001) and older age (AOR 1.007, 95% CI 1.003-1.009; *P*<.001) predicted barriers to EPP access (Table 4). Similarly, only low eHealth literacy level and age were associated with predicting

XSL•FO
RenderX

a technology discomfort–related barrier (low eHealth literacy, AOR 1.25, 95% CI 1.11-1.40; *P*<.001; age, AOR 1.004, 95% CI 1.002-1.007; *P*<.001; Table 4) or a preference-related barrier (low eHealth literacy, AOR 1.33, 95% CI 1.17-1.51; *P*<.001; age, AOR 1.004, 95% CI 1.00-1.01; *P*=.01; Table 4). The ADI was not associated with predicting overall barriers (*P*=.59), preference-related barriers (*P*=.35), or technology discomfort–related barriers (*P*=.76; Table 4).

Table . Regression of patient characteristics associated with self-reporting at least one barrier to electronic patient portal use (any barrier type, preference barrier, or technology discomfort barrier).

Characteristic	Regression 1: any barrier type[a]		Regression 2: preference barriers[b]		Regression 3: technology discomfort barriers[c]	
	AOR[d] (95% CI)	*P* value	AOR (95% CI)	*P* value	AOR (95% CI)	*P* value
Age	1.007 (1.003-1.009)	<.001	1.004 (1.00-1.01)	.01	1.004 (1.002-1.007)	<.001
Race or ethnicity						
Hispanic or Latino	0.99 (0.78-1.25)	.92	1.07 (0.88-1.31)	.49	1.04 (0.86-1.24)	.71
Non-Hispanic Black	1.09 (0.94-1.27)	.26	1.09 (0.95-1.24)	.21	1.04 (0.92-1.17)	.52
Non-Hispanic White	Reference	—[e]	Reference	—	Reference	—
Other identity or preferred not to answer	1.02 (0.79-1.32)	.88	1.05 (0.85-1.30)	.63	0.95 (0.78-1.15)	.56
National ADI[f]	1.00 (1.00-1.00)	.59	1.00 (1.00-1.00)	.35	1.00 (1.00-1.00)	.76
eHealth literacy level (eHEALS[g])						
High eHealth literacy	Reference	—	Reference	—	Reference	—
Low eHealth literacy	1.32 (1.13-1.54)	<.001	1.33 (1.17-1.51)	<.001	1.25 (1.11-1.40)	<.001
Subspecialty clinic						
Hand	Reference	—	Reference	—	—	—
Foot and ankle	1.01 (0.88-1.16)	.91	1.02 (0.91-1.15)	.73	—	—
Trauma	1.18 (0.99-1.41)	.07	1.17 (1.00-1.37)	.05	—	—
Highest educational level attained						
High school or lower	1.06 (0.90-1.25)	.65	—	—	—	—
Some college	0.96 (0.82-1.13)	.51	—	—	—	—
College or higher	Reference	—	—	—	—	—
Annual income bracket (US $)						
≤30,000	1.11 (0.94-1.30)	.22	—	—	—	—
30,001-50,000	1.05 (0.89-1.24)	.59	—	—	—	—
>50,000	Reference	—	—	—	—	—

[a]Regression model 1: 88.5% (254/287) of the patients were included after participants with missing values were excluded (outcome size: 124/254, 48.8% reported any barrier; df=12; pseudo-R^2=0.21).

[b]Regression model 2: 96.5% (277/287) of the patients were included after participants with missing values were excluded (outcome size: 78/277, 28.2% reported preference barriers; df=8; pseudo-R^2=0.16).

[c]Regression model 3: 96.5% (277/287) of the patients were included after participants with missing values were excluded (outcome size: 55/277, 19.9% reported technology discomfort barriers; df=6; pseudo-R^2=0.13).

[d]AOR: adjusted odds ratio.

[e]Not applicable.

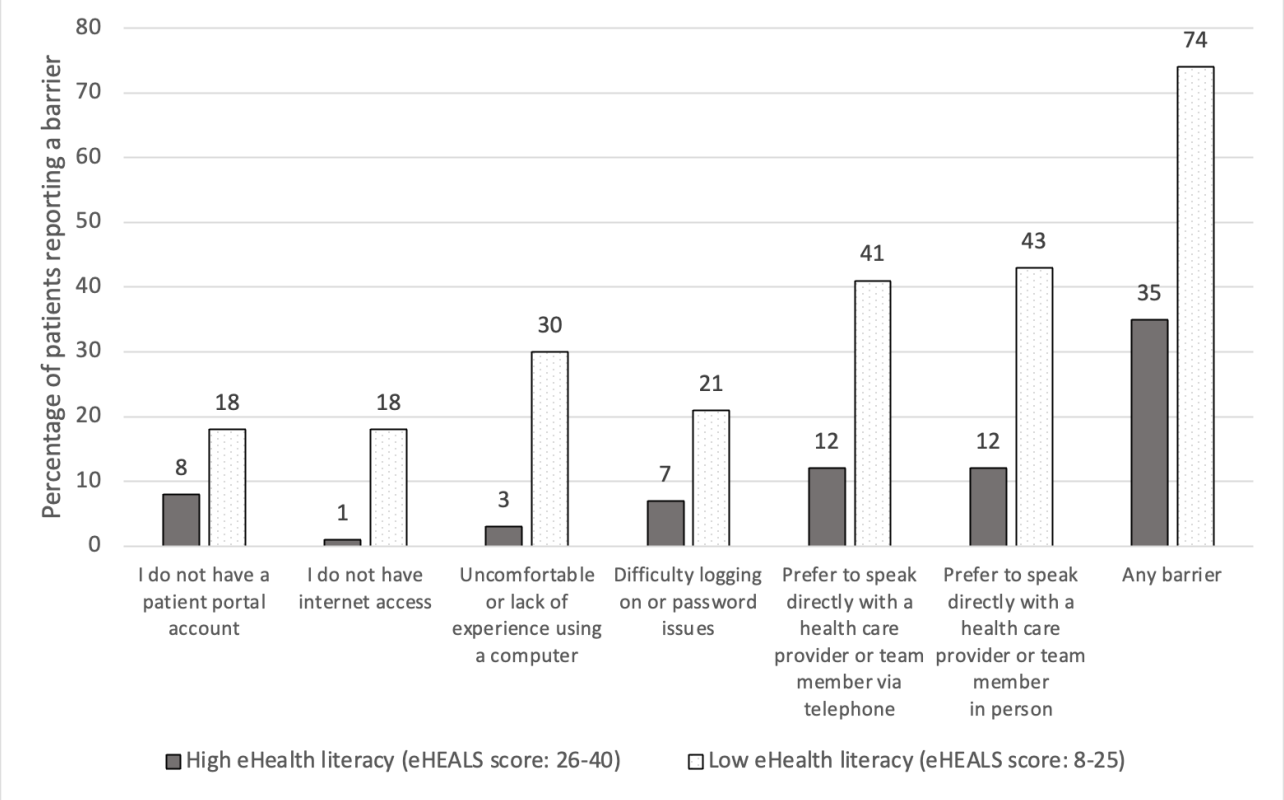
[f]ADI: Area Deprivation Index.

[g]eHEALS: eHealth Literacy Scale.

XSL•FO
RenderX

The most prevalent patient-reported barrier to EPP access was a preference to speak directly to the health care provider or team member in person or via telephone (52/287, 18.1%). A lack of comfort with a computer was cited as a barrier to EPP access by 29.5% (18/61) of patients with low eHealth literacy (Figure 1). Additionally, 17.8% (51/287) of study participants reported poor or fair ability to use a computer, tablet, or smartphone to find information that they needed on the internet. One or more barriers to EPP use were indicated by 80.4% (41/51) of patients with poor or fair self-ratings, compared to 35.3% (83/235) of patients with good, very good, or excellent self-ratings (P<.001).

Figure 1. Frequency and percentages of patient self-reported barriers to electronic patient portal (EPP) access stratified by eHealth literacy level. The percentages of patients who reported a particular barrier are reported over the denominator of patients with either high eHealth literacy scores (n=224) or low eHealth literacy scores (n=61). Percentages do not necessarily add up to 100 across barrier categories as patients could indicate more than one barrier. The barriers "I do not need a patient portal account," "I have multiple patient portals," and "I have privacy concerns" are not depicted separately due to lower frequency of observations.



Discussion

Principal Findings

Equitable implementation of digitized health tools relies on efforts from clinicians, researchers, and policymakers alike. This study assessed patient-reported barriers to use within an expanded framework of individual- and neighborhood-level factors. Low eHealth literacy level was the most significant determinant of overall barriers to EPP use for orthopedic surgery, followed by older age, as compared to other demographic factors and measures of neighborhood-level disadvantage. Contrary to expectation, neighborhood-level disadvantage as measured via the ADI had no mediating effect on barriers after adjusting for eHealth literacy level. Patients with physical access barriers did have appreciably higher ADIs; however, few patients reported physical access barriers overall. These findings build on prior work that showed that older age, among other patient demographics, was associated with reduced EPP enrollment in orthopedic surgery [3,4,8]. This also builds on prior work in the hospital medicine setting that showed that lower eHealth literacy correlated with decreased awareness and use and less favorable attitudes toward use of EPPs [13].

Barriers of physical access, such as lack of internet access, were infrequent compared to barriers related to preference or discomfort with technology, including lack of experience using a computer or difficulty logging in. This finding contrasts with previous work that found internet access to be a significant determinant of portal use for orthopedic surgery [6]. Additionally, among the 17.8% (51/287) of the participants with lower self-ratings of ability to use a computer, tablet, or smartphone to find information that they needed on the internet, 80.4% (41/51) indicated barriers to accessing their EPPs. The final rule of the 21st Century Cures Act alleviated many barriers related to physical access (eg, computer access or broadband coverage) by requiring an interoperability standard that any EPP programming interface be compatible with smartphone apps [29]. However, this legislation does not address barriers experienced by patients who may technically have the physical and digital tools to access their EPPs but not the self-efficacy or the skill sets to effectively use these tools.

These findings may support a second digital divide being dependent on disparities in skill sets rather than physical access [13,30]. With regard to EPPs, this is clinically significant as most institutions introduce EPPs via email with time-sensitive

XSL•FO
RenderX

links and sign-up instructions. This may be ineffective at best in promoting EPP adoption or postenrollment use in populations who may have internet access yet lack the internet experience and skills to navigate setting up a digital account with time-pressure activation codes for protected health information. Moreover, the finding that over a quarter of all patients (52/287, 18.1%) perceived their preference to speak with a health care provider in person or via telephone as an actual barrier to using the EPP suggests that patients may view the EPP as a substitute for health care provider communication missing the personal element rather than as an adjunct to improve communication and transparency, as it was intended. This finding also suggests that it may be important to document patient preference for EPP versus non-EPP communication even if a patient does have an EPP as patients with activated EPPs may not be active users.

Importantly, while most patients were enrolled in the EPP, a significant portion of enrolled patients reported barriers to access and use of their portal. This is significant as prior studies of EPP use for orthopedic surgery have either only assessed EPP activation status as a surrogate for use and without assessing barriers to use [3,4,8] or qualitatively analyzed optional comments regarding nonuse in a small fraction of the study cohort (38 of 150 patients) [6]. These findings are clinically relevant as prior efforts to reduce disparities in EPP use have also focused on enrollment [9]. However, simply enrolling patients does little to address the underlying barriers of certain patients to using their EPPs. Enrollment will not address individual-level factors such as eHealth literacy, which may constitute a larger underlying barrier to sustained EPP use after enrollment (eg, patients' technological capabilities and skills rather than physical access to the technology itself).

Additionally, this study substantiates that older adults are a vulnerable population that may be left behind in a digitized health system. This is particularly critical to perioperative care in orthopedic surgery. Older patients may have more complex discharge needs, including perioperative medication changes and rehabilitation requiring close postoperative communications, and addressing them within the EPP may be ineffective. Older patients have intersecting factors that impede their access. While age and eHealth literacy were noncollinear in this study, age has well-known associations with traditional health and eHealth literacy [31]. Addressing deficiencies in these skill sets may improve lower levels of self-efficacy to adopt and use EPPs previously reported in older individuals [32]. This group may benefit from a proactive staff-level intervention that supports an in-person EPP enrollment option followed by an initial lesson on how to use the EPP, which Bhashyam et al [33] previously showed may be beneficial to improving postoperative follow-up. Notably, while older patients may have an elevated sense of caution in using online platforms due to counseling from groups such as the American Association of Retired Persons, patients infrequently noted privacy concerns as a barrier to EPP use in this study.

Patients with physical access barriers also had appreciably higher ADIs despite this analysis not meeting statistical significance. These patients may also benefit from routine touchpoints with staff outside of the EPP. Ensuring effective follow-up is especially important in these patients as, in addition to worse ADIs predicting worse comorbid chronic disease outcomes, simply living in a disadvantaged neighborhood confers similar readmission risk as having a chronic lung disease and higher risk than having a chronic condition such as diabetes [34]. At an informatics design level, a widget within the electronic medical record could be implemented to automatically yield an address-generated ADI analogously to how BMI is automatically calculated based on a patient's weight and height as this may generate similarly important contextual information to a patient's overall health, especially in a perioperative context.

Limitations

First, this was a single-institution study conducted across several orthopedic surgery subspecialty clinics not including adult reconstruction. Hence, the results may not be generalizable across other settings. Moreover, while eHEALS is a commonly used, validated screener for eHealth literacy in outpatient settings [18-20], it has not been validated in our specific patient population. Additionally, self-reporting via a survey may be limited by response bias; however, we felt that this method of examination was critical to include the patient perspective. Importantly, this study did not include non–English-speaking patients, who may experience additional barriers to EPP access; however, this study did include the perspectives of patients with limited reading and writing skills.

Additionally, level of use was dichotomized as routine and nonroutine, similar to the study by Maroney et al [17], without an option for "as needed," which assumes regular use. The National Cancer Institute's 2020 HINTS showed that only 40% of those with EPPs used it every year: 65% felt that they did not need to use it every year [11]. An additional checkbox option for "use as needed" may better capture this nuance; however, this could introduce indeterminate subjectivity.

Notably, the sample size may have been underpowered to detect small effect sizes in demographic differences and in novel outcomes such as the ADI. The sample size calculation was predicated on the 2020 HINTS, which analyzed barriers to EPP enrollment and did not include patients with activated EPPs who still experienced barriers (259/287, 90.2% of our study population) [11]. Moreover, no relevant existing literature has assessed the ADI. Finally, the secondary barrier categories were inductively coded by our research team after data collection based on natural groupings in which we were interested. This rendered our initial sample size calculation insufficient to perform a secondary regression analysis for the access-related barrier category, which had a small outcome size. Additionally, this classification was not validated, which may introduce bias but, importantly, allowed for discovery of new insights that may not have been generated through precoding.

Conclusions

Routine use of EPPs for online scheduling, patient communication, and telehealth continues to be a critical aspect of care. It is necessary to understand existing disparities in barriers to EPP access to not only improve access to care for all patients but also to continue building patients' toolbox and self-efficacy to take on active roles in their care. Future research

XSL•FO
RenderX

should establish whether interventions, education, and improved eHealth literacy may overcome these barriers.

Acknowledgments

The authors would like to thank Jennifer Wolf, MD, and Megan Conti Mica, MD, for helping facilitate data collection within their outpatient clinics. The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

Funding

The authors did not receive support from any organization for the submitted work. No funding was received to assist with the preparation of this manuscript. No funding was received for conducting this study.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to institutional review board constraints but are available from the corresponding author on reasonable request.

Authors' Contributions

ALL, NL, KH, JS, and JGS contributed equally to the design, data collection, and writing of the manuscript.

Each author certifies that there are no funding or commercial associations (consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article related to the author or any immediate family members.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of survey questions regarding patient portal.
[DOCX File, 15 KB - periop_v9i1e72035_app1.docx]

Multimedia Appendix 2
Summary of eHealth Literacy Scale items.
[DOCX File, 15 KB - periop_v9i1e72035_app2.docx]

References

1. Dendere R, Slade C, Burton-Jones A, Sullivan C, Staib A, Janda M. Patient portals facilitating engagement with inpatient electronic medical records: a systematic review. J Med Internet Res 2019 Apr 11;21(4):e12779. [doi: 10.2196/12779] [Medline: 30973347]
2. Davis Giardina T, Menon S, Parrish DE, Sittig DF, Singh H. Patient access to medical records and healthcare outcomes: a systematic review. J Am Med Inform Assoc 2014;21(4):737-741. [doi: 10.1136/amiajnl-2013-002239] [Medline: 24154835]
3. Varady NH, d'Amonville S, Chen AF. Electronic patient portal use in orthopaedic surgery is associated with disparities, improved satisfaction, and lower no-show rates. J Bone Joint Surg Am 2020 Aug 5;102(15):1336-1343. [doi: 10.2106/JBJS.19.01080] [Medline: 32769600]
4. Vallurupalli N, Lawrence KW, Habibi AA, Bosco JA, Lajam CM. Socioeconomic disparities in online patient portal utilization among total knee arthroplasty recipients. J Arthroplasty 2024 Sep;39(9S2):S95-S99. [doi: 10.1016/j.arth.2024.04.040] [Medline: 38670173]
5. Plate JF, Ryan SP, Bergen MA, Hong CS, Attarian DE, Seyler TM. Utilization of an electronic patient portal following total joint arthroplasty does not decrease readmissions. J Arthroplasty 2019 Feb;34(2):211-214. [doi: 10.1016/j.arth.2018.11.002] [Medline: 30497899]
6. Kolodychuk NL, Wong M, Chimento G, Adams T, Gastanaduy M, Waddell BS. Web-based patient portal access in an orthopedic adult reconstruction patient population. Arthroplast Today 2019 Feb 22;5(1):83-87. [doi: 10.1016/j.artd.2019.01.004] [Medline: 31020029]
7. Fatehi A, Gonzalez A, Ring DC, Queralt M. Psychosocial factors are associated with electronic portal registration. Clin Orthop Relat Res 2020 Sep;478(9):2079-2084. [doi: 10.1097/CORR.0000000000001278] [Medline: 32332246]
8. Bell K, Warnick E, Nicholson K, et al. Patient adoption and utilization of a web-based and mobile-based portal for collecting outcomes after elective orthopedic surgery. Am J Med Qual 2018;33(6):649-656. [doi: 10.1177/1062860618765083] [Medline: 29562769]

XSL•FO
RenderX

9. Richwine C, Johnson C, Patel V. Disparities in patient portal access and the role of providers in encouraging access and use. J Am Med Inform Assoc 2023 Jan 18;30(2):308-317. [doi: 10.1093/jamia/ocac227] [Medline: 36451262]
10. Anthony DL, Campos-Castillo C, Lim PS. Who isn't using patient portals and why? Evidence and implications from a national sample of US adults. Health Aff (Millwood) 2018 Dec;37(12):1948-1954. [doi: 10.1377/hlthaff.2018.05117] [Medline: 30633673]
11. Johnson C, Richwine C, Patel V. Individuals' access and use of patient portals and smartphone health apps, 2020. In: ASTP Health IT Data Brief [Internet]: Office of the Assistant Secretary for Technology Policy; 2021.
12. Nishii A, Campos-Castillo C, Anthony D. Disparities in patient portal access by US adults before and during the COVID-19 pandemic. JAMIA Open 2022 Dec;5(4):ooac104. [doi: 10.1093/jamiaopen/ooac104] [Medline: 36540762]
13. Deshpande N, Arora VM, Vollbrecht H, Meltzer DO, Press V. eHealth literacy and patient portal use and attitudes: cross-sectional observational study. JMIR Hum Factors 2023 Jan 27;10:e40105. [doi: 10.2196/40105] [Medline: 36705947]
14. Lyles CR, Nelson EC, Frampton S, Dykes PC, Cemballi AG, Sarkar U. Using electronic health record portals to improve patient engagement: research priorities and best practices. Ann Intern Med 2020 Jun 2;172(11 Suppl):S123-S129. [doi: 10.7326/M19-0876] [Medline: 32479176]
15. Lyles CR, Nguyen OK, Khoong EC, Aguilera A, Sarkar U. Multilevel determinants of digital health equity: a literature synthesis to advance the field. Annu Rev Public Health 2023 Apr 3;44:383-405. [doi: 10.1146/annurev-publhealth-071521-023913] [Medline: 36525960]
16. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
17. Maroney K, Curtis LM, Opsasnick L, et al. eHealth literacy and web-based patient portal usage among kidney and liver transplant recipients. Clin Transplant 2021 Feb;35(2):e14184. [doi: 10.1111/ctr.14184] [Medline: 33278846]
18. Norman CD, Skinner HA. eHEALS: the eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27. [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
19. Lee J, Lee EH, Chae D. eHealth literacy instruments: systematic review of measurement properties. J Med Internet Res 2021 Nov 15;23(11):e30644. [doi: 10.2196/30644] [Medline: 34779781]
20. Cherid C, Baghdadli A, Wall M, et al. Current level of technology use, health and eHealth literacy in older Canadians with a recent fracture-a survey in orthopedic clinics. Osteoporos Int 2020 Jul;31(7):1333-1340. [doi: 10.1007/s00198-020-05359-3] [Medline: 32112118]
21. Richtering SS, Hyun K, Neubeck L, et al. eHealth literacy: predictors in a population with moderate-to-high cardiovascular risk. JMIR Hum Factors 2017 Jan 27;4(1):e4. [doi: 10.2196/humanfactors.6217] [Medline: 28130203]
22. Vollbrecht H, Arora VM, Otero S, Carey KA, Meltzer DO, Press VG. Measuring eHealth literacy in urban hospitalized patients: implications for the post-COVID world. J Gen Intern Med 2021 Jan;36(1):251-253. [doi: 10.1007/s11606-020-06309-9] [Medline: 33078294]
23. Kind AJ, Buckingham WR. Making neighborhood-disadvantage metrics accessible - the Neighborhood Atlas. N Engl J Med 2018 Jun 28;378(26):2456-2458. [doi: 10.1056/NEJMp1802313] [Medline: 29949490]
24. The Python language reference. Python. URL: https://docs.python.org/3/reference/ [accessed 2022-08-31]
25. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020 Mar;17(3):261-272. [doi: 10.1038/s41592-019-0686-2] [Medline: 32015543]
26. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python. Presented at: Proceedings of the 9th Python in Science Conference; Jun 28 to Jul 3, 2010. [doi: 10.25080/Majora-92bf1922-011]
27. El-Toukhy S, Méndez A, Collins S, Pérez-Stable EJ. Barriers to patient portal access and use: evidence from the health information national trends survey. J Am Board Fam Med 2020;33(6):953-968. [doi: 10.3122/jabfm.2020.06.190402] [Medline: 33219074]
28. Lockwood MB, Dunn-Lopez K, Pauls H, Burke L, Shah SD, Saunders MA. If you build it, they may not come: modifiable barriers to patient portal use among pre- and post-kidney transplant patients. JAMIA Open 2018 Jul 10;1(2):255-264. [doi: 10.1093/jamiaopen/ooy024] [Medline: 31984337]
29. ONC's Cures Act Final Rule. HealthIT.gov. URL: https://www.healthit.gov/topic/oncs-cures-act-final-rule [accessed 2025-12-24]
30. van Deursen A, van Dijk J. Internet skills and the digital divide. New Media Soc 2010;13(6):893-911. [doi: 10.1177/1461444810386774]
31. Chesser AK, Keene Woods N, Smothers K, Rogers N. Health literacy and older adults: a systematic review. Gerontol Geriatr Med 2016 Mar 15;2:2333721416630492. [doi: 10.1177/2333721416630492] [Medline: 28138488]
32. Price-Haywood EG, Harden-Barrios J, Ulep R, Luo Q. eHealth literacy: patient engagement in identifying strategies to encourage use of patient portals among older adults. Popul Health Manag 2017 Dec;20(6):486-494. [doi: 10.1089/pop.2016.0164] [Medline: 28384076]
33. Bhashyam AR, Bansal M, McGovern MM, van der Vliet QMJ, Heng M. In-hospital enrollment into an electronic patient portal results in improved follow-up after orthopedic surgery: cluster randomized controlled trial. JMIR Perioper Med 2022 Aug 11;5(1):e37148. [doi: 10.2196/37148] [Medline: 35969452]

XSL•FO
RenderX

34. Kind AJ, Jencks S, Brock J, et al. Neighborhood socioeconomic disadvantage and 30-day rehospitalization: a retrospective cohort study. Ann Intern Med 2014 Dec 2;161(11):765-774. [doi: 10.7326/M13-2946] [Medline: 25437404]

Abbreviations

ADI: Area Deprivation Index
AOR: adjusted odds ratio
eHEALS: eHealth Literacy Scale
EPP: electronic patient portal
HINTS: Health Information National Trends Survey
HIPAA: Health Insurance Portability and Accountability Act
REDCap: Research Electronic Data Capture
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by N Rohatgi; submitted 03.Feb.2025; peer-reviewed by A Tsehay, M Rincon, X Ding; revised version received 06.Dec.2025; accepted 09.Dec.2025; published 07.Jan.2026.

Please cite as:
Litvak AL, Lin N, Hynes K, Strelzow J, Stepan JG
Assessing the Effects of eHealth Literacy and the Area Deprivation Index on Barriers to Electronic Patient Portal Use for Orthopedic Surgery: Cross-Sectional Observational Study
JMIR Perioper Med 2026;9:e72035
URL: https://periop.jmir.org/2026/1/e72035
doi:10.2196/72035

© Audrey Lynn Litvak, Nicholas Lin, Kelly Hynes, Jason Strelzow, Jeffrey G Stepan. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 7.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Automated Identification of Surgical Site Infections From Electronic Medical Records: Retrospective Observational Predictive Modeling Study

Arjun Chakraborty[1], PhD; Peter Tarczy-Hornoch[1,2,3], MD; Dustin Long[4], MD; Meliha Yetisgen[1], PhD

[1]Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington, 222 15th Street SW, Rochester, MN, United States

[2]Department of Pediatrics, School of Medicine, University of Washington, Seattle, WA, United States

[3]Department of Computer Science and Engineering, Paul G. Allen School of Computer Science and Engineering, College of Engineering, University of Washington, Seattle, WA, United States

[4]Department of Anesthesiology and Pain Medicine, School of Medicine, University of Washington, Seattle, WA, United States

Corresponding Author:
Arjun Chakraborty, PhD
Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington, 222 15th Street SW, Rochester, MN, United States

Abstract

Background: Surgical site infections (SSIs) affect 160,000 to 300,000 patients annually, increasing postoperative mortality, causing significant complications, and incurring US $3.5 to US $10 billion in excess costs each year. Effective SSI surveillance can inform strategies to mitigate these outcomes. Traditional SSI surveillance methods, primarily manual chart reviews, are costly and labor-intensive.

Objective: This study aimed to evaluate whether an automated SSI surveillance system built using newer natural language processing methods and deep learning could outperform previous approaches and whether such an approach could enable more efficient infection surveillance.

Methods: Our dataset comprised approximately 30,000 surgical cases from the University of Washington Medical Center (UWMC) and Harborview Medical Center (HMC). Data from UWMC were captured for the National Surgical Quality Improvement Program, and data from HMC were captured for the National Healthcare Safety Network. Electronic health record (EHR) data for each surgical case included structured data pertaining to surgical procedure characteristics, laboratory values, and antibiotic administration, as well as clinical text notes for a surgical case from 7 days before to 90 days after surgery. Using this data, we applied a myriad of machine learning approaches to the task of SSI prediction. We reported the following performance metrics: F_1-score, precision (positive predictive value), recall (sensitivity), area under the precision-recall curve, and precision at 0.9 recall for each machine learning approach.

Results: In a cohort of 5996 surgical cases, incorporating multimodal EHR information—including contextual information from clinical text and temporal information from laboratory values—improved SSI prediction performance. Models using structured data and clinical text outperformed structured data alone (F_1=0.68, 95% CI 0.68 - 0.69 vs F_1=0.55, 95% CI 0.54 - 0.56; P<.001). Adding temporal features further improved performance (F_1=0.70, 95% CI 0.69 - 0.71; P<.001). Deep learning approaches leveraging large language models also outperformed the state-of-the-art rule-based system (F_1=0.70, 95% CI 0.69 - 0.71 vs F_1=0.43; P<.001). The optimal approach combined foundation models for text summarization with deep learning methods for clinical text and temporal data processing. This system achieved a precision of 0.38 at 0.9 recall, demonstrating its potential for efficient, data-driven SSI surveillance.

Conclusions: Automated surveillance approaches—particularly deep learning approaches—in combination with voluminous, multimodal data from the EHR, can enable more efficient infection surveillance processes. This has the potential to increase the quantity of SSI surveillance data available to guide interventions aimed at reducing SSI rates.

(JMIR Perioper Med 2026;9:e87896) doi:10.2196/87896

KEYWORDS

electronic health record; machine learning; natural language processing; surgical wound infection; epidemiology; diagnosis; neural networks computer; deep learning; large language models

XSL•FO
RenderX

Introduction

Background

Surgical site infections (SSIs) are postsurgical complications that increase the risk of death and other secondary complications while substantially increasing overall health care costs [1,2]. SSIs extend the average length of hospital stay by 9.7 days and incur excess costs of approximately $20,000 per case [3]. The total cost of managing SSIs in the United States each year is estimated at $3.5 to $10 billion [4].

Surveillance of SSIs is essential for identifying infection trends, guiding prevention strategies, and benchmarking health care quality across institutions [5,6]. However, current surveillance practices rely heavily on manual chart abstraction, in which trained infection prevention personnel review electronic health records (EHRs) to determine whether a surgical case resulted in an SSI. This process is time-consuming, expensive, and difficult to scale, resulting in only a small proportion of surgical procedures being monitored [7]. Automated surveillance systems have been proposed to address this limitation by screening surgical cases using computational approaches prior to manual review, allowing infection prevention staff to focus on cases most likely to involve infection [8]. Most existing automated systems rely on rule-based algorithms or conventional machine learning (ML) methods applied to either structured clinical data (such as laboratory values) or unstructured clinical notes [9-13]. While these approaches can achieve high recall, their precision remains relatively low, meaning that multiple charts must still be reviewed to identify each true SSI case.

Recent advances in deep learning and natural language processing (NLP) offer opportunities to improve automated SSI surveillance. Clinical text and structured EHR data contain complementary signals relevant to infection detection, and temporal patterns in clinical data—such as changes in laboratory values over time—can provide additional diagnostic context. However, many prior approaches either exclude one of these modalities or rely on extensive manual feature engineering [9-13]. Deep learning models, such as convolutional neural networks (CNNs) and long short-term memory networks (LSTMs), can learn complex representations from multimodal and sequential data without extensive rule design. In addition, recent large language models (LLMs) and generalist foundation models have demonstrated strong performance across clinical NLP tasks, including summarizing clinical documentation. In this study, we developed an automated SSI surveillance system that integrates structured EHR data, clinical text, and temporal information using deep learning models and foundation model–based text summarization. We evaluated whether this multimodal, data-driven approach could improve the precision of automated SSI detection while maintaining high recall, thereby reducing the burden of manual chart review.

Study Objectives

This study introduces and evaluates a multimodal, data-driven framework that integrates structured EHR data, temporal clinical signals, and foundation model–based clinical text representations using deep learning to improve the automated detection of SSIs. The focus of this study was on the retrospective identification rather than the prospective prediction of SSIs, with the purpose of improving the efficiency of surveillance methods. The main objectives of this study were as follows:

1. Apply one of the largest clinical datasets used to date for this task
2. Develop and assess a framework for how state-of-the-art NLP approaches, in combination with deep learning, can be used in a data-driven way to combine structured clinical text and temporal data to build highly performant post hoc SSI identification models (primary study emphasis)
3. Perform one of the first evaluations, to our knowledge, on the performance of approaches using foundational LLMs (eg, Llama 3) for this task
4. Produce a model that is more performant than previous methods on the task of post hoc SSI identification, potentially allowing for a substantial reduction in the workload of infection prevention staff.

Methods

Dataset

Our dataset included a subset of all surgical cases performed at the University of Washington Medical Center (UWMC) and Harborview Medical Center (HMC) between 2010 and 2020. The dataset comprised 29,980 surgical cases (n=1206, 4% SSI-positive cases and n=28,774, 96% SSI-negative cases) from 2 registries: data from UWMC were captured for the National Surgical Quality Improvement Program (NSQIP), and data from HMC were captured for National Healthcare Safety Network (NHSN). A total of 17,403 (58%) surgical cases came from UWMC and 12,577 (42%) from HMC.

To address differences in data capture frameworks (targeted vs broader surveillance), surveillance windows, and potential variation in documentation practices between NSQIP and NHSN, we standardized postoperative observation to a 90-day window and trained models on multimodal EHR data rather than clinical notes alone. Although integrating data across institutions and surveillance programs may reduce observed performance, this approach was intended to improve generalizability by incorporating broader clinical variability.

Our dataset represents one of the largest studies of its kind to date, featuring a diverse range of 7 procedure classes, including craniotomy, spine, orthopedic, general, gynecological, cardiothoracic, and vascular surgeries. Data from UWMC consisted of NSQIP “target” procedures in cardiothoracic, vascular, general, and gynecological surgeries (cardiac surgery, vascular surgery, pancreatectomy, hepatectomy, esophagectomy, colectomy, proctectomy, appendectomy, ventral hernia repair, bariatric surgery, gynecologic reconstructive surgery, hysterectomy, cystectomy, and thyroidectomy). Data from HMC consisted of colorectal surgery, abdominal hysterectomy, spinal fusion, craniotomy, ventriculoperitoneal shunt placement, and instrumented hip and lower-extremity procedures. Cases with infections present at the time of surgery were excluded. All classes of SSI (superficial, deep, or organ/space) within the procedure-specific 30- or 90-day surveillance period were included. We included transplant cases and other instances

XSL•FO
RenderX

where patients may develop infections, despite having normal laboratory values and vital signs (eg, normal white blood cell [WBC] counts) due to immunosuppression. Each surgical case was assigned a binary value, indicating the occurrence of SSI according to NHSN or NSQIP criteria. We used EHR data to represent each case, incorporating static structured and temporal variables, as well as unstructured clinical notes.

Data Representation

Structured Data Representation

Structured data elements included (1) surgical procedure characteristics, (2) laboratory values (eg, WBC counts and microbiology orders), (3) postoperative fever (temperature), (4) antibiotic administration events, and (5) postoperative consultations with infectious disease teams or interventional radiology for drain placement. The surgical procedure type was one-hot encoded as "Surgical Procedure Class," and each case was labeled based on whether a reoperation occurred within 90 days ("Reoperation").

Consultation variables were represented as binary indicators for postoperative infectious disease consultation ("ID Consult") and interventional radiology drain placement ("IR Drain") within 90 days. Additional binary indicators captured whether a culture order ("Culture Ordered") or SSI-relevant antibiotic administration ("Antibiotic Administration") occurred during the same period.

Continuous laboratory values and vitals (WBC counts and temperature) were summarized using their maximum values within the first 90 postoperative days. Outliers (eg, >100 k/mm^3 for WBC or >50 °C for temperature) and missing values were replaced with the mean from the training set. All features were then min-max normalized prior to model input. For additional details on our structured data representation, see Multimedia Appendix 1.

Unstructured Clinical Note Representation

The clinical text corpus included all notes associated with each surgical case from 7 days before surgery to 90 days afterward (3,193,094 notes total). The most common note types included inpatient nursing notes, telephone encounters, progress notes, operative reports, and discharge summaries. Patient portal messages were not included in the unstructured notes. For each surgical case, all notes were concatenated into a single pseudodocument representing the case's text data.

Preprocessing included lowercasing and, for discrete vectorization approaches, the removal of stopwords, numbers, and punctuation [14]. A key challenge was that SSI-related information is sparse and distributed across diverse note types and time points. To address this, we developed representations that selectively extract the information most relevant to SSI ("information extracted").

We evaluated 4 text representation approaches:

1. Unigrams: Each word was treated as a feature using a bag-of-words representation with term frequency-inverse document frequency (TF-IDF) weighting [15]. Only words appearing in at least 5 documents were included. The top N features were selected using an ANOVA *F* test prior to model training, with N determined through a hyperparameter search.
2. Unified Medical Language System (UMLS) disease and chemical concepts: Disease and chemical concepts were defined using UMLS semantic groups [16] and identified with scispaCy (Allen Institute for AI) [17]. We retained the 200 concepts with the highest ANOVA *F* values relative to the SSI label. These concepts were represented using a bag-of-words TF-IDF approach, where document frequency was defined by the number of surgical cases containing the concept. The remaining TF-IDF implementation details matched the unigram approach.
3. Sentences containing UMLS disease and chemical concepts (information extracted): To capture contextual information, we extracted sentences containing identified disease or chemical UMLS concepts. Sentences were ordered temporally by note date and by their position within each note. To reduce noise, only sentences containing concepts appearing in fewer than 15% of pseudodocuments were retained. The resulting sentence sequences were vectorized using word embeddings trained on our dataset with the word2vec continuous bag-of-words architecture [18]. In a separate approach, we also used embeddings derived from ClinicalBERT, which incorporates domain-specific knowledge through pretraining and fine-tuning on clinical corpora [19].
4. Clinical note summaries (information extracted): We used Llama-3-70B-Instruct (Meta AI), accessed through the Hugging Face Transformers framework, to summarize clinical notes. The average length of each clinical note in our dataset was 1772 words. We condensed each note into a summary of 100 words. The manuscript authors, 2 of whom are clinicians, jointly reviewed the generated summaries to ensure they were consistent, reproducible, accurate, and without hallucinations. Using parallel graphics processing unit inference (4 parallel graphics processing units), the average processing time was approximately 1 second per note, and the full summarization process for our 3,193,094 notes was completed in about 9 days. These summaries for each clinical note were concatenated temporally in order of note date. Our summarization prompt was as follows: "Summarize the following clinical note emphasizing development of infection and administration of antibiotics for surgical site infection. Limit the summary to less than 100 words." Our constrained prompt ("emphasizing development of infection and administration of antibiotics") was designed to limit hallucinations. To capture information more salient to SSI prediction, we only summarized notes containing disease or chemical UMLS concepts. We vectorized the sequence of clinical note summaries using word embeddings. For additional details on our clinical text representation, see Multimedia Appendix 1.

Temporal Data Representation

To incorporate temporal dynamics, we extended the structured data by including time-stamped laboratory values (WBC and absolute neutrophil count) and vitals (temperature) from 7 days

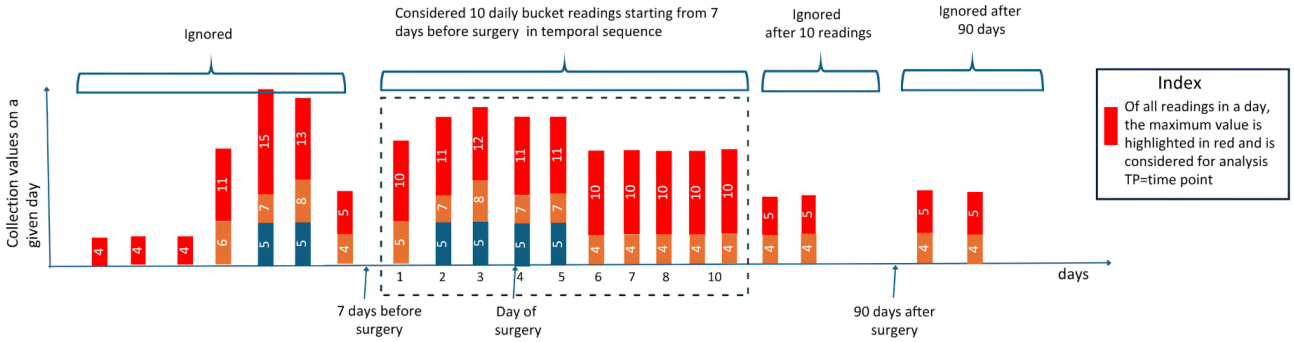
XSL•FO
RenderX

preoperatively to 90 days postoperatively. For temporal models, these variables were excluded from the static structured features.

Daily maximum values were computed and organized into fixed-length sequences, with shorter sequences padded with zeros. The sequence length (10 time points) was determined through hyperparameter tuning to balance model complexity with available data while capturing the clinically relevant perioperative window. Daily maxima were selected to emphasize peak physiological abnormalities associated with infection rather than averaging values, which may attenuate clinically meaningful signals.

Zero padding was used to standardize sequence lengths and was informed by observed differences between SSI-positive and SSI-negative cases (Figure 1). In this dataset, SSI-negative cases generally exhibited fewer recorded measurements and lower overall values. As such, zero padding serves as a proxy for the absence of clinically significant events, rather than introducing artificial trends. The models were trained to interpret these padded values accordingly.

Figure 1. Method for constructing our temporal representation. Srl: serial TP: time point.



Srl #	Data or feature	TP #1	TP #2	TP #3	TP #4	TP #5	TP #6	TP #7	TP # 8	TP #9	TP # 10
1	Collection value	10	11	12	11	11	10	10	10	10	10
2	Collection days from surgery	−6	−3	−2	0	2	3	29	40	71	75
3	Feature #1 (Actual value)	10	11	12	11	11	10	10	10	10	10
4	Feature #2 (Days since last date of collection)	1 =−6− (−7)	3 =−3− (−6)	1 =−2− (−3)	2 =0-(−2)	2 =2−0	1 =3−2	26 =29−3	11 =40−29	31 =71−40	4 =75−71
5	Feature #3 (Days since date of surgery)	−6	−3	−2	0	2	3	29	40	71	75
6	Feature #4 (Binary value if actual value is upper limit of range=10.5)	0	1	1	1	1	0	0	0	0	0

For each variable, 4 derived features were constructed: (1) fold change relative to the upper limit of normal, (2) time between consecutive measurements, (3) time since surgery, and (4) an indicator of whether the value exceeded the normal range. These features were generated from temporally ordered sequences and combined to form inputs to CNN and LSTM models. For additional details on our temporal representation, see Multimedia Appendix 1.

SSI Classification Approach

We defined our SSI identification problem as a binary classification task. We defined 2 baselines and proposed several neural or deep learning–based approaches to solve the problem.

Baseline 1 (Rule-based algorithm): As the first baseline system, we used a rule-based system [9] that forms the basis of a published gold standard contemporary automated surveillance system [9]. This previously described rule-based system is available online [9]. It uses a text-processing pipeline with components including a section detector, named entity recognizer, and context detector to analyze clinical notes associated with surgical cases. We input each pseudodocument into this pipeline. The system produced a binary classification decision for each clinical text sample (pseudodocument), indicating whether SSI occurred or not after surgery.

Baseline 2 (Random forest): As the second baseline system, we developed a random forest (RF) model, which used either structured features alone or a combination of structured and unigram text features. RF was a strong baseline for this task, as many published works in the past have used this modeling approach [10,13].

Deep learning models: We compared the performance of the 2 baseline models against our neural or deep learning approaches. We constructed our static structured representation by inputting static structured features into a dense layer. In our feed-forward neural network models, we used this static structured representation alone or combined it with unigram features or UMLS concept features. Our deep learning models took UMLS concept sentences or Llama 3–generated clinical text summaries as input. These sentences or summaries were vectorized using either locally trained word embeddings or those generated by ClinicalBERT. The latent representations derived by different deep learning layers (CNN and bidirectional long short-term memory network [BiLSTM]) were concatenated with static structured data representations within our deep learning framework. In models incorporating temporal information, we

XSL•FO
RenderX

input our temporal data representation into different deep learning layers (CNN and LSTM) to generate latent temporal representations. We concatenated these latent temporal representations with our best-performing (highest F_1-score) text representation and our static structured data representation for training. Our neural and deep learning models included a dense layer and a final output layer following the concatenation layer.

The dataset was split into training, validation, and test sets in a 7:1:2 ratio. We used stratified training or validation or test splits to address the class imbalance in our dataset, in which only 4% (1206/29,980) of cases were SSI-positive. Because SSI is a relatively infrequent event (<5%), a class-weight parameter was applied during model training to account for the imbalance between SSI-positive and SSI-negative cases. We also explored alternative strategies to address class imbalance, including up-sampling SSI-positive cases through duplication and down-sampling SSI-negative cases through random removal. Interim analyses indicated that these approaches did not meaningfully affect model performance, and they were therefore not used in the final model development.

We systematically tested a range of hyperparameters for different model types using a grid-search technique as part of the tuning process to derive the optimal values. Hyperparameter tuning aimed to optimize the F_1-score or the precision at 0.9 recall metric. We evaluated performance in each experiment using F_1-score, precision (positive predictive value), recall (sensitivity), and area under the precision-recall curve. To measure performance, we also used a clinical utility metric (precision at 0.9 recall), which more closely reflects usefulness in a clinical use setting. One-sample 95% CIs based on the *t* distribution for each metric were reported.

We compared the performance achieved by an individual representation or strategy with another individual representation or strategy using a 2-tailed *t* test for the means of 2 independent samples. To control the family-wise error rate across multiple comparisons, we applied the Bonferroni correction, multiplying each *P* value by the total number of comparisons performed (10 comparisons for SSI prediction performance and 2 comparisons for clinical significance).

Ethical Considerations

This study was reviewed and approved by the University of Washington Institutional Review Board (approval number: STUDY00005987). The requirement for informed consent was waived because the research involved a retrospective analysis of existing EHR data and posed minimal risk to participants. Access to the data was granted under the institutional review board approval, and all analyses were conducted in accordance with the University of Washington policies on human participants research and data protection.

Only authorized study personnel had access to identifiable information during data extraction. No direct patient contact occurred, and no participant compensation was involved. Privacy and confidentiality were protected through secure data storage on institutional, access-controlled servers, with all procedures following the Health Insurance Portability and Accountability Act (HIPAA) [20] and relevant institutional data-handling policies. All researchers and annotators completed human participants training and were authorized to handle data containing protected health information.

Results

Composition of Procedure Types and SSI Event Rates

The total number of cases and total SSI-positive cases for each procedure type are provided in Table 1. The SSI positivity rate in our dataset was 1206 out of 29,980 (4%).

Table . Surgery counts and infection rate.

Procedure class	Case count	SSI[a] events rate, n (%)
Spine	3724	196 (5.3)
Orthopedic (nonspine)	2739	105 (3.8)
Neurosurgery (nonspine)	4990	185 (3.7)
General surgery	14,200	585 (4.1)
Gynecologic surgery	3211	120 (3.7)
Cardiothoracic surgery	328	8 (2.2)
Vascular surgery	788	7 (0.9)
Total	29,980	1206 (4.0)

[a]SSI: surgical site infection.

Contributions of Model Type, Architecture, and Text Representation to Model Performance

We assessed the impact of using different feature sets, text representations, and modeling strategies on the task of SSI prediction. Table 2 summarizes the precision, recall, overall performance (F_1-score), and area under the precision-recall curve of our different baseline comparators, feature sets, data representation methods, and modeling strategies. Our best-performing deep learning models (Table 2: Exp 11) outperformed RF models (Table 2: Exp 1 and 3). Integration of clinical text and structured data enhances the performance for both RF (Table 2: Exp 1 vs Exp 3) and neural models (Table 2: Exp 2 vs Exp 4 and 5). This supports our hypothesis that the incorporation of clinical text data adds new signals relevant to SSI prediction and thereby can enhance the performance of SSI

XSL•FO
RenderX

prediction models. Using Llama 3 summaries outperformed using UMLS concept sentences (Table 2: Exp 7 vs Exp 9).

XSL•FO
RenderX

Table . Surgical site infection (SSI) prediction performance.

Experiment number	Model[a]	Features	Precision, mean (95% CI)	Recall, mean (95% CI)	F_1, mean (95% CI)	AUPRC[b], mean (95% CI)	*P* value	Comparison with experiment number
0	Rule-based model (baseline 1)	—[c]	0.34[d]	0.59[d]	0.43[d]	—	—	—
Structured only								
1	RF[e] (baseline 2)	Structured only	0.64 (0.63-0.65)	0.42 (0.42-0.43)	0.51 (0.50-0.51)	0.53 (0.53-0.53)	—	—
2	FFNN[f]	Structured only	0.76 (0.76-0.76)[g]	0.42 (0.41-0.44)	0.54 (0.52-0.56)	0.57 (0.57-0.57)	.22	1
Multimodal								
3	RF	Structured+unigrams	0.69 (0.69-0.69)	0.47 (0.47-0.48)	0.57 (0.56-0.57)[h]	0.64 (0.64-0.64)	.005	1
4	FFNN	Structured+unigrams	0.58 (0.58-0.58)	0.67 (0.67-0.67)	0.62 (0.62-0.62)[h]	0.64 (0.64-0.65)	<.001	3
5	FFNN	Structured+UMLS[i] concepts	0.60 (0.59-0.60)	0.67 (0.66-0.68)	0.63 (0.63-0.63)[h]	0.65 (0.64-0.65)	<.001	4
6	CNN[j]	Structured+UMLS sentences	0.62 (0.60-0.64)	0.64 (0.63-0.64)	0.63 (0.62-0.63)	0.63 (0.63-0.64)	—	5
7	BiLSTM[k]	Structured+UMLS sentences	0.62 (0.59-0.65)	0.67 (0.64-0.69)	0.64 (0.64-0.64)[h]	0.67 (0.67-0.67)	<.001	5
8	ClinicalBERT	Structured+UMLS sentences	0.61 (0.61-0.61)	0.66 (0.66-0.66)	0.63 (0.63-0.63)	0.63 (0.63-0.63)	—	5
9	BiLSTM	Structured+Llama 3 summaries	0.68 (0.68-0.69)	0.69 (0.69-0.70)	0.68 (0.68-0.69)[h]	0.74 (0.74-0.75)	<.001	5
Temporal methods								
10	BiLSTM or CNN	Structured+Llama 3summaries+temporal	0.66 (0.65-0.66)	0.66 (0.66-0.66)	0.66 (0.66-0.66)[h]	0.72 (0.72-0.72)	<.001	9
11	BiLSTM or LSTM[l]	Structured+Llama 3 summaries+temporal	0.69 (0.68-0.69)	0.71 (0.69-0.73)[g]	0.70 (0.69-0.71)[h,g]	0.78 (0.78-0.79)[g]	<.001	9

[a]Column values are formatted as text representation or temporal representation.

[b]AUPRC: area under the precision-recall curve.

[c]Not applicable.

[d]A 95% confidence interval is not applicable in this case, as the method is a nonstochastic rule-based system and therefore yields identical performance across runs, with no sampling variability to estimate uncertainty.

[e]RF: random forest.

[f]FFNN: feed-forward neural network.

[g]Highest precision, recall, F_1, and AUPRC values.

[h]F_1 values with statistical significance (P<.05).

[i]UMLS: Unified Medical Language System.

[j]CNN: convolutional neural network.

XSL•FO
RenderX

[k]BiLSTM: bidirectional long short-term memory network.

[l]LSTM: short-term memory network.

Using an LSTM layer to process temporal data outperforms using a CNN layer. The incorporation of temporal data with an LSTM layer thus improves overall performance, primarily driven by improvements in recall (Table 2: Exp 9 vs Exp 11). A plausible explanation for this is that the incorporation of temporal information on laboratory values and vitals added signals pertaining to SSI (eg, elevated laboratory values). Thus, it led to the correct classification of some previously false-positive cases that exhibit such signals.

Factors Driving the Performance of Clinical Text Summaries Versus UMLS Concept Sentences

Our error analyses revealed that the following factors resulted in Llama 3–generated summaries (Table 2: Exp 9) outperforming UMLS concept sentences (Table 2: Exp 7):

1. Focusing our text representation on signals relevant to SSI: Our targeted summarization prompt ("emphasizing development of infection and administration of antibiotics for surgical site infection") focused the clinical text summaries on information relevant to SSI prediction (eg, "treated with antibiotics for Klebsiella, *E. coli* …," "the patient developed an abdominal wound infection after hernia repair surgery," and "elevated white blood cell count"). This led to the model being able to focus on more salient information for SSI prediction (eg, signs and symptoms of, treatment for, and diagnosis of SSI) rather than being confused by less relevant information present in UMLS concept sentences (eg, "IVPB discontinued due to pain").
2. Clinically irrelevant concept selection based on SSI correlation (part of the UMLS concept sentences approach): The UMLS concepts approach restricted sentence extraction to the top 200 concepts with the highest ANOVA *F* value between the concept and the SSI positivity label. This strategy helped avoid including sentences related to commonly recorded but irrelevant concepts, such as repeated mentions of the patient receiving a saline bolus. However, focusing solely on highly correlated concepts resulted in sentences enriched in positive cases but lacking clinical relevance to SSI (eg, "Renal insufficiency. Cr range of" and "chew Vitamin supplement"), which reduced the precision of Approach 2 for post hoc SSI prediction. In contrast, Llama 3, with its broad human biology knowledge and advanced reasoning capabilities, produced clinical note summaries rich in information essential for accurate post hoc SSI prediction.
3. Building more coherent text representations: Because our clinical text summaries were temporally ordered (by note date) and each contained information about one note, the resulting text representation was more coherent than one constructed using sentences containing UMLS concepts:

> *Summary: A [age] year old female patient...The patient had an abdominal exam worrisome for surgical site infection, with tenderness to palpation and erythema around an incision site ct abdomen/pelvis showed large amounts of abdominal subcutaneous gas, concerning for possible necrotizing soft tissue infection. ...summary: A [age] year old female patient was admitted to the hospital with symptoms of confusion, lethargy, and abdominal pain...a ct scan revealed free intra-abdominal air, extensive subcutaneous emphysema, and a possible abscess or leak.*

- This led to our BiLSTM models being able to process the information more effectively. In contrast, in the case of UMLS concept sentences, our BiLSTM models may have been confounded by abrupt chains of disjointed sentences with little connection to each other and lacking an overall logical message:

> *Allergies tylenol NSAIDs aztreonam Medications Nystatin topical as needed. Nystatin topical application Topical BID A CT was performed showing concern for abdominal wall abscess which was I&Ded initially with associated cellulitis. She was taken back to the OR today due to jejunostomy anastamosis leak and abdominal cellulitis that was repaired and debrided with no evidence of fasciitis.* [Problems Interventions Education Sepsis]

We have provided examples of false negatives and false positives to further illustrate these points in Multimedia Appendix 2Multimedia Appendix 33.

Clinical Significance

Assessment of the overall performance of our various approaches is not meaningful in actual practice if it does not reflect performance in a real-world clinical setting. In such a setting, our automated surveillance approach is likely to be used as a screening tool prior to manual chart review to rule out clearly negative cases. Thus, to assess the actual clinical importance of our various SSI prediction approaches, we assessed the precision of our models at a high (90%) recall (Table 3).

XSL•FO
RenderX

Table . Clinical significance of the various data representation and modeling approaches.

Experiment number	Model[a]	Features	Precision at 0.9 recall, mean (95% CI)	Number of charts reviewed per SSI[b]	Comparison with	*P* value
M	Manual review	—[c]	0.03[d]	33[d]	—	—
3	RF[e]	Structured+unigrams	0.17 (0.15-0.19)	5.9	—	—
5	FFNN[f]	Structured+UMLS[g] sentences	0.26 (0.25-0.27)	3.8	—	—
9	BiLSTM[h]	Structured+Llama 3 summaries	0.35 (0.35-0.36)[i]	2.9	Exp 5	<.001
11	BiLSTM or LSTM[j]	Structured+Llama 3 summaries+temporal	0.38 (0.37-0.38)[k]	2.6	Exp 9	<.001

[a]Column values are formatted as text representation or temporal representation.

[b]SSI: surgical site infection.

[c]Not applicable.

[d]CIs are not reported for the manual review estimates, as these values are derived from published average processing times aggregated across multiple institutions.

[e]RF: random forest.

[f]FFNN: feed-forward neural network.

[g]UMLS: Unified Medical Language System.

[h]BiLSTM: bidirectional long short-term memory networks.

[i]Precision at 0.9 recall values with statistical significance ($P<.05$).

[j]LSTM: short-term memory networks.

[k]Highest precision at 0.9 recall values.

Overall, we found that clinical utility tracked overall performance well. Our best deep learning model (Table 3: Exp 11) achieved higher clinical utility than our best RF models (Table 3: Exp 3). The addition of temporal information led to a slight increase in clinical utility (Table 3: Exp 11 vs Exp 9). Neural models achieved higher clinical utility than published gold-standard conventional ML models (Table 3: Exp 5 vs Exp 3). However, harnessing the ability of LLMs to understand complex contextual underpinnings is required to achieve substantial performance gains over conventional approaches (Table 3: Exp 11 vs Exp 3). This would entail a 1.5-fold improvement (2.6 charts reviewed per SSI detected with our method vs 4 charts reviewed per SSI detected with published methods [12,13]) in the efficiency of surveillance over published gold-standard automated surveillance systems. It would also enable approximately a 13-fold increase (2.6 charts reviewed per SSI detected with our method vs 33 charts reviewed per SSI detected with manual chart review [13]) in the number of SSIs that can be detected and the effective number of charts that can be reviewed in the same time over manual surveillance alone.

Discussion

Principal Findings

In this study, we developed and evaluated an automated SSI surveillance system using multimodal EHR data from approximately 30,000 surgical procedures across 2 health care facilities. Our approach integrated structured clinical data, temporal information, and clinical text using deep learning models, while leveraging generalist foundation models to summarize clinical documentation. The best-performing architecture combined foundation model–based text summarization with recurrent neural networks (BiLSTM and LSTM) that captured contextual patterns in clinical notes and temporal dependencies in structured clinical data. This approach achieved a precision of 0.38 at a recall of 0.9. Additionally, inclusion of data from multiple hospitals and surveillance programs may improve the generalizability of the model by exposing it to broader clinical variability.

These findings suggest that multimodal, data-driven deep learning approaches can improve the performance of automated SSI surveillance systems compared with previously reported rule-based or conventional ML methods. Two key implications follow from this observation. First, improvements in performance appear to scale with the availability of large clinical datasets. Over the past 2 decades, health care systems have generated rapidly increasing volumes of digital clinical data, including structured EHR data and clinical documentation [21]. Data-driven approaches that leverage these large datasets may therefore enable increasingly effective surveillance systems capable of monitoring a larger proportion of surgical procedures.

Second, deep learning approaches reduce the need for labor-intensive rule design and manual feature engineering, which have historically been required for rule-based or traditional ML surveillance systems [22]. By learning patterns directly from multimodal data, these models can simplify system development and potentially reduce the cost and effort required to maintain surveillance tools as clinical workflows evolve.

XSL•FO
RenderX

Recent advances in LLMs and generalist foundation models also create new opportunities for extracting meaningful information from clinical text. Previous research has shown that the use of LLMs can improve the efficiency of SSI surveillance methods for some procedure types [23-25]. These models contain broad linguistic and domain knowledge that allow them to perform many NLP tasks with minimal task-specific training [26]. Our results suggest that using foundation models selectively for tasks where they demonstrate strong capabilities—such as summarizing clinical documentation—can enhance downstream prediction tasks when combined with deep learning architectures designed to model temporal clinical data.

Comparison With Prior Work

Automated SSI surveillance systems are most useful when deployed as a screening step prior to manual chart review, allowing infection prevention personnel to focus on cases most likely to involve infection. In this setting, systems must maintain high recall to ensure that true SSI cases are not missed while maximizing precision to minimize the number of charts requiring manual review.

Our best-performing model achieved a precision of 0.38 at a recall of 0.9, representing approximately a 1.5-fold improvement in precision compared with previously reported automated surveillance approaches, which achieved precision levels of 0.24 to 0.28 at similar recall thresholds. In practical terms, this improvement would substantially reduce the number of charts that infection prevention staff must review to identify each SSI case.

When applied at scale, improvements in precision of this magnitude could significantly expand the number of surgical cases that can be monitored through surveillance programs while simultaneously reducing the manual review workload. Estimates based on national surgical volumes suggest that such improvements could translate into substantial reductions in annual manual review hours for infection prevention personnel (see Multimedia Appendix 4 for detailed calculations).

Limitations

Our study has the following limitations. First, although our dataset comprises data from 2 facilities, both are affiliated with the same health system, University of Washington Medicine. However, the procedure types included at each hospital differ substantially, and therefore the 2 sites are not directly comparable for the purposes of stratified performance reporting. As a result, we did not report model performance stratified by hospital. This distinction also limits our ability to assess site-level variability in model performance within this study.

More broadly, this also constrains the generalizability of our findings. Because the data were derived from a single health system, the model may reflect shared characteristics such as documentation practices, clinical workflows, and EHR infrastructure. Accordingly, the performance of our model in external settings—particularly in institutions with different EHR vendors, surveillance protocols, or documentation patterns—remains uncertain and will require validation in independent cohorts. Future studies should evaluate the transportability of this approach across diverse health care systems and EHR environments.

We did not leverage prompt engineering approaches with LLMs, which would enable completely zero-shot approaches requiring no training samples. In our approaches, we used BiLSTM and LSTM models as part of our architecture, which require training. Future work could explore the performance of such approaches on SSI prediction. In addition, technologies like retrieval-augmented generation might improve intrainstitutional performance. Finally, we excluded imaging data as a modality, though it could enhance performance by confirming SSI-negative cases or identifying non-SSI infections. Future work will focus on standardizing image formats and extracting relevant features to improve predictions [27]. Recently, vision-aligned LLMs may also hold promise in this regard and will be explored as well. Lastly, it is important to mention that while automated surveillance methods can mitigate certain aspects of bias in the manual review process, they can potentially also introduce new bias caused by biases in the training data used to train such models and other factors.

Conclusions

Automated surveillance approaches using ML can enable more efficient SSI surveillance. This can enable more robust analyses of surveillance data and more informed quality improvement initiatives targeting SSI rates.

Acknowledgments

The authors thank the University of Washington Perioperative & Pain Initiatives in Quality Safety Outcome Center for support with clinical data extraction and informatic support.

No generative artificial intelligence tools were used at any stage in the preparation of this manuscript. The authors take full responsibility for the accuracy, originality, and integrity of all content, including references and citations.

Funding

The authors acknowledge the support provided by the National Institutes of Health or National Institute of Arthritis and Musculoskeletal and Skin Diseases Mentored Patient-Oriented Research Career Development Award (5K23AR080209-02, DL). This work was supported by the National Library of Medicine Biomedical and Health Informatics Training Program at the University of Washington (T15LM007442, AC).

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are not publicly available due to their large volume and the inclusion of sensitive patient information. Access to the data may be granted upon reasonable request and with institutional approval under a signed Data Use Agreement.

Authors' Contributions

Conceptualization: DL, MY
Data curation: DL
Formal analysis: AC
Investigation: AC
Methodology: AC
Supervision: DL, MY, PT-H
Validation: AC
Writing – original draft: AC
Writing – review & editing: AC, DL, MY, PT-H
All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Expanded data representation.
[DOCX File, 21 KB - periop_v9i1e87896_app1.docx]

Multimedia Appendix 2
Examples of clinical text excerpts which drove model predictions towards a positive or negative classification for models using UMLS concept sentences.
[PNG File, 167 KB - periop_v9i1e87896_app2.png]

Multimedia Appendix 3
Examples of clinical text excerpts which drove model predictions towards a positive or negative classification for models using summarized clinical notes.
[PNG File, 242 KB - periop_v9i1e87896_app3.png]

Multimedia Appendix 4
Calculation details on how surveillance efficiency improvement with our study was calculated.
[PNG File, 155 KB - periop_v9i1e87896_app4.png]

References

1. Anderson DJ, Perl TM. Basics of surgical site infection: surveillance and prevention. In: Practical Healthcare Epidemiology: Cambridge University Press; 2018:147-161. [doi: 10.1017/9781107153165.015]
2. Magill SS, O'Leary E, Janelle SJ, et al. Changes in prevalence of health care-associated infections in U.S. hospitals. N Engl J Med 2018 Nov 1;379(18):1732-1744. [doi: 10.1056/NEJMoa1801550] [Medline: 30380384]
3. Pinchera B, Buonomo AR, Schiano Moriello N, Scotto R, Villari R, Gentile I. Update on the management of surgical site infections. Antibiotics (Basel) 2022 Nov 11;11(11):1608. [doi: 10.3390/antibiotics11111608] [Medline: 36421250]
4. Seidelman JL, Mantyh CR, Anderson DJ. Surgical site infection prevention: a review. JAMA 2023 Jan 17;329(3):244-252. [doi: 10.1001/jama.2022.24075] [Medline: 36648463]
5. Petherick ES, Dalton JE, Moore PJ, Cullum N. Methods for identifying surgical wound infection after discharge from hospital: a systematic review. BMC Infect Dis 2006 Nov 27;6(1):170. [doi: 10.1186/1471-2334-6-170] [Medline: 17129368]
6. Khuri SF, Daley J, Henderson WG. The comparative assessment and improvement of quality of surgical care in the Department of Veterans Affairs. Arch Surg 2002 Jan;137(1):20-27. [doi: 10.1001/archsurg.137.1.20] [Medline: 11772210]
7. Selby LV, Sjoberg DD, Cassella D, et al. Comparing surgical infections in National Surgical Quality Improvement Project and an institutional database. J Surg Res 2015 Jun 15;196(2):416-420. [doi: 10.1016/j.jss.2015.02.072] [Medline: 25840487]
8. Rosenthal R, Weber WP, Marti WR, et al. Surveillance of surgical site infections by surgeons: biased underreporting or useful epidemiological data? J Hosp Infect 2010 Jul;75(3):178-182. [doi: 10.1016/j.jhin.2009.10.028] [Medline: 20227139]

XSL•FO
RenderX

9. Bucher BT, Shi J, Ferraro JP, et al. Portable automated surveillance of surgical site infections using natural language processing: development and validation. Ann Surg 2020 Oct;272(4):629-636. [doi: 10.1097/SLA.0000000000004133] [Medline: 32773639]
10. Grundmeier RW, Xiao R, Ross RK, et al. Identifying surgical site infections in electronic health data using predictive models. J Am Med Inform Assoc 2018 Sep 1;25(9):1160-1166. [doi: 10.1093/jamia/ocy075] [Medline: 29982511]
11. Hu Z, Simon GJ, Arsoniadis EG, Wang Y, Kwaan MR, Melton GB. Automated detection of postoperative surgical site infections using supervised methods with electronic health record data. Stud Health Technol Inform 2015;216:706-710. [Medline: 26262143]
12. Karhade AV, Bongers MER, Groot OQ, et al. Natural language processing for automated detection of incidental durotomy. Spine J 2020 May;20(5):695-700. [doi: 10.1016/j.spinee.2019.12.006] [Medline: 31877390]
13. Shi J, Liu S, Pruitt LCC, et al. Using natural language processing to improve EHR structured data-based surgical site infection surveillance. AMIA Annu Symp Proc 2019;2019:794-803. [Medline: 32308875]
14. Sarica S, Luo J. Stopwords in technical language processing. PLoS One 2021;16(8):e0254937. [doi: 10.1371/journal.pone.0254937] [Medline: 34351911]
15. Juluru K, Shih HH, Keshava Murthy KN, Elnajjar P. Bag-of-words technique in natural language processing: a primer for radiologists. Radiographics 2021;41(5):1420-1426. [doi: 10.1148/rg.2021210025] [Medline: 34388050]
16. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D267-D270. [doi: 10.1093/nar/gkh061] [Medline: 14681409]
17. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. Presented at: Proceedings of the 18th BioNLP Workshop and Shared Task; Florence, Italy p. 319-327. [doi: 10.18653/v1/W19-5034]
18. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv. Preprint posted online on Jan 16, 2013. [doi: 10.48550/arXiv.1301.3781]
19. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T. Publicly available clinical BERT embeddings. Presented at: Proceedings of the 2nd Clinical Natural Language Processing Workshop; Jun 7, 2019. [doi: 10.18653/v1/W19-1909]
20. U.S. Department of Health and Human Services. PART 164—SECURITY AND PRIVACY. Code of Federal Regulations. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164 [accessed 2026-06-17]
21. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med 2022 Sep;28(9):1773-1784. [doi: 10.1038/s41591-022-01981-2] [Medline: 36109635]
22. Alzubaidi L, Zhang J, Humaidi AJ, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 2021;8(1):53. [doi: 10.1186/s40537-021-00444-8] [Medline: 33816053]
23. Morgan DJ, Goodman KE, Branch-Elliman W, et al. Using generative artificial intelligence for healthcare-associated infection surveillance. Clin Infect Dis 2026 Mar 17;82(3):473-479. [doi: 10.1093/cid/ciaf619] [Medline: 41222316]
24. Pierce L, Pak C, Stanley K, et al. Retrospective validation study of a large language model approach to screening for intraabdominal surgical site infections for quality and safety reporting. BMJ Qual Saf 2026 May 18;35(6):424-427. [doi: 10.1136/bmjqs-2025-019493] [Medline: 41423338]
25. Badia JM, Casanova-Portoles D, Membrilla E, Rubiés C, Pujol M, Sancho J. Evaluation of ChatGPT-4 for the detection of surgical site infections from electronic health records after colorectal surgery: a pilot diagnostic accuracy study. J Infect Public Health 2025 Feb;18(2):102627. [doi: 10.1016/j.jiph.2024.102627] [Medline: 39740340]
26. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Presented at: NIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems; Dec 6-12, 2020 URL: https://dl.acm.org/doi/abs/10.5555/3495724.3495883 [accessed 2026-06-06]
27. Sawyer RG, Evans HL, Hedrick TL. Technological advances in clinical definition and surveillance methodology for surgical site infection incorporating surgical site imaging and patient-generated health data. Surg Infect (Larchmt) 2019 Oct;20(7):541-545. [doi: 10.1089/sur.2019.153] [Medline: 31460834]

Abbreviations

BiLSTM: bidirectional long short-term memory network
CNN: convolutional neural network
EHR: electronic health record
HIPAA: Health Insurance Portability and Accountability Act
HMC: Harborview Medical Center
LLM: large language model
LSTM: long short-term memory network
ML: machine learning
NHSN: National Healthcare Safety Network
NLP: natural language processing
NSQIP: National Surgical Quality Improvement Program

XSL•FO
RenderX

RF: random forest
SSI: surgical site infection
TF-IDF: term frequency-inverse document frequency
UMLS: Unified Medical Language System
UWMC: University of Washington Medical Center
WBC: white blood cell

Edited by N Rohatgi; submitted 21.Nov.2025; peer-reviewed by D Chrimes, P Tekkali; revised version received 04.Apr.2026; accepted 25.May.2026; published 26.Jun.2026.

Please cite as:
Chakraborty A, Tarczy-Hornoch P, Long D, Yetisgen M
Automated Identification of Surgical Site Infections From Electronic Medical Records: Retrospective Observational Predictive Modeling Study
JMIR Perioper Med 2026;9:e87896
URL: https://periop.jmir.org/2026/1/e87896
doi:10.2196/87896

© Arjun Chakraborty, Peter Tarczy-Hornoch, Dustin Long, Meliha Yetisgen. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 26.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

In-House 3D Planning for Corrective Osteotomy of the Distal Radius: Prospective Cohort Study With 1-Year Radiological and Clinical Follow-Up

Charlotte Stor Swinkels[1,2,3], MSc; Peter Axelsson[1,3], MD, PhD; Katleen Libberecht[1,3], MD; Per Fredrikson[1,3], MD; Emilia Gryska[1,3], PhD; Anders Björkman[1,3], MD, PhD

[1]Department of Hand Surgery, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Medicinaregatan 3, Göteborg, Västra Götaland, Sweden

[2]Department of Medical Physics and Biomedical Engineering, Sahlgrenska University Hospital, Gothenburg, Sweden

[3]Department of Hand Surgery, Sahlgrenska University Hospital, Mölndal, Sweden

Corresponding Author:
Charlotte Stor Swinkels, MSc
Department of Hand Surgery, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Medicinaregatan 3, Göteborg, Västra Götaland, Sweden

Abstract

Background: Distal radius malunions occur in 5% to 17% of fractures and may lead to pain and functional impairment. While 3D surgical planning with patient-specific guides has demonstrated high accuracy and reduced fluoroscopy use compared with conventional 2D techniques, its routine clinical implementation remains limited, particularly for in-house hospital-based workflows.

Objective: This study aimed to evaluate the precision of corrective osteotomy for distal radius malunions using in-house–designed 3D patient-specific surgical guides 1 year after surgery. Secondary objectives were to assess clinical and patient-reported outcomes and document intraoperative fluoroscopy use.

Methods: Corrective osteotomies of extra-articular distal radius malunions were performed in 16 consecutive patients using in-house 3D surgical planning and patient-specific surgical guides. Accuracy was evaluated by comparing preoperative plans with 1-year postoperative computed tomography–based 3D models, assessing ulnar variance, volar tilt, and radial inclination. Acceptable error margins were defined as 5° or less for volar tilt and 2 mm or less for ulnar variance. Clinical outcomes were evaluated using patient-reported outcome measures and range of motion, lift, grip, and torque strength measurements.

Results: The mean absolute error was 1 (SD 0.8) mm in ulnar variance, 4.9° (SD 2.7°) in volar tilt, and 1.9° (SD 3.6°) in radial inclination. The corrections were within predefined equivalence bounds of the virtual plans in terms of ulnar variance but not volar tilt. Significant improvements were observed in Patient-Rated Wrist Evaluation (PRWE) and Disabilities of the Arm, Shoulder, and Hand (DASH) scores; pain during activity; pain at rest; lift strength; wrist flexion; and ulnar deviation. In total, 12.5% (2/16) of the patients experienced complications following concomitant ulnar shortening procedures and were transferred to standard care. The median intraoperative fluoroscopy duration was 56 (IQR 26 - 72) seconds.

Conclusions: In-house 3D planning was feasible and associated with generally favorable radiological and clinical outcomes at 1 year. However, the predefined equivalence criterion was not achieved for volar tilt, and the findings should be interpreted considering the small sample size and observed complications. In-house 3D planning supports continuous surgical refinement and quality improvement.

(JMIR Perioper Med 2026;9:e101764) doi:10.2196/101764

KEYWORDS

3D; virtual surgical planning; surgical guides; patient specific; distal radius; in-house

Introduction

Distal radius fractures are the most common type of fractures in the upper extremities. In 2024, wrist fractures made up 17% of all fractures registered in the Swedish Fracture Register [1]. While most distal radius fractures heal with satisfactory alignment [2], approximately 5% to 17% result in malunion, which may cause functional impairment and pain [3-5]. Previous studies have demonstrated that accurate anatomical correction of distal radius malunions leads to better clinical outcomes [5-8].

Traditional corrective techniques using 2D radiographic planning and freehand osteotomy are limited in their ability to achieve precise restoration of the anatomy [5,9]. To improve surgical precision, 3D surgical planning and patient-specific

XSL•FO
RenderX

surgical guides are increasingly used in complex skeletal reconstructions. Studies applying this approach to distal radius malunions report high precision with small residual errors [10-17] and reduced use of intraoperative fluoroscopy [6]. Although 3D planning has been in clinical use for over 15 years and evidence supporting its effectiveness continues to grow, its routine integration into clinical practice is currently limited.

Two main models exist for the clinical implementation of 3D virtual planning and patient-specific surgical guides. The standard approach involves outsourcing the planning and production of guides to commercial providers. However, this model requires transferring patient data to external partners and restricts direct collaboration between surgeons and engineers, with high costs being an additional limitation. An alternative is to establish an in-house 3D planning unit within the hospital. This model facilitates close, continuous collaboration between surgeons and engineers, potentially improving planning accuracy and cost efficiency. However, it also requires dedicated resources, staff training, and a robust quality management system to ensure clinical reliability and safety. In our previous preclinical study [18], the same patient cases underwent independent in-house and commercially provided 3D planning. The resulting surgical guides were evaluated on 3D-printed bone models, and the in-house workflow demonstrated noninferiority to the commercial workflow. The present study extends these findings by evaluating the surgical accuracy and 1-year clinical and radiological outcomes after clinical implementation of the same in-house planning workflow.

The primary aim of this study was to assess the precision of corrective osteotomy of malunited distal radius fractures 1 year after surgery using in-house–designed 3D patient-specific surgical guides. Secondary aims were to assess clinical and patient-reported outcome measures (PROMs) to determine whether using patient-specific surgical guides leads to improvements in function and/or satisfaction 1 year after surgery compared to before surgery. An additional aim was to document the use of intraoperative fluoroscopy.

Methods

Study Design and Population

A total of 16 consecutive adult patients requiring corrective osteotomy for a symptomatic extra-articular distal radius malunion were prospectively enrolled over a 2-year period from July 2021 to July 2023. The median age of the case series (n=14, 87.5% women and n=2, 12.5% men) was 61 (range 22 - 77) years. Exclusion criteria were bilateral distal radius deformity, cognitive impairment, psychiatric illness, active substance abuse, inability to read and understand the language, and any condition known to impair bone healing. The indication for surgery was pain in most patients (n=11, 68.8%). A total of 18.8% (n=3) of the patients experienced loss of range of motion (ROM), but only one of them had this as the only indication for surgery. In addition to corrective osteotomy of the distal radius, 18.8% (n=3) of the patients underwent concomitant ulnar shortening to avoid excessive distraction of the distal radius. The ulnar shortening procedures were also performed using 3D surgical planning and guides. A total of 6.3% (n=1) of the patients withdrew before surgery, and an additional 12.5% (n=2) of the patients were excluded from the final analysis after developing nonunion following a concurrent ulnar shortening procedure. These patients were transferred to standard clinical care and underwent revision fixation of the ulna using longer plates. Because they did not complete the study-specific follow-up, they were excluded from the final analysis.

Ethical Considerations

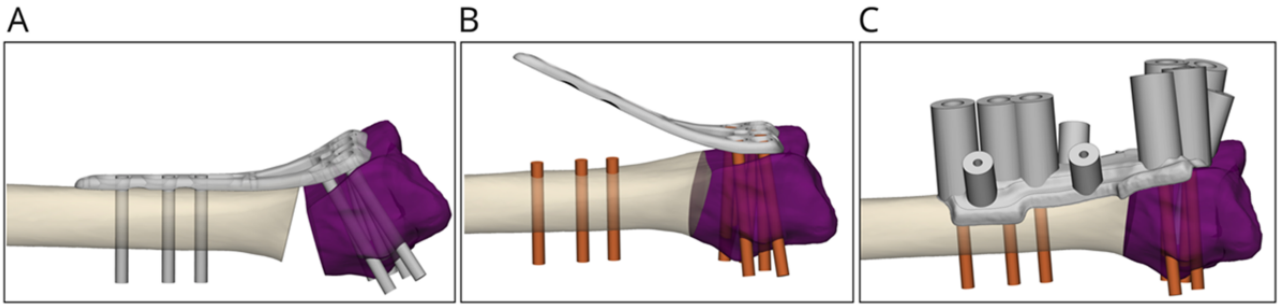
This study was approved by the Swedish Ethical Review Authority (2021-01974) and performed in accordance with the Declaration of Helsinki. All patients received written and oral information about the study and provided written informed consent prior to enrollment (all participants were adults; therefore, a legal guardian was not required).

3D Surgical Planning and Design of Patient-Specific Surgical Guides

Each patient underwent bilateral high-resolution computed tomography of the entire forearm using a standardized scanning protocol (100 kVp, 0.625-mm slice thickness, and convolution kernel: bone). Bilateral 3D virtual models of the forearm bones were created by segmenting the patients' Digital Imaging and Communications in Medicine images using the Mimics Medical software (Materialise NV). The uninjured forearm was mirrored, aligned at the proximal part of the radius, and used as a template. The maximum planned distraction was limited to approximately 12 to 14 mm. This threshold was used as a technical planning constraint based on clinical experience and published recommendations suggesting that concomitant ulnar shortening should be considered in patients with radial shortening exceeding approximately 12 mm [19]. It should not be interpreted as a formally validated cutoff. The amount of ulnar shortening was then planned to achieve bone contact in the radius osteotomy after fixation. Surgical planning and guide design were conducted in-house by a hand surgeon (KL) and a clinical engineer (CSS) using the 3-matic Medical software (Materialise NV) according to previously described methods [18,20] (Figure 1). Digital files of the in-house–designed guides were sent to a subcontractor (Prototal UK) to be 3D printed in accordance with the medical device manufacturing standard ISO 13485. The printing technology used was selective laser sintering on a 3D printer (EOS P 396; EOS GmbH) with a biocompatible polyamide raw material (PA 2200; EOS GmbH). The guides were quality controlled by the hospital's clinical engineers upon delivery from the subcontractor, repacked, labeled, and sterilized using autoclave according to the hospital's clinical routines (134°; 4 minutes). The complete manufacturing process was conducted in accordance with the European Union Medical Device Regulation 2017/745.

XSL•FO
RenderX

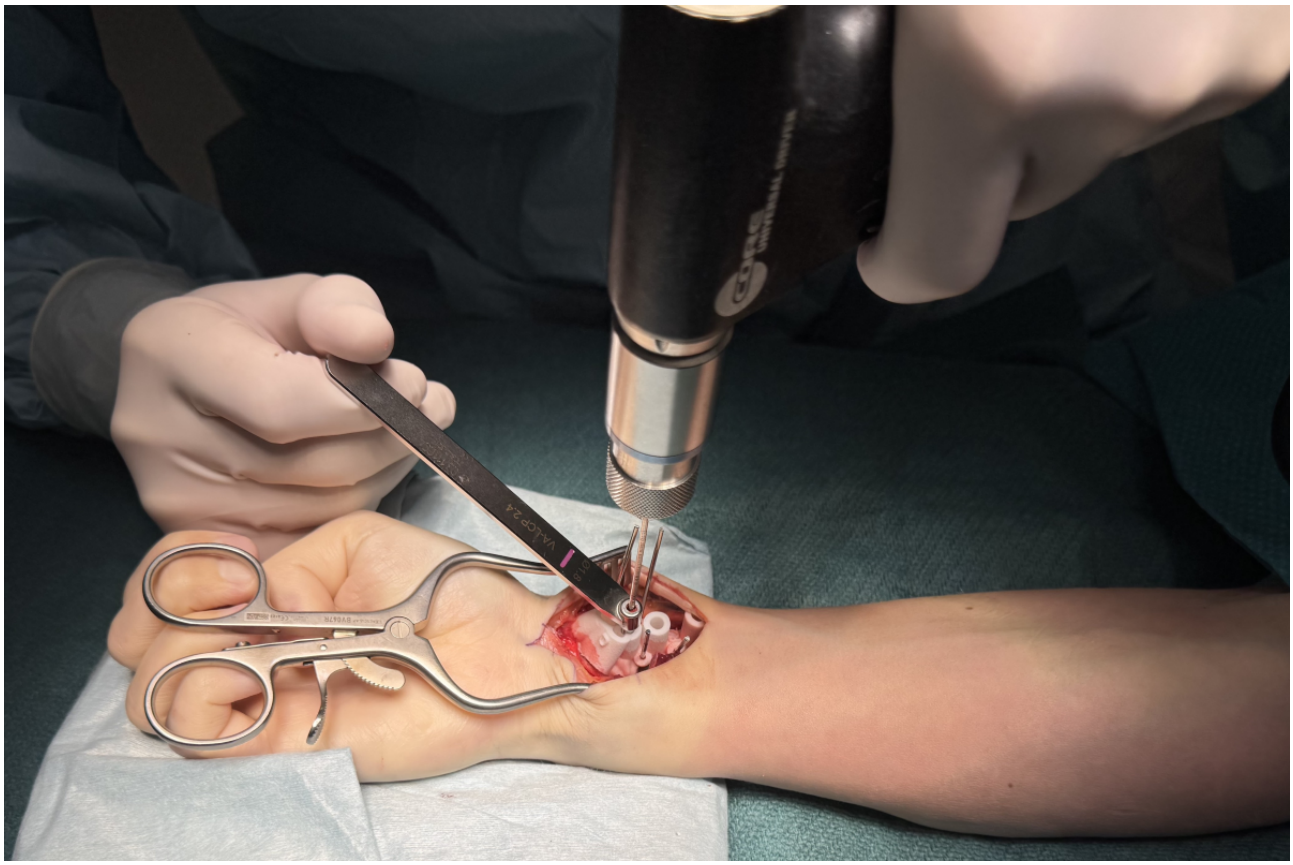
Figure 1. Construction of patient-specific surgical guides: virtual surgical planning (A); the distal bone fragment and osteotomy plate, including distal screws, reversed to the preoperative position (B); and guide design based on planned screw positioning (orange; C).



Surgical Procedure

All surgeries were performed by 2 senior hand surgery consultants (PA and KL) with experience level 4 according to Tang and Giddins [21]. Both surgeons had prior experience using externally purchased surgical guides. Patients were operated on under general anesthesia, regional anesthesia (brachial plexus block), or a combination of both. After exposure of the volar surface of the radius, the first drilling guide was placed on the bone and fixed with four 1.5-mm K-wires. Proper placement of the first guide was verified using a flat elevator to confirm continuous contact between the guide and the bone surface. The screw holes were predrilled using the guide's cylindrical sleeves (Figure 2).

Figure 2. Intraoperative photograph of predrilling screw holes using the surgical guide's cylindrical sleeves.



When all holes were drilled, the 2 parallel K-wires were left in the bone, and the other 2 K-wires were removed to allow for removal of the drill guide by sliding it off the remaining K-wires. The cutting guide was then positioned by sliding it down to the bone over the 2 parallel K-wires, and the osteotomy was carried out using a power saw. To fixate the bone, either a Variable Angle LCP Two-Column Plate (DePuy Synthes) or a VariAx 2 Distal Radius Plate (Stryker) was used depending on best anatomical fit during the 3D planning phase. The plate was fixated to the distal fragment first with locking screws and then reduced to the radius shaft using the previously drilled holes. If the osteotomy gap lacked bone contact, it was filled with iliac crest graft or residual bone from a concomitant ulnar shortening. If volar bone contact was present, no bone graft was used. Following wound closure, the wrist was immobilized for 2 weeks in a plaster cast. The duration of fluoroscopy use (seconds) was recorded by the radiographic fluoroscopy unit in the surgical room.

Rehabilitation

Two weeks postoperatively, the cast was replaced with a removable wrist brace, and patients initiated gentle wrist ROM

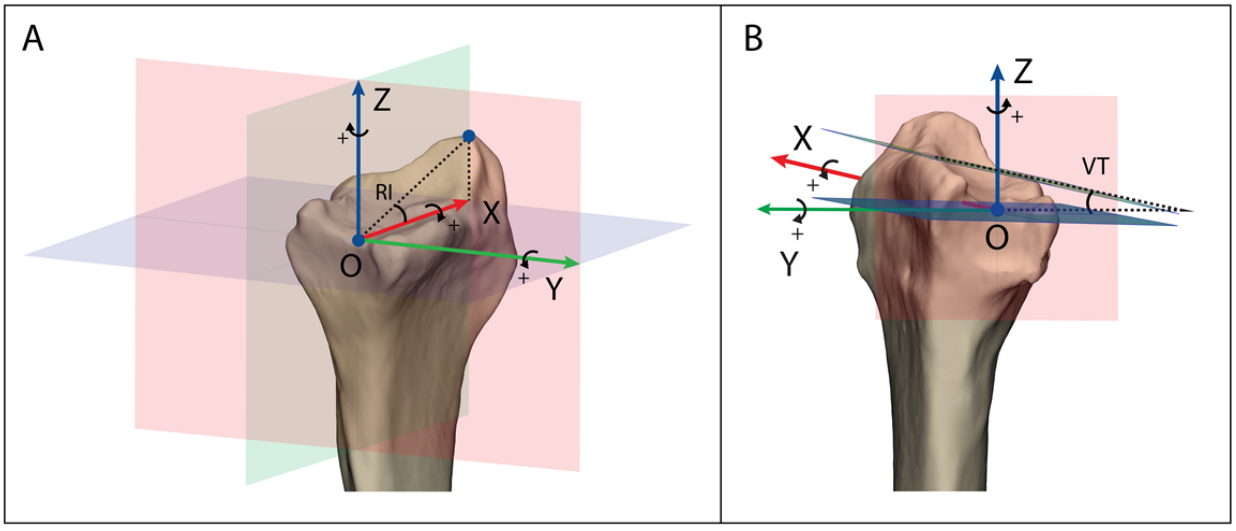
XSL•FO
RenderX

exercises 4 times daily. Light loading was permitted beginning at 4 weeks. After 6 weeks, brace use was typically discontinued, and full wrist loading was allowed provided that radiographic evidence confirmed initial bone healing.

Primary Outcome: Radiological Measurements

The primary outcome we report is the discrepancy between the 1-year postoperative computed tomography scan and the 3D virtual plan. Volar tilt, ulnar variance, and radial inclination were selected as clinically relevant outcome parameters. On the basis of prior literature, acceptable error margins were predefined as 5° or less for volar tilt and 2 mm or less for ulnar variance [3-5,22]. No specific error margins have been defined for radial inclination as previous studies did not find a significant correlation between radiographic measurements and clinical outcomes [23,24]. Using the 3-matic Medical software, an anatomical reference system was established to ensure clinical interpretability (Figures 3A and 3B) and described in our previous study [18]. Briefly, the origin (0, 0, 0) was positioned at the center of the rim between the lunate fossa and the sigmoid notch, similar to the central reference point used in standard radiographic assessments [18]. The central axis of the radius was defined as the intersection of 2 planes created using a datum-plane function (“two points, perpendicular to view”). The 2 points were manually placed on the radial shaft, one at 30 mm and one at 50 mm proximal to the central reference point using the anteroposterior and lateral views, respectively. The z-axis was aligned parallel to the radial shaft axis at the origin, the x-axis was directed toward the projected tip of the radial styloid and perpendicular to the z-axis, and the y-axis was defined as perpendicular to the x-z plane. Within this reference framework, rotation around the x-axis represented volar tilt, rotation around the y-axis corresponded to radial inclination, and translation along the z-axis represented ulnar variance.

Figure 3. Anatomical reference system and measurement of clinical parameters. “RI” refers to the angle between the x-y plane and the line from O to the radial styloid (A). “VT” refers to volar tilt, the angle between the x-y plane and the plane through the dorsal and volar rim tips (B). The system was adapted for left and right anatomy.



The postoperative and planned radius models were registered in 2 stages. First, the proximal radius was used as the stable reference. Corresponding regions covering approximately the proximal half of the radius were manually selected on the planned and postoperative models using a brush selection tool. The postoperative model was then registered to the planned model using an iterative closest-point algorithm restricted to the selected proximal surfaces.

Following this proximal registration, the coordinate system was reset to the anatomical reference frame of the virtual plan. Corresponding articular regions of the distal fragment were then manually selected on both models, and a second iterative closest-point registration was performed. The transformation between the planned and postoperative positions of the distal fragment was exported as a homogeneous transformation matrix from which the translational and rotational errors were calculated. A transformation matrix shows translation and rotation differences between the planned and postoperative 3D models in all 3 spatial dimensions (x, y, and z).

Secondary Outcomes: Patient-Reported and Clinical Outcome Measurements

PROM data were collected using the Disabilities of the Arm, Shoulder, and Hand (DASH) questionnaire [25]; the Patient-Rated Wrist Evaluation (PRWE); and the 100-mm visual analog scale for pain at rest and pain during activity. The total resulting scores from both the DASH and PRWE measures range from 0 to 100, where 0 indicates no disability or pain and 100 indicates the worst possible pain or disability [26].

Clinical outcome measurements were assessed preoperatively and 1 year postoperatively. Wrist ROM in extension, flexion, supination, pronation, radial deviation, and ulnar deviation was assessed using a goniometer. Grip strength was measured using a dynamometer (Jamar STD) [27]. Torque (ie, strength during forearm rotation) was measured in kilograms and converted into newton-meters using a dynamometer with a handle attached to a wall using a connection plate [28]. Lifting strength with one arm was measured in kilograms in a neutral wrist position using a hanging scale dynamometer (HCB 50K20; Kern & Sohn GmbH) [28]. Before strength testing, patients were allowed to

XSL•FO
RenderX

familiarize themselves with the dynamometer and the testing procedure. One measurement of each ROM and strength outcome was recorded at each assessment. The same measurement equipment was used throughout the study. Measurements were not repeated or averaged, and intraobserver reliability was therefore not assessed. The measurements were performed by a highly experienced hand surgeon (PA; level 4) [21] following the Swedish national manual for measurements in hand surgery [27].

Data Analysis

The equivalence bounds were defined as mean volar tilt error within –5° to +5° and mean ulnar variance error within –2 mm to +2 mm, representing the clinically acceptable margins based on previous literature. To assess whether postoperative radiological outcomes were equivalent to the preoperative virtual plans (ideal=0° or 0 mm), an equivalence test was performed using the two one-sided tests (TOST) procedure using JASP (University of Amsterdam). Equivalence testing uses TOST at an α value of .05 to assess whether the effect is neither too low nor too high. Combining these 2 tests yields a 90% CI, representing values that satisfy both 1-sided tests of 5%. Equivalence could be claimed if both TOST were significant, meaning that the entire 90% CI of the mean postoperative deviation lay within the defined intervals for volar tilt (−5°, +5°) and ulnar variance (−2 mm, +2 mm). Equivalence testing was performed using the signed differences between the planned and achieved outcomes to assess whether the mean directional error was contained within the predefined lower and upper equivalence bounds. In addition, mean absolute errors were calculated and reported descriptively to quantify the magnitude of the deviations irrespective of their direction (positive or negative).

Due to the small sample size, the secondary outcomes were reported using medians and IQRs, and a nonparametric Wilcoxon signed-rank test was applied to assess the differences between preoperative and postoperative measurements. A *P* value below .05 was considered statistically significant. No corrections for multiple comparisons were applied to secondary outcomes as they were considered exploratory.

Results

Radiological Patient Outcomes

For the radiological analysis, of the initial 16 patients, 13 (81.3%) were included (n=1, 6.3% withdrew prior to surgery, and n=2, 12.5% were moved to standard clinical care due to complications). A supplementary table summarizing the available sample size for each outcome and the reasons for missing data is provided in Multimedia Appendix 1, and a study flowchart is shown in Figure 4. An equivalence test was calculated based on the signed values and showed that the error in ulnar variance was considered equivalent to the preoperative plan. Both the lower (–2 mm) and upper (+2 mm) equivalence bounds were statistically significant (*P*<.001), and the 90% CI for the mean error was included in the equivalence boundaries. The signed mean error in volar tilt was not considered equivalent to the preoperative plan. The test against the upper bound (+5°) was significant (*P*<.001), but the test against the lower equivalence bound (−5°) was not (*P*=.33). Thus, the 90% CI for the mean deviation extended beyond the equivalence boundaries. Patient-level errors in ulnar variance and volar tilt, together with the cohort-level mean errors and their 90% CIs used for the equivalence analyses, are shown in Figures 5A and 5B.

XSL•FO
RenderX

Figure 4. Study flowchart of patient inclusion and data collection. CT: computed tomography; DASH: Disabilities of the Arm, Shoulder, and Hand; PRWE: Patient-Rated Wrist Evaluation.

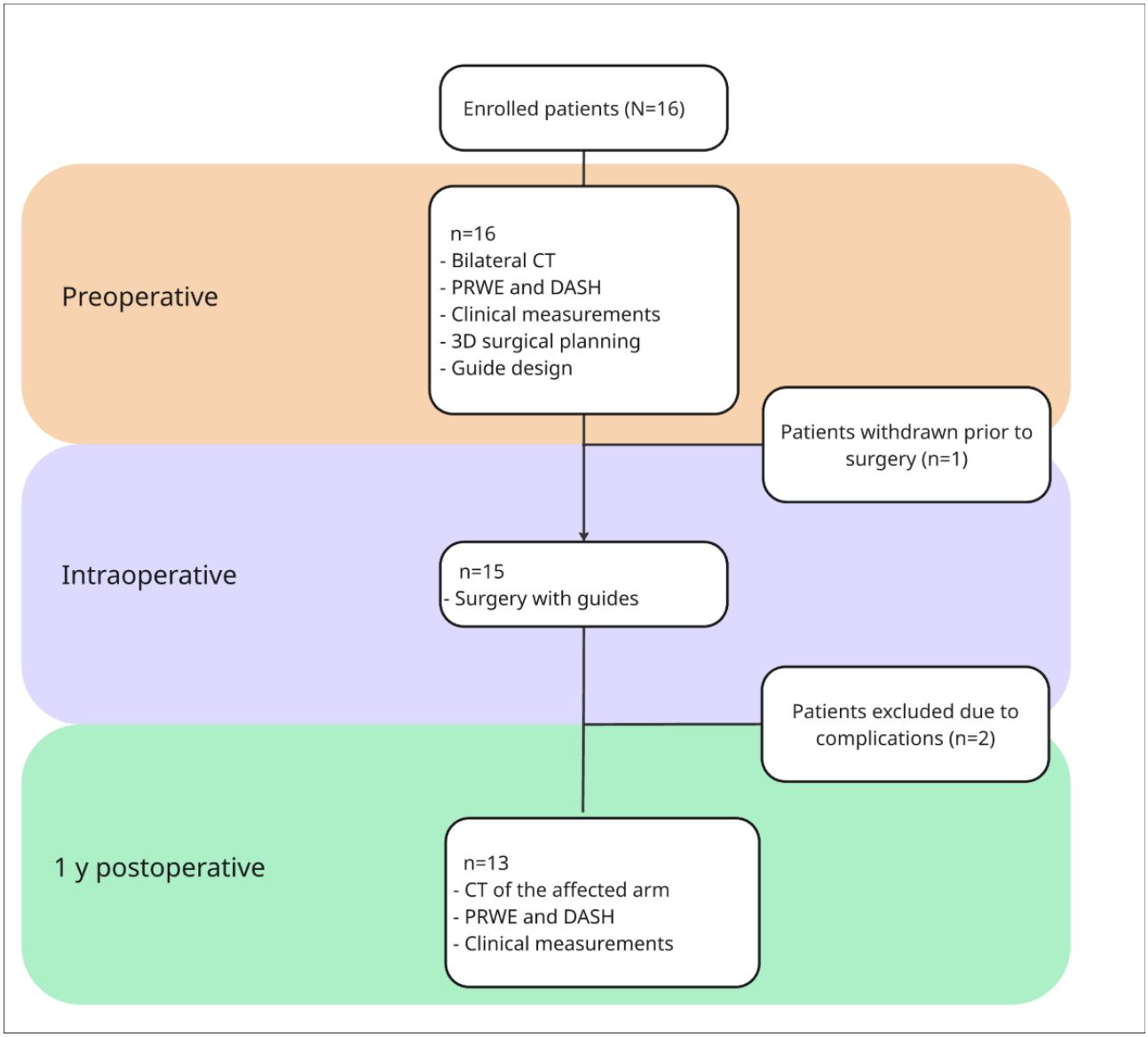
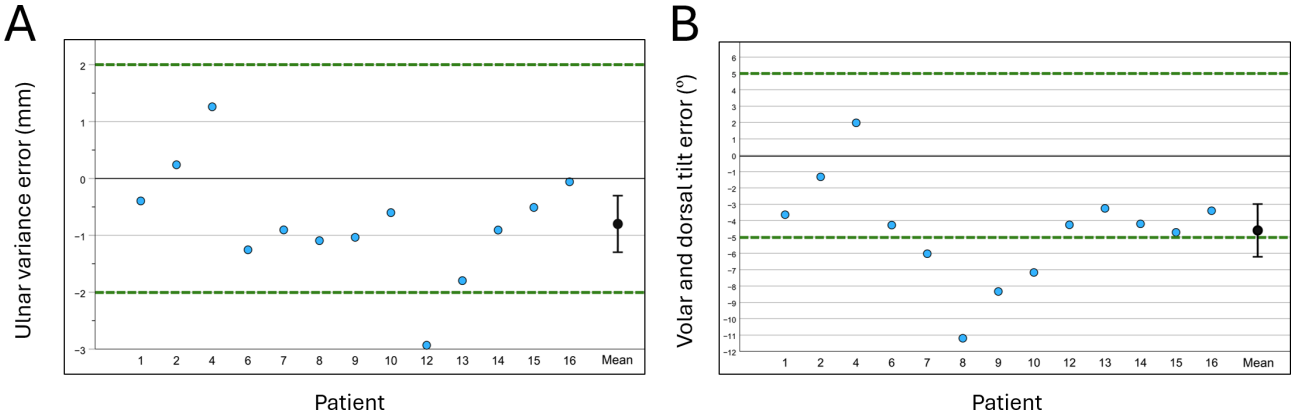


Figure 5. Signed ulnar variance error (A) and volar and dorsal tilt error (B) per patient case and mean (with 90% CI) based on the equivalence test. Clinically relevant and equivalence error margins are marked with a green dashed line (n=13).



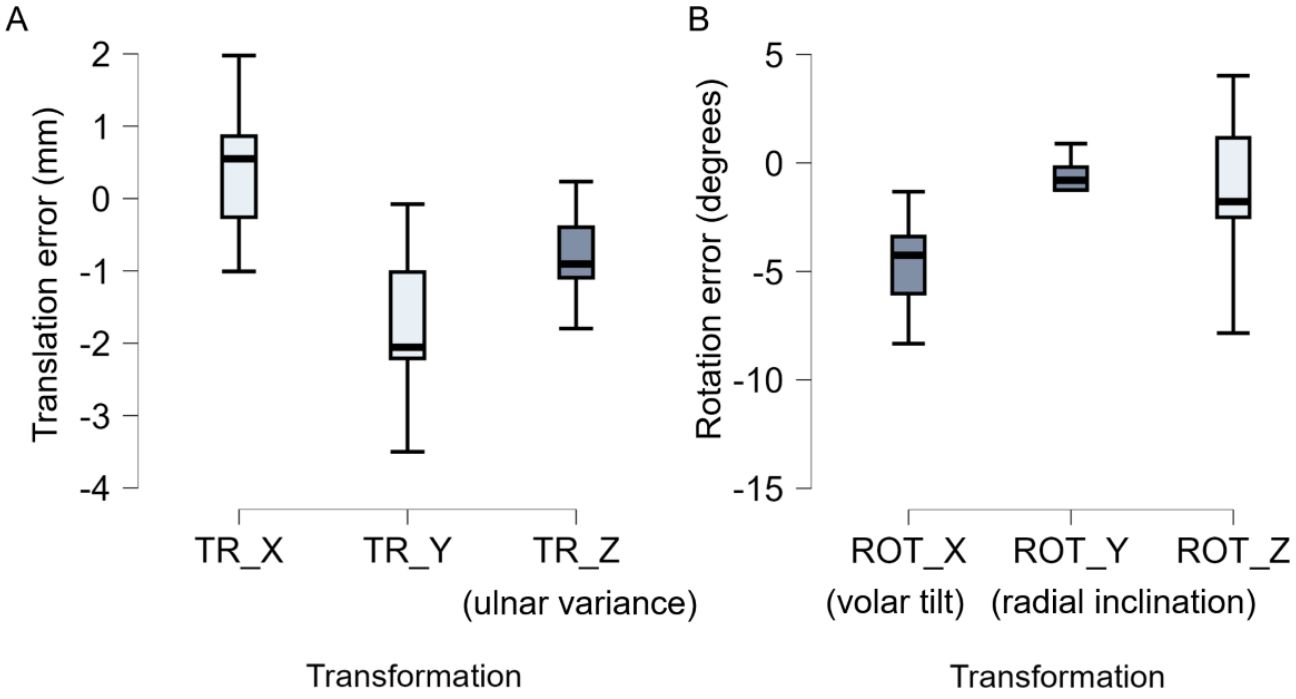
Mean absolute errors, representing the magnitude of deviation irrespective of direction, were analyzed for all translational and rotational components. The mean absolute translational errors were below 2 mm in all directions. Among the rotational components, the largest mean absolute error was observed for rotation corresponding to volar tilt. Detailed results for all 6 components are shown in Table 1. Box plots of the signed mean errors in all transformation directions are shown in Figures 6A and 6B.

XSL•FO
RenderX

Table . Mean absolute errors in all directions between the virtual plan and the postoperative models.

Absolute translational and rotational errors	Values, mean (SD)
Translation X-direction (mm)	0.8 (0.5)
Translation Y-direction (mm)	1.8 (1.0)
Ulnar variance (mm)	1.0 (0.8)
Volar tilt (degrees)	4.9 (2.7)
Radial inclination (degrees)	1.9 (1.7)
Rotation Z-direction (degrees)	4.0 (3.6)

Figure 6. Box plots of the signed mean error between the virtual plan and the postoperative models: translation error (TR) in the x, y, and z directions (A) and rotation error (ROT) in the x, y, and z directions (B; n=13).



PROMs

Preoperative and postoperative PROMs are shown in Table 2. PRWE and DASH scores were analyzed for the same patients included in the radiological analysis (n=13). Pain at rest and pain during activity were analyzed in 76.9% (10/13) of the patients because preoperative data were missing for an additional 23.1% (3/13). Numerical improvements from baseline were observed in PRWE, pain at rest, and pain at activity in all patients except 1 (9/10, 90%). DASH scores improved in all patients except 2 (8/10, 80%). These counts represent any reduction in score and should not necessarily be interpreted as clinically meaningful improvement. Statistically significant improvements were noted in all measurements (P<.05 in all cases).

Table . Medians and IQRs of the pre- and postoperative patient-reported outcome measurements.

	Patients, n	Preoperative, median (IQR)	Postoperative, median (IQR)	*P* value
PRWE[a] (0-100)	13	70 (60 - 77)	38 (11-57)[b]	.003
DASH[c] (0-100)	13	49 (27 - 58)	23 (8-36)[b]	.03
Pain at rest (0-100)	10	51 (15 - 76)	14 (1-20)[b]	.02
Pain at activity (0-100)	10	72 (53 - 82)	36 (11-47)[b]	.007

[a]PRWE: Patient-Rated Wrist Evaluation.

[b]The difference was statistically significant (P<.05).

[c]DASH: Disabilities of the Arm, Shoulder, and Hand.

XSL•FO
RenderX

Clinical Outcome Measurements

Preoperative and postoperative clinical outcome measurements are shown in Table 3. Lifting and grip strength were analyzed for the same patients included in the radiological analysis (n=13). Torque measurements were analyzed in 92.3% (12/13) of these patients because preoperative values were missing for 7.7% (1/13). A statistically significant improvement in lifting strength was observed postoperatively compared to preoperatively in 66.7% (8/12) of the patients. The remaining strength measurements were not statistically significant.

The ROM data for the affected arm both pre- and postoperatively are shown in Table 4.

Table . Medians and IQRs of the pre- and postoperative clinical outcome measurements.

	Patients, n	Preoperative, median (IQR)	Postoperative, median (IQR)	*P* value
Lifting strength in a neutral position (kg)	13	6.2 (4.2 - 9.1)	9.9 (6.2-12.3)[a]	.04
Grip strength (kg)	13	14.0 (6.0 - 23.0)	22.0 (12.5 - 30.0)	.15
Torque supination (Nm)	12	2.9 (1.9 - 4.4)	3.8 (2.9 - 4.5)	.13
Torque pronation (Nm)	12	2.0 (1.2 - 2.4)	2.2 (1.4 - 2.8)	.78

[a]The difference was statistically significant (*P*<.05).

Table . Range of motion of the affected arm pre- and postoperatively (n=13).

	Preoperative, median (IQR)	Postoperative, median (IQR)	*P* value
Extension (degrees)	60 (45 - 70)	60 (53 - 68)	.44
Flexion (degrees)	45 (30 - 60)	60 (50-63)[a]	.003
Supination (degrees)	75 (60 - 90)	70 (65 - 88)	.61
Pronation (degrees)	65 (60 - 70)	70 (63 - 80)	.33
Radial deviation (degrees)	17 (15 - 25)	20 (15 - 30)	.22
Ulnar deviation (degrees)	30 (25 - 35)	40 (40-48)[a]	.008

[a]The difference was statistically significant (*P*<.05).

Significant improvements in flexion and ulnar deviation were observed postoperatively relative to preoperatively. All patients except 1 (11/12, 91.7%) showed numerical improvement in these measurements. No significant improvement was observed in extension, supination, pronation, and radial deviation.

Intraoperative Fluoroscopy

The median intraoperative fluoroscopy duration was 56 (IQR 26 - 72) seconds (mean 54.9, SD 9.3 seconds) to allow for comparison with other studies. Fluoroscopy data were available for 12 procedures. Two patients who underwent a concomitant ulnar shortening procedure were excluded from the fluoroscopy analysis to avoid including fluoroscopy attributable to the additional procedure. The final fluoroscopy analysis therefore included 10 patients.

Complications

In our cohort including 15 patients (1/16, 6.3% withdrew prior to surgery), 4 (26.7%) required additional or revision surgery. Two of these patients developed nonunion following the concomitant ulnar shortening procedure and underwent revision fixation of the ulna using longer plates. Because they were transferred to standard clinical care and did not complete the study-specific follow-up, they were excluded from the final analysis of the primary radiological outcome. Nonunion occurred in the ulna for 2 out of 3 patients treated with concomitant ulnar shortening. In one of these patients, screw loosening in the distal radius also occurred, leading to the partial recurrence of radial deformity.

Implant removal and median nerve decompression were performed in one patient, with relief of symptoms. Another patient had persistent ulnar-sided wrist pain and mild symptoms of carpal tunnel syndrome. Resection of the ulnar styloid, arthroscopic shaving of a central triangular fibrocartilage complex tear, and median nerve decompression yielded good results. Additionally, one patient had transient increased chronic pain and returned to the same pain level as prior to surgery.

Discussion

This prospective study demonstrates that corrective osteotomy of distal radius malunions using in-house 3D surgical planning and patient-specific surgical guides can be performed with generally small deviations between the planned and postoperative correction. However, because the study did not include a control group, no conclusions can be drawn regarding whether this approach is more accurate than conventional 2D planning or other surgical techniques. The results showed equivalence in ulnar variance, whereas differences remained in volar tilt. The mean absolute alignment errors 1 year after surgery compared with the preoperative virtual plans were very low for ulnar variance (1, SD 0.8 mm) and radial inclination (1.9°, SD 3.6°), whereas a higher mean error was observed for volar tilt (4.9°, SD 2.7°), indicating less consistent correction.

XSL•FO
RenderX

The additional translational and rotational error components were also evaluated because deviations outside the conventional radiographic planes may have clinical consequences. Translation in the radioulnar and dorsovolar directions may influence mediolateral and anteroposterior positioning of the distal fragment, with potential effects on cortical alignment, distal radioulnar joint congruity, tendon clearance, and implant positioning. Rotation around the longitudinal axis of the radius may alter the rotational alignment of the distal fragment and potentially affect joint congruity and forearm mechanics. In the present study, the mean absolute translational errors were below 2 mm in all directions. Among the rotational components, the largest mean absolute error was observed for the component corresponding to volar tilt. However, clinically meaningful thresholds have not been established for these individual 3D error components, including radial inclination, and their clinical significance therefore cannot be determined definitively from the present data. Further studies should relate these errors to functional outcomes, patient-reported outcomes, and clinically relevant complications.

Upon closer case-level analysis, 38.5% (5/13) of the patients had errors outside the predefined margins. The ulnar variance error exceeded the predefined margin in 7.7% (1/13) of the patients, whereas the volar tilt error exceeded the margin in 30.8% (4/13) of the patients included in the analysis. A possible explanation for the volar tilt errors is the challenge of achieving the planned offset angle between the plate and the bone while using variable-angle screws in patients with reduced bone mineral density. As covered in earlier publications [11,20], an offset between the bone and the plate may be required to achieve anatomical correction using standard volar plates due to suboptimal fit on the malunited distal radius.

All cases except one had less volar tilt than planned. Contributing factors to the negative volar tilt error include soft tissue tension and settling of the radius on the distal locking screws. Volar plates achieve fixation primarily through the volar cortical bone [29] as screws provide limited purchase in the cancellous or dorsal cortex. Consequently, restoration and maintenance of volar tilt depend largely on a single point of fixation in the rotational plane. Our assumption is supported by the results of our prior surgical simulations on plastic bone models, which showed no errors outside the predefined margins in terms of volar tilt and radial inclination [18]. These discrepancies highlight the complexity of surgical procedures. Plastic models cannot replicate the compression forces, soft tissue involvement, and accessibility in the wound.

When comparing our results to those reported in the literature, the mean alignment errors are in line with those of similar studies. In a randomized controlled trial, Buijze et al [6] reported statistically significant improvement of volar tilt and radial inclination correction with 3D vs 2D planning. The residual errors after 3D planning were similar to ours, with residual deformity relative to the contralateral side (volar tilt: 4.1°; radial inclination: 2.2°; ulnar variance: 1.5 mm) [6]. However, their measurements were performed using conventional 2D radiographs. Compared to conventional 2D planning, the mean alignment errors using 3D planning were notably smaller. A study using 3D measurements of 2D planned osteotomies by Vroemen et al [30] found higher errors (volar tilt: −6.2°; radial inclination: 0.3°; ulnar variance: −2.2 mm). The 2D planning group in the study by Buijze et al [6] reported significantly higher errors in volar tilt and radial inclination but similar ulnar variance (volar tilt: 7.4°; radial inclination: 4.9°; ulnar variance: 1.2 mm) [6]. A study using conventional 2D planning by von Campe et al [22] reported that only 6 out of 15 patients satisfied the criteria of a correction within 5° angular deformity and 2-mm ulnar variance compared to the uninjured wrist.

The mean fluoroscopy duration in the present study was 54.9 (SD 9.3) seconds, which was similar to the 57.8 (SD 37.7) seconds reported for the 3D planned group by Buijze et al [6] and lower than the 140.3 (SD 101.4) seconds reported for their conventional 2D planned group. These findings suggest that the in-house 3D planning workflow did not increase fluoroscopy requirements and may reduce fluoroscopy use compared with conventional planning. This indicates that ionizing radiation exposure is reduced for both patients and medical personnel when 3D virtual planning and patient-specific surgical guides are used during corrective osteotomies—a finding that has also been demonstrated in several previous studies [6,31,32].

We observed postoperative improvements in PROM scores, lifting strength, flexion, and ulnar deviation, which were consistent with results reported in other studies [6,8,14]. The median improvements in PRWE and DASH scores were both above the minimum clinically important difference of 11.5 and 14 points, respectively [33]. The significantly improved PROM scores, despite the modest improvement in ROM, could be explained by a significant decrease in pain at rest and pain during activity, which allowed the patients to perform their everyday activities with less pain.

A total of 12.5% (2/16) of the patients required additional surgery, and 12.5% (2/16) required revision surgery due to complications related to ulnar shortening, which is an additional procedure carrying its own risk of complications. After in-depth incident analysis, it was concluded that the ulnar shortening failures were a result of suboptimal fixation in combination with pronounced shortening (5 and 10 mm). Although the complications were not directly related to 3D planning, we decided to return to conventional techniques for performing concomitant ulnar shortening while continuing to use 3D planning for distal radius malunions.

In this study, the cost of in-house planning and 3D-printed surgical guides was approximately €1100 per case (including costs for clinical engineers, surgeons, software licenses, and manufacturing; €1=US $1.17 as of August 27, 2026). The comparable cost from an external commercial provider was €2700 per case. However, the reported costs reflect only the planning and manufacturing workflow and do not include broader treatment-related costs, such as those associated with additional imaging, readmission, implant removal, treatment of nonunion, or revision surgery. These findings should therefore not be interpreted as a comprehensive analysis of total treatment costs.

This study has several limitations. It was not designed as a randomized trial, and it did not include a control group treated with conventional 2D methods. However, comparisons could

XSL•FO
RenderX

be made with other studies using 2D and 3D planning methods. Furthermore, the small sample size and missing outcome data reduced the precision and generalizability of the secondary findings and precluded meaningful subgroup analyses, such as comparisons between younger and older patients. The varying number of observations across outcomes may also have introduced attrition bias. Although all procedures were corrections of extra-articular distal radius fracture malunions, there was some heterogeneity as ulnar shortening was performed in 18.8% (3/16) of cases.

Future studies could assess outcomes following technical adjustments, including strategies to compensate for predictable deviations such as loss of volar tilt. By analyzing prior residual errors relative to the planned outcomes of larger datasets, we could better estimate the target volar tilt needed to achieve the desired correction. In addition, as osteoporotic bone tolerates less distraction than nonosteoporotic bone, future research should focus on identifying acceptable distraction limits according to bone quality.

In conclusion, the use of 3D virtual planning and patient-specific surgical guides in corrective osteotomies enables precise evaluation of outcomes by comparing preoperative plans with postoperative results. Residual alignment errors and complications in some cases underscore that, even with advanced planning tools, surgical outcomes remain influenced by patient-specific anatomical and biomechanical factors. Surgeons should be aware of potential insufficient corrections with this technique, particularly in volar tilt, and consider integrating anticipated adjustments during the planning phase. The results underscore the value of an in-house 3D planning team, which facilitates continuous quality improvement and refinement of surgical technique.

Acknowledgments

The authors thank Nesrin Vurgun, scientific editor at Sahlgrenska Academy, for the help with language editing and Johan Andersson at Sahlgrenska University Hospital for creating the animated ToC image. We used ChatGPT Edu, a generative AI tool, solely for final language editing.

Funding

This study was financed by grants from the Swedish state under the agreement between the Swedish government and the county councils, the ALF agreement (ALFGBG-966260).

Data Availability

Data are available on reasonable request.

Authors' Contributions

KL, PA, and AB conceived the study. CSS, PA, KL, and EG researched the literature. CSS, KL, and AB were involved in developing the study protocol and obtaining ethics approval. PA, PF, and AB were responsible for patient recruitment. CSS, KL, and EG performed in-house segmentation, surgical planning, and guide design. PA performed all clinical measurements. CSS and PA collected the data. CSS conducted data analysis and was lead writer of the manuscript. All authors reviewed and edited the manuscript and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Available sample size for each analysis and reasons for missing data.
[DOCX File, 16 KB - periop_v9i1e101764_app1.docx]

References

1. The Swedish Fracture Register [Article in Swedish]. Svenska Frakturregistret. URL: https://sfr.registercentrum.se/in-english/the-swedish-fracture-register/p/HyEtC7VJ4 [accessed 2026-08-28]
2. Raudasoja L, Vastamäki H, Aspinen S, Ryhänen J, Hulkkonen S. Distal radial fractures: a nationwide register study on corrective osteotomies after malunion. J Hand Surg Eur Vol 2024 Mar;49(3):329-333. [doi: 10.1177/17531934231193849] [Medline: 37694946]
3. Aro HT, Koivunen T. Minor axial shortening of the radius affects outcome of Colles' fracture treatment. J Hand Surg Am 1991 May;16(3):392-398. [doi: 10.1016/0363-5023(91)90003-t] [Medline: 1861016]
4. McQueen M, Caspers J. Colles fracture: does the anatomical result affect the final function? J Bone Joint Surg Br 1988 Aug;70(4):649-651. [doi: 10.1302/0301-620X.70B4.3403617] [Medline: 3403617]
5. Prommersberger KJ, Van Schoonhoven J, Lanz UB. Outcome after corrective osteotomy for malunited fractures of the distal end of the radius. J Hand Surg Br 2002 Feb;27(1):55-60. [doi: 10.1054/jhsb.2001.0693] [Medline: 11895348]

XSL•FO
RenderX

6. Buijze GA, Leong NL, Stockmans F, et al. Three-dimensional compared with two-dimensional preoperative planning of corrective osteotomy for extra-articular distal radial malunion: a multicenter randomized controlled trial. J Bone Joint Surg Am 2018 Jul 18;100(14):1191-1202. [doi: 10.2106/JBJS.17.00544] [Medline: 30020124]
7. van Es EM, Dijkhof M, Souer JS, et al. Forearm rotation improves after corrective osteotomy in patients with symptomatic distal radius malunion. Heliyon 2024;10(9):e29570. [doi: 10.1016/j.heliyon.2024.e29570] [Medline: 38765076]
8. Andreasson I, Kjellby-Wendt G, Fagevik-Olsén M, Aurell Y, Ullman M, Karlsson J. Long-term outcomes of corrective osteotomy for malunited fractures of the distal radius. J Plast Surg Hand Surg 2020 Apr;54(2):94-100. [doi: 10.1080/2000656X.2019.1693392] [Medline: 31762349]
9. Bauer DE, Zimmermann S, Aichmair A, et al. Conventional versus computer-assisted corrective osteotomy of the forearm: a retrospective analysis of 56 consecutive cases. J Hand Surg Am 2017 Jun;42(6):447-455. [doi: 10.1016/j.jhsa.2017.03.024] [Medline: 28434832]
10. Ma B, Kunz M, Gammon B, Ellis RE, Pichora DR. A laboratory comparison of computer navigation and individualized guides for distal radius osteotomy. Int J Comput Assist Radiol Surg 2014 Jul;9(4):713-724. [doi: 10.1007/s11548-013-0966-8] [Medline: 24323402]
11. Roner S, Carrillo F, Vlachopoulos L, Schweizer A, Nagy L, Fuernstahl P. Improving accuracy of opening-wedge osteotomies of distal radius using a patient-specific ramp-guide technique. BMC Musculoskelet Disord 2018 Oct 15;19(1):374. [doi: 10.1186/s12891-018-2279-0] [Medline: 30322393]
12. Benayoun M, Langlais T, Laurent R, et al. 3D planning and patient-specific surgical guides in forearm osteotomy in children: radiographic accuracy and clinical morbidity. Orthop Traumatol Surg Res 2022 Oct;108(6):102925. [doi: 10.1016/j.otsr.2021.102925] [Medline: 33845175]
13. Oka K, Murase T, Moritomo H, et al. Accuracy of corrective osteotomy using a custom-designed device based on a novel computer simulation system. J Orthop Sci 2011 Jan;16(1):85-92. [doi: 10.1007/s00776-010-0020-4] [Medline: 21359511]
14. de Muinck Keizer RJ, Lechner KM, Mulders MA, Schep NW, Eygendaal D, Goslings JC. Three-dimensional virtual planning of corrective osteotomies of distal radius malunions: a systematic review and meta-analysis. Strategies Trauma Limb Reconstr 2017 Aug;12(2):77-89. [doi: 10.1007/s11751-017-0284-8] [Medline: 28444580]
15. Miyake J, Murase T, Moritomo H, Sugamoto K, Yoshikawa H. Distal radius osteotomy with volar locking plates based on computer simulation. Clin Orthop Relat Res 2011 Jun;469(6):1766-1773. [doi: 10.1007/s11999-010-1748-z] [Medline: 21203873]
16. Schweizer A, Fürnstahl P, Nagy L. Three-dimensional correction of distal radius intra-articular malunions using patient-specific drill guides. J Hand Surg Am 2013 Dec;38(12):2339-2347. [doi: 10.1016/j.jhsa.2013.09.023] [Medline: 24189159]
17. Stockmans F, Dezillie M, Vanhaecke J. Accuracy of 3D virtual planning of corrective osteotomies of the distal radius. J Wrist Surg 2013 Nov;2(4):306-314. [doi: 10.1055/s-0033-1359307] [Medline: 24436834]
18. Stor Swinkels C, Libberecht K, Gryska E, Axelsson P, Fredrikson P, Björkman A. Preclinical paired noninferiority study comparing in-house and commercially available 3D planning for corrective osteotomy of the distal radius. Sci Rep 2025 Apr 12;15(1):12652. [doi: 10.1038/s41598-025-97788-5] [Medline: 40221557]
19. Prommersberger KJ, Pillukat T, Mühldorfer M, van Schoonhoven J. Malunion of the distal radius. Arch Orthop Trauma Surg 2012 May;132(5):693-702. [doi: 10.1007/s00402-012-1466-y] [Medline: 22294090]
20. Gryska E, Libberecht K, Stor Swinkels C, Axelsson P, Fredrikson P, Björkman A. Improving standard volar plate fixation in 3D-guided corrective osteotomy of the distal radius: evaluation of a shim instrument. J Plast Surg Hand Surg 2024 May 15;59:46-52. [doi: 10.2340/jphs.v59.39839] [Medline: 38747532]
21. Tang JB, Giddins G. Why and how to report surgeons' levels of expertise. J Hand Surg Eur Vol 2016 May;41(4):365-366. [doi: 10.1177/1753193416641590] [Medline: 27076602]
22. von Campe A, Nagy L, Arbab D, Dumont CE. Corrective osteotomies in malunions of the distal radius: do we get what we planned? Clin Orthop Relat Res 2006 Sep;450:179-185. [doi: 10.1097/01.blo.0000223994.79894.17] [Medline: 16721354]
23. Mulders MA, Detering R, Rikli DA, Rosenwasser MP, Goslings JC, Schep NW. Association between radiological and patient-reported outcome in adults with a displaced distal radius fracture: a systematic review and meta-analysis. J Hand Surg Am 2018 Aug;43(8):710-719.e5. [doi: 10.1016/j.jhsa.2018.05.003] [Medline: 29908929]
24. Schmidt V, Gordon M, Tägil M, Sayed-Noor A, Mukka S, Wadsten M. Association between radiographic and clinical outcomes following distal radial fractures: a prospective cohort study with 1-year follow-up in 366 patients. J Bone Joint Surg Am 2023 Aug 2;105(15):1156-1167. [doi: 10.2106/JBJS.22.01096] [Medline: 37172109]
25. Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand) [corrected]. The Upper Extremity Collaborative Group (UECG). Am J Ind Med 1996 Jun;29(6):602-608. [doi: 10.1002/(SICI)1097-0274(199606)29:6<602::AID-AJIM4>3.0.CO;2-L] [Medline: 8773720]
26. MacDermid JC, Tottenham V. Responsiveness of the Disability of the Arm, Shoulder, and Hand (DASH) and Patient-Rated Wrist/Hand Evaluation (PRWHE) in evaluating change after hand therapy. J Hand Ther 2004;17(1):18-23. [doi: 10.1197/j.jht.2003.10.003] [Medline: 14770134]
27. Manual för rörelse- och styrkemätning av armbåge, underarm och hand. : HAKIR; 2026 URL: https://rcsyd.se/hakir/wp-content/uploads/sites/23/2026/03/HAKIR-matmanual-2026-1.pdf [accessed 2026-09-01]

XSL•FO
RenderX

28. Axelsson P, Kärrholm J. New methods to assess forearm torque and lifting strength: reliability and validity. J Hand Surg Am 2018 Sep;43(9):865.e1. [doi: 10.1016/j.jhsa.2018.01.009] [Medline: 29454600]
29. Orbay JL, Touhami A. Current concepts in volar fixed-angle fixation of unstable distal radius fractures. Clin Orthop Relat Res 2006 Apr;445:58-67. [doi: 10.1097/01.blo.0000205891.96575.0f] [Medline: 16505728]
30. Vroemen JC, Dobbe JG, Strackee SD, Streekstra GJ. Positioning evaluation of corrective osteotomy for the malunited radius: 3-D CT versus 2-D radiographs. Orthopedics 2013 Feb;36(2):e193-e199. [doi: 10.3928/01477447-20130122-22] [Medline: 23380014]
31. Zheng W, Su J, Cai L, et al. Application of 3D-printing technology in the treatment of humeral intercondylar fractures. Orthop Traumatol Surg Res 2018 Feb;104(1):83-88. [doi: 10.1016/j.otsr.2017.11.012] [Medline: 29248764]
32. Chen C, Cai L, Zhang C, Wang J, Guo X, Zhou Y. Treatment of die-punch fractures with 3D printing technology. J Invest Surg 2018 Oct;31(5):385-392. [doi: 10.1080/08941939.2017.1339150] [Medline: 28722489]
33. Walenkamp MM, de Muinck Keizer RJ, Goslings JC, Vos LM, Rosenwasser MP, Schep NW. The minimum clinically important difference of the patient-rated wrist evaluation score for patients with distal radius fractures. Clin Orthop Relat Res 2015 Oct;473(10):3235-3241. [doi: 10.1007/s11999-015-4376-9] [Medline: 26040969]

Abbreviations

DASH: Disabilities of the Arm, Shoulder, and Hand
PROM: patient-reported outcome measure
PRWE: Patient-Rated Wrist Evaluation
ROM: range of motion
TOST: two one-sided tests

Edited by AAD Arcaya; submitted 21.May.2026; peer-reviewed by C Smees, MD Savran; revised version received 11.Jul.2026; accepted 21.Aug.2026; published 09.Sep.2026.

Please cite as:
Stor Swinkels C, Axelsson P, Libberecht K, Fredrikson P, Gryska E, Björkman A
In-House 3D Planning for Corrective Osteotomy of the Distal Radius: Prospective Cohort Study With 1-Year Radiological and Clinical Follow-Up
JMIR Perioper Med 2026;9:e101764
URL: https://periop.jmir.org/2026/1/e101764
doi:10.2196/101764

© Charlotte Stor Swinkels, Peter Axelsson, Katleen Libberecht, Per Fredrikson, Emilia Gryska, Anders Björkman. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 9.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Benchmark Development for Fundamental Arthroscopic Skills Using a Simulation-Based Training Program: Observational Study

Eric Davis[1], MBA; Brianna Caraet[2], MD; Robert Pedowitz[3], MD, PhD; Gregg Nicandri[1], MD

[1]Department of Orthopaedic Surgery, University of Rochester, 601 Elmwood Avenue, Rochester, NY, United States

[2]Kaiser Permanente Orthopedic Sports Medicine, San Diego, CA, United States

[3]Department of Orthopaedic Surgery, University of California, Los Angeles, CA, United States

Corresponding Author:
Eric Davis, MBA
Department of Orthopaedic Surgery, University of Rochester, 601 Elmwood Avenue, Rochester, NY, United States

Abstract

Background: Surgical education has shifted from the traditional Halstedian apprenticeship model toward incorporating simulation due to work-hour restrictions, increasing case complexity, and economic and liability pressures. Building on the success of the Fundamentals of Laparoscopic Surgery program for general surgery, the Fundamentals of Arthroscopic Surgery Training (FAST) program was developed to establish proficiency benchmarks for orthopedic trainees in basic arthroscopic skills.

Objective: We aimed to establish benchmarks for 5 FAST workstation modules.

Methods: Sports medicine fellowship–trained faculty members were given instructions on the modules and 2 minutes of practice time, and they then performed each task 3 times with both their dominant and nondominant hand. For each module, mean faculty performance was used to establish an efficiency benchmark (time) and precision benchmark (errors).

Results: The Probing module should be completed in less than 95 seconds with no errors. The Ring Transfer module should be completed in less than 134 seconds with no more than 1 error. The Maze module should be completed in less than 99 seconds with no errors. The Meniscectomy module should be completed in less than 68 seconds with no more than 1 error. Lastly, the Suture Passing module should be completed in less than 195 seconds with no more than 1 error.

Conclusions: The FAST workstation can be used as a proficiency-based learning tool for residents to safely and effectively develop arthroscopic skills outside of the operating room. These benchmarks were established via a method previously validated in surgical simulation and balance precision and efficiency for skills that are considered generalizable and transferable to arthroscopic surgeries.

(JMIR Perioper Med 2026;9:e82723) doi:10.2196/82723

KEYWORDS

surgical education; orthopedic surgery; arthroscopy; simulation training; skills assessment; resident education

Introduction

Historically, surgical education has been based upon the Halstedian apprenticeship model, in which surgical trainees primarily acquire technical skills in the operating room while assisting and later performing procedures on live patients. In recent years, there has been a paradigm shift away from this approach in orthopedic surgery [1]. This can be attributed to multiple factors, including but not limited to resident work-hour restrictions, increased case complexity and variety, the risk of liability, and social and economic pressure. Simulation in surgical education has been suggested as a potential avenue to fill the gap in training, allowing trainees opportunities to obtain valuable experience in a low-risk, more cost-effective environment. In concordance with this, the Accreditation Council for Graduate Medical Education now requires that residency programs include surgical skills training in their curricula; this mandate has also been echoed by the American Board of Orthopaedic Surgery and the Residency Review Committee.

Simulation as an educational tool has been validated extensively in the field of general surgery. The Fundamentals of Laparoscopic Surgery program, developed in 1997 by the Society of American Gastrointestinal and Endoscopic Surgeons, combines cognitive and physical skills components shown to improve a given student's laparoscopic abilities over the course of the curriculum [2]. The program has been studied rigorously and validated since its release [3-9]. Currently, a passing score is required by the American Board of Surgery as prerequisite to sit for the certification exam. This has been well received by trainees, surgeons, and the public. As a result of this successful program in general surgery, the field of orthopedics has explored the development of a similar program.

XSL•FO
RenderX

In 2011, the Arthroscopy Association of North America (AANA), the American Board of Orthopaedic Surgery, and the American Academy of Orthopaedic Surgeons came together in a joint effort to establish the Fundamentals of Arthroscopic Surgery Training (FAST) program. The program is based on the tenet that fundamental surgical skills are best obtained sequentially, and that proficiency can be established through the successful completion of a basic skills curriculum. A curriculum and simulator have been developed for the FAST program and proficiency benchmarks for knot tying have been established.

The aim of this study was to develop time- and error-based proficiency benchmarks for 5 fundamental arthroscopic skills modules within a simulation-based training program.

Methods

Overview

The AANA offered a voluntary skills training course for residents and faculty at the Orthopaedic Learning Center in Chicago, Illinois, on December 7 to 9, 2017, and January 18 to 20, 2018. During downtime in the course, participants were recruited and provided with an information sheet outlining participation in this study. Two separate cohorts of 20 faculty each completed orthopedic simulations to improve their skill. For each module, the participant was given a list of written instructions and allowed to practice for a period of 2 minutes. Then, each participant performed the task 3 times with their dominant hand and 3 times with their nondominant hand. The order of dominant/nondominant completion was randomized so that exactly half of the faculty started with their nondominant hand. The time it took each participant to complete the module as well as the number of errors committed was recorded. Cohort 1 completed the Probing, Ring Transfer, and Maze Navigation modules. Cohort 2 completed the Meniscectomy and Suture Passing modules.

All modules were completed using camera-based visualization rather than live arthroscopes. Depending on the module, visualization was provided using either a USB camera system or a stationary camera, consistent with the simulation equipment available in the skills laboratory. This approach was used to standardize task execution across participants and to support reproducibility in a controlled training environment.

Five simulation-based arthroscopic skills modules were evaluated, including Probing, Ring Transfer, Maze Navigation, Partial Meniscectomy, and Suture Passage. Each module assessed fundamental psychomotor skills relevant to arthroscopy, with efficiency measured by task completion time and precision measured by predefined error criteria. Detailed descriptions of each module are provided in Multimedia Appendix 1.

Descriptive statistics were used for demographic variables, experience, handedness, efficiency (time) score, and precision (error) score. Mean performance time for all participants was determined. Any value lying more than 2 SDs from the mean was excluded. Exclusion of values greater than 2 SDs from the mean was performed to minimize the influence of extreme outliers that may reflect momentary lapses, equipment issues, or deviations from standardized task execution rather than true expert performance. This approach has been used in prior simulation-based benchmarking studies to derive stable proficiency thresholds that reflect typical expert performance rather than maximal or anomalous values. Outlier exclusion affected a small proportion of observations and did not meaningfully alter the relative ordering of module performance or the final benchmark thresholds. The trimmed mean was used as the basis for determining the proficiency level benchmark for each task. Efficiency benchmarks were set at the minimum threshold of the mean time of the expert cohort. Precision benchmarks were set at the mean number of errors committed by participants rounded to the nearest whole number. The benchmark was established as the mean performance when data from both hands were aggregated. Aggregation of dominant and nondominant hand performance was chosen a priori to reflect the ambidextrous demands of arthroscopic surgery, in which instrument laterality varies by procedure and portal placement. Use of the mean performance across both hands was intended to establish benchmarks that promote bilateral skill development while remaining feasible and generalizable for trainee progression.

Ethical Considerations

This study was reviewed by the University of Rochester Research Subjects Review Board and was granted exempt status with a waiver of documentation of informed consent. All participants were provided with an information sheet outlining the purpose, procedures, and voluntary nature of the study prior to participation. All data collected were deidentified. Each participant was assigned a unique study identification number that contained no personally identifiable information. No names or direct identifiers were recorded at any point in data collection or analysis. To protect participant privacy and confidentiality, data were recorded using study IDs only and stored securely in accordance with institutional research guidelines. Only members of the research team had access to the data. Results are presented in aggregate form to prevent identification of individual participants. Participation in this study was voluntary. Participants were not provided with any financial compensation or incentives for participation, and there were no costs associated with participation.

Results

Demographics

A total of 40 sports medicine fellowship–trained faculty surgeons were included in the analysis. The mean age of the participants was 43.9 (SD 9.0) years. They had an average of 10.5 (SD 8.3) years in practice and performed approximately 277 (SD 125) arthroscopic cases per year. All were male and 3 (8%) were left-handed. There were no significant differences in demographic data between cohorts from the 2 courses (Table 1).

XSL•FO
RenderX

Table . Demographics.

	Cohort 1 (n=20)	Cohort 2 (n=20)	Overall (n=40)	*P* value[b]
Age, mean (SD)	43.1 (8.38)	44.7 (9.22)	43.9 (9.0)	.59
Male gender, n (%)	20 (100)	20 (100)	40 (100)	—[a]
Percent fellowship trained, n (%)	20 (100)	20 (100)	40 (100)	—
Years in practice, mean (SD)	9.5 (7.7)	11.4 (8.5)	10.5 (8.3)	.48
Cases per year (n), mean (SD)	275 (130.1)	279 (117.3)	277 (125)	.94
Right-handed, n (%)	18 (90)	19 (95)	37 (93)	.55

[a]Not applicable.

[b]For age, number of years in practice, and number of cases per year, 2-sided *t* tests were used to test for statistical significance. For handedness, the χ^2 test was used. No differences were statistically significant.

Probing

For the Probing module, mean completion times were 92.5 (SD 17.6) seconds for the dominant hand and 97.3 (SD 13.6) seconds for the nondominant hand, with no errors recorded and no significant difference between hands (*P*=.35); the benchmark was <95 seconds with no errors.

Ring Transfer

For the Ring Transfer module, mean completion times were 117.5 (SD 26.7) seconds for the dominant hand and 151.3 (SD 38.1) seconds for the nondominant hand, with mean error rates of 0.6 (SD 0.8) and 1.1 (SD 0.9) dropped rings, respectively; dominant and nondominant hand performance differed significantly (*P*<.001), and the benchmark was <134 seconds with no more than 1 dropped ring.

Maze

For the Maze module, mean completion times were 89.1 (SD 24.3) seconds for the dominant hand and 108.6 (SD 25.0) seconds for the nondominant hand, with mean error rates of 0.2 (SD 0.4) and 0.1 (SD 0.3) lost balls, respectively; dominant and nondominant hand performance differed significantly (*P*=.02), and the benchmark was <99 seconds with no errors.

Meniscectomy

For the Meniscectomy module, mean completion times were 73.2 (SD 20.7) seconds for the dominant hand and 63.7 (SD 23.6) seconds for the nondominant hand, with mean error rates of 0.1 (SD 0.3) over- or underresection events for both hands; no significant difference was observed between dominant and nondominant hand performance (*P*=.14), and the benchmark was <68 seconds with no more than 1 area of over- or underresection.

Suture Passing

For the Suture Passing module, mean completion times were 177.5 (SD 41.2) seconds for the dominant hand and 211.9 (SD 40.9) seconds for the nondominant hand, with mean error distances of 0.5 (SD 0.6) mm and 0.6 (SD 0.8) mm from the target, respectively; dominant and nondominant hand performance differed significantly (*P*<.001), and the benchmark was <195 seconds with no more than 1 mm from the target and no suture anchor unload.

For all modules, benchmark time was defined as the mean performance across dominant and nondominant hands to reflect bilateral task demands. For example, if the right hand took 60 seconds and the left hand took 30 seconds, the benchmark time was 45 seconds. Benchmarks for all modules are summarized in Table 2.

Table . Benchmarks.

Module	Benchmark time (seconds)[a]	Benchmark errors
Probing module	<95	No errors
Ring Transfer module	<134	No more than 1 dropped ring
Maze module	<99	No balls off platform
Meniscectomy module	<68	Only 1 area of over- or underresection
Suture Passing module	<195	<1 mm from the target area and no suture anchor unloads

[a]Benchmark time is the mean performance for both hands. For example, if the right hand takes 60 seconds, and the left hand takes 30 seconds, the time to compare against the benchmark is 45 (90/2) seconds.

XSL•FO
RenderX

Discussion

Principal Findings

In this study, we established objective time- and error-based proficiency benchmarks for 5 fundamental arthroscopic skills modules using performance data from fellowship-trained sports medicine faculty members. These benchmarks were derived using a standardized, previously validated simulation-based methodology and were designed to balance efficiency and precision across tasks of increasing technical complexity. Defining proficiency thresholds based on expert performance involves an inherent trade-off. Benchmarks that are too lenient may allow progression without adequate skill consolidation, whereas overly stringent thresholds may discourage practice or be impractical within residency training constraints. By deriving benchmarks from aggregated expert performance and excluding extreme outliers, the present thresholds were designed to balance feasibility with meaningful skill assessment. Continued evaluation of trainee outcomes and future transfer-validity studies will be important to further calibrate these benchmarks.

The FAST program consists of a curriculum, a simulator, and a series of tests, which enable trainees to iteratively practice fundamental skills outside of the operating room and ideally before performing surgery on patients. Training such motor skills outside of the OR is crucial in arthroscopy, which has a steep learning curve [10]. The FAST program's hands-on design is in alignment with the Institute of Medicine's recommendations for graduate medical education, which aim to transition from process-driven training to a more outcome-driven approach. Instead of the assumption that residents acquire their needed skills after completing a certain number of procedures or after a certain amount of time in training, the institute recommends evaluating proficiency in such skills before advancement [11]. The FAST program uses this proficiency-based progression approach to training to encourage "deliberate practice" to improve motor skills [12]. With the establishment of benchmarks for each FAST module, a resident is able to receive real-time objective feedback on their performance and is encouraged to strive to be comparable to experienced arthroscopy surgeons. With the accessibility of the workstations, residents are able to obtain additional practice repetitions and adjust their technique as needed to meet the established standard. Recent evaluations of the FAST workstation have demonstrated measurable differences in performance across postgraduate training levels and institutions, further supporting the need for standardized, objective benchmarks [13].

The FAST program can be integrated as early as the intern year in orthopedic training so that young surgeons can develop foundational psychomotor skills before even touching an arthroscope in the operating room. This early exposure to arthroscopic instruments and techniques can instill confidence in a new resident and their abilities, help them develop good habits early, and potentially decrease patient morbidity. Studies have demonstrated that medical students and residents who train with simulators perform better in the operating room than their counterparts without simulator training, with recent arthroscopy-specific work demonstrating transfer validity to diagnostic knee arthroscopy and meniscectomy performance [14,15]. With improved resident performance in the operating room, a faculty member walking a resident through the case would likely feel more comfortable increasing resident autonomy, and valuable operating room time can be spent teaching and instructing on the unique nuances of a particular case instead of on basic skills.

The FAST workstations are easily accessible, cost-effective, and efficient. They require a simple set-up, and at our institution, they are available in the skills laboratory for all residents to use at any time. Residents can easily independently set up the workstation in the laboratory as long as they have an arthroscopy tower or USB camera with which to practice. Currently, we have each of the testing modules set up in front of a computer monitor with all the instruments available that are necessary to complete the task. We have a QR code that links to a video of the task and instructions mounted on the wall above each of the modules (Figure 1). We use iPevo cameras when a stationary light source is sufficient (Knot Tying) and Sawbones USB cameras when a replica of an arthroscope is required (Ring Transfer). We have found that having easy-to-access instructions and all modules already set up increases use of the simulator and maximizes resident time. We typically introduce FAST during a 2-to-3-hour session with all interns that is proctored by an instructor during their skills month. This familiarizes them with the simulator and modules. They are expected to practice to proficiency and then use the simulators to refine their skills prior to their arthroscopic rotations. The modular design of the FAST program also supports scalability across institutions with varying resource availability and learner profiles, as the workstations can be implemented using differing levels of visualization technology and integrated at multiple stages of training. This flexibility allows programs to adapt FAST implementation to local constraints while maintaining standardized proficiency benchmarks.

XSL•FO
RenderX

Figure 1. Simulation laboratory set-up.



Prior papers have aimed to establish clear quantitative benchmarks for simulation training based on the mean performance of an experienced surgeon cohort [16]. One such paper, by Pedowitz et al [17], established the benchmark for the Knot Tying module in the FAST program, which is the last module to be completed in the series. It evaluated the performance of 50 faculty members in attendance at AANA resident arthroscopy courses. The 2 faculty cohorts had an average of 19.3 and 19.9 years of experience in practice. A benchmark was established using data from the more successful of the 2 faculty cohorts and established proficiency as equal to 2 knot failures or less out of 5 knot attempts when using the knot tester workstation. This randomized, prospective study then evaluated a training group of 44 postgraduate year 4 or 5 orthopedic residents. They were divided into 3 subgroups: group A received standard didactic training, group B was allowed additional knot-tying workstation practice, and group C received proficiency-based progression training with the knot-tying workstation. While the aggregate resident knot failure rate of 26% was higher than the 22% knot failure rate of the faculty, resident group C had only an 11% knot failure rate with 94% of residents in this group passing the threshold. This paper both demonstrated a system to establish a benchmark standard and proved that proficiency-based progression training improves the likelihood of meeting said benchmarks. Recent FAST-based educational interventions have demonstrated improvements in objective performance metrics, including arthroscopic knot integrity, when guided, proficiency-based training is used [18]. The arthroscopic knot-tying station is the last in the sequence of the FAST workstation modules as it tests one of the more complex skills in arthroscopy. The purpose of the current investigation was to establish objective benchmarks for the remaining 5 FAST workstation modules: Probing, Ring Transfer, Maze Navigation, Meniscectomy, and Suture Passing.

These benchmarks derived from the performance of fellowship-trained sports medicine faculty members were designed to be both realistic and achievable for residents. The 6 modules progress from basic skills to more complex tasks. The first module, Probing, allows residents to establish basic arthroscopic skills of horizontal control, telescoping, periscoping, and triangulation in order to feel comfortable performing a diagnostic arthroscopy. Gaining familiarity with the arthroscope in this module lays the foundation for establishing visualization in the remainder of the tasks and in the operating room. The second module, Ring Transfer, reinforces these skills and allows the resident to develop proficiency with the arthroscopic grasper, which is used when they are tasked with performing an arthroscopic loose-body removal. The dexterity with the instruments developed in this module is applicable to using arthroscopic instruments to manipulate tissues or implants in all arthroscopic procedures. The next module, Maze Navigation, builds on prior probing skills from the first module and develops tracking skills with nonstationary objects. This skill is necessary to develop before interventional arthroscopy can be performed. While the first 3 modules focus more on skills generalization, the final 3 modules correlate even more closely with specific surgical interventions, thereby focusing more on skills transfer. The fourth module directly simulates completing a partial meniscectomy with a benchmark designed to encourage an adequate amount of resection. During this practice, residents develop more familiarity with a biter and the ability to maneuver more deliberately and precisely. Resection of a paper “meniscus” is a safer alternative to a new resident damaging cartilage with the biter or over-resecting and destabilizing a meniscus tear because they are using a biter for the first time in an actual patient’s knee. The second to last task, Suture Passing, integrates the use of other new instruments, including an antegrade suture passer, a piercing suture passer/retriever, and a suture lasso or suture shuttle. This task’s surgical correlate is an arthroscopic rotator cuff repair or Bankart repair, which are among the more complex arthroscopic procedures. Completion of this task relies on the mastery and coordination of skills obtained in the previous modules. The last module, Knot Tying, allows residents to evaluate the biomechanical integrity of their arthroscopic knots in a setting safer than discovering loose knots in a failed arthroscopic suture repair. In conclusion, a resident’s ability to meet the time constraints and limit their errors for each task

XSL•FO
RenderX

translates into more efficient and accurate performance in the operating room, which promotes patient safety.

In establishing the benchmarks, there was a statistically significant difference in performance between the faculty surgeons' dominant and nondominant hands with the Ring Transfer, Maze, and Suture Passing modules. However, surgeons have to be ambidextrous. Which hand holds the arthroscope and which holds the remaining instruments is often based on the laterality of the procedure. Previous studies have shown that experienced arthroscopic surgeons are more ambidextrous than novices [19]. Specifically, experts demonstrate significantly smaller dominant-nondominant differences in task completion time and error rates on simulated arthroscopic tasks compared with novices, supporting bilateral performance assessment as a marker of skill acquisition. Therefore, the FAST program requires all tasks to be completed with the right hand and the left hand in separate attempts in order to develop skills in both hands regardless of hand dominance. Establishing separate benchmarks for the dominant and the nondominant hand based on the differences seen in the expert cohorts was considered. Aggregated performance metrics are commonly used in established simulation benchmarking frameworks, such as the Objective Structured Assessment of Technical Skill (OSATS) and the Fundamentals of Laparoscopic Surgery program, which define proficiency using composite measures rather than limb-specific performance. This supports the use of bilateral performance means when establishing reproducible and generalizable training benchmarks. That said, the scientific ideal needs to be weighed equally with practicality in this case. The FAST Program puts multiple new demands on trainees and faculty in the programs that adopt it. Ultimately, in order to be successfully adopted, it needs to be easy to implement. Based upon our experience facilitating many FAST courses at the Orthopaedic Learning Center in Chicago and at multiple residency programs, we determined that a single benchmark for each task was necessary.

We also discussed whether to take the mean performance of both hands or to set the benchmark at the level of the dominant hand. We chose the mean performance. This was the best compromise between ensuring adequate ambidexterity while also recognizing that differences exist, even in expert faculty members. The FAST program is designed to enhance surgical skills and promote deliberate, repetitive practice for benchmarks. We were concerned that setting the benchmark as the performance of the dominant hand of the faculty members would have made it significantly harder to achieve for the trainees. The goal was for residents to be able to pass the FAST program by the end of their third year in training, and we decided that the likely significant increase in repetitions required to meet the more stringent benchmark would be frustrating and time consuming for residents who already have many other competing demands. Passing the FAST program at the end of the third year does not mean that residents are proficient at arthroscopy. It means that they have sufficiently practiced and likely have improved their fundamental arthroscopic skills by meeting a minimum threshold of performance. They still have additional years in residency to build upon the skills acquired in the FAST program. Future validation of FAST benchmarks could be pursued through prospective randomized or longitudinal studies comparing proficiency-based FAST training with standard curricula. Studies should be powered to detect differences in operative performance and stratified by postgraduate year to assess differential benefit by training stage. Translation to operating room performance may be measured using objective metrics such as task-specific error rates, time to task completion, and validated global rating scales (eg, OSATS) during comparable arthroscopic procedures. This benchmarking is foundational work that is required to enable such an investigation.

The development of the FAST program benchmarks is one of the factors that allow this to be a self-sufficient progression learning tool, setting it apart from other skills programs. In 2007, the Carolinas Medical Center integrated a laparoscopic skills curriculum into their general surgery residency program [20]. A skills laboratory coordinator needed to be present to record the duration of participation in each task and keep track of errors. In contrast, the FAST program is designed as a web-based program residents can access independently at a time convenient for them. A stopwatch is built into the program to record time to complete a task, and each task has a section to input the number of errors. These factors determine progression to the next module, all of which can be done at the resident's own pace. The paper about the Carolinas Medical Center program by Stefanidis et al [20] describes feedback occurring after each training session or when a resident is called in by the coordinator. With the FAST program, instant feedback is available with the clear benchmarks established for each task.

Limitations and Future Studies

A limitation of this study was that the included participants were exclusively faculty members teaching at the AANA Fundamentals of Arthroscopic Surgery Residents Course. All participants were male, sports medicine fellowship–trained faculty members, which may limit generalizability across genders, subspecialties, and training environments. Ultimately, this was a convenience sample due to the significant logistical hurdles to obtaining data with another method from a large number of expert orthopedic surgeons. This may have introduced a selection bias for surgeons, and they may not be fully representative of a wider population of sports fellowship–trained surgeons. The included cohort were deemed experts by virtue of their completed training, but there were outliers in performance among the participants. There were 21 (of 320) outlier data points, and these were subsequently excluded from the dataset before formulating the benchmarks.

Another limitation of this study is that it did not demonstrate the feasibility of the FAST program or its transfer validity, necessitating further studies. Further investigation is also needed to determine the average number of repetitions and practice time required to meet the proficiency benchmarks. These results could then be subanalyzed based on training level. This would provide a clearer estimated duration of FAST programs for other orthopedic residency programs interested in integrating the FAST program into their curricula. Further application of the FAST program could also be used to assess whether those who reach proficiency benchmarks perform better than those with

XSL•FO
RenderX

no simulation training on other simulated surgical procedures and in the operating room for comparable surgical procedures. Lastly, further studies are needed to investigate which subset of residents achieve the most benefit from the FAST program as determined by the greatest improvement in arthroscopic skills. This could be used to determine in which year of residency training the FAST program should be integrated.

Benchmarks in this study were derived on a module-specific basis using different expert cohorts and should be interpreted as reference values for individual skills rather than curriculum-level proficiency standards. This design limits direct cross-module comparison but reflects practical constraints of expert data collection and supports initial benchmark development for discrete arthroscopic tasks. Evaluation of learners across the full FAST curriculum will be necessary to assess proficiency progress and program-level validity.

Conclusions

FAST workstations can be used as proficiency-based learning tools for residents to safely and effectively develop arthroscopic skills outside of the operating room. These benchmarks were established via a method previously validated in surgical simulation and balance precision and efficiency for skills that are considered generalizable and transferable to arthroscopic surgeries.

Funding

This study was supported by an educational grant from the Arthroscopy Association of North America (AANA) in the amount of $44,000. Study activities were conducted in part at the AANA Orthopaedic Learning Center. The funding source had no role in the design of the study; analysis; interpretation of the data; or in the writing of the manuscript or decision to submit for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed procedural descriptions.
[DOCX File, 2779 KB - periop_v9i1e82723_app1.docx]

References

1. Pedowitz RA, Marsh JL. Motor skills training in orthopaedic surgery: a paradigm shift toward a simulation-based educational curriculum. J Am Acad Orthop Surg 2012 Jul;20(7):407-409. [doi: 10.5435/JAAOS-20-07-407] [Medline: 22751159]
2. Derossis AM, Fried GM, Abrahamowicz M, Sigman HH, Barkun JS, Meakins JL. Development of a model for training and evaluation of laparoscopic skills. Am J Surg 1998 Jun;175(6):482-487. [doi: 10.1016/s0002-9610(98)00080-4] [Medline: 9645777]
3. McCluney AL, Vassiliou MC, Kaneva PA, et al. FLS simulator performance predicts intraoperative laparoscopic skill. Surg Endosc 2007 Nov;21(11):1991-1995. [doi: 10.1007/s00464-007-9451-1] [Medline: 17593434]
4. Soper NJ, Fried GM. The fundamentals of laparoscopic surgery: its time has come. Bull Am Coll Surg 2008 Sep;93(9):30-32. [Medline: 19469354]
5. Scott DJ, Bergen PC, Rege RV, et al. Laparoscopic training on bench models: better and more cost effective than operating room experience? J Am Coll Surg 2000 Sep;191(3):272-283. [doi: 10.1016/s1072-7515(00)00339-2] [Medline: 10989902]
6. Seymour NE, Gallagher AG, Roman SA, et al. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Ann Surg 2002 Oct;236(4):458-463. [doi: 10.1097/00000658-200210000-00008] [Medline: 12368674]
7. Hyltander A, Liljegren E, Rhodin PH, Lönroth H. The transfer of basic skills learned in a laparoscopic simulator to the operating room. Surg Endosc 2002 Sep;16(9):1324-1328. [doi: 10.1007/s00464-001-9184-5] [Medline: 11988802]
8. Korndorffer JR Jr, Dunne JB, Sierra R, Stefanidis D, Touchard CL, Scott DJ. Simulator training for laparoscopic suturing using performance goals translates to the operating room. J Am Coll Surg 2005 Jul;201(1):23-29. [doi: 10.1016/j.jamcollsurg.2005.02.021] [Medline: 15978440]
9. Sroka G, Feldman LS, Vassiliou MC, Kaneva PA, Fayez R, Fried GM. Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room-a randomized controlled trial. Am J Surg 2010 Jan;199(1):115-120. [doi: 10.1016/j.amjsurg.2009.07.035] [Medline: 20103076]
10. Atesok K, Mabrey JD, Jazrawi LM, Egol KA. Surgical simulation in orthopaedic skills training. J Am Acad Orthop Surg 2012 Jul;20(7):410-422. [doi: 10.5435/JAAOS-20-07-410] [Medline: 22751160]
11. Committee on the Governance and Financing of Graduate Medical Education, Board on Health Care Services, Institute of Medicine. In: Eden J, Berwick DM, Wilensky GR, editors. Graduate Medical Education That Meets the Nation's Health Needs: National Academies Press (US); 2014. [Medline: 25340242]
12. Thomas GW, Johns BD, Marsh JL, Anderson DD. A review of the role of simulation in developing and assessing orthopaedic surgical skills. Iowa Orthop J 2014;34:181-189. [Medline: 25328480]

XSL•FO
RenderX

13. Chapek M, Otlans PT, Buuck T, et al. Resident performance on the Fundamentals of Arthroscopic Surgery Training workstation does not predictably improve with postgraduate year. Arthrosc Sports Med Rehabil 2024 Feb;6(1):100866. [doi: 10.1016/j.asmr.2023.100866] [Medline: 38318395]
14. Palter VN, Grantcharov T, Harvey A, Macrae HM. Ex vivo technical skills training transfers to the operating room and enhances cognitive learning: a randomized controlled trial. Ann Surg 2011 May;253(5):886-889. [doi: 10.1097/SLA.0b013e31821263ec] [Medline: 21394017]
15. Tronchot A, Casy T, Vallee N, et al. Virtual reality simulation training improve diagnostic knee arthroscopy and meniscectomy skills: a prospective transfer validity study. J Exp Orthop 2023 Dec 14;10(1):138. [doi: 10.1186/s40634-023-00688-8] [Medline: 38095746]
16. Gallagher AG, O'Sullivan GC. Fundamentals of Surgical Simulation: Principles and Practice: Springer; 2011.
17. Pedowitz RA, Nicandri GT, Angelo RL, Ryu RKN, Gallagher AG. Objective assessment of knot-tying proficiency with the Fundamentals of Arthroscopic Surgery Training program workstation and knot tester. Arthroscopy 2015 Oct;31(10):1872-1879. [doi: 10.1016/j.arthro.2015.06.021] [Medline: 26298642]
18. Farooq H, Gaetano A, Salazar D, Garbis N. Guided education and the Fundamentals of Arthroscopy Surgery Training (FAST) workstation improve surgical resident knot-tying skills. Arthrosc Sports Med Rehabil 2025 Jun;7(3):101110. [doi: 10.1016/j.asmr.2025.101110] [Medline: 40692921]
19. Pedowitz R, Nicandri G, Tuchschmid S. Asymmetry in dominant / non-dominant hand performance differentiates novices from experts on an arthroscopy virtual reality serious game. Stud Health Technol Inform 2016;220:289-294. [Medline: 27046594]
20. Stefanidis D, Acker CE, Swiderski D, Heniford BT, Greene FL. Challenges during the implementation of a laparoscopic skills curriculum in a busy general surgery residency program. J Surg Educ 2008;65(1):4-7. [doi: 10.1016/j.jsurg.2007.11.009] [Medline: 18308276]

Abbreviations

AANA: Arthroscopy Association of North America
FAST: Fundamentals of Arthroscopic Surgery Training
OSATS: Objective Structured Assessment of Technical Skill

Edited by N Rohatgi; submitted 25.Aug.2025; peer-reviewed by D Demirel, W Li; revised version received 22.Feb.2026; accepted 25.Feb.2026; published 08.Apr.2026.

Please cite as:
Davis E, Caraet B, Pedowitz R, Nicandri G
Benchmark Development for Fundamental Arthroscopic Skills Using a Simulation-Based Training Program: Observational Study
JMIR Perioper Med 2026;9:e82723
URL: https://periop.jmir.org/2026/1/e82723
doi:10.2196/82723

© Eric Davis, Brianna Caraet, Robert Pedowitz, Gregg Nicandri. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 8.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Machine Learning for Predicting Critical Postoperative Interventions: Proof-of-Concept Study Using the INSPIRE Dataset

Manan Shukla[1], BS; Paul Fodor[2], PhD; Suresh Yelika[3], MD; Nicholas Ahn[3], MD

[1]Renaissance School of Medicine, Stonybrook University Hospital, 118 Leeward Lane, Port Jefferson, NY, United States

[2]Department of Computer Science, Stony Brook University, Stony Brook, NY, United States

[3]Colorectal Surgery, Stony Brook Medicine, Stony Brook, NY, United States

Corresponding Author:
Manan Shukla, BS
Renaissance School of Medicine, Stonybrook University Hospital, 118 Leeward Lane, Port Jefferson, NY, United States

Abstract

Background: Postoperative complications contribute significantly to patient morbidity and mortality. Early prediction of such complications could enable the care team to intervene promptly and improve patient outcomes. Existing surgical risk scores are easy to use but lack accuracy and do not provide individualized risk or guidance for clinical decision-making.

Objective: This study aimed to develop and evaluate machine learning models using the INSPIRE perioperative dataset to predict whether an individual patient will require critical postoperative interventions—ventilatory support, extracorporeal membrane oxygenation (ECMO), an intra-aortic balloon pump (IABP), or continuous renal replacement therapy (CRRT).

Methods: Four artificial neural network models and 4 random forest classifiers were trained and tested using the publicly available INSPIRE dataset (131,000 surgical cases). Preoperative laboratory data, medication use, surgery type, and intraoperative vital signs were used as inputs. The dataset was randomly divided into a 70% training set (91,000 cases) and a 30% test set (38,998 cases), with a temporal split to prevent leakage. Model performance was assessed using accuracy, sensitivity, specificity, positive and negative predictive values, and area under the curve (AUC).

Results: The artificial neural network models achieved high predictive performance for ECMO (accuracy 98.9%; AUC 0.992), ventilatory support (accuracy 97.7%; AUC 0.995), IABP (accuracy 98.0%; AUC 0.992), and CRRT (accuracy 94.9%; AUC 0.983). In the test set of 38,998 patients, the neural network models correctly classified 38,569 cases for ECMO, 38,097 for ventilatory support, 38,230 for IABP, and 37,021 for CRRT, corresponding to accuracies of 98.9%, 97.7%, 98.0%, and 94.9%, respectively. The corresponding random forest classifiers achieved accuracies of 98.0%, 99.0%, 98.2%, and 96.5%, respectively, with similar AUC values. These results indicate that the models can reliably identify patients who will or will not require lifesaving postoperative interventions despite the marked class imbalance. Our work began in April 2023, with the research concept taking shape over the following 1 to 2 months. Model development spanned approximately 3 to 4 months and was followed by 1 month of testing and validation conducted by other computer scientists.

Conclusions: Machine learning models trained and tested on the INSPIRE dataset can provide individualized risk estimates for critical postoperative interventions. By supplying accurate predictions before or during surgery, these models can support proactive perioperative planning (eg, intensive care unit bed allocation and device readiness).

(JMIR Perioper Med 2026;9:e65327) doi:10.2196/65327

KEYWORDS

postoperative complication prediction; surgical risk; machine learning; extracorporeal membrane oxygenation; ECMO; postoperative ventilation; intra-aortic balloon pump; continuous renal replacement therapy; CRRT; patient management

Introduction

Prevention and early detection of complications are often the best treatments in medicine. This is especially true in surgery, where particularly life-threatening complications can occur acutely. As a result, a significant amount of research has focused on how postoperative complications can be accurately predicted at an earlier period and how proactive management based on this information may improve patient outcomes. Over the past few decades, this has mostly occurred with the development of scores or scales that can potentially predict morbidity and mortality risk. A clear example of this is the Surgical Apgar Score (SAS), which uses estimated blood loss, mean arterial pressure, and heart rate to predict patient morbidity or mortality [1]. While such types of scales are easy to use, they happen to have an overall lower accuracy [2]. After conducting a randomized controlled experiment, Manstein et al [2] found the surgical APGAR's receiver operating characteristic curve to be only 0.661. The authors concluded that "the ability of the SAS

XSL•FO
RenderX

to accurately predict risk of postoperative SC [serious complication] at the patient level was limited." We believe that this limitation is due to the fact that the biomarkers, laboratory values, and vital signs that are clinically relevant predictors of complications for one patient may differ from those that are most informative for another. Therefore, while a scoring system such as the SAS can likely predict postoperative morbidity or mortality in a large percentage of the patient population, there is always a certain subset of patients in which the scale will fail to predict morbidity or mortality (which is where the scoring system fails). Therefore, there is a need to develop a prediction system that can more accurately predict postoperative morbidity and mortality at the individual level and, consequently, at the population level.

Machine learning (ML) has proved to be very proficient in solving such types of problems. Defined as a study of algorithms designed to learn from data and generate insights from unseen data, ML can be used to potentially predict complications based on data such as laboratory results, vital signs, and medication information. Many models that can successfully predict postoperative morbidity (such as postoperative kidney injury [3] or delirium [4]) and mortality have been developed based on this concept. However, although they may be useful for a clinician in predicting a future diagnosis, such models do not provide insights on patient management. By limiting a model's training to just diagnosis, the utility of the model is also thereby constrained. Knowledge that a patient may develop postoperative kidney injury may encourage a surgeon to be more careful and vigilant during and before the surgery, but planning postoperative management would still be very difficult (especially in time-sensitive situations).

In this work, we aimed to expand upon the current state of the art by creating a proof-of-concept model that encourages proactivity by not only successfully predicting postoperative complications but also providing the physician with clinical decision support to best manage the patient if or when complications occur. We achieve this by creating a set of models that may predict whether a patient will require a certain treatment in the postoperative setting, such as a lifesaving medical device (eg, a ventilator). This can also serve as a proxy for certain management decisions (eg, whether a patient should be transferred to the intensive care unit [ICU] instead of the postanesthesia care unit postoperatively). We then evaluated the model in a dataset comprising 131,000 surgical cases to evaluate the accuracy and robustness of the system.

Methods

Overview

At its core, the proposed system attempts to predict whether a patient will require the use of lifesaving devices or services in the postoperative period. The overall architecture of the model and design decisions regarding engineering of the model are discussed below.

The engineered model is an artificial neural network composed of 4 fully connected (dense) layers with SoftMax activation, followed by a final sigmoid activation layer. This 4-layer architecture yielded the highest performance across accuracy, precision, and area under the curve (AUC) metrics. The neural network and activation function was decided upon based on hyperparameter optimization and data type (continuous and dichotomous data). Preoperative laboratory data, medication data, and surgery type served as the input to the model, along with intraoperative vital signs. To clarify, our model uses similar input variables as existing risk scores but uses a novel ML architecture and tuning. The model was developed to predict various different binary labels: extracorporeal membrane oxygenation (ECMO), mechanical ventilation, continuous renal replacement therapy, and intra-aortic balloon pump (IABP). These 4 parameters were chosen due to 3 specific reasons. First, medical devices often require the most amount of time (eg, compared to a medication delivery) to find and set up, especially if their use is not expected. Consequently, the utility of the model is most significantly improved in this setting. A similar rationale applies to predicting postoperative ICU admission; if hospital staff can anticipate the need for ICU care after surgery, appropriate preparations can be made in advance, such as reserving a bed. Note that while we do not model for postoperative ICU admissions (due to dataset constraints), all of the devices that serve as labels in this work require an ICU level of care. Second, the medical devices involved in this model can often dramatically improve a patient's outcome if used in a timely manner, hence enhancing the utility of the model. Finally, the treatment can serve as a proxy for the patient's diagnosis. For example, a patient requiring a ventilator or ECMO postoperatively indicates some risk of respiratory failure after surgery. This can potentially enhance a physician's ability to prevent a complication from happening in the first place. Note that the model is designed to theoretically predict use of other potentially lifesaving devices as well. However, as this is a proof-of-concept study, the scope was limited to the devices described above due to dataset constraints.

The model was trained (defined as the process of feeding data into the model to enable pattern recognition) and evaluated using a dataset named INSPIRE [5], which comprises data from 131,000 patients undergoing surgery from various surgical specialties. Specifically, vital signs from the operating room, ward, and ICU; laboratory results from 6 months before and after discharge; and medication data and general patient characteristics (such as age, sex, and diagnosis) were included. Additionally, complications such as hospital or ICU length of stay, in-hospital death, and use of certain lifesaving medical devices were also included in the dataset. These input data were encoded with a numerical format (male=0.0, female=1.0, etc). Continuous variables were included and not encoded. Patients who required lifesaving medical devices preoperatively were excluded. Missing data regarding a patient were also excluded from the final dataset. In total, 1002 patients were excluded based on the abovementioned exclusion criteria. The postoperative observation period during which the interventions were considered was 30 days before the operation. The dataset was used due to the high patient volume and the presence of postoperative treatment (which matches our specific use case) data. At the same time, inclusion of data from numerous surgical specialties increases the robustness of the model. Specific

XSL•FO
RenderX

demographic and operation data from the dataset are shown in Table 1.

Table . Summary of patient demographics and clinical characteristics (N=131,000).

Characteristic	Values
Age at operation (years), median (IQR)	60 (45 - 70)
ASA-PS[a] classification 1 or 2, n (%)	115,280 (88)
ASA-PS classification >2, n (%)	15,720 (12)
Emergency operations, n (%)	13,100 (10)
Department—general surgery, n (%)	34,715 (26.5)
Department—orthopedic surgery, n (%)	17,423 (13.3)
ICU[b] admissions, n (%)	14,971 (11.4)
In-hospital mortality, n (%)	1581 (1.2)
ECMO[c] use (perioperative), n (%)	166 (0.17)
IABP[d] use (perioperative), n (%)	180 (0.18)
CRRT[e] use (perioperative), n (%)	855 (0.86)

[a]ASA-PS: American Society of Anesthesiologists Physical Status.

[b]ICU: intensive care unit.

[c]ECMO: extracorporeal membrane oxygenation.

[d]IABP: intra-aortic balloon pump.

[e]CRRT: continuous renal replacement therapy.

It is important to note that 1 model will not be used to predict all 4 labels as the labels are not mutually exclusive. Instead, 4 models are generated (each with the same hyperparameters), with a different binary label (rather than a single model with multiple classes). Therefore, when making a prediction, all 4 models will be run to predict the 4 different management possibilities.

The dataset was split into training and test sets in a 70:30 ratio (91,000 to 38,998 patients, respectively) and used a binary cross-entropy loss function due to the binary nature of output labels. All models were trained and evaluated exclusively on the INSPIRE dataset; no local hospital data were used in this study. This ensures that the reported results arise solely from open-source perioperative data and that there was no inadvertent leakage of information from our institution's clinical records. A temporal split was done to prevent confounding between preoperative and postoperative data. The model was run for 100 epochs, with a batch size of 1000 and a learning rate of 0.0001. To mitigate the severe class imbalance for each intervention label, we applied random oversampling of the positive class in the training data only. Specifically, minority (positive) cases were sampled with replacement until the number of positives matched the number of negatives (1:1). No oversampling or class weighting was applied to the validation or test split; all performance metrics were computed on the original, imbalanced test set. This procedure improved minority class recall while preserving real-world prevalence during evaluation. While it is also possible to decrease the "0" labeled cases to equal the number of cases with the "1" label, this decision was abandoned due to the fact that model accuracy will decrease substantially (due to lack of sample size). After repeated trials, this hyperparameter configuration led to the best learning result. A random forest classification model (which classifies data using an ensemble of decision trees and can help explain model decisions) was also used to better understand model characteristics, such as variable importance (through a Shapley additive explanation [SHAP] analysis), and the decision-making criteria used by the model.

Ethical Considerations

This study used only the publicly available INSPIRE dataset and involved a secondary analysis of deidentified data. No local hospital data were used, no direct participant contact occurred, and no new informed consent was required for this analysis.

Results

Model accuracy for both the binary classification model and the random forest model across the 4 features is shown in Table 2. We chose accuracy as our primary metric to ensure that the model is clinically viable. AUC analysis was done to determine the relationship between the false positive and true positive rate. Other methods of clinical validation are also included, such as sensitivity, specificity, negative predictive value, and positive predictive value.

XSL•FO
RenderX

Table . Model accuracy and area under the curve (AUC) analysis (N=38,998).

Metric	ECMO[a]	Ventilation	IABP[b]	CRRT[c]
Binary classification accuracy, n (%)	38,569 (98.90)	38,097 (97.69)	38,230 (98.03)	37,021 (94.93)
Random forest accuracy, n (%)	38,218 (98.00)	38,608 (99.00)	38,296 (98.20)	37,614 (96.45)
Sensitivity (recall)	0.99	0.99	0.99	0.97
Specificity	0.99	0.98	0.99	0.97
Positive predictive value	0.99	0.98	0.99	0.97
Negative predictive value	0.99	0.99	0.99	0.97
AUC	1.00	0.99	0.99	0.98

[a]ECMO: extracorporeal membrane oxygenation.

[b]IABP: intra-aortic balloon pump.

[c]CRRT: continuous renal replacement therapy.

For each model, we also examined absolute counts of correct vs incorrect classifications. In the test set of 38,998 patients, the artificial neural network models correctly classified 38,569 cases for ECMO (429 misclassifications), 38,097 for ventilatory support (901 misclassifications), 38,230 for IABP (768 misclassifications), and 37,021 for continuous renal replacement therapy (1977 misclassifications). Presenting both percentages and counts facilitates understanding of the clinical significance of the reported accuracies.

While it is difficult to compare the above metrics with the current state of the art (due to slightly different labels, input features, etc), it is clear that the model is able to predict each of the 4 features shown above with a high accuracy based on both evaluation methods. As explained earlier, risk for false negative test results is high in this use case due to the rare likelihood that a patient requires the use of postoperative lifesaving medical devices. While random oversampling was done prior to training to prevent this complication, there are limitations to this technique. Therefore, AUC analysis was done to evaluate both the specificity and sensitivity of the model. This analysis plots the false positive rate against the true positive rate and determines how well a model can distinguish between classes. As shown in Table 2, it is clear that the AUC is substantially above 0.5 (which would indicate that the model is making an entirely random choice or "guessing"), indicating that the models properly identify true positive and true negative results. Given the pronounced class imbalance (rare positive events), the near-perfect AUCs (0.98 - 1.00) should be interpreted alongside sensitivity, specificity, positive predictive value or negative predictive value, and absolute correct or incorrect counts.

Further analysis through a random forest classification model reveals insights into the model fit and the features important for the model's decisions.

Figure 1 presents a SHAP graph (a chart that demonstrates how much a certain variable contributes to the prediction of the ML model) for the ECMO ML model. Analysis of the chart reveals 2 important points. First, the most important variables used by the model are clinically relevant. The fraction of inspired oxygen (FiO_2) feature played a significant role in the decision-making process of the model, while the other features are deemed less important by the model. Indeed, current literature confirms the importance of FiO_2 in ECMO indications. In a review paper by Rabah et al [6], the authors list 8 indications for ECMO device use. Of this list, the specific FiO_2 pressure serves as 3 of these indications. Given that the mean accuracy increases when FiO_2 is added to the feature set, this is correctly reflected in the ML model. This is an important result because it reveals that the decisions made by the model are due to physiologically relevant reasons. This finding is further supported by prior studies showing that lower arterial partial pressure of oxygen to FiO_2 ratios are strongly predictive of ECMO need and correlate with poorer outcomes in critically ill patients [7].

Second, although certain features are not as important as others, they still play a role in the complication prediction for a subset of the patient population. We can see this in the SHAP analysis of the ventilator ML model.

Figure 2 shows that outside of the "gcs_m," "gcs_e," and "fio2" variables, accuracy increases only slightly when added into the feature set. This indicates that a certain feature is important in predicting complications in a small subset of the population (rather than the overall patient population), which supports the need for risk analysis at a patient-specific level. If one decides to predict a patient's complication risk based on 4 or 5 of the most important features, this individual will run the risk of creating a false prediction for a certain group of patients. This can be confirmed by the IABP ML model decision tree shown in Figure 3 (refer to Multimedia Appendix 1 for the expanded version).

XSL•FO
RenderX

Figure 1. Shapley additive explanation (SHAP) graph of the extracorporeal membrane oxygenation machine learning model. aPTT: activated partial thromboplastin time; AST: aspartate aminotransferase; BUN: blood urea nitrogen; CK: creatine kinase; CK-MB: creatine kinase myocardial band; CRP: C-reactive protein; FiO_2 : fraction of inspired oxygen; GCS-E: Glasgow Coma Scale eye response; GCS-M: Glasgow Coma Scale motor response; HR: heart rate; NIBP-MBP: noninvasive blood pressure mean blood pressure; NIBP-SBP: noninvasive blood pressure systolic blood pressure; PaO_2 : partial pressure of arterial oxygen; PT-INR: prothrombin time international normalized ratio; SaO_2 : arterial oxygen saturation.

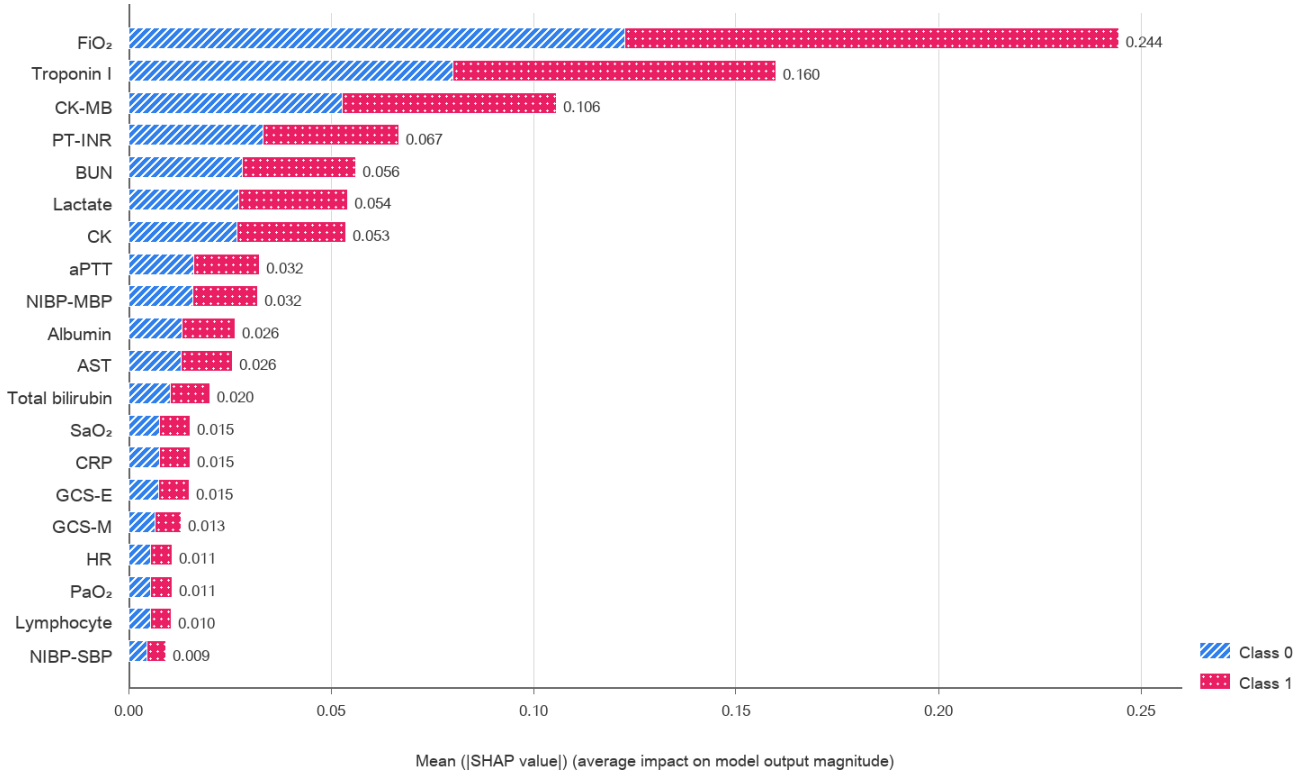
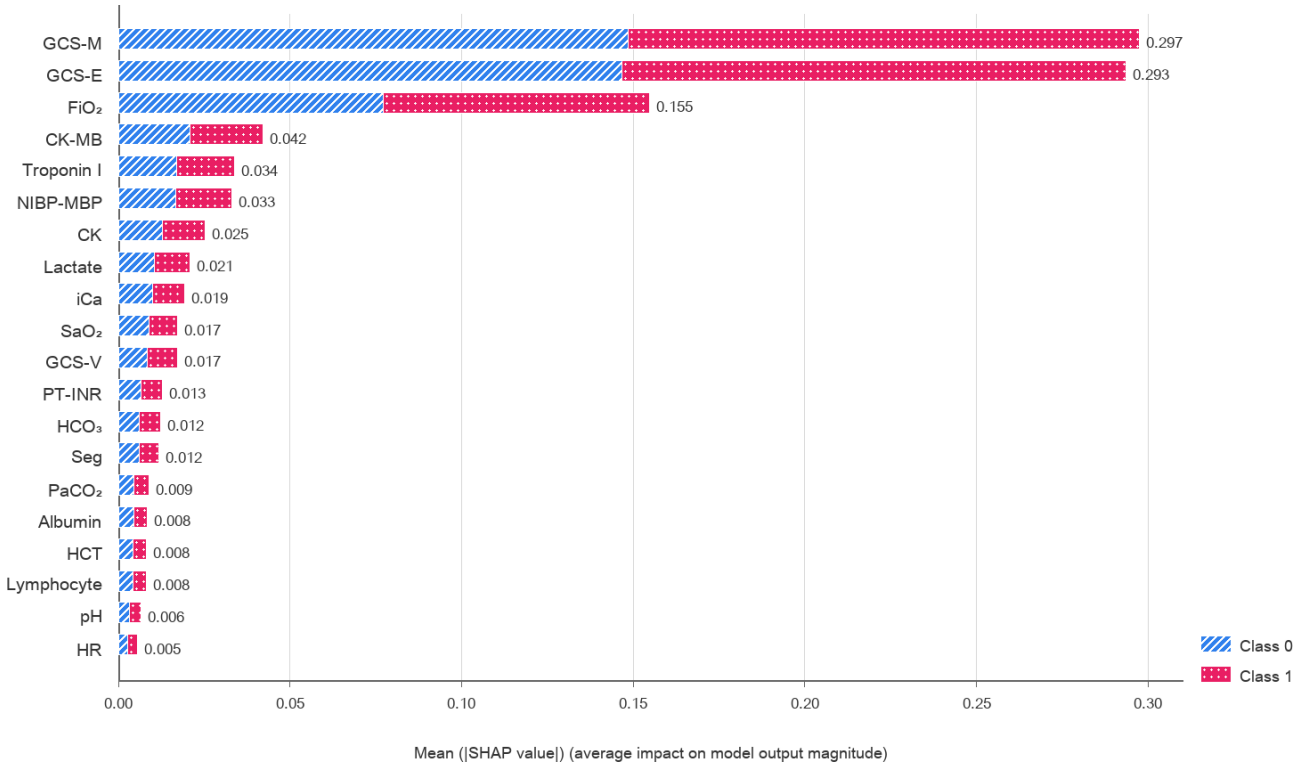
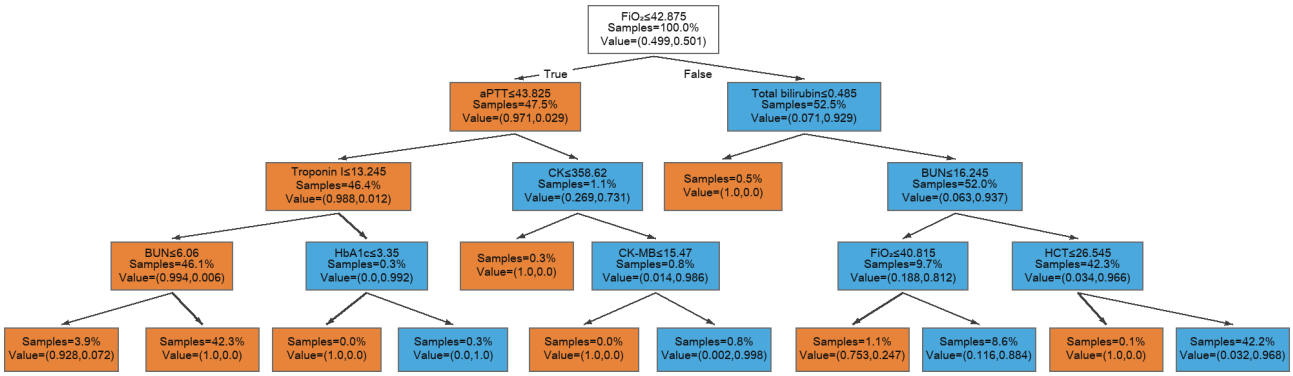


Figure 2. Shapley additive explanation (SHAP) graph of the ventilator machine learning model. CK: creatine kinase; CK-MB: creatine kinase myocardial band; FiO_2 : fraction of inspired oxygen; GCS-E: Glasgow Coma Scale eye response; GCS-M: Glasgow Coma Scale motor response; GCS-V: Glasgow Coma Scale verbal response; HCO_3 : bicarbonate; HCT: hematocrit; HR: heart rate; iCa: ionized calcium; NIBP-MBP: noninvasive blood pressure mean blood pressure; $PaCO_2$: partial pressure of arterial carbon dioxide; pH: potential of hydrogen; PT-INR: prothrombin time international normalized ratio; SaO_2 : arterial oxygen saturation; Seg: segmented neutrophils.



XSL•FO
RenderX

Figure 3. Decision tree graph for the intra-aortic balloon pump model. aptt/aPTT: activated partial thromboplastin time; bun: blood urea nitrogen; ck: creatine kinase; ckmb/CK-MB: creatine kinase-MB, also called creatine kinase myocardial band; fio2/FiO2: fraction of inspired oxygen; hba1c/HbA1c: hemoglobin A1c/glycated hemoglobin; hct: hematocrit; total_bilirubin: total bilirubin; troponin_i: troponin I.



The balanced pattern of the tree (which is similar to the trees seen in the other 4 models) indicates that no group of features explains complication risk in an entire patient population. For example, in the last row of the chart, though a certain group of features explains 42.2% (n=54,860) and 42.3% (n=54,990) of the population, there is a certain percentage of patients whose complications (or lack of complications) are not explained by those features. This confirms the need for risk assessment at a patient level rather than at a population level. The alternative reasoning is that the model is overfitting, but we believe that this is unlikely due to 2 reasons. First, the model shows high accuracy and AUC on the test data subset, which the model will have not seen before. Second, the random forest classification was limited to a maximum of 4 layers, which prevents pure memorization of the entire patient population. This 4-layer architecture also yielded the highest performance while preventing overfitting based on the evaluation metrics used above.

Project work began in April 2023, with the research concept taking shape over the following 1 to 2 months. Model development spanned approximately 3 to 4 months and was followed by 1 month of testing and validation conducted by other computer scientists.

Discussion

The main findings of this study are that it is possible to use ML models to predict whether lifesaving medical treatments would be necessary in postoperative care. In this section, we will discuss the clinical and research benefits of our work and how it can be further improved upon.

Despite significant advances in surgical technique and technology, postoperative mortality still accounts for 0.5% of elective surgical procedures (varying by procedure type and urgency). As a result, a significant amount of research and development has been done to reduce this mortality after a patient's surgery. With an understanding of an individual's risk of complications, clinicians can identify complications earlier, better plan postoperative care, or take steps to prevent complications from occurring. Currently, health care guidelines encourage the use of various scales and scores to predict risk of morbidity and mortality (such as the SAS) [2]. While these scores may be efficacious at a population level, they do not predict postoperative mortality at an individual level—an aspect that is often necessary in improving patient outcomes. This is confirmed by the random forest model in the Results section. Moreover, while ML models have been developed to predict postoperative morbidity, we believe that models that improve physician proactivity can better improve patient outcomes.

As a result, we have developed a model that aims to provide physicians with insight on whether a lifesaving medical device is necessary in the postoperative setting. The treatment serves as a proxy for the patient's diagnosis (eg, a ventilator used in the postoperative setting indicates potential respiratory failure). This provides 2 ideal benefits: better postoperative management of a patient in a potentially critical condition and better preoperative decision-making that can potentially prevent the problem in question. Given the time-sensitive situation of many of these postoperative complications, the practical aspects of patient management, such as alerting the necessary ICU staff beforehand or reserving a lifesaving medical device (if applicable in certain hospital settings), can also be facilitated through this model. As shown in the Results section, the model has a very high accuracy and correctly differentiates true negatives and true positives despite a significantly unbalanced dataset. The random forest classification model reveals a balanced decision tree with an importance graph that reflects current literature.

In practice, such predictions could be integrated into perioperative dashboards to flag high-risk cases before surgery, prompting anesthesiology and ICU teams to prepare necessary equipment and staff. Postoperatively, continuous data streaming from monitors could update risk in real time, allowing automated alerts when a patient's likelihood of device intervention increases. As this can likely happen in real time through EHR integration, the need for apps or calculators to calculate patient risk scores is also bypassed.

Beyond technical integration, it is essential to demonstrate that the models perform well on the actual patient population. As our proof-of-concept study was built exclusively on the INSPIRE dataset, no data from our own hospital were used for training or testing. We are therefore preparing a retrospective validation study using electronic health record data from our institution. This real-world evaluation will allow us to compare model outputs with actual postoperative interventions, assess generalizability across different demographics and surgical

XSL•FO
RenderX

practices, and identify whether recalibration or retraining is necessary. This step is critical for ensuring patient safety before deploying the tool as part of routine perioperative workflows.

Further development is necessary before attempting to use these models clinically. Primarily, we acknowledge that demographic data and laboratory data vary significantly from country to country and even from hospital to hospital, which can potentially change the relevance and real-world accuracy of the system. However, the aim of this study is to develop a proof-of-concept system that can be further retrained and fine-tuned for use in a specific hospital (based on patient data from that specific hospital). This allows for a prediction that is highly accurate for a specific patient population that the hospital serves. Therefore, while we have used a dataset from a university hospital to develop the baseline model, we plan to develop a pipeline that allows ML models to become integrated in other hospital systems, regardless of changes in demographics. Second, as mentioned earlier, there is a very high prevalence of "negative" labels, where the patient did not require the use of these medical device services. Therefore, in a real-world setting, false negatives can occur at a much higher rate. While the positive labels were oversampled to prevent false negative predictions, overfitting of the model is still possible. While evidence of this was not seen in the current state of the work, a different dataset may reveal different results.

Therefore, it is important to test this model in a clinical study to determine both the changes in efficacy and the true clinical relevance of the model. Second, the model should be extended to more predictor variables. While we selected 4 variables for this proof-of-concept study, additional variables can be added as necessary by the physician. By addressing these limitations, we aim to deploy this model in both elective and emergent settings and develop a tool that not only predicts clinical problems but also potentially supports management solutions.

Acknowledgments

The authors did not use generative artificial intelligence tools in the research, writing, or preparation of this manuscript. All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee support for the publication of this paper.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The INSPIRE dataset analyzed in this study is publicly available. The code used in this study is available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: MS
Methodology: MS
Software: MS
Formal analysis: MS
Validation: MS
Visualization: MS
Project administration: MS
Writing—original draft: MS
Writing—review and editing: PF, SY, NA
Supervision: PF, SY, NA

Conflicts of Interest

None declared.

Multimedia Appendix 1
Expanded decision tree.
[PNG File, 162 KB - periop_v9i1e65327_app1.png]

References

1. Gawande AA, Kwaan MR, Regenbogen SE, Lipsitz SA, Zinner MJ. An Apgar score for surgery. J Am Coll Surg 2007 Feb;204(2):201-208. [doi: 10.1016/j.jamcollsurg.2006.11.011] [Medline: 17254923]

XSL•FO
RenderX

2. Manstein SM, Goel N, Ward WH, et al. How well does the surgical Apgar score predict risk of serious complications after cancer surgery? Academic Surgical Congress. 2018. URL: https://www.asc-abstracts.org/abs2018/11-13-how-well-does-the-surgical-apgar-score-predict-risk-of-serious-complications-after-cancer-surgery/ [accessed 2024-05-02]
3. Dong JF, Xue Q, Chen T, et al. Machine learning approach to predict acute kidney injury after liver surgery. World J Clin Cases 2021 Dec 26;9(36):11255-11264. [doi: 10.12998/wjcc.v9.i36.11255] [Medline: 35071556]
4. Lee DY, Oh AR, Park J, et al. Machine learning-based prediction model for postoperative delirium in non-cardiac surgery. BMC Psychiatry 2023 May 4;23(1):317. [doi: 10.1186/s12888-023-04768-y] [Medline: 37143035]
5. Lim L, Lee H, Jung CW, et al. INSPIRE, a publicly available research dataset for perioperative medicine. Sci Data 2024 Jun 21;11(1):655. [doi: 10.1038/s41597-024-03517-4] [Medline: 38906912]
6. Rabah H, Rabah A. Extracorporeal membrane oxygenation (ECMO): what we need to know. Cureus 2022 Jul;14(7):e26735. [doi: 10.7759/cureus.26735] [Medline: 35967165]
7. Scherer C, Lüsebrink E, Joskowiak D, et al. Mortality in cardiogenic shock patients is predicted by Pao2/Fio2 (Horowitz Index) measured on ICU after venoarterial extracorporeal membrane oxygenation implantation. Crit Care Explor 2021 Sep;3(10):e0540. [doi: 10.1097/CCE.0000000000000540] [Medline: 34604783]

Abbreviations

AUC: area under the curve
CRRT: continuous renal replacement therapy
ECMO: extracorporeal membrane oxygenation
FiO_2 : fraction of inspired oxygen
IABP: intra-aortic balloon pump
ICU: intensive care unit
ML: machine learning
SAS: Surgical Apgar Score
SHAP: Shapley additive explanation

Edited by N Rohatgi; submitted 12.Aug.2024; peer-reviewed by AG Rao, Y Fuse; revised version received 01.Nov.2025; accepted 11.Nov.2025; published 28.Jul.2026.

Please cite as:
Shukla M, Fodor P, Yelika S, Ahn N
Machine Learning for Predicting Critical Postoperative Interventions: Proof-of-Concept Study Using the INSPIRE Dataset
JMIR Perioper Med 2026;9:e65327
URL: https://periop.jmir.org/2026/1/e65327
doi:10.2196/65327

© Manan Shukla, Paul Fodor, Suresh Yelika, Nicholas Ahn. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 28.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Perioperative Medicine, is properly cited. The complete bibliographic information, a link to the original publication on http://periop.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

XSL•FO
RenderX