Contents

Original Papers

A Neural Network Model Using Pain Score Patterns to Predict the Need for Outpatient Opioid Refills Following Ambulatory Surgery: Algorithm Development and Validation (e40455)
Rodney Gabriel, Sierra Simpson, William Zhong, Brittany Burton, Soraya Mehdipour, Engy Said. .......................................................... 3

The Association Between Borderline Dysnatremia and Perioperative Morbidity and Mortality: Retrospective Cohort Study of the American College of Surgeons National Surgical Quality Improvement Program Database (e38462)
Jacob Cole, Krista Highland, Scott Hughey, Brendan O'Shea, Thomas Hauert, Ashton Goldman, George Balazs, Gregory Booth. .............. 16

Toward Enhanced Clinical Decision Support for Patients Undergoing a Hip or Knee Replacement: Focus Group and Interview Study With Surgeons (e36172)
Sabrina Grant, Emma Tonkin, Ian Craddock, Ashley Blom, Michael Holmes, Andrew Judge, Alessandro Masullo, Miquel Perello Nieto, Hao Song, Michael Whitehouse, Peter Flach, Rachael Gooberman-Hill. .......................................................... 26

The Accuracy of Wrist-Worn Photoplethysmogram–Measured Heart and Respiratory Rates in Abdominal Surgery Patients: Observational Prospective Clinical Validation Study (e40474)
Jonna van der Stam, Eveline Mestrom, Jai Scheerhoorn, Fleur Jacobs, Simon Nienhuijs, Arjen-Kars Boer, Natal van Riel, Helma de Morree, Alberto Bonomi, Volker Scharnhorst, R Bouwman. .......................................................... 62

Prediction of Pelvic Organ Prolapse Postsurgical Outcome Using Biomaterial-Induced Blood Cytokine Levels: Machine Learning Approach (e40402)
Mihyun Waugh, Nicholas Boltin, Lauren Wolf, Jane Goodwin, Patti Parker, Ronnie Horner, Matthew Hermes, Thomas Wheeler, Richard Goodwin, Melissa Moss. ........................................................................................................... 72

Early Warning Scores to Support Continuous Wireless Vital Sign Monitoring for Complication Prediction in Patients on Surgical Wards: Retrospective Observational Study (e44483)
Mathilde van Rossum, Robin Bekhuis, Ying Wang, Johannes Hegeman, Ellis Folbert, Miriam Vollenbroek-Hutten, Cornelis Kalkman, Ewout Kouwenhoven, Hermie Hermens. .......................................................... 82

A Real-Time Mobile Intervention to Reduce Sedentary Behavior Before and After Cancer Surgery: Pilot Randomized Controlled Trial (e41425)
Carissa Low, Michaels Danko, Krina Durica, Julio Vega, Meng Li, Abhineeth Kunta, Raghu Mulukutla, Yiyi Ren, Susan Sereika, David Bartlett, Dana Bovbjerg, Anind Dey, John Jakicic. .......................................................... 102

A Mobile App for Postoperative Pain Management Among Older Veterans Undergoing Total Knee Arthroplasty: Mixed Methods Feasibility and Acceptability Pilot Study (e50116)
Jessica Morgan, Caitlin Rawlins, Steven Walther, Andrew Harvey, Annmarie O'Donnell, Marla Greene, Troy Schmidt. .......................................................... 112
Perioperative Risk Assessment of Patients Using the MyRISK Digital Score Completed Before the Preanesthetic Consultation: Prospective Observational Study (e39044)
Fabrice Ferré, Rodolphe Laurent, Philippine Furelau, Emmanuel Doumard, Anne Ferrier, Laetitia Bosch, Cyndie Ba, Rémi Menut, Matt Kurrek, Thomas Geeraerts, Antoine Piau, Vincent Minville. ................................................................. 119

Remote Home Monitoring of Continuous Vital Sign Measurements by Wearables in Patients Discharged After Colorectal Surgery: Observational Feasibility Study (e45113)
Jobbe Leenen, Vera Ardesch, Gijsbert Patijn. ................................................................. 135

Reducing Alcohol Use Before and After Surgery: Qualitative Study of Two Treatment Approaches (e42532)
Lyndsay Chapman, Tom Ren, Jake Solka, Angela Bazzi, Brian Borsari, Michael Mello, Anne Fernandez. ................................................................. 147

Dashboard of Short-Term Postoperative Patient Outcomes for Anesthesiologists: Development and Preliminary Evaluation (e47398)
Rama Sreepada, Ai Chang, Nicholas West, Jonath Sujan, Brendan Lai, Andrew Poznikoff, Rebecca Munk, Norbert Froese, James Chen, Matthias Görges. ................................................................. 157

An Ensemble Learning Approach to Improving Prediction of Case Duration for Spine Surgery: Algorithm Development and Validation (e39650)
Rodney Gabriel, Bhavya Harjai, Sierra Simpson, Austin Du, Jeffrey Tully, Olivier George, Ruth Waterman. ................................................................. 170

Assessing Barriers to Implementation of Machine Learning and Artificial Intelligence–Based Tools in Critical Care: Web-Based Survey Study (e41056)
Eric Mlodzinski, Gabriel Wardi, Clare Viglione, Shamim Nemati, Laura Crotty Alexander, Atul Malhotra. ................................................................. 183

Viewpoint

Patient Safety of Perioperative Medication Through the Lens of Digital Health and Artificial Intelligence (e34453)
Jiancheng Ye. ................................................................. 46

Review

The Role of Wearable Technology in Measuring and Supporting Patient Outcomes Following Total Joint Replacement: Review of the Literature (e39396)
Gregory Iovanel, David Ayers, Hua Zheng. ................................................................. 53
A Neural Network Model Using Pain Score Patterns to Predict the Need for Outpatient Opioid Refills Following Ambulatory Surgery: Algorithm Development and Validation

Rodney Allanigue Gabriel¹, MD, MAS; Sierra Simpson¹, PhD; William Zhong¹; Brittany Nicole Burton², MD; Soraya Mehdipour¹, MD; Engy Tadros Said¹, MD

¹Department of Anesthesiology, University of California San Diego, La Jolla, CA, United States
²Department of Anesthesiology and Perioperative Medicine, University of California Los Angeles, Los Angeles, CA, United States

Corresponding Author:
Rodney Allanigue Gabriel, MD, MAS
Department of Anesthesiology
University of California San Diego
9400 Campus Pt Dr
La Jolla, CA, 92037
United States
Phone: 1 8586637747
Email: ragabriel@health.ucsd.edu

Abstract

Background: Expansion of clinical guidance tools is crucial to identify patients at risk of requiring an opioid refill after outpatient surgery.

Objective: The objective of this study was to develop machine learning algorithms incorporating pain and opioid features to predict the need for outpatient opioid refills following ambulatory surgery.

Methods: Neural networks, regression, random forest, and a support vector machine were used to evaluate the data set. For each model, oversampling and undersampling techniques were implemented to balance the data set. Hyperparameter tuning based on k-fold cross-validation was performed, and feature importance was ranked based on a Shapley Additive Explanations (SHAP) explainer model. To assess performance, we calculated the average area under the receiver operating characteristics curve (AUC), F1-score, sensitivity, and specificity for each model.

Results: There were 1333 patients, of whom 144 (10.8%) refilled their opioid prescription within 2 weeks after outpatient surgery. The average AUC calculated from k-fold cross-validation was 0.71 for the neural network model. When the model was validated on the test set, the AUC was 0.75. The features with the highest impact on model output were performance of a regional nerve block, postanesthesia care unit maximum pain score, postanesthesia care unit median pain score, active smoking history, and total perioperative opioid consumption.

Conclusions: Applying machine learning algorithms allows providers to better predict outcomes that require specialized health care resources such as transitional pain clinics. This model can aid as a clinical decision support for early identification of at-risk patients who may benefit from transitional pain clinic care perioperatively in ambulatory surgery.

(JMIR Perioper Med 2023;6:e40455) doi:10.2196/40455

KEYWORDS
opioids; ambulatory surgery; machine learning; surgery; outpatient; pain medication; pain; pain management; patient needs; predict; algorithms; clinical decision support; pain care

Introduction

Opioids play an essential role in acute perioperative pain management. Increased attention to pain management as a quality metric has brought to light an overuse of prescription opioids contributing to an epidemic across the United States. The United States has had increased opioid prescriptions filled in the immediate postoperative period; a study reported that the mean dose of opioids prescribed for most surgical procedures in the United States was higher than that prescribed in other
Persistent opioid prescribing is often postsurgical [4], in which as many as 3% of opioid-naive patients required opioids for more than 90 days after a major elective surgery [5]. One potential service that may help curb outpatient opioid use after surgery is the transitional pain clinic, which consists of a team of providers who implement multidisciplinary opioid-sparing approaches such as pharmacological, nonpharmacological, and psychological interventions with the goal of weaning patients from opioids postoperatively as outpatients [6,7]. Transitional pain clinics have been shown to reduce opioid use postoperatively, symptoms of anxiety and depression, pain catastrophizing, and postsurgical pain [8,9]. Given the increased resources required to provide this type of service, not all surgical patients may realistically receive postoperative care from transitional clinics. Currently, the criteria for recommendation to transitional pain services for surgical patients are not uniformly defined; thus, accurate predictive methods for patients who may benefit from transitional pain clinics are needed. Less work has been done on patients undergoing ambulatory surgery and on the identification of patients who may likely require more opioids as an outpatient. In such populations, machine learning may be used to identify postoperative opioid use in the recovery room [10]. In addition, some studies have described the risk factors for using outpatient opioids after ambulatory surgery [11-13].

The objective of this study was to develop machine learning–based predictive models that may aid in the identification of patients likely to require opioid refills after their initial discharge prescription. Specifically, pain score patterns were incorporated (ie, trends in reported pain scores in the recovery room) into the models. We focused on patients who underwent ambulatory surgery, which included orthopedic surgery (eg, joint arthroscopy, forearm/hand surgery), nonmastectomy breast surgery, urology (eg, cystoscopy), minimally invasive surgery (eg, cholecystectomy, hernia repair), colorectal surgery (eg, hemorrhoidectomy), and gynecology (dilation and curettage/evacuation, hysteroscopy). We hypothesized that the use of neural networks that incorporate various features, including recovery room pain phenotypes, may identify patients at higher risk. The pain phenotypes included patterns in patient-reported pain scores in the recovery room, including trajectory of pain and median and maximum pain scores.

Methods

Study Population

Data were retrospectively collected from the electronic medical records of patients who underwent outpatient surgery from January to July 2020 at a single ambulatory surgery center. The outpatient surgeries included in the analysis included orthopedic surgery (eg, joint arthroscopy, forearm/hand surgery), nonmastectomy breast surgery, urology (eg, cystoscopy), minimally invasive surgery (eg, cholecystectomy, hernia repair), colorectal surgery (eg, hemorrhoidectomy), and gynecology (dilation and curettage/evacuation, hysteroscopy).

Ethics Approval

Our institutional review board (Human Research Protections Program) waived the consent requirement and approved this retrospective study (protocol 210099).

Primary Outcome and Features

The primary outcome of interest was a binary variable (response range, 0 or 1), in which 0 was defined as “no opioid refill” and 1 was defined as the patient “needed to refill their outpatient opioid prescriptions” within 2 weeks after surgery (no opioid refill vs opioid refill). This was captured retrospectively from the electronic medical record by review of the following: (1) any telephone note describing patient calling in for opioid renewal; (2) any progress/office visit note from primary care provider, pain medicine specialist, or surgical provider describing the need for opioid refill; or (3) any renewal order in the medication list for opioids within this time frame. On postanesthesia care unit (PACU) discharge, all patients were prescribed up to 5 days of opioids. For perioperative multimodal analgesia, all patients received preoperative acetaminophen unless contraindicated. For a subset of surgical procedures, regional nerve blocks were routinely offered preoperatively (ie, shoulder, hand/forearm, and knee surgeries). Intraoperatively, patients may have received fentanyl, hydromorphone, ketamine, ketorolac, and dexmedetomidine at the discretion of the anesthesiologist. In the PACU, patients were given oxycodone, fentanyl, and hydromorphone, as needed.

Features that were integrated into the model were collected retrospectively from the electronic medical record system. The data included age (years), sex (male vs female), body mass index (kg/m²), English-speaking, comorbidities, regional nerve block performance, general anesthesia, intraoperative ketamine, intraoperative total intravenous anesthesia, opioid consumption, and pain scores (11-point numeric rating scale [NRS] from 0 to 10). These features were included, as they were determined to be relevant to postoperative opioid use based on clinical judgement and previous research [14,15]. Opioid consumption, defined as total opioids consumed intraoperatively and in the PACU, was measured in intravenous morphine equivalents (MEQ). Pain scores were captured as preoperative pain score, median pain score in the PACU, maximum pain score in the PACU, and slope of pain score trajectory in the PACU. Preoperative pain scores were collected by nurses upon arrival for preoperative check-in. PACU pain scores were captured every 5-15 minutes and recorded in the electronic medical record. A negative value for the pain score slope was defined as an overall decrease in pain scores throughout the PACU stay. A positive value for the pain score slope was defined as an overall increase in the pain scores throughout the PACU stay. A zero value of the pain score slope was defined as no change overall increase in the pain scores throughout the PACU stay. A positive value for the pain score slope was defined as an overall increase in the pain scores throughout the PACU stay. A zero value of the pain score slope was defined as no change in the overall trend of pain scores throughout the PACU stay.

Statistical Analysis

Python (v3.10.1) was used for all statistical analyses. Patient and surgical characteristics were compared with chi-squared test (categorical) and Wilcoxon rank sum test (continuous). A
generalized linear mixed model fit by maximum likelihood was implemented to model the features to the primary outcome of opioid refill. The random effect in this model was the surgical procedure. All features were included in the model, and their association with the outcome was reported by their respective odds ratios (ORs), 95% CI, and P values. A neural network model to predict the need for opioid refills following surgery was constructed. Logistic regression, random forest, and support vector machine classifiers were implemented for comparison. For all models, patient data were divided into training and test data sets with a 70:30 split by using a stratified randomized splitter—the train_test_split method from the sci-kit learn library. K-fold cross-validation on the training set was used to tune the hyperparameters and to optimize oversampling techniques as well as to calculate the average sensitivity, specificity, F1-score, and area under the curve (AUC) for the receiver operating characteristic curve. The final version of each model was then validated on the test set and the AUC was reported. Feature importance from the neural network was ranked based on Shapley Additive Explanations (SHAP).

**Data Balancing**

Synthetic Minority Oversampling Technique (SMOTE) for Nominal and Continuous algorithm and random undersampling were both implemented using the imblearn library [16]. These tools were used to achieve a balanced class distribution with minimal difference between positive and negative outcomes. A data set with a large difference between positive and negative outcomes was considered unbalanced and may make it difficult for predictive machine learning models to draw useful conclusions, given the uneven classification of data.

Random undersampling of the majority outcome is frequently used to reduce the impact of imbalanced data sets; however, SMOTE oversamples were used to create synthetic minority class examples to balance the minority class with the majority class. SMOTE uses samples from the minority class and a set number of nearest neighbors—in this case, 5—to generate synthetic cases from the sample class. Combining the 2 techniques as outlined yielded positive outcomes. Both techniques were only applied on the training set. Different combinations for proportions of minority to majority class were analyzed, ranging from 0.25 to 1.00. After performing k-fold cross-validation, the parameter “sampling_strategy” for the SMOTE class from imblearn was optimal when set to 0.24 and the parameter “sampling_strategy” for the RandomUnderSampler class from imblearn was optimal when set to 0.94. Optimal results were based on which hyperparameters produced the highest performance metrics for the model (eg, AUC, F1-score, sensitivity, specificity).

**Machine Learning Models**

Four different machine learning classification models were evaluated: neural network, logistic regression, random forest classifier, and support vector classifier. For each model, the following sampling methods were compared: oversampling the training set via SMOTE, undersampling the majority class in the training set, a combination of SMOTE and undersampling of the majority classes, and no oversampling or undersampling technique. The results from the sampling method that provided the optimal results were reported. For each model, all features were included as inputs. One-hot encoding was used for categorical features.

**Multilayer Perceptron Neural Network**

Using the Keras interface for the TensorFlow library, a shallow feedforward neural network was constructed. The rectified linear unit function was used as the activation function. The final output layer used the sigmoid activation function, and the overall model used the Adam optimizer. Repeated k-fold cross-validation was used to tune the hyperparameters to find the optimal parameter values for the number of hidden layers (1), number of neurons per hidden layer (100), maximum number of iterations (300), batch size (16), and learning rate (0.0001).

**Logistic Regression**

The logistic regression classifier predicts the probabilities of the different outcome possibilities based on the input. A newton-cg solver regression model was implemented without specifying individual class weights. This model provided a baseline score and helped make the case for improvement over the evaluation metrics. Repeated k-fold cross-validation was used to tune the hyperparameters to find the optimal parameter value for C (the strength of the regularization is inversely proportional to C), which was 0.3.

**Random Forest Classifier**

The random forest is an ensemble approach, which has been proven effective for a variety of classification problems. To tune the hyperparameters, we performed repeated k-fold cross-validation to find the optimal parameter values for maximum depth (75), minimal samples required to be at the leaf node (4), minimal samples required to split an internal node (4), and number of estimators (100) (ie, number of trees).

**Support Vector Classifier**

A support vector classifier maps the data onto an n-dimensional space (n being the number of features) and then identifies the hyperplane decision boundary that best separates the data into 2 classes by maximizing the distance between the hyperplane and the nearest data point in either class. K-fold cross-validation was used to tune the hyperparameters to identify the optimal parameter value for C, which was 130.

**Performance Metrics**

The primary performance metric of interest was the AUC for the receiver operating characteristic curve. In addition, we reported F1-scores, sensitivity, and specificity.

**K-Fold Cross-Validation**

To effectively tune the hyperparameters of the models, stratified k-fold cross-validation was implemented on the training set to observe the sensitivity, specificity, F1-score, and AUC scores, for 10 splits. For each iteration, the data set was split into 10 groups (folds). One fold served as the test set with the other 9 serving as the training set. When assessing the effectiveness of SMOTE and random undersampling, only the training folds were changed. This process was repeated until each fold served as the test set once. Every model exhibited improved
performance metrics when SMOTE and random undersampling were applied.

Results

Study Cohort Characteristics

There were 1333 patients and 28 unique surgical procedures included in the final analysis, and 144 (10.8%) patients refilled their opioid prescription within 2 weeks after outpatient surgery. Univariate analysis revealed that patients who required opioid refills were more likely to be smokers (32/144, 22.2% vs 156/1189, 13.1%, respectively; \( P = .005 \)) and had a regional nerve block performed (87/144, 60.4% vs 440/1189, 37%, respectively; \( P < .001 \)). Those who required opioid refills had higher total perioperative opioid consumption (intraoperative and PACU opioid use; \( P < .001 \)), preoperative pain scores \( (P < .001) \), maximum PACU pain scores \( (P < .001) \), and median PACU pain scores \( (P < .001) \). Table 1 lists the differences between the opioid refill and non-opioid refill cohorts represented in the study population in order to provide information regarding the baseline characteristics. All surgical procedures included in our analysis are listed in Table 1.
Table 1. Baseline characteristics of the study cohorts.a

<table>
<thead>
<tr>
<th>Feature</th>
<th>No opioid refill (n=1189)</th>
<th>Required opioid refill (n=144)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Surgical procedure, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arthrodesis (finger)</td>
<td>10 (0.8)</td>
<td>3 (2.1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Arthroscopy (hip)</td>
<td>36 (3)</td>
<td>5 (3.5)</td>
<td></td>
</tr>
<tr>
<td>Arthroscopy (knee)</td>
<td>99 (8.3)</td>
<td>12 (8.3)</td>
<td></td>
</tr>
<tr>
<td>Arthroscopy (shoulder)</td>
<td>61 (5.1)</td>
<td>19 (13.2)</td>
<td></td>
</tr>
<tr>
<td>Arthroscopy (shoulder, with rotator cuff repair)</td>
<td>45 (3.8)</td>
<td>13 (9)</td>
<td></td>
</tr>
<tr>
<td>Arthroscopy (wrist)</td>
<td>9 (0.8)</td>
<td>4 (2.8)</td>
<td></td>
</tr>
<tr>
<td>Breast lumpectomy</td>
<td>56 (4.7)</td>
<td>7 (4.9)</td>
<td></td>
</tr>
<tr>
<td>Transperineal prostate biopsy</td>
<td>17 (1.4)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Laparoscopic cholecystectomy</td>
<td>29 (2.4)</td>
<td>1 (0.7)</td>
<td></td>
</tr>
<tr>
<td>Cystoscopy</td>
<td>43 (3.6)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Dilation and curettage of the uterus</td>
<td>76 (6.4)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Dilation and evacuation of the uterus</td>
<td>36 (3)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Anorectal examination under anesthesia</td>
<td>29 (2.4)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Condyloma excision</td>
<td>13 (1.1)</td>
<td>3 (2.1)</td>
<td></td>
</tr>
<tr>
<td>Lesion excision of the head and neck</td>
<td>31 (2.6)</td>
<td>1 (0.7)</td>
<td></td>
</tr>
<tr>
<td>Lesion excision of the upper extremity</td>
<td>78 (6.6)</td>
<td>4 (2.8)</td>
<td></td>
</tr>
<tr>
<td>Extracorporeal shockwave lithotripsy</td>
<td>20 (1.7)</td>
<td>1 (0.7)</td>
<td></td>
</tr>
<tr>
<td>Anal fistulectomy</td>
<td>73 (6.1)</td>
<td>1 (0.7)</td>
<td></td>
</tr>
<tr>
<td>Hemorrhoidectomy</td>
<td>29 (2.4)</td>
<td>8 (5.6)</td>
<td></td>
</tr>
<tr>
<td>Inguinal herniorrhaphy</td>
<td>16 (1.3)</td>
<td>2 (1.4)</td>
<td></td>
</tr>
<tr>
<td>Hysteroscopy</td>
<td>47 (4)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Incision and drainage of the upper extremity</td>
<td>12 (1)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>ORIF&lt;sup&gt;b&lt;/sup&gt;, distal radius fracture</td>
<td>71 (6)</td>
<td>15 (10.4)</td>
<td></td>
</tr>
<tr>
<td>ORIF, scaphoid fracture</td>
<td>11 (0.9)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>ORIF, hand</td>
<td>70 (5.9)</td>
<td>10 (6.9)</td>
<td></td>
</tr>
<tr>
<td>Ligament reconstruction with tendon interposition, upper extremity</td>
<td>30 (2.5)</td>
<td>9 (6.3)</td>
<td></td>
</tr>
<tr>
<td>Open carpal tunnel release</td>
<td>50 (4.2)</td>
<td>2 (1.4)</td>
<td></td>
</tr>
<tr>
<td>Arthroscopic anterior crucial ligament repair</td>
<td>92 (7.7)</td>
<td>24 (16.7)</td>
<td></td>
</tr>
<tr>
<td><strong>Patient characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years), median (quartiles)</td>
<td>43 (32, 60)</td>
<td>46.5 (33.0, 58.25)</td>
<td>.67</td>
</tr>
<tr>
<td>Male sex, n (%)</td>
<td>551 (46.3)</td>
<td>70 (48.6)</td>
<td>.59</td>
</tr>
<tr>
<td>Body mass index (kg/m&lt;sup&gt;2&lt;/sup&gt;), median (quartiles)</td>
<td>26.1 (23.0, 30.2)</td>
<td>26.3 (23.9, 31.6)</td>
<td>.14</td>
</tr>
<tr>
<td>English speaker, n (%)</td>
<td>1106 (93)</td>
<td>129 (89.6)</td>
<td>.19</td>
</tr>
<tr>
<td><strong>Comorbidities, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preoperative opioid use</td>
<td>55 (4.6)</td>
<td>8 (5.6)</td>
<td>.77</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>87 (7.3)</td>
<td>13 (9)</td>
<td>.57</td>
</tr>
<tr>
<td>Alcohol use</td>
<td>53 (4.5)</td>
<td>7 (4.9)</td>
<td>.99</td>
</tr>
<tr>
<td>Active smoker</td>
<td>156 (13.1)</td>
<td>32 (22.2)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Substance abuse</td>
<td>43 (3.6)</td>
<td>10 (6.9)</td>
<td>.09</td>
</tr>
<tr>
<td>Hypertension</td>
<td>255 (21.4)</td>
<td>34 (23.6)</td>
<td>.63</td>
</tr>
</tbody>
</table>
Mixed Effects Logistic Regression Model

The need for opioid refill within 2 weeks after ambulatory surgery was modeled utilizing a mixed effects logistic regression analysis fit by maximum likelihood, in which the random effect was the surgical procedure (Table 2). Features that were statistically significantly associated with higher odds of need for opioid refill were active smokers (OR 1.99, 95% CI 1.19-3.31; \( P = .009 \)), substance abuse history (OR 2.34, 95% CI 1.02-5.37; \( P = .04 \)), regional block performed (OR 2.81, 95% CI 1.62-4.88; \( P < .001 \)), total opioid consumption (MEQ mg) intraoperatively and in PACU (OR 1.03, 95% CI 1.01-1.06; \( P = .008 \)), and median PACU pain score (NRS) (OR 1.19, 95% CI 1.04-1.36; \( P = .01 \)). A feature that was significantly associated with decreased odds of opioid refill was English-speaking patients (OR 0.47, 95% CI 0.24-0.93; \( P = .03 \)).
Table 2. Mixed effects logistic regression modeling of refills.a

<table>
<thead>
<tr>
<th>Feature</th>
<th>Odds ratio (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>0.99 (0.98-1.01)</td>
<td>.85</td>
</tr>
<tr>
<td>Male sex</td>
<td>0.96 (0.63-1.46)</td>
<td>.84</td>
</tr>
<tr>
<td>Body mass index (kg/m²)</td>
<td>1.01 (0.98-1.01)</td>
<td>.48</td>
</tr>
<tr>
<td>English speaker</td>
<td>0.47 (0.24-0.93)</td>
<td>.03</td>
</tr>
</tbody>
</table>

**Comorbidities**

- Preoperative opioid use: 0.70 (0.25-1.98)  .51
- Diabetes mellitus: 1.44 (0.67-3.09)  .36
- Alcohol use: 0.82 (0.33-2.03)  .66
- Active smoker: 1.99 (1.19-3.31)  .009
- Substance abuse: 2.34 (1.02-5.37)  .04
- Hypertension: 1.26 (0.73-2.18)  .41
- Depression/anxiety: 1.07 (0.66-1.73)  .79
- Coronary artery disease: 1.23 (0.30-5.00)  .78
- Dementia: 1.33 (0.09-18.80)  .84
- Renal insufficiency: 0.79 (0.24-2.58)  .99
- Chronic obstructive pulmonary disease: 0.99 (0.19-5.17)  .99
- Asthma: 0.99 (0.53-1.86)  .99
- Obstructive sleep apnea: 0.62 (0.26-1.44)  .27
- Chronic pain: 1.31 (0.62-2.79)  .48
- Congestive heart failure: 0  .99

**Anesthesia/perioperative medications**

- Regional block performed: 2.81 (1.62-4.88)  <.001
- General anesthesia: 1.05 (0.56-1.94)  .88
- Intraoperative ketamine used: 1.17 (0.51-2.65)  .71
- Total intravenous anesthetic: 1.21 (0.46-3.18)  .71
- Total (intraoperative and PACUb) opioid consumption (MEQc): 1.03 (1.01-1.06)  .008

**Pain (numeric rating scale 0-10)**

- Preoperative pain score: 0.95 (0.86-1.05)  .35
- Maximum PACU score: 1.02 (0.92-1.13)  .67
- Median PACU pain score: 1.19 (1.04-1.36)  .01
- Slope of PACU pain: 0.78 (0.56-1.09)  .16

---

aResults of mixed effects logistic regression modeling need for opioid refill after ambulatory surgery. The random effect in this model was the surgical procedure.
bPACU: postanesthesia care unit.
cMEQ: morphine equivalents.

**Neural Network Approach to Predicting Opioid Refills**

Hyperparameter tuning via grid search cross-validation was implemented to identify the best architecture of the multilayer perceptron neural network, which consisted of 1 hidden layer, 100 neurons within the hidden layer, 300 maximum iterations for learning, a batch size of 16, and a learning rate of 0.0001. Based on this architecture, the average AUC calculated from k-fold cross-validation was 0.71 (95% CI 0.68-0.74). The final model was then validated on the test set, which yielded an AUC of 0.75 (Figure 1). The features with the highest impact on model output for the neural network based on the absolute SHAP values were performance of a regional nerve block, maximum pain score in the PACU, median pain score in the PACU, active smoking history, and total opioid consumption (intraoperative and PACU) (Figure 2).
Next, various other machine learning–based models were implemented to predict the need for opioid refills after ambulatory surgery. Based on k-fold cross-validation, the average AUCs from models with optimized hyperparameters were identified for support vector machine (0.64, 95% CI 0.57-0.71), random forest (0.66, 95% CI 0.60-0.71), and logistic regression (0.69, 95% CI 0.66-0.74) (Table 3). The final models for each machine learning approach were then validated on a separate test set when SMOTE was not applied versus when SMOTE was applied—the calculated AUCs were identified for support vector machine (0.65), random forest (0.68), and logistic regression (0.73). SMOTE improved the performance of each model.

**Figure 1.** The calculated area under the curve of each machine learning model when trained on 70% of the data and validated on the remaining 30%. The models predict the need for opioid refill within 2 weeks following ambulatory surgery. AUC: area under the curve.

![Figure 1](image1.png)

**Figure 2.** Shapley Additive Explanations (SHAP) values (impact on model output) of the top features used in the neural network predicting the need for opioid refill within 2 weeks following ambulatory surgery. MEQ: morphine equivalents; NRS: numeric rating scale; ORIF: open reduction and internal fixation; PACU: postanesthesia care unit.

![Figure 2](image2.png)
Discussion

Principal Findings

We demonstrated that a shallow feedforward neural network and other machine learning approaches that integrated pain score patterns had adequate performance to predict the need for opioid refills within 2 weeks following ambulatory surgery. The features with the highest impact on model output were active smoking history, intraoperative opioid consumption, PACU opioid consumption, regional nerve block utilization, as well as maximum and median pain scores in the PACU. The importance of pain score patterns (ie, median and maximum pain scores) in predicting opioid refills is interesting and highlights the association of PACU analgesia and opioid consumption with the requirement for more opioids following the initial prescription. This neural network may be useful in identifying patients at risk who require a longer duration of opioid use so that the limited hospital resources can be better utilized in a precise manner.

Comparison to Prior Work

Previous studies have reported the utilization of machine learning for predicting postoperative opioid use in ambulatory surgery [10,17]. Nair et al [10] reported the accuracies of regression, naïve Bayes, neural networks, random forest, and extreme gradient boosting in predicting postoperative opioid use in the recovery room and showed that random forest performed best when using only preoperative features. Anderson et al [17] utilized models to predict prolonged opioid use specifically in patients who underwent anterior cruciate ligament repair by using regression, Bayesian belief network, gradient boosting, and random forest. They found that gradient boosting was able to achieve an AUC of 0.77. In our study, we reviewed multiple types of ambulatory surgeries and focused on predicting the need for additional outpatient opioid refills weeks after surgery. Four computational approaches were used to determine the best model for our data set, and all had similar performances. Random forest, logistic regression, and support vector machine tools did not perform as well as the neural network, though the random forest model had increased specificity compared to the neural network. Both the support vector classifier and neural network can increase the dimensionality of the data to find a solution, but given the time and training, neural networks usually outperform support vector classifiers. Random forest and neural networks approach data inversely, as random forest decision trees are independent and neural network neurons are dependent on other neurons. Logistic regression is the standard approach but often does not perform well in multidimensional data sets. By surveying multiple models, the benefits of each can be identified and evaluated to improve the validity of the predicted features [18].

Opioids remain the cornerstone for acute postoperative pain management, and the perioperative period is often the patients’ first introduction to prescription opioids. Our study’s patient cohort was primarily opioid-naïve; only 4.6% (55/1189) of the patients in the non-refill group and 5.6% (81/144) of the patients in the refill group reported preoperative opioid use. Studies have shown surgical procedure as an independent risk factor for prolonged opioid use [4,19,20]. Other risk factors include preoperative opioids, tobacco use, gender, and mood disorders [21-26]. Although efforts are in place to standardize postoperative opioid prescriptions per surgical procedure [27], there continues to be a wide variety in the amount and duration of opioids prescribed and often in excess [1,28-31].

An estimated 67%-92% of the prescribed opioids for postoperative pain remain unused [1,32], leaving great potential for diversion and misuse. An increasing number of heroin users reported first being introduced to opioids via prescription and then resorting to heroin for cost and availability factors [33]. Likewise, Bartels et al [34] report that 80% of the opioid prescriptions remain unused with limited and challenging disposal options. Similarly, a 2017 systematic review reports that patients took only 29%-58% of the prescribed opioid pills [32]. Over the decades, we have learned that excess opioids do not necessarily reduce persistent postsurgical pain or any other pain-related outcomes [35]. As Porter and colleagues [36] demonstrated, sometimes less is more—patient-centered opioid discharge prescription guidelines satisfied 93% of the patients, with 99% in the 0 morphine milligram equivalents group [36]. Although there may be some procedures that do not require postoperative opioids, we must also find a balance and prescribe opioids as necessary to meet individual patients’ pain needs [37]. For these reasons, risk stratification can be a helpful tool for guiding the process of postoperative opioid prescribing.

The use of regional anesthesia was associated with opioid refilling. It is important to note that there is no causality that may be drawn from these results but rather an association. It may be that the use of regional anesthesia was associated with surgeries that were more painful in nature, and despite pain

### Table 3. Model performance for each machine learning model calculated by k-fold cross-validation.

<table>
<thead>
<tr>
<th>Classification model</th>
<th>Area under the curve</th>
<th>F1-score</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No SMOTEa</td>
<td>SMOTE</td>
<td>No SMOTEa</td>
<td>SMOTE</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>0.516</td>
<td>0.698</td>
<td>0.033</td>
<td>0.335</td>
</tr>
<tr>
<td>Neural network</td>
<td>0.509</td>
<td>0.711b</td>
<td>0.067</td>
<td>0.347b</td>
</tr>
<tr>
<td>Support vector classifier</td>
<td>0.500</td>
<td>0.643</td>
<td>0.001</td>
<td>0.291</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.500</td>
<td>0.655</td>
<td>0.002</td>
<td>0.317</td>
</tr>
</tbody>
</table>

aSMOTE: Synthetic Minority Oversampling Technique.
bRepresents the best performance for the given metric.
scores being likely lower in the PACU, this group would more likely require additional opioids as outpatients when compared to other surgical procedures that are less likely to receive regional anesthesia for pain management. Other potential limitations include the variability in surgery type, which may range in pain level, both during surgery and during recovery, as well as the subjective nature of pain scores. Despite these limitations, the features that have been identified are actionable and trackable in future studies.

**Limitations**

There are several limitations in this study—mainly due to the inherent limitations of a retrospective analysis. First, the primary outcome (opioid refills) may potentially be underestimated, as we captured this data based on clinical notes and orders in the electronic medical record system. It is possible that the need for opioid refills was missed in some patients who sought care outside of our health care system (and thus not recorded in our records). However, we extracted the data via a manual clinician review to optimize accuracy as much as possible. A prospective study would be needed to assess the incidence of postsurgical opioid refills more accurately. Second, an issue of generalizability is also of concern, as this is single-institution data. Model performance (eg, AUC) could decrease in a surgical population outside of this institution. To avoid the issue of overfitting and, thus, limited generalizability, we calculated the metrics from k-fold cross-validation and furthermore used a holdout data set for validation. What is needed is a high-quality prospective study that can more accurately capture the features and outcomes from each patient and, subsequently, be validated at external institutions.

**Future Directions**

Early identification of at-risk patients prior to their elective surgical procedure is the key. These patients can then be referred to establish care with a dedicated and comprehensive transitional pain program. Built on solid evidence-based medicine, this multidisciplinary transitional pain service includes anesthesiology, pharmacy, psychiatry, and physical therapy. Patients are often evaluated preoperatively to help manage expectations regarding anticipated postoperative pain and offer preoperative weaning when appropriate. This anesthesiologist-led team makes recommendations about intraoperative and immediate postoperative pain management [38], including predischarge and postdischarge tapering plans, if applicable. After the discharge, the transitional pain service can continue to manage these patients by using a multimodal approach with nonopioid medication, interventional procedures such as regional peripheral nerve blocks [39] or cryoanalgesia [40], as well as provide necessary psychological support [41]. Transitional pain clinics have been shown to reduce opioid use postoperatively, symptoms of anxiety and depression, pain catastrophizing, and pain [7,8]. Early identification of these clinical predictors, in conjunction with knowing the typical pain trajectories and patterns of common surgical procedures [42], can serve as the foundation for the basis of prescribing the right regimen and duration for the opioid prescription. Anesthesiology as a specialty, and especially in the setting of a dedicated acute pain service, is well positioned to take the lead in defining personalized pain medicine through all 3 phases of perioperative care [43].

**Conclusions**

Applying machine learning algorithms to electronic health data allows providers to develop models to predict more accurately and therefore appropriately allocate the limited health care resources (ie, transitional pain clinics). In this study, we showed that the need for regional anesthesia, high intraoperative opioid consumption, increased PACU pain scores, and opioid consumption were important features in models predicting outpatient opioid refills. Although providers are aware of the potential risk factors of opioid misuse, it remains challenging to accurately predict patients that will benefit from services as an outpatient. This prediction model serves as an example of a model that could be formalized into clinical decision support tools to help us better understand which patients will benefit from transitional pain clinics following ambulatory surgery.

**Conflicts of Interest**

SS is the Chief Executive Officer and founder of BrilliantBiome. RAG’s institution has received funding and products for other research projects from Epimed International, Infultronics, and SPR Therapeutics for the following authors: RAG and ETS. The University of California San Diego is a consultant for Avanos, which is represented by RAG. This work was funded by the Division of Perioperative Informatics at the University of California San Diego.

**References**


Abbreviations

AUC: area under the curve
MEQ: morphine equivalents
OR: odds ratio
PACU: postanesthesia care unit
SHAP: Shapley Additive Explanations
SMOTE: Synthetic Minority Oversampling Technique
Abstract

Background: Hyponatremia and hypernatremia, as conventionally defined (<135 mEq/L and >145 mEq/L, respectively), are associated with increased perioperative morbidity and mortality. However, the effects of subtle deviations in serum sodium concentration within the normal range are not well-characterized.

Objective: The purpose of this analysis is to determine the association between borderline hyponatremia (135-137 mEq/L) and hypernatremia (143-145 mEq/L) on perioperative morbidity and mortality.

Methods: A retrospective cohort study was performed using data from the American College of Surgeons National Surgical Quality Improvement Program database. This database is a repository of surgical outcome data collected from over 600 hospitals across the United States. The National Surgical Quality Improvement Program database was queried to extract all patients undergoing elective, noncardiac surgery from 2015 to 2019. The primary predictor variable was preoperative serum sodium concentration, measured less than 5 days before the index surgery. The 2 primary outcomes were the odds of morbidity and mortality occurring within 30 days of surgery. The risk of both outcomes in relation to preoperative serum sodium concentration was modeled using weighted generalized additive models to minimize the effect of selection bias while controlling for covariates.

Results: In the overall cohort, 1,003,956 of 4,551,726 available patients had a serum sodium concentration drawn within 5 days of their index surgery. The odds of morbidity and mortality across sodium levels of 130-150 mEq/L relative to a sodium level of 140 mEq/L followed a nonnormally distributed U-shaped curve. The mean serum sodium concentration in the study population was 139 mEq/L. All continuous covariates were significantly associated with both morbidity and mortality (P<.001). Preoperative serum sodium concentrations of less than 139 mEq/L and those greater than 144 mEq/L were independently associated with increased morbidity probabilities. Serum sodium concentrations of less than 138 mEq/L and those greater than 142 mEq/L were associated with increased mortality probabilities. Hypernatremia was associated with higher odds of both morbidity and mortality than corresponding degrees of hyponatremia.

Conclusions: Among patients undergoing elective, noncardiac surgery, this retrospective analysis found that preoperative serum sodium levels less than 138 mEq/L and those greater than 142 mEq/L are associated with increased morbidity and mortality, even within currently accepted “normal” ranges. The retrospective nature of this investigation limits the ability to make causal determinations for these findings. Given the U-shaped distribution of risk, past investigations that assume a linear relationship
between serum sodium concentration and surgical outcomes may need to be revisited. Likewise, these results question the current definition of perioperative eunatremia, which may require future prospective investigations.

**KEYWORDS**

hypernatremia; hyponatremia; perioperative care; postoperative complications; reference values; sodium; morbidity; mortality; database; data; cohort; surgery; sodium; preoperative; serum

**Introduction**

Abnormal preoperative sodium levels are associated with multiple adverse outcomes, including increased risk of venous thromboembolism, major bleeding and return to the operating room, perioperative coronary events, wound infection, and prolonged postoperative length of hospital stay [1-6]. Both hyponatremia and hypernatremia are associated with an increased risk of perioperative mortality [2,4,5]. Past investigations in nonsurgical populations suggest that optimizing sodium intake may reduce the risk of mortality [7,8]. While these studies provide a clinical rationale for intervention in the presence of hyponatremia or hypernatremia, the granularity of results has been limited due to broad categorizations of hyponatremia and hypernatremia.

Many previous studies investigating patient outcomes categorize sodium levels as hyponatremic (serum sodium concentration less than 135 mEq/L), eunatremic, and hypernatremic (serum sodium concentration greater than 145 mEq/L) [1,9-11]. Some studies also identified an increased risk of in-hospital and 1-year mortality in hospitalized patients with mild hyponatremia (125-134 mEq/L) and hypernatremia (146-150 mEq/L) [12,13]. Such evidence indicates that there are gradations of risk per sodium level outside of the eunatremic range, but it is unknown if such gradations of risk occur within the eunatremic range. Therefore, a more granular resolution is needed to determine if there is an increased risk of poor postoperative outcomes in patients within the range of serum sodium concentrations that are currently accepted as normal.

The culmination of research to date indicates that the role of sodium in morbidity and mortality risk is broad across a variety of surgeries, including hip arthroplasty [7,8], lower extremity arthroplasty [14], cervical spinal fusion [15], and cardiac surgery [9,16]. Moreover, risk prediction models, including those based on the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) data, indicate that sodium level, when categorized (eg, hyponatremia, eunatremic, and hypernatremia), is an important indicator of postsurgical morbidity and mortality in a large surgically diverse sample [17]. Such risk models do not allow clinicians to delineate an ideal target for clinical intervention. Taken together, there is a need to provide clinically informative research that evaluates the nonnormally distributed relationship between sodium levels, morbidity, and mortality across a large surgical population. Therefore, the purpose of this investigation was to explore the potential nonlinear relationship between preoperative sodium levels, modeled as a continuous predictor, and the odds of 30-day postoperative morbidity and mortality in patients undergoing elective, noncardiac surgery. We hypothesized that preoperative serum sodium concentration was independently associated with increased odds of both postoperative morbidity and mortality when modeled as a continuous variable, assuming a reference normal serum sodium concentration of 140 mEq/L.

**Methods**

**Ethical Considerations**

This study is a retrospective cohort design and was approved by the Naval Medical Center Portsmouth’s Institutional Review Board (NMCP.2021.0054).

**Study Design and Data Source**

Data from the ACS NSQIP database during the years 2015-2019 were obtained. These data come from over 700 hospitals and are collected using well-described methods to assure a high level of validity [18]. Noncardiac surgical procedures were included using current procedural terminology (CPT) codes 10000-32999 and 34000-69999. Patients undergoing cardiac surgery were excluded from this analysis due to the unique risks associated with that patient population, including the risks associated with cardiopulmonary bypass. Similar to previous investigations [19], we excluded minor surgeries such as endoscopies (CPT 43200-43272, 45300-45392, 46600-46608) and minor musculoskeletal procedures (CPT 29000-29750). Additionally, patients were excluded if they underwent emergency surgery.

The following demographic and health data were collected for each patient: CPT code, age, race, ethnicity, height, weight, sex assigned in the medical record, functional status, American Society of Anesthesiologists (ASA) Physical Score, sodium level, hematocrit, creatinine, steroid use, ascites, sepsis or septic shock, ventilator dependence, disseminated cancer, diabetes, hypertension, weight loss (at least 10% in the past year), congestive heart failure (CHF), dyspnea, smoking, chronic obstructive pulmonary disease, and dialysis. Patient records were included based on the following criteria: sodium, hematocrit, and creatinine assessment <5 days prior to surgery; BMI of >12 and <60; ages 18 to 89 years; hematocrit of >21% and <50%; sodium level of ≥130 mEq/L and ≤150 mEq/L; creatinine level of ≥0.5 mg/dL and ≤4.0 mg/dL; and undergoing surgery under a primary CPT listed in at least 50 patient records.

**Exposure**

The primary exposure was the preoperative sodium level. A priori, the serum sodium level of 140 mEq/L was empirically determined to be the reference value for the development of statistical models.
Outcomes

The 2 primary outcomes were defined as aggregate morbidity within 30 days of index surgery and mortality within 30 days of index surgery. Aggregate morbidity included any of the following: cardiac arrest, myocardial infarction, cerebrovascular accident, deep vein thrombosis, pulmonary embolism, postoperative sepsis or septic shock, renal insufficiency or failure, reintubation, failure to wean from the ventilator, pneumonia, wound dehiscence, or surgical site infection (including superficial, deep, or organ space). Details regarding the standardized definitions of these variables have been previously published [19].

Statistical Analysis

Univariate and Bivariate Analyses

First, nonparametric analyses (eg, chi-square, Kruskal-Wallis rank sum test, and Mann-Whitney U tests) examined differences between patient records that were and were not included in the analyses. Next, bivariate analyses evaluated differences in demographic characteristics and medical comorbidities by morbidity and mortality status. Bivariate analyses were performed using the TableOne R package (R Foundation) [20]. Due to the elevated likelihood of rejecting the null hypothesis (P<.05) in large samples and because the information rendered by the P value does not describe the strength of differences, both the P value and the standardized mean difference are reported for bivariate analyses. Standardized mean difference is reported specifically to describe the effect size of the included demographic characteristics and medical comorbidities on the outcomes of morbidity and mortality.

Inverse Probability Weights

Given the potential for selection bias in this analysis, outcome models included weights corresponding to the inverse probability of meeting inclusion criteria. This previously validated method accounts for selection bias due to missing predictor data [21]. Inverse probability weights were constructed through a multistep process. First, a generalized additive model (GAM) was conducted using the mgcv R package [22] to estimate the propensity of record inclusion. GAMs allowed for the modeling of nonlinear relationships between continuous predictors and the outcomes (smooth effects). In the GAM, the binary outcome was recorded as exclusion (0) versus inclusion (1), and the predictors were covariates associated with included versus excluded status. Sodium, creatinine, and hematocrit were not used in this analysis, as the lack of preoperative laboratory data was indicative of an excluded status. To account for the role of primary CPT in the propensity to be included, the proportion (%) of included patients per primary CPT was calculated. This proportion was included in the GAM as an additional covariate. The predicted and fitted values indicated the propensity of record inclusion given demographic characteristics, medical comorbidities, and primary CPT. Lastly, the propensity scores were transformed into inverse probability weights through the following formula: Inverse probability weight = (Included status / Propensity score) + ((1 Included status) / (1 Propensity score)). These weights were used to control for potential selection bias in subsequent outcome models [23].

Generalized Additive Models

The previously described factors associated with morbidity and mortality within the NSQIP database were included as covariates in 2 separate GAMs. One model was generated to predict aggregate morbidity, and the other to predict mortality. If missing data in the included sample was >1%, multiple imputations were planned. To assess the degree of multicollinearity, the performance R package was used to compute the variance inflation factor of each fixed covariate; a variance inflation factor <5 indicated acceptable levels of multicollinearity. GAM results were extracted using the sjPlot R package [24]. Estimated conditional means (95% CI) were calculated using the ggeffects R package [25]. Both the adjusted odds ratios (95% CI) and adjusted relative risks (RRs, 95% CI) of morbidity and mortality at sodium levels 130-150 mEq/L, relative to the a priori defined reference of 140 mEq/L, were calculated as well. The ggplo2 [26] and ggpubr [27] R packages were used to construct customized plots of model results. Statistical significance was indicated by P<.05.

Sensitivity Analysis

Sensitivity analyses were performed using E-values [28] and stratification of the included sample by the previously calculated propensity scores. The EValue R package [29] was used to calculate E-values corresponding to each RR of sodium levels 130-150 mEq/L. E-values indicate the strength a confounding variable would need to have on both the predictor (sodium) and outcome, beyond the effects of covariates already included in the model, to render the effect of sodium on the outcome null [30]. As such, E-values provide an assumption-free means of evaluating the robustness of model results [28]. For comparison purposes, the RR (95% CI) of fixed effects was also calculated. Within the included sample, propensity scores corresponding to the propensity to be included in analyses were divided into terciles. The outcome GAMs were replicated without the weights in the subsample of included records with the lowest tercile of propensity scores. Sensitivity analyses were graphically rendered for comparison purposes.

Results

Sample Description

Of the 4,551,726 patient records available, 1,003,956 met all inclusion criteria. Most patient records were excluded due to laboratory assessments occurring more than 4 days from surgery or not at all (n=3,388,178), continuous variables outside of the prespecified ranges (n=145,458), and a primary CPT that was not represented in at least 50 patient records (n=14,134). Bivariate analyses indicated that those included versus excluded differed across all identified demographic characteristics and medical comorbidities (Multimedia Appendix 1). In the included sample, 15,474 (0.3%) patient records had missing data; therefore, no imputation was performed. Morbidity and mortality rates in the included cohort were 8.5% and 1.3%, respectively. Descriptive statistics are reported (Table 1). Morbidity (Table 2) and mortality status (Table 3) are also reported. Bivariate
test results indicated that all demographic characteristics and medical comorbidities were associated with morbidity and mortality status. As such, all of these factors were included as covariates in the GAMs.

Table 1. Descriptive statistics of the overall sample (N=977,343).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), median (IQR)</td>
<td>60.0 (46.0-71.0)</td>
</tr>
<tr>
<td>Sex, n (%)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>452,054 (45.0)</td>
</tr>
<tr>
<td>Female</td>
<td>551,884 (55.0)</td>
</tr>
<tr>
<td>Race and ethnicity, n (%)</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>654,377 (65.2)</td>
</tr>
<tr>
<td>American Indian and Alaska Native</td>
<td>6366 (0.6)</td>
</tr>
<tr>
<td>Asian</td>
<td>27,927 (2.8)</td>
</tr>
<tr>
<td>Black</td>
<td>111,166 (11.1)</td>
</tr>
<tr>
<td>Latino</td>
<td>73,748 (7.3)</td>
</tr>
<tr>
<td>Native Hawaiian and Pacific Islander</td>
<td>3575 (0.4)</td>
</tr>
<tr>
<td>Other</td>
<td>2451 (0.2)</td>
</tr>
<tr>
<td>Unknown</td>
<td>124,346 (12.4)</td>
</tr>
<tr>
<td>BMI, median (IQR)</td>
<td>28.66 (24.69-33.67)</td>
</tr>
<tr>
<td>ASA\textsuperscript{a} physical status, n (%)</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>56,585 (5.7)</td>
</tr>
<tr>
<td>II</td>
<td>387,503 (38.7)</td>
</tr>
<tr>
<td>III</td>
<td>477,321 (47.7)</td>
</tr>
<tr>
<td>IV</td>
<td>79,712 (8.0)</td>
</tr>
<tr>
<td>Presence of comorbidities, n (%)</td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>481,752 (48.0)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>191,078 (19.6)</td>
</tr>
<tr>
<td>COPD\textsuperscript{b}</td>
<td>56,487 (5.6)</td>
</tr>
<tr>
<td>History of smoking</td>
<td>200,591 (20.0)</td>
</tr>
<tr>
<td>Chronic steroid use</td>
<td>48,421 (4.8)</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>14,385 (1.4)</td>
</tr>
<tr>
<td>Active cancer diagnosis</td>
<td>40,880 (4.1)</td>
</tr>
<tr>
<td>Sepsis or septic shock</td>
<td>37,231 (3.8)</td>
</tr>
<tr>
<td>Preoperative laboratory values, median (IQR)</td>
<td></td>
</tr>
<tr>
<td>Sodium (mEq/L)</td>
<td>139 (137-141)</td>
</tr>
<tr>
<td>Hematocrit (%)</td>
<td>39.2 (35.2-42.5)</td>
</tr>
<tr>
<td>Creatinine (mg/dL)</td>
<td>0.84 (0.70-1.01)</td>
</tr>
<tr>
<td>Percent CPT\textsuperscript{c} morbidity (IQR)</td>
<td>5.30 (2.63-11.28)</td>
</tr>
<tr>
<td>Percent CPT mortality (IQR)</td>
<td>0.24 (0.08-1.47)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}ASA: American Society of Anesthesiologists.
\textsuperscript{b}COPD: chronic obstructive pulmonary disease.
\textsuperscript{c}CPT: current procedural terminology.
Table 2. Aggregate morbidity outcomes status.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>No morbidity (N=918,385)</th>
<th>Morbidity (N=85,571)</th>
<th>P value</th>
<th>SMD&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age (years), median (IQR)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>410,367 (44.7)</td>
<td>41,687 (48.7)</td>
<td>&lt;.001</td>
<td>0.08</td>
</tr>
<tr>
<td>Female</td>
<td>508,002 (55.3)</td>
<td>43,882 (51.3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Race and ethnicity, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>597,599 (65.1)</td>
<td>56,778 (66.4)</td>
<td>&lt;.001</td>
<td>0.1</td>
</tr>
<tr>
<td>American Indian and Alaska Native</td>
<td>5814 (0.6)</td>
<td>552 (0.6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>25,943 (2.8)</td>
<td>1984 (2.3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>100,829 (11.0)</td>
<td>10,337 (12.1)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>Latino</td>
<td>69,192 (7.5)</td>
<td>4556 (5.3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native Hawaiian and Pacific Islander</td>
<td>3303 (0.4)</td>
<td>272 (0.3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>2281 (0.2)</td>
<td>170 (0.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>113,424 (12.4)</td>
<td>10,922 (12.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>BMI, median (IQR)</strong></td>
<td>28.69 (24.74-33.67)</td>
<td>28.69 (24.74-33.67)</td>
<td>&lt;.001</td>
<td>0.03</td>
</tr>
<tr>
<td><strong>ASA&lt;sup&gt;b&lt;/sup&gt; physical status, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>55,391 (6.0)</td>
<td>1194 (1.4)</td>
<td>&lt;.001</td>
<td>0.57</td>
</tr>
<tr>
<td>II</td>
<td>369,550 (40.3)</td>
<td>17,953 (21.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>426,405 (46.6)</td>
<td>50,916 (59.8)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>64,588 (7.1)</td>
<td>15,124 (17.8)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td><strong>Presence of comorbidities, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>430,631 (46.9)</td>
<td>51,121 (59.7)</td>
<td>&lt;.001</td>
<td>0.26</td>
</tr>
<tr>
<td>Diabetes</td>
<td>168,658 (18.4)</td>
<td>22,420 (26.2)</td>
<td>&lt;.001</td>
<td>0.21</td>
</tr>
<tr>
<td>COPD&lt;sup&gt;c&lt;/sup&gt;</td>
<td>47,235 (5.1)</td>
<td>9252 (10.8)</td>
<td>&lt;.001</td>
<td>0.21</td>
</tr>
<tr>
<td>History of smoking</td>
<td>181,440 (19.8)</td>
<td>19,151 (22.4)</td>
<td>&lt;.001</td>
<td>0.06</td>
</tr>
<tr>
<td>Chronic steroid use</td>
<td>41,464 (4.5)</td>
<td>6957 (8.1)</td>
<td>&lt;.001</td>
<td>0.15</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>11,224 (1.2)</td>
<td>3161 (3.7)</td>
<td>&lt;.001</td>
<td>0.16</td>
</tr>
<tr>
<td>Active cancer diagnosis</td>
<td>33,660 (3.7)</td>
<td>7220 (8.4)</td>
<td>&lt;.001</td>
<td>0.20</td>
</tr>
<tr>
<td>Sepsis or septic shock</td>
<td>31,324 (3.4)</td>
<td>5907 (6.9)</td>
<td>&lt;.001</td>
<td>0.17</td>
</tr>
<tr>
<td><strong>Preoperative laboratory values, median (IQR)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sodium (mEq/L)</td>
<td>139 (137-141)</td>
<td>139 (137-141)</td>
<td>&lt;.001</td>
<td>0.12</td>
</tr>
<tr>
<td>Hematocrit (%)</td>
<td>39.4 (35.6-42.6)</td>
<td>37.0 (32.0-41.0)</td>
<td>&lt;.001</td>
<td>0.41</td>
</tr>
<tr>
<td>Creatinine (m)g/dL</td>
<td>0.83 (0.70-1.00)</td>
<td>0.88 (0.70-1.11)</td>
<td>&lt;.001</td>
<td>0.22</td>
</tr>
<tr>
<td>Percent CPT&lt;sup&gt;d&lt;/sup&gt; morbidity (IQR)</td>
<td>5.02 (2.63-10.26)</td>
<td>12.38 (5.93-19.54)</td>
<td>&lt;.001</td>
<td>0.81</td>
</tr>
<tr>
<td>Percent CPT mortality (IQR)</td>
<td>0.22 (0.08-1.15)</td>
<td>1.25 (0.33-3.05)</td>
<td>&lt;.001</td>
<td>0.51</td>
</tr>
</tbody>
</table>

<sup>a</sup>SMD: standardized mean difference.

<sup>b</sup>ASA: American Society of Anesthesiologists.

<sup>c</sup>COPD: chronic obstructive pulmonary disease.

<sup>d</sup>CPT: current procedural terminology.
### Table 3. Mortality outcome status.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>No mortality (N=991,327)</th>
<th>Mortality (N=12,629)</th>
<th>P value</th>
<th>SMD&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), median (IQR)</td>
<td>60.00 (46.00-71.00)</td>
<td>75.00 (66.00-82.00)</td>
<td>&lt;.001</td>
<td>1.02</td>
</tr>
<tr>
<td>Sex, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>445,213 (44.9)</td>
<td>6841 (54.2)</td>
<td>&lt;.001</td>
<td>0.19</td>
</tr>
<tr>
<td>Female</td>
<td>546,096 (55.1)</td>
<td>5788 (45.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race and ethnicity, n (%)</td>
<td></td>
<td></td>
<td>&lt;.001</td>
<td>0.24</td>
</tr>
<tr>
<td>White</td>
<td>645,010 (65.1)</td>
<td>9367 (74.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Indian &amp; Alaska Native</td>
<td>6315 (0.6)</td>
<td>51 (0.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>27,682 (2.8)</td>
<td>245 (1.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>109,799 (11.1)</td>
<td>1367 (10.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latino</td>
<td>73,256 (7.4)</td>
<td>492 (3.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native Hawaiian and Pacific Islander</td>
<td>3546 (0.4)</td>
<td>29 (0.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>2430 (0.2)</td>
<td>21 (0.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>123,289 (12.4)</td>
<td>1057 (8.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMI, median (IQR)</td>
<td>28.69 (24.74-33.73)</td>
<td>25.99 (22.20-30.99)</td>
<td>&lt;.001</td>
<td>0.35</td>
</tr>
<tr>
<td>ASA&lt;sup&gt;b&lt;/sup&gt; physical status, n (%)</td>
<td></td>
<td></td>
<td>&lt;.001</td>
<td>1.25</td>
</tr>
<tr>
<td>I</td>
<td>56,574 (5.7)</td>
<td>11 (0.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>386,890 (39.1)</td>
<td>613 (4.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>470,707 (47.6)</td>
<td>6614 (53.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>74,518 (7.5)</td>
<td>5194 (41.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presence of comorbidities, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>472,832 (47.7)</td>
<td>8920 (70.6)</td>
<td>&lt;.001</td>
<td>0.48</td>
</tr>
<tr>
<td>Diabetes</td>
<td>187,320 (18.9)</td>
<td>3758 (29.8)</td>
<td>&lt;.001</td>
<td>0.29</td>
</tr>
<tr>
<td>COPD&lt;sup&gt;c&lt;/sup&gt;</td>
<td>54,107 (5.5)</td>
<td>2380 (18.8)</td>
<td>&lt;.001</td>
<td>0.42</td>
</tr>
<tr>
<td>History of smoking</td>
<td>198,133 (20.0)</td>
<td>2458 (19.5)</td>
<td>&lt;.15</td>
<td>0.01</td>
</tr>
<tr>
<td>Chronic steroid use</td>
<td>47,044 (4.7)</td>
<td>1377 (10.9)</td>
<td>&lt;.001</td>
<td>0.23</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>13,072 (1.3)</td>
<td>1313 (10.4)</td>
<td>&lt;.001</td>
<td>0.39</td>
</tr>
<tr>
<td>Active cancer diagnosis</td>
<td>38,451 (3.9)</td>
<td>2429 (19.2)</td>
<td>&lt;.001</td>
<td>0.50</td>
</tr>
<tr>
<td>Sepsis or septic shock</td>
<td>35,342 (3.6)</td>
<td>1889 (1.5)</td>
<td>&lt;.001</td>
<td>0.42</td>
</tr>
<tr>
<td>Preoperative laboratory values, median (IQR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sodium (mEq/L)</td>
<td>139 (137-141)</td>
<td>138 (136-141)</td>
<td>&lt;.001</td>
<td>0.19</td>
</tr>
<tr>
<td>Hematocrit (%)</td>
<td>39.3 (35.3-42.5)</td>
<td>33.0 (28.8-38.0)</td>
<td>&lt;.001</td>
<td>0.88</td>
</tr>
<tr>
<td>Creatinine (mg/dL)</td>
<td>0.83 (0.70-1.01)</td>
<td>1.00 (0.76-1.44)</td>
<td>&lt;.001</td>
<td>0.54</td>
</tr>
<tr>
<td>Percent CPT&lt;sup&gt;d&lt;/sup&gt; morbidity (IQR)</td>
<td>5.30 (2.63-10.90)</td>
<td>12.54 (9.08-18.46)</td>
<td>&lt;.001</td>
<td>0.91</td>
</tr>
<tr>
<td>Percent CPT mortality (IQR)</td>
<td>0.23 (0.08-1.37)</td>
<td>2.91 (1.25-5.09)</td>
<td>&lt;.001</td>
<td>1.05</td>
</tr>
</tbody>
</table>

<sup>a</sup>SMD: standardized mean difference.
<sup>b</sup>ASA: American Society of Anesthesiologists.
<sup>c</sup>COPD: chronic obstructive pulmonary disease.
<sup>d</sup>CPT: current procedural terminology.

**GAM Results**

In both outcome GAMs, all continuous covariates (age, BMI, sodium, hematocrit, creatinine, and percent CPT morbidity or mortality) were modeled as smooth terms and were substantially associated with both morbidity and mortality. Across both models, patients assigned male in the medical record with an elevated ASA status, steroid use, sepsis or septic shock, cancer, a positive smoking status, chronic obstructive pulmonary disease, or a history of heart failure were more likely to experience either morbidity or mortality.
disease, renal failure, and CHF were more likely to experience morbidity and mortality compared to their reference counterparts (Multimedia Appendix 2). When controlling for other demographic characteristics and medical comorbidities, patients whose race and ethnicity were listed as Asian or Latino had a lower probability of morbidity and mortality relative to White patients. Similarly, White patients had a greater probability of morbidity relative to Native Hawaiian and Pacific Islander patients, but White patients had a lower probability of morbidity than patients of unknown race and ethnicity. Patients with diabetes and hypertension had a greater probability of morbidity relative to those without these conditions. The odds of morbidity and mortality across sodium levels of 130-150 mEq/L relative to a sodium level of 140 mEq/L followed a U-shaped curve (Figure 1).

![Figure 1. Odds ratios (95% CI) of morbidity (left) and mortality (right).](image)

**Sensitivity Analyses**

The E-values corresponding to the RR of morbidity and mortality across sodium levels are shown in Figures S1A and S1B in Multimedia Appendix 3. For example, at a sodium level of 135 mEq/L, the morbidity (RR 1.07, 95% CI 1.07-1.07) and mortality (RR 1.30, 95% CI 1.30-1.30) would be rendered null if an unmeasured confounder was associated with both sodium and the outcome by a RR of 1.35-fold (lower 95% CI 1.35) and 1.93-fold (lower 95% CI 1.92), respectively. For reference, these E-values are similar to the effects of ASA I versus III on morbidity (RR 1.33, 95% CI 1.32-1.33) and CHF on mortality (RR 1.97, 95% CI 1.94-1.99). The RR (95% CI) for fixed effects on both outcomes are reported in Multimedia Appendix 2.

Lastly, GAMs evaluating morbidity and mortality were conducted on a subsample of the included group with the lowest tercile propensity scores (n=333,701). Model results were similar to the main analysis, such that the effect of sodium was significant (P<.001), and the nonlinear pattern followed a U-shape (Figures S1C and S1D in Multimedia Appendix 3).

**Discussion**

**Principal Results**

This exploratory analysis calls into question the current understanding of the “normal” range of serum sodium levels (135-145 mEq/L) within the context of perioperative care, as values of serum sodium concentration within this range of normal values were associated with 30-day aggregate morbidity and mortality. By examining preoperative sodium levels in over 1 million patients as a continuous variable instead of the commonly used categories (eg, hyponatremic, eunatremic, and hypernatremic), this study provides improved granularity on the association between small deviations in sodium and perioperative outcomes. As such, what is considered “normal” sodium values in the general population may not be normal in patients undergoing elective noncardiac surgery.

**Comparison With Prior Work**

As health care shifts to value-based care, these findings may also play a role in evaluating value-based perioperative practices. For example, recent evidence using NSQIP data indicates that preoperative laboratory assessment is not associated with the odds of postoperative complications and readmission in patients undergoing ambulatory surgery with an ASA I or II status, thereby suggesting the low value of preoperative laboratory assessment [31]. However, such findings may be premised on clinician practices that are contingent on a definition of “normal” that is, per these findings, associated with increased risk of aggregate morbidity and mortality (eg, ~135 mEq/L). Given the potential impact of these findings, combined with the lack of causal assumptions that can be made, future work is needed to assess whether clinical intervention addressing high- and low-normal sodium serum concentrations improves clinical outcomes and value-based care.

**Strengths and Limitations**

This study possessed several strengths. Though the main variable of interest was serum sodium concentrations, models included many demographic characteristics and medical comorbidities that have previously been shown to be substantially associated with aggregate morbidity and mortality risk. These factors included other laboratory values (eg, creatinine and hematocrit) that may also warrant further inspection, given their relationship with postoperative outcomes. By controlling for these covariates and using a weighted approach based on the inverse probability of record inclusion, the results of this study are likely
generalizable to adult patients undergoing any elective, noncardiac surgery in the United States. We restricted records to those with laboratory results collected less than 5 days before surgery, thereby increasing the likelihood that the recorded values actually reflected serum sodium levels at the time of surgery.

This study was tempered by several limitations. First, no causal conclusions can be drawn from the study due to the retrospective, associative nature of the study design and analytic approach. Additionally, there may be several covariates, including specific health conditions, medication receipt (both in the days leading up to surgery and perioperatively), preoperative recommendations (eg, fasting), and prior health care received, that are neither collected in the NSQIP database nor included in the analysis but could be associated with morbidity and mortality. While this database is a robust and extensive collection of surgical outcome data in the United States [18], the inclusion of other covariates mentioned above could serve to refine this model and provide more specific areas of research to explore. Examples of other potential confounders include medications, preoperative fasting, and certain comorbidities, which themselves may be associated with abnormal sodium levels. When considering the potential impact of missing confounders on model results, E-values indicated that any confounder would need to surpass the strength of most fixed covariates within our models and account for unique variance not otherwise accounted for by current covariates to render the effect of sodium null.

Conclusions
This analysis indicated that both preoperative hyponatremia and preoperative hypernatremia were associated with an increased risk of 30-day aggregate morbidity and mortality. The relationship was nonlinear, such that the risk increased with further deviation from a serum sodium concentration of 140. While prior investigations have demonstrated that dysnatremia is a modifiable risk factor and optimization of preoperative serum sodium levels may represent an opportunity for a reduction in both perioperative morbidity and mortality [31], this study suggests that preoperative serum sodium levels that are within the currently accepted upper and lower limits of normal are likely indicative of elevated risk. As such, future prospective studies are needed to better confer sodium level ranges associated with optimized outcomes after surgery, as well as the potential to directly alter patients’ serum sodium concentrations to improve postoperative outcomes.

Disclaimer
The views expressed are solely those of the authors and do not reflect the official policy or position of the Uniformed Services University, US Army, US Navy, US Air Force, the Department of Defense, the US Government, or the Henry M Jackson Foundation for the Advancement of Military Medicine, Inc. We are military service members. This work was prepared as part of our official duties. Title 17 U.S.C. 105 provides that “Copyright protection under this title is not available for any work of the United States Government.” Title 17 U.S.C. 101 defines a United States Government work as a work prepared by a military service member or employee of the United States Government as part of that person’s official duties. The American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) and the hospitals participating in the ACS NSQIP are the sources of the data used herein; they have not verified and are not responsible for the statistical validity of the data analysis or the conclusions derived by the authors.

Authors’ Contributions
JHC, SBH, and GJB were responsible for the initial conceptualization of this investigation. JHC, SBH, AHG, GCB, and GJB developed the formal methodology that was used in this study. JHC, BJO, TH, and GJB all participated in the background investigation for this work. KBH was instrumental in the software development, formal analysis, and data curation of this investigation. JHC, SBH, BJO, TH, and GJB all assisted in the creation of the original draft, while JHC, KBH, SBH, AHG, GCB, and GJB all facilitated the review and editing process.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Bivariate differences between excluded versus included records.
[XLSX File (Microsoft Excel File), 13 KB - periop_v6l1e38462_app1.xlsx]

Multimedia Appendix 2
Risk ratios (95% CI) of fixed covariates across outcomes.
[XLSX File (Microsoft Excel File), 10 KB - periop_v6l1e38462_app2.xlsx]

Multimedia Appendix 3
Sensitivity analyses of sodium effects. (A,B) Adjusted relative risk (95%CI) and E-values (lower 95% CI) of morbidity and mortality, respectively. (C,D) Odds ratios (95%CI) of morbidity and mortality, respectively.

References


Abbreviations

ACS: American College of Surgeons
ASA: American Society of Anesthesiologists
CHF: congestive heart failure
CPT: current procedural terminology
GAM: generalized additive model
NSQIP: National Surgical Quality Improvement Program
RR: relative risk
Toward Enhanced Clinical Decision Support for Patients Undergoing a Hip or Knee Replacement: Focus Group and Interview Study With Surgeons

Sabrina Grant1*, BSc, MSc, PhD; Emma Tonkin2*, MPhys, PhD; Ian Craddock2, CEng, PhD; Ashley Blom3, MD, PhD; Michael Holmes3, PhD; Andrew Judge1, PhD; Alessandro Masullo2, PhD; Miquel Perello Nieto2, MSc; Hao Song2, PhD; Michael Whitehouse1, MBChB, PhD; Peter Flach3, PhD; Rachael Gooberman-Hill1, PhD

1Musculoskeletal Research Unit, University of Bristol, Southmead Hospital, Bristol Medical School, Bristol, United Kingdom
2Digital Health, Faculty of Engineering, Bristol, United Kingdom
3Faculty of Medicine, Dentistry and Health, University of Sheffield, Sheffield, United Kingdom
4Intelligent Systems Laboratory, Department of Computer Science, University of Bristol, Bristol, United Kingdom
*these authors contributed equally

Corresponding Author:
Emma Tonkin, MPhys, PhD
Digital Health
Faculty of Engineering
University of Bristol
Bristol, BS8 1TH
United Kingdom
Phone: 44 117 42 82343 ext 82343
Email: cselt@bristol.ac.uk

Abstract

Background: The current assessment of recovery after total hip or knee replacement is largely based on the measurement of health outcomes through self-report and clinical observations at follow-up appointments in clinical settings. Home activity-based monitoring may improve assessment of recovery by enabling the collection of more holistic information on a continuous basis.

Objective: This study aimed to introduce orthopedic surgeons to time-series analyses of patient activity data generated from a platform of sensors deployed in the homes of patients who have undergone primary total hip or knee replacement and understand the potential role of these data in postoperative clinical decision-making.

Methods: Orthopedic surgeons and registrars were recruited through a combination of convenience and snowball sampling. Inclusion criteria were a minimum required experience in total joint replacement surgery specific to the hip or knee or familiarity with postoperative recovery assessment. Exclusion criteria included a lack of specific experience in the field. Of the 9 approached participants, 6 (67%) orthopedic surgeons and 3 (33%) registrars took part in either 1 of 3 focus groups or 1 of 2 interviews. Data were collected using an action-based approach in which stimulus materials (mock data visualizations) provided imaginative and creative interactions with the data. The data were analyzed using a thematic analysis approach.

Results: Each data visualization was presented sequentially followed by a discussion of key illustrative commentary from participants, ending with a summary of key themes emerging across the focus group and interview data set.

Conclusions: The limitations of the evidence are as follows. The data presented are from 1 English hospital. However, all data reflect the views of surgeons following standard national approaches and training. Although convenience sampling was used, participants’ background, skills, and experience were considered heterogeneous. Passively collected home monitoring data offered a real opportunity to more objectively characterize patients’ recovery from surgery. However, orthopedic surgeons highlighted the considerable difficulty in navigating large amounts of complex data within short medical consultations with patients. Orthopedic surgeons thought that a proposed dashboard presenting information and decision support alerts would fit best with existing clinical workflows. From this, the following guidelines for system design were developed: minimize the risk of misinterpreting data, express a level of confidence in the data, support clinicians in developing relevant skills as time-series data are often unfamiliar, and consider the impact of patient engagement with data in the future.

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2018-021862
Introduction

Background

Hip and knee replacements are major surgical procedures that aim to improve function and reduce pain related to joint diseases, particularly osteoarthritis. During hip or knee replacement, the affected joint is removed and replaced with an artificial joint. In 2019 in the United Kingdom, the National Joint Registry recorded 101,651 hip replacements and 108,713 knee replacements [1]; in the United States, >1 million total hip and knee replacement procedures are performed each year [2]. These surgical procedures are increasingly common, and numbers are projected to increase as a result of aging populations and increasing prevalence of obesity [3].

UK clinical guidelines for follow-up after hip and knee replacement surgery usually include face-to-face consultation, radiographs, and an assessment of health outcomes through telephone or web-based patient-reported outcome measures (PROMs) [4]. PROMs are designed to assess patients’ own views of their health and outcomes without interpretation by clinicians or others [5]. Of these, generic measures such as the 12-item Short Form Survey [6] and EQ-5D [7] aim to assess all important dimensions of health-related quality of life [8]. The Oxford Hip Score (OHS) [9] and Oxford Knee Score (OKS) [10] are additional disease-specific PROMs used by orthopedic surgeons in the United Kingdom. As validated instruments, PROMs are valuable sources of information for clinicians and researchers. However, several practicalities must be considered when implementing PROMs: missing or incomplete data; potential burden for patients; and cost, time, and administrative labor-intensiveness [11-13]. A recent review found that PROMs were prone to several types of bias: bias because of collection mode; nonresponse bias; proxy or caregiver response bias; recall bias (eg, bias because of the quality of patient recollection of past states); language bias (eg, semantic ambiguity); timing bias, representing a limited number of snapshots; and fatigue bias [13,14]. The OKS and OHS in particular may also fail to stratify activity level across a younger, more active population as they are not designed for this purpose. Instead, other instruments such as the Knee Injury and Osteoarthritis Outcome Score may be used, which are specifically designed for young and physically active patients, capturing additional domains of sport and recreation function and knee-related quality of life such that it has greater responsiveness as an outcome measure [15]. Taken together, these issues mean that, although PROMs are extremely valuable sources of information for clinicians and researchers, particularly because of their standardized and validated status, it is worthwhile to consider other methods to assess outcomes after joint replacement [16,17] that could be used in parallel to PROMs to support decision-making.

In this study, we used a qualitative approach to explore how time-based data may be used by clinicians to supplement PROMs. Qualitative methods were used to explore inductively what matters to busy clinical staff and develop initial guidelines for a future system. Any system developed using these guidelines could be evaluated in future studies.

Measuring Activity in a Joint Replacement Population

An objective method of activity assessment—step counting—has been accurately used to monitor changes in gait and activity in musculoskeletal disorders and diseases affecting gait, including hip and knee arthritis [18,19]. The current objective method used to measure function is accelerometry via wearable sensors. These are inexpensive and easy to wear. However, the data currently derived from these sensors have some limitations, particularly when measuring complex activities and movements that are common in activities of daily living [17,20]. To capture the daily variation in a patient’s functional abilities in their real living environment, it is necessary to move to automated measurement in the patient’s home as well as toward analysis techniques that more directly reflect performance in activities of daily living. For example, cameras can be used to study the kinematics of the transition from sitting to standing [21].

The Sensor Platform for Healthcare in a Residential Environment (SPHERE) Interdisciplinary Research Collaboration has developed a technology comprising an integrated platform of low-power sensors that can measure information continuously about the home (eg, temperature, energy consumption, and humidity) as well as information about people in the home (eg, location, how active they are, and extent of movement) and their health-related behaviors [22]. Data capture has been demonstrated over months or years [23], and so the continuous time-series data collected by SPHERE offer a potentially useful source of data to supplement conventional methods such as PROMs.

This study considers real time-series data generated by SPHERE systems monitoring patient activity in the home before and after total hip or knee replacement. The types of data available from the SPHERE system in each home include metrics derived from Bluetooth-based indoor localization of the patient [18], continuous estimation of posture and ambulatory activities using a wrist-worn accelerometer [24], and silhouette data generated using a depth-sensing video camera [25]. Although the overall system was developed by SPHERE in the absence of equivalent commercial systems, the capabilities of such a system would readily be within the reach of several companies in the consumer smart home market. The costs of systems of this kind vary according to implementation decisions as different use cases may benefit from the deployment of different sensors. The patient burden is likely to be minimal once the system is successfully in place. In comparison with cross-sectional PROMs, the costs of continuous time-series data monitoring lie primarily in maintenance following initial installation, and hence, time-series approaches may be more practical for...
longer-term observation of patients’ symptoms. Therefore, the findings of this study are a good guide to the strengths, weaknesses, and potential clinical utility of a near product.

**Objectives**

The main objectives of this study were to (1) introduce surgeons to continuous home data by visualizing time-series sensor data, (2) understand how these data could assist in postoperative clinical decision-making, and (3) identify design recommendations arising from clinician feedback.

The study departs from previous literature on the use of data in clinical decision support, which is largely focused on data from clinical environments such as intensive care [26,27]. To date, studies that have presented data from community settings to surgeons have focused principally on manually clinician-reported data [28] and laboratory outcomes [29], such as those commonly stored in electronic health records, or patient self-reported data [30] such as PROMs [31]. Where sensor data are sampled, this is often at a relatively low sample rate (e.g., a daily measurement or 12 measurements per day) or over a relatively short period, from a few minutes [32] to a week or month [33]. Herein, we consider the challenges of how busy surgeons would make sense of thousands of data points a day over periods of up to 3 months—as would be easily within the capability and requirements of a home-based sensor system [34,35] monitoring recovery from major surgery [36].

**Methods**

**Participants and Recruitment**

From October 2018 to May 2019, orthopedic surgeons at a hospital in South West England, United Kingdom, were invited to take part in a focus group study. Participant demographics were collected (sex and level of experience performing hip or knee replacement surgical procedures). Identification of potential participants was conducted using convenience and snowball sampling. During this process, surgeons known to the study team were asked to identify other potential participants. Participants were initially screened against the inclusion criteria (a minimum of 2 years of experience performing total hip and knee replacement surgical procedures). The exclusion criteria were a lack of experience in joint replacement specific to the hip or knee. Potential participants were emailed invitations, and those who agreed to consider taking part were invited to attend focus groups. Individual interviews were offered if the focus group timings were not suitable. Those who were contacted were also asked to nominate other potential participants—a snowball sampling approach.

A total of 9 participants (surgeons who were either working as consultants or registrars [residents]) took part. Several potential participants declined because of their clinical workloads and time constraints. At the start of each focus group, the study was discussed with the potential participants, who were invited to ask any questions about the study. Before the focus group started, they provided their written informed consent to participate, including to the publication of anonymous quotations. Focus groups were held at a clinical research center on the same site as the hospital to make it as straightforward as possible for busy surgeons to attend [37]. Face-to-face interviews at the surgeons’ places of work were offered where attendance to the focus group was impractical because of time or distance. In total, most of the surgeons (7/9, 78%) attended 3 focus groups, and 22% (2/9) attended one-to-one interviews. The sample size was considered adequate as enough information was collected to clearly demonstrate concepts or ideas related to the topic addressed and with sufficient repetition of those concepts [38].

**Ethics Approval**

Ethics approval was provided by Southwest – Central Bristol National Health Service (NHS) Research Ethics Committee (17/SW/0121) on June 22, 2017.

**Topic Guide and Procedure**

A structured topic guide (Multimedia Appendix 1) was developed by the research team, which comprised an interdisciplinary group of health researchers and psychologists (SG and RGH), orthopedic surgeons (AB and MW), data scientists (IC, HS, ET, MP, AM, MH, and PF), and a translational statistician (AJ).

In part 1, a scenario (Textbox 1) was used as a tool to explore the current clinical systems in orthopedic care.

In part 2, a series of visualizations (Figures 1-9) were presented based on real participant data from the SPHERE 100 Homes study [23] (in which the system was deployed in homes of the general public) and the Hip and Knee Study of a Sensor Platform for Healthcare in a Residential Environment [39] (in which the same system was deployed in the homes of orthopedic patients).

**Textbox 1. Scenario 1—Joyce (aged 63 years).**

- Joyce is a 63-year-old lady who lives in a large three storey house with her daughter and daughter’s fiancé. Joyce is a self-employed therapist and runs her practice from her home. She has a second part-time role at the local University as an administrator.
- Joyce previously had trouble walking distances. Because of a limp she uses a walking aid at times and reports significant hip pain.
- Joyce has recently had her left hip replaced.

Many metrics can be generated using home sensor data. For the purposes of this study, a series of target metrics were generated based on the literature on hip and knee studies. Metrics referenced in manually administered survey instruments such as the OHS and OKS [10] and the Pittsburgh Sleep Quality Index [40] were considered useful targets. Once this step was complete, a series of sample visualizations was generated using these metrics. These visualizations were first proposed and improved over multiple discussions and careful analysis of real patients by members of the Hip and Knee Study of a Sensor Platform for Healthcare in a Residential Environment.
Platform for Healthcare in a Residential Environment—mainly data scientists and health researchers. After several iterations, the resulting visualizations were used as examples to provide during the focus groups with clinicians. The detailed rationale behind the development of these figures has been published separately [24]. The use of realistic (eg, noisy and incomplete) prototype data from real homes and real patients was considered desirable throughout this study to ensure that the feedback was related to the characteristics of achievable systems that could plausibly be developed for clinical use in the future.

Scenario-Based Exploration

An action research approach was used in which participants were seen as able to identify value in context when encouraged to take initiative and identify possibilities for improvements [41].

Participating orthopedic surgeons were presented with a fictitious but realistic orthopedic patient scenario (Textbox 1); the narrative nature of the scenario approach is known to be a useful tool in the design process [41]. Surgeons were asked to focus on the 6- to 8-week postoperative consultation for this hypothetical patient. This creates a familiar and meaningful context [42] in which they are well placed to imagine whether new forms of data would assist them in carrying out their professional responsibilities.

To maximize discussion and allow participants to write thoughts and views independently, participants were provided with printouts of presented visualizations for use within idea generation sessions.

Focus Groups and Interviews

A total of 3 focus groups (with 2-3 surgeons in each group) and 2 interviews were conducted to explore the data visualizations. Each focus group was facilitated by 2 researchers and lasted approximately 1 to 1.5 hours. The interviews lasted approximately 45 to 60 minutes. The focus groups and interviews were digitally audio recorded and transcribed verbatim.

Part 1 of each data collection phase discussed the scenario (Textbox 1) exploring the assessment of recovery for patients after surgery.

Part 2, led by data analysts (MH and MPN), was a structured exploration in which the surgeons were presented with a selection of visualizations based on data generated from the homes of 2 orthopedic patients who had been recovering from total hip replacement. Participants were asked to consider the use of the visualizations as a way of assessing patient outcome and recovery after surgery. Participants were provided with paper copies of each visualization for any further thoughts or comments that were not captured by discussion—these were reviewed during the coding process.

Data Analysis

Qualitative data from the focus groups and interviews were analyzed using an inductive thematic approach [43]. The initial labeling generated a list (a “frame”) that was then systematically applied to the data and refined as the analysis progressed [44]. Members of the research team from clinical and nonclinical disciplines were allocated 10% of these transcripts to label independently. After collaborative discussions, further labels were identified, defined, and grouped into themes. This process of investigator triangulation increases internal validity [44]. Excerpts of data were placed on charts according to themes. All data were managed using NVivo software (version 12.0; QSR International).

This qualitative study focuses on the views expressed by the surgeons. Quantitative data were presented to surgeons to elicit those views, but the quantitative data themselves were not the subject of this study. A brief description of the method used to generate the quantitative data and visualizations is provided in the Results section, and the interested reader is referred to the study by Holmes et al [24] for further details.

Once participant feedback was evaluated and themes were identified, feedback was presented to the interdisciplinary research team consisting of researchers, surgeons, machine learners, and interface engineers. This step was intended to facilitate the integration of these findings into future iterations of the sensor and data analysis platform. This step resulted in the development of a series of guidelines that integrate findings from the participants with insights from the interdisciplinary team. This asynchronous codevelopment approach offers an opportunity for participant surgeon feedback and guidance to be made available for future engineering and design processes through the provision of guidelines.

Results

Participating surgeons. Quantitative data were presented to surgeons to elicit those views, but the quantitative data themselves were not the subject of this study. A brief description of the method used to generate the quantitative data and visualizations is provided in the Results section, and the interested reader is referred to the study by Holmes et al [24] for further details.

Once participant feedback was evaluated and themes were identified, feedback was presented to the interdisciplinary research team consisting of researchers, surgeons, machine learners, and interface engineers. This step was intended to facilitate the integration of these findings into future iterations of the sensor and data analysis platform. This step resulted in the development of a series of guidelines that integrate findings from the participants with insights from the interdisciplinary team. This asynchronous codevelopment approach offers an opportunity for participant surgeon feedback and guidance to be made available for future engineering and design processes through the provision of guidelines.

Outcomes From Focus Groups and Interviews

Individual commentary from participants is presented regarding each of the visualizations (Figures 1-9), followed by a series of broader themes from the focus groups and interviews.

Visualization 1 (Figure 1) presented a series of summary statistics calculated using patient indoor location and accelerometer data. These included room-level occupancy data,
transitions between rooms, activity predictions generated via machine learning, and actigraphy analysis. The sample data given here were intended to be representative, not exhaustive, and it is possible to identify many other summary statistics that could be relevant.

**Figure 1.** Visualization 1—a screenshot of a dashboard displaying summary statistics, including activity, location, room transitions, floor transitions, and sleep routine.

Participants stated that, although the tabulated data looked very informative, rapid extraction of relevant information was not straightforward given the constraints of a busy 10-minute clinic consultation. Population norms or other references would be required to assess any change or improvement. Visuals would be improved if areas were highlighted to provide a focus for surgeons during a consultation:

So I don’t know where the mobility is on total hours walking, 76, I don’t know what that means, is 76 a lot, is it not a lot? Obviously if you had pre-op and post-op data then that’s great because you could get the data just to show you which are better, worse, whatever, but that I think I wouldn’t look at because it would be too hard to navigate. [#0046]

Participants reflected on how current methods of assessing patient health outcomes using the OHS or OKS lacked some of the valuable temporal information contained within the SPHERE time-series data. It was suggested that this insight into function over time could help them better understand the recovery process.

Most participants (7/9, 78%) suggested that the existing routine (face-to-face) clinic follow-up appointment presented a similarly rich opportunity for assessing recovery through movement. See the Themes Arising From Data Analysis section for discussion of this point:

> So I think actually the bits that I think are important on here all come back to Oxford Hip Score. So people getting up and down stairs, people on the move, people sleeping you know, these are all things that are kind of covered in one way and that, but this [the sensor data] gives you more detail, it’s not just “Yes, good, very good” or whatever. So for me I think, although like the moving one [visualization 7] is quite cool...I don’t see how me looking at that no matter how many hours I had to look at the patient, that when they come in and they stand up and sit down I've made my judgment whether they’ve got a problem or not. [#0050]

I suppose I’m expecting a patient to be compared against other patients that I’ve seen before. Usually, when I see them, the first thing I do is watch them walk. I watch them walk into the consultation room and check whether or not they’re using walking aids. I then take a good history about how their recovery is going and whether that’s meeting their expectations as well as mine. [#0051]

Visualizations 2 and 3 (Figures 2 and 3) illustrated the recovery of 2 patients over weeks 1 and 6 after surgery. The top and bottom figures correspond to weeks 1 and 6, respectively. Patient 1 was an example of a “good recovery,” and patient 2 was an example of a “poor recovery.” “Good” and “poor” recovery were determined descriptively from analysis of PROM data on function and sleep and qualitatively from interviewing the patients, which were data collected as part of the wider program of study [45,46].
Figure 2. Visualization 2—activity levels of a patient recovering well after surgery.
The visualizations made use of accelerometer magnitude data, preprocessed to establish the SD, which is an approximate measure of physical activity [47]. These data were then visualized as a series of horizontally stacked axes, each representing a day of the week. Data for 1 week were available in each chart. Stacked charts are a commonly used approach to time-series visualization, with known limitations; notably, line length is visually easier to compare than position [48], and hence, comparison of multiple series is challenging using this type of chart.

Some general suggestions were offered to improve visualization 2. Displaying the mean value of that week rather than daily values was suggested as easier to use:

*Because if you put all that together you would get a trace that resembles that [visualization 2/3] and you’d be able to say straightaway that at six weeks, they’re doing great and then at one [week], they’re not.*

However, it was noted that striking the right balance of information using the mean would be a challenge:

*The problem is, if you average everything out, then you lose the detail don’t you...but if you present all of the detail, then it becomes impossible.*

Adding a reference value was suggested by some as useful within the weekly charts to demonstrate a “typically” poor recovery and where the patient sits in relation to that. Visually representing this recovery process for patients was again considered a better outlet for a conversation regarding the expectations of surgery:

*I guess in the second slide [visualization 2-3] if you were able to put some points to say, this is low activity at time of sleep which is what you’re expecting...*

However, there was a level of disagreement between some surgeons about using population-based comparators with patients, that is, comparing a patient’s data with population norms or averages:

*No I think it would depersonalize it in my opinion...Your [the patient’s own] starting point...may be very, very much lower than another patient. So I...*
would just work on an individual improvement. [#0050]

The assessment of sleep patterns across a 12-week period is beyond the conventional self-reported assessment of sleep. Within these focus groups, participants viewed long-term changes in sleep as a new approach to understanding recovery. However, as it was an unfamiliar metric, there were varying opinions about how useful this could be for making clinical decisions. On the one hand, it may be helpful for more tailored advice:

Postop [after surgery] we tell them, they’ve got to sleep on their back and most of them have got a bad back, and they hate it, so that’s why they’re doing this but this would be so interesting if in time we changed to give them advice to sleep any which way they like...you might notice a real difference. [#0049]

In contrast, some felt that the minutiae of sleep were affected by several different factors following surgery and, therefore, could not be assessed or considered alone:

There are so many other factors that are going to affect sleep other than the joint replacement, especially in this particular demographic...There are potentially too many confounding factors in there for us to be able to use it [visualization 4] usefully. [#0051]

Visualization 4 (Figure 4) graphically represents intraday variance of sleep patterns drawn from raw actigraphy data.

---

Patient 1

(reCOVERS well according to PROMS)

```plaintext
Inter-day variance in sleep patterns decreases across 12-week period, indicating good recovery
```

Patient 2

(poor recovery according to PROMS)

```plaintext
Inter-day variance in sleep patterns remains high at 12 weeks, indicating poor recovery
```

Visualization 5 (Figure 5) presented a spiral representation of patient physical activity (derived from accelerometer data). Each complete ring represented a week of data—the spiral map was chosen in response to the observation by Weber et al [49] that spiral charts permit visualization of lengthy time series and that the circular representation is appropriate for time series with high periodicity (in this case, weekly periodicity). Each chart represented approximately 2 to 3 months of data.

---

**Figure 4.** Visualization 4—sleep trend data. PROM: patient-reported outcome measure.

**Figure 5.** Visualization 5—long-term interpatient comparison of movement. PROM: patient-reported outcome measure.
The color scheme applied identified absence of data as pure white and represented increasingly strenuous activity using darkening shades of blue. Hence, regular and restful sleep patterns were reflected using light blue “striping” through each period of least activity, with the most active periods represented as dark areas. Periods outside the home resulted in white striping (absent data) on the chart as the project did not collect data outside the home. Known limitations of the spiral format include difficulty reading “older” data at the center of the spiral because of the small area of the central cells. Facilitation involved some narrative running alongside the presentation of visualizations 5 and 6, which explained the unusual visualization to participants.

Across the focus groups, opinions varied on the use of these representations.

In contrast to tabulated data illustrated in visualizations 2 and 3 (Figures 2-3), participants felt that this visualization provided a layer of depth to understanding activity beyond a conventional table:

*I quite like...[visualization 5] because I think that at least gives you a bit more depth to the data rather than just a bog-standard table form...visually the patient can understand it with a similar explanation.* [###0052]

...you want it to be as simple as possible, the patients want to see [physical] models, they want to see x-rays, they want to see a simple form of data that shows that they’ve done better, or they are improving. [###0050]

Although interesting to some surgeons, most participants (6/9, 67%) thought that simple line graphs would be more user-friendly.

Furthermore, comparing data from week to week seemed to be favored as a tool for discussion with patients, primarily for illustrating any improvements to them:

...again we want to have less explanation to the patient as possible, so something visual that they can see, this was my activity level, a percentage even, pre [surgery], this is what is was post [surgery]...the patients just want to know has it made any difference, has it improved from their pre-operative state? [###0050]

Presenting data in this way stimulated new ways of thinking about activity for the participants—specifically, looking at variation in activity levels over time rather than absolute magnitudes:

*I think it’s brilliant what it’s capturing in the house!* [###0049]

Participants indicated that the variability visible in the charts was of interest but that it was difficult to interpret:

*This is an unusual way to display data.* [###0051]

Making incorrect inferences from the data within a consultation was a concern for the participants, and therefore, guidance or training would be needed for surgeons to use these unfamiliar visualizations:

*During the investigation with the patient I would not use the wheels [visualization 5/6]...because it would take twenty minutes to explain to them and half of them still wouldn’t understand. It’s quite a difficult concept.* [###0049]

Nevertheless, there was broad agreement that such visualizations would be useful for surgeons to use ahead of the consultation but not to explain to patients:

*I can see that that [circular plot 1 in visualization 5] is regular and yeah, great, and I can see that [circular plot 2 in visualization 5] is somewhere in between...but I wouldn’t be able to interpret what the hell that means.* [###0046]

*But looking at that and understanding it, I like it but having it explained, having that as a visual reference with the patient in clinic, it would take too long.* [###0049]

Visualization 6 (Figure 6) presented a spiral chart [49] designed on the same principles as visualization 5, representing a summary of room occupancy information drawn from 14 weeks of a patient’s recovery. Estimates of the average least active times (ie, “L5” in actigraphy terms) and most active times in the participant’s day were drawn from actigraphy data [50] and applied to the chart as an overlay to guide the eye to sections that were expected to have similar characteristics. Empty cells indicated that the participant was not present in the home at that time.
Visualization 6 captures trends in participant behavior within the home. Although clinicians thought that it was useful to describe these trends, they also considered that the data were too complex:

*The trends are useful, but although you can see the trends on there, they don’t jump off the page.* [#0052]

However, identifying anomalies or information that conflicted with any predicted outcomes or expectations was useful:

*If you saw that and it shows that they’re not spending any time in the kitchen because they’re immobile because they can’t walk, then that’s useful.* [#0051]

However, the length of time to arrive at these conclusions was an issue in a short consultation per patient:

*It’s really interesting but it takes a long time specifically to, I guess, actually put meaning to it...it needs to highlight people that are struggling or not getting on, rather than presenting really intricate data about what they’re doing which is interesting but...I couldn’t sit down and look at that with every post-op patient.* [#0046]

Sharing this information with the patient was further highlighted as a challenge if patients interpreted it incorrectly:

*That is just an absolute bombardment of colour and data to a patient. It would take you ten minutes to explain activity levels, trends, patterns.* [#0051]

Accurate assessment of activity using concentric circles was also an issue for some clinicians:

*I have a little bit of a problem with it being displayed as a circle because the radius of the circle increases. The surface area of each block increases as you go out. I think it looks like that amount of time is less because you’re looking at something closer in.* [Participant 1 in 0051]

*It feels a little bit as though we’re exaggerating the good bits on the outside.* [Participant 2 in 0051]

Capturing change over time was viewed as an essential component of the data, and this was not met by this visualization:

*I mean, largely, our job is looking at change over time. I don’t think you can interpret change over time very easily on that, I would say (Participant 1 in 0051)...I think we want something that quickly conveys the information that is most important to what your clinical decision making will be at that point.* [Participant 2 in 0051]

The sit-to-stand movement is used to assess patients in clinics and in research. It can be evaluated using a variety of metrics, including the speed of the motion [51]. Figure 7 is a screenshot of a video showing multiple sit-to-stand transitions collected over the progression of the patient’s recovery (this was an early result from the project, and the video has since been much improved). The data are ordered from left to right, with more recent data to the right.

SPHERE does not store videos captured in people’s homes, and hence, the visualizations presented are “silhouettes.” The use of silhouettes was designed to ensure privacy and acceptability (including acceptability to nonpatient household members and visitors to the home), and some similar processing is a likely feature of any commercial product developed for a similar purpose.

Although the moving images were pleasing, participants least preferred this visualization. Most surgeons expressed that this transition is, in most cases, informally assessed by them as a matter of routine as the patient comes into the first follow-up consultation:

*If they walk into your room and they can’t stand up or sit down you know the answer. You’ve actually clocked it before they’ve even got into your room because you’ve watched them get up in the waiting room and walk towards you and whether they’ve got
sticks and things. So yeah and you don’t really care about the trend, in that one it’s the absolute. [#0052]

Participants recognized that visualization 7 (Figure 7) presented an early prototype of movement data display and not a final outcome:

> I don’t think there’s enough resolution there to understand the sit to stand process...we’re primarily concerned about are they flexing too deeply in the early phases, are they rotating and you can’t really see rotation there. [#0049]

Sharing this information with patients was also problematic:

> It’s too many different images flashing at the same time for a patient to make heads or tails of it. [#0050]

Figure 7. Visualization 7—movement data in long-term sit-to-stand transitions over 16 weeks.

Visualization 8 (Figure 8) presented the progression of a quantitative sit-to-stand metric—average speed—over several weeks of recovery time. This is automatically extracted from the data presented in visualization 7 and, hence, constitutes a simplified view of those data.

Figure 8. Visualization 8—trends in sit-to-stand speed. PROM: patient-reported outcome measure.

The speed of transitions over time was viewed as helpful, although an extension of this line of inquiry would be to look at daily habits changing over time:

> But not necessarily the sit to stand times so much as actually I was thinking “Oh it’d be interesting to know whether they get down the corridor to the kitchen quicker” because most people having a hip replacement will have a seat they sit in during the day and they will go and make cups of tea and it would be interesting to know are they faster at getting to their kettle over six weeks. [#0052]

The speed of the sit-to-stand transition over time is an existing objective test that is sometimes used in clinics. Therefore, this measurement is not in itself an advance in the state of the art in assessing recovery. Rather, the innovation in this case is its use as an in-home metric collected daily. Surgeons proposed that a
broader range of metrics could be used to account for the observed range of patient behaviors:

Your data over time from sitting to rising [visualization 8] is useful. There are a couple of other tests so that’s very useful, [an] easy graph for patients to understand. But that’s just one specific activity that you’re looking at. [#0050]

Yeah, but it’s a pity it doesn’t measure how far they walk, and is it possible to capture activity data outside the home? Because we do get patients who will unnecessarily restrict themselves, particularly older patients, doing their exercises in the house and you’ll get patients at week one or week two, [who] are going round the block. [#0049]

Visualization 9 (Figure 9) presents a screenshot of a prototype decision support tool. Sample notifications are presented that make use of available environmental and participant localization information (such as that displayed in Figure 6) to generate responses to 2 example tests: patient bathing or showering (top) and the suitability of the environmental conditions within the home in comparison with standard guidelines (bottom). In practice, it is likely that many unitary tests of this nature exist. Hence, to avoid a “busy” interface, the results would be filtered in accordance with clinical decision support recommendations in a real-world context of use [52].

Figure 9. Visualization 9—decision support data notifications and alerts.

This final visualization stimulated lively discussions in the focus groups and interviews. The presentation of a notification dashboard aligned with several of the clinicians’ previous expectations of how the data might look. Participants consistently mentioned that clinical tools should capture key pieces of information that can be interpreted rapidly and accurately by both the clinician and patient:

I think that’s what most surgeons would use, and they’d have ten minutes in the clinic and they go. [#0049]

I really like the dashboard, “your patient can’t sleep,” “your patient can’t do the stairs,” “the humidity suggests they haven’t had a shower for two weeks,” you know, that kind of data is really helpful—“they can’t cook.” [#0046]

This is a very good thing, because you don’t want the surgeons interpreting their own way because they might interpret very differently. [#0049]

Therefore, the presentation of this visualization best met the expectations for a tool that could be used in current clinical workflows:

When we go back to that one [visualization 9] that’s quite useful is someone automatically telling us, exactly as you said there, a flag saying “the patient isn’t going out as much as they used to” or “the patient is going out more than they used to, the patient doesn’t appear to be sleeping as much as they used to” or “sleep patterns are still irregular at 6 weeks.” I think those notifications would be good, but I could imagine there being a fairly hefty list of them. [#0052]

You know, much as we get ten minutes per patient, you will get the occasional patient who’ll take forty minutes to sort out with a ten-minute slot. And then you’re playing catch up for the next five patients, and that’s the point where you switch to the notification screen, right, is there anything standing out that I need to know about. [#0049]

Textual summaries such as these notifications alerting the surgeon to relevant features within the data were perceived as aligning well with a 6-week follow-up consultation appointment routinely offered to patients during which the surgeon could include this information in their existing review procedures. Central to this appointment was the opportunity for surgeons to identify any early warnings of surgical complications from the patient’s perspective:

I think it is useful to have these notifications definitely, the point of the six week follow up is to identify patients who’ve got a problem that we need to do something about and those are principally early infection and dislocation or an early fracture. That’s the point of that six week appointment and so that’s what we need these tools to tell us. Or [alternatively]...
the other point of it would be do we need to offer more social support or physiotherapy support in this patient’s home to prevent them falling over or help them? [0051]

At six weeks, its actually quite useful because if they’re not actually showering, they might think that’s because they can’t get the wound wet, patients do have concerns about that. [0049]

Participants broadly liked the idea of a top-level decision-making process integrated within a dashboard, primarily as this removed the need to interpret data for every patient and enabled integration more easily with current clinical management systems:

So you mentioned dashboard, so if I was on BlueSpier [a clinical management system]...if I went in to the patient...if it came up as a red alert, then I’d have a look at it and then go into the data a little bit closer and speak to them about it. [0046]

Some participants (3/9, 33%) felt that similar reports could usefully be distributed in paper form (or presumably by email) to the patient, suggesting that patients might engage with this information outside appointments:

If there was an available one [printout] absolutely, patients would be able to take it home and have a look at it. [0050]

Of the visualizations presented, visualizations 2 to 8 were considered better suited as research tools, whereas the suggested dashboard and notification system presented in visualization 9 (Figure 9) was more appealing for clinical use:

As a clinical tool, I think the notifications are very helpful. I think what would be useful is if you actually provided it to us and gave it to a few surgeons and test it. [0049]

Themes Arising From Data Analysis

Overview

Several general themes were identified common to all visualizations. These were generated by thematic analysis as described in the Data Analysis subsection of the Methods section. Four themes across focus groups and interviews were (1) home data represent a more objective measurement of activity, (2) home data provide a stimulus for discussion in a consultation with a patient, (3) there is interest in the use of home data for clinical research purposes, and (4) there is a need to meet clinicians’ requirements in the development of visualizations.

Data From the Home Can Give a More Objective Measurement of Activity

Assessment of a patient following surgery is mostly done via a face-to-face clinical appointment approximately 6 to 8 weeks after the operation. In this appointment, through questioning, surgeons routinely assess how patients are getting on at home with their activities of daily living and general independence. Participants widely felt that the SPHERE home data presented an advancement from this current practice:

What would be quite good with this is that you get an objective measure so, you know, can you cook? “Oh yeah, much better.” But you’re not cooking, so it’s not better, maybe physically you can’t cook but the times you have managed to get into the kitchen and it didn’t hurt so you remember it as being better but you’ve only cooked one meal in the week. [0049]

So assessing your patient, you take them on their word really as to how they’re doing. So you ask about how they’re getting on at home, activities, are they still independent, is someone else doing the shopping, can they manage stairs, have they moved, so it’s all those sorts of things which we take on their word. So I suppose this would give you objective information as to whether that’s true, not saying that they’re not telling the truth but it would just give you another side of things. [0050]

The only thing that this offers that we struggle with in clinic is the typical, stoic...farmer that says “No, I’ve not had any help, I’m absolutely fine,” he leaves the room and his wife goes “Yeah, he can’t get to the toilet on his own” and this potentially picks up those problem patients because we do get those patients, not infrequently. And so this is a way of potentially flagging it up if it can do that. [0052]

Some participants expressed that the data could help avoid the common problem that reports by patients of function improvement are often not accurate as they are masked by pain:

I mean if you look at the Oxford Hip Score, 80% of the effects in the Oxford Hip Score are due to pain, so pain dominates in terms of what you see...That’s why they tend to improve because you’ve reliably improved the pain...but this is the difference, this not reported function, this is real function and the two are quite different. “Can you go up and down the stairs?” is not the same as are you going up and down stairs? That’s what’s useful about this, isn’t it? [0049]

However, a participant identified a dilemma if the home data contradict the patient’s own account, potentially damaging the patient-clinician relationship:

You can’t break that trust that you have to have, if someone says this is what I do, then I have to take that at face value, regardless of whether I believe it or not. [0046]

Data From the Home Provide a Stimulus for Discussion in a Consultation With a Patient

Although participants indicated that visualizations 4 to 8 were suitable only for use by clinicians, some visualizations were considered to be a good basis for discussion with patients:

The other thing that I think it would be really useful for is providing information to patients after the operation. So you could monitor a cohort of total hips, total knees, hip fracture, different patients and say “We can expect that your sleep will have returned to normal after eight weeks” or “You will be leaving...
the house more back the way you were at six weeks” and that would be really useful. So there are some things we get from patients and we tell them that we’re told that on average you get back to bowls [note: the sport of lawn bowls] at six weeks, but actually having a bit of an evidence base to say “...most [patients] were sleeping through the night by six weeks.” That would be a nice thing to be able to say to patients. [0052]

If they say, “I’m not sleeping well,” and you still look and you say, well although you’re not sleeping perfectly, you’ve definitely improved over the last six weeks—do you see what I mean? [Be]cause they don’t always remember that. [0049]

Surgeons felt that it would be feasible for them to use a decision support system in consultations with patients. Given the inevitable complexity of data derived from people’s daily lives, automated processing of data was preferred over a presentation of relatively raw data. Surgeons found the breadth of possible patient information fascinating. However, many said that the need for speed in their necessarily brief consultations left little room to conduct anything other than the “essentials”:

It’s got huge amount of potential, I just don’t know what to do with all these lovely graphs and figures really. [0052]

I think it’s fascinating to see and I think but the reality at the coalface is that in a clinical situation, you just need to do essentials as quickly as possible. I’m struggling to see how that could happen in the ordinary, everyday situation because in this early phase, patients’ recovery trajectories will vary very much...during this early phase, depending on their co-morbidities and everything else, there’s a very different speed of achieving certain milestones. [0051]

I do think that there is a time element there when you’re using the data...if you had a summary page and that was compared to what a normal recovery would be, like a traffic light system. It’s way too much information to process in a clinic. [0051]

Sadly I can’t get past the fact for routine follow up of post-op [hip replacement] patients, we’re already cutting back how many we see and what we do, because they all tend to do so well and so giving us more information is probably not helpful. [0050]

Use of Home Data for Research Purposes

A possibility was that the data could be developed into outcome measures for research purposes, with the overall aim to be able to consult such information when addressing individual patient cases:

Certainly from a research point of view if you’re wanting to follow something up like a new hip prosthesis and you wanted to know whether this was making any difference in this early phase...this could be very useful in supporting that. [0051]

Using the data as an outcome measure for research was felt to have considerable potential:

I think the power of this is on a clinical basis, we could do more pilot stuff, you can correlate that with your interviews with the patients. [0046]

Meeting Surgeons’ Requirements in the Development of Visualizations

Most participants (8/9, 89%) identified concerns regarding existing visualizations and proposed a way to address them. Challenges included the difficulty of representing large amounts of time-based data without losing detail, accessibility of visualizations to patients, the time required to interpret the visualization, and the provision of excessive detail. To address these issues, surgeons suggested that goal-focused visualizations that solve a small number of competency questions would be of value. For instance, charts showing “a simple form of data” could more easily support clinical evaluation of “one specific activity.”

Discussion

Principal Findings

On the basis of an action research approach [41], this paper reports the findings of scenario-based focus groups and interviews. This study aimed to provide insights into the presentation of time-series data as a way of assessing recovery after surgery and to what extent the data supported clinical workflows. Participants generally noted that the data offered a more objective assessment of patient recovery than current methods used in their clinical practice.

Of the visualizations presented, a dashboard comprising specific notifications and alerts seemed to be the best fit for existing workflows. Automation of clinical decisions based on “moment-to-moment quantification of individual level data” [53] and rapidly condensing large amounts of data into meaningful information aligned with the 10-minute appointment time that NHS surgeons have with patients at follow-up.

The tabular and circular data visualizations spanning longer periods were considered useful by surgeons for identifying trends and changes ahead of the consultation. Furthermore, the granular detail of patients’ movement trajectory immediately before and after surgery was considered useful within a consultation, in which the surgeon and patient could address expectations of outcomes after surgery and longer-term follow-up.

It was noted that such discussions would require assurance of sensitive and accurate interpretation of any data beforehand to avoid any negative impact of patient engagement with the data. Furthermore, it would be necessary to decide whether to measure the patient’s progress in absolute terms with reference to a population mean or purely relative to their own initial baseline.

Surgeons are not accustomed to visualizing and conceptualizing time-series data from the home, and as with any new form of clinical data, undoubtedly training would be a prerequisite for the introduction of this type of data into clinical practice. The
participants in this study were interested in the complexity of the granular data and were aware that insights would be lost if they were summarized or averaged. However, they had not been provided with professional development training in interpretation of the data. It is reasonable to suppose that training and familiarity would unlock more of the value in the data and lessen some of the legitimate concerns about the data being confusing or time-consuming to use. The challenge of finding intuitive ways of presenting weeks of continuous data to clinicians for use in a 10-minute clinical appointment would be a good area for future research.

**Comparison With Prior Work**

The UK National Joint Registry recently introduced a patient decision support tool that aimed to enhance patients’ understanding of their own risks and the potential benefits of having joint replacement surgery [54]. Innovative tools may empower patients to have informed conversations with their physicians about treatment options, and such tools can support evidence-based choices, moving closer toward personalized medicine. Our findings suggest that a clinical decision support system that tracks and interprets activity at home could supplement such information, further enhancing a patient’s choices about treatment options and postrecovery options after surgery. Furthermore, by collecting data before and after surgery, there is the chance to compare outcomes after surgery with presurgical ability and help in communication about expectations before surgery and whether those expectations have been met.

Clinical decision support systems make use of appropriate data analytics and visualization methods to provide advice and guidance to aid health care providers’ problem-solving and decision-making [55]. Potential benefits of designing clinical decision support systems include improving consistency in decision-making, increasing efficiency, and reducing task interruptions and the corresponding fatigue [56]. A recent randomized controlled trial that evaluated the use of a patient decision aid and preference report (ie, a summary of patient clinical and decisional data) by surgeons performing joint replacement [57] found that this supports shared decision-making between clinician and patient and that there was significant improvement in decision quality when such aids were used. The findings of our study are in accordance with those in other contexts, and all UK surgeons follow national approaches and training. More importantly, there may be differences between the findings of our study and those that are relevant in other countries; although the surgical procedure is similar in different contexts, patients’ expectations and the resources available to surgeons may vary. Furthermore, in the United Kingdom and internationally, professionals other than surgeons are involved in the provision of care to patients undergoing knee or hip replacement. For instance, specialist physiotherapists are involved in assessments before surgery and provide care afterward. We did not include their professional experience in this study, and this could be a topic of further research; however, in practice, most health professionals face similar challenges related to time pressures on consultations and the need to collect and convey clear and relevant information.

Everyday practice following the COVID-19 pandemic has required adjustment to deal with service backlogs, such as a move toward day case surgery as well as decreasing length of stay and adoption of remote assessment of postoperative recovery status. It is not yet clear to what extent what proportion of sites has moved to this model and what proportion of patients are affected by this change. There is also a move toward patient-initiated follow-up. However, this is at an early stage, and it remains to be seen what the benefits and shortcomings might be for patients and participants.

**Conclusions and Recommendations**

**Overview**

In line with an action research approach [41], we propose the following 4 guidelines for further design and development of home activity monitoring systems. Each guideline draws on the findings described previously and was developed by the conventional follow-up assessments following total hip and knee replacement surgery.

**Strengths**

A key strength of this study was the use of an action research approach, which included an exploratory phase followed by discussion of a proposed model of data presentation using real patients’ stories. The triangulation of patterns detected in the quantitative data with real patient participants’ accounts from qualitative interviews contributed to a robust analysis of the data with a good degree of accuracy. A qualitative analysis of patients’ experiences has been reported elsewhere [45]. Finally, this study uniquely explores surgeons’ views of data visualization from novel sensing technology, which is not currently commercially available but could be put on the market in the near future if desired. Insights from this study can help inform research and design directions for products in this space.

**Limitations**

The sample comprised surgeons from 1 UK hospital and, as such, only reflects experience in 1 setting. Convenience sampling was used, which may limit the ability to generalize from this sample. However, in practice, participants’ background, skills, and experience were heterogeneous, as were their age and sex. The experiences described are likely to be consonant with those in other contexts, and all UK surgeons follow national approaches and training. More importantly, there may be differences between the findings of our study and those that are relevant in other countries; although the surgical procedure is similar in different contexts, patients’ expectations and the resources available to surgeons may vary. Furthermore, in the United Kingdom and internationally, professionals other than surgeons are involved in the provision of care to patients undergoing knee or hip replacement. For instance, specialist physiotherapists are involved in assessments before surgery and provide care afterward. We did not include their professional experience in this study, and this could be a topic of further research; however, in practice, most health professionals face similar challenges related to time pressures on consultations and the need to collect and convey clear and relevant information.
interdisciplinary group of coauthors in light of the findings. As such, the guidelines consolidate the views of surgeons and the thematic areas developed in this study and provide recommendations for next steps, including how best to support surgeons—or other health care professionals—and how best to design and deliver a user-appropriate system. Each guideline reflects the content of more than one thematic area. Our aim was to build on surgeons’ views to provide concrete recommendations to support future developments in the collection, visualization, and use of data on recovery or other health changes.

**Guideline 1: Minimize the Risk of Misinterpreting Data**
Surgeons demonstrated consensus on the importance of reducing the risk of misinterpretation of data and the associated variability of interpretations between surgeons. To minimize the risk of misinterpretation, clear summary statistics are recommended. Explainable design principles appropriate for each visualization or presentation of data should be applied to clarify the meaning and limitations of the data and the associated findings. It would be misguided to promise absolute objectivity as the activities of data acquisition, data analysis, and machine learning frequently result in the reproduction of bias present in source data or in the unconscious predispositions held by data analysts themselves [60].

**Guideline 2: Express the Level of Confidence in the Data**
Surgeons expressed a preference for simple and unambiguous metrics. However, electronics and sensor systems in the home inevitably experience many challenges to reliability, such as device failure or wireless network failures; therefore, data from such an uncontrolled environment must always be interpreted with caution, and a level of confidence would need to accompany any data analysis.

**Guideline 3: Improve Familiarity With Time-Series Data**
Exploratory methods of accessing big data are a poor fit with constraints on surgeons’ time. Efficient, rapidly accessible representations of home data requiring minimal expert knowledge are recommended in the first instance. For example, data summarization can facilitate the interpretation of complex data, removing outliers and supporting existing clinical consultation activities. The 2019 Topol Review [61] indicates that training and digital literacy are key to making the most of digital health technologies, particularly artificial intelligence and machine learning. Identifying an understanding of confidence and probability is a necessary prerequisite for interpreting these data and is a required skill. We suggest that familiarity with time-series representations of data acquired through training may increasingly be an advantageous skill for clinical purposes.

**Guideline 4: Consider the Impact of Patient Engagement**
Data are of interest to surgeons as a resource that they can use to assist in their communications with patients. Future developments such as interfaces that support patients in examining their own data may offer a level of empowerment. Greater patient empowerment is positively associated with adherence to treatments and improved outcomes [62]. It also supports the UK NHS commitment to person-centered care, in which patients are encouraged to be actively involved in their own care [63]. Therefore, patient-centered design practices are a substantial component in the development of systems that use home data to support patient-clinician interactions. The time constraints experienced by surgeons limit their opportunities to have direct overview of time-series home data. A patient-centric approach could support patients in monitoring changes in their own condition, potentially facilitating conversations with clinicians. Finally, the schedule by which surgeons or other clinicians review data is not a close fit with the potential for “just-in-time” alerting systems, suggesting that some of the potential of home data may rely on structural innovation and integration with wider support teams.

**Acknowledgments**
This work was carried out under the Sensor Platform for Healthcare in a Residential Environment Interdisciplinary Research Collaboration, funded by the UK Engineering and Physical Sciences Research Council, grant EP/K031910/1.

**Authors' Contributions**
All authors contributed to the design of the study. MH, AM, MPN, HS, and ET developed the visualizations and provided explanatory material. SG, MH, and MPN conducted the participant interviews, and SG analyzed the interview data with support from AB, AJ, and RGH. SG, ET, and IC wrote the manuscript. All authors reviewed the final manuscript.

**Conflicts of Interest**
RGH is the co-chair of the UK Committee on Research Integrity. The authors have no further conflicts to declare.

Multimedia Appendix 1
Topic guides used during this study.
[PDF File (Adobe PDF File), 89 KB - periop_v6i1e36172_app1.pdf]

**References**


34. De Rosis S, Pennucci F, Lungu DA, Manca M, Nuti S. A continuous PREMs and PROMs Observatory for elective hip and knee arthroplasty: study protocol. BMJ Open 2021 Sep 21;11(9):e049826 [FREE Full text] [doi: 10.1136/bmjopen-2021-049826] [Medline: 34548358]


Abbreviations

- **NHS:** National Health Service
- **OHS:** Oxford Hip Score
- **OKS:** Oxford Knee Score
- **PROM:** patient-reported outcome measure
- **SPHERE:** Sensor Platform for Healthcare in a Residential Environment
©Sabrina Grant, Emma Tonkin, Ian Craddock, Ashley Blom, Michael Holmes, Andrew Judge, Alessandro Masullo, Miquel Perello Nieto, Hao Song, Michael Whitehouse, Peter Flach, Rachael Gooberman-Hill. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 24.04.2023. 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.
Patient Safety of Perioperative Medication Through the Lens of Digital Health and Artificial Intelligence

Jiancheng Ye¹, PhD
Feinberg School of Medicine, Northwestern University, Chicago, IL, United States

Corresponding Author:
Jiancheng Ye, PhD
Feinberg School of Medicine
Northwestern University
633 N Saint Clair St
Chicago, IL, 60611
United States
Phone: 1 312 503 3690
Email: jiancheng.ye@u.northwestern.edu

Abstract

Perioperative medication has made significant contributions to enhancing patient safety. Nevertheless, administering medication during this period still poses considerable safety concerns, with many errors being detected only after causing significant physiological disturbances. The intricacy of medication administration in the perioperative setting poses specific challenges to patient safety. To address these challenges, implementing potential strategies and interventions is critical. One such strategy is raising awareness and revising educational curricula regarding drug safety in the operating room. Another crucial strategy is recognizing the importance of redundancy and multiple checks in the operating room as a hallmark of medication safety, which is not a common practice. Digital health technologies and artificial intelligence (AI) also offer the potential to improve perioperative medication safety. Computerized physician order entry systems, electronic medication administration records, and barcode medication administration systems have been proven to reduce medication errors and improve patient safety. By implementing these strategies and interventions, health care professionals can enhance the safety of perioperative medication administration and improve patient outcomes.

(JMIR Perioper Med 2023;6:e34453) doi:10.2196/34453

KEYWORDS
perioperative medicine; patient safety; anesthesiology; human factors; medication errors; digital health; health information technology

Introduction

Medication errors represent a critical patient safety problem, arising from failures in completing required actions or using the wrong plan or action to achieve patient care aims [1]. These errors can be classified by type, including incorrect dose, substitution, omission, repetition, insertion, and unattended drug use [2]. In the perioperative setting, the administration of intravenous (IV) medications, or the “medication use process” [3], presents unique patient safety obstacles. The medication use process consists of several steps, including requesting, dispensing, preparing, administering, documenting, and monitoring patients for the effects of medication [4]. However, compared to almost any other hospital setting, medication administration in the operating room lacks most standard and accepted safety checks [1]. Unlike nurses, who require a physician to place an order for a medication, the pharmacy to prepare the medication, and a second nurse to verify the medication prior to administration, the anesthesiologist working alone in the operating room can determine the need for a medication (ie, diagnose and prescribe), draw up the medication (ie, prepare), administer and monitor the effects of the medication, and record events without any verification check for safety and accuracy [5]. Furthermore, the often fast-paced and high-stress environment of the operating room can further increase the likelihood of medication errors. Thus, it is not surprising that medication errors are common in this setting [6]. To address this issue, it is essential to implement strategies and interventions that improve the safety and accuracy of the process of medication use in the operating room.
Patient Safety Problems Associated With Perioperative Medication Errors

Medication errors can be classified based on their potential for patient harm and whether they result in an adverse drug event or not. These classifications include errors with no potential for harm (near miss), those with little potential for patient harm, those with potential for adverse drug events, and those resulting in adverse drug events [1]. An adverse drug event is defined as any patient injury resulting from medication [7]. However, it is essential to note that adverse drug events can occur even without medication errors; for example, in the case of an allergic reaction. Nanji et al [1] further classified medication errors and adverse drug events by their severity (significant, serious, life-threatening, and fatal) and preventability (definitely preventable, probably preventable, probably not preventable, and definitely not preventable), and they found that out of the 193 (of a total of 3671, 5.3%) identified medication errors and adverse drug events, 153 (79.3%) were preventable. Additionally, 32 (20.9%) of these medication errors had little potential for harm, 70 (45.8%) had the potential for patient harm, and 51 (33.3%) resulted in an adverse drug event. The errors were further classified as serious (n=99, 64.7%), significant (n=51, 33.3%), and life-threatening (n=3, 2%), with no fatalities attributable to medication errors. In a separate study by Cooper et al [2], the authors identified 52 medication errors, resulting in no harm in 24 patients, minor harm in 15 patients, and harm in 13 patients.

Risk Factors

Medication errors have been a concern since the 1970s, but the exact prevalence of these errors is still unknown due to underreporting [3]. According to Cooper et al [8], human factors are the primary cause of medication errors. These factors include failure to check, poor labeling, syringe swaps, decreased vigilance, fatigue, and production pressure [9]. Distraction, pressure to proceed, and unreadable labels were found to be the top 3 factors contributing to medication errors. Nanji et al [1] identified the following as the 3 most common medication errors: wrong dose, improper labeling, and failure to deliver the appropriate medication. High-acuity medications such as propofol, phenylephrine, and fentanyl were the most common perioperative medications involved in these errors [10].

Despite increased awareness and emphasis on perioperative medication safety by the Anesthesia Patient Safety Foundation, little progress has been made in addressing the human factor constraints that lead to medication errors. These constraints include lack of standardized labels; varied drug vial sizes, shapes, and colors; poorly designed medication carts and drug dispensing machines; and look-alike and sound-alike medications. The labeling and packaging of medications were found to be contributors to almost one-third of voluntarily reported medication errors leading to fatalities in the 1990s, according to the Institute of Safe Medication Practices’ Medication Errors Reporting Program [11]. The Food and Drug Administration estimates that suboptimal labeling and packaging contribute to approximately 20% of medication errors. Moreover, drug shortages may force pharmacies to source unfamiliar substitute medications that can lead to errors, as reported by 21% of hospital pharmacists in a 2017 survey [12].

The high rate of medication errors in perioperative medication administration can be attributed to several factors, but perhaps the most significant is the anesthesia culture itself. The norm of an anesthesiologist performing the medication administration process with no oversight is deeply ingrained in professional and organizational culture. Prielipp et al [13] discussed the concept of “normalized deviance” in anesthesia practice, where departure from correct behavior becomes so ingrained in work culture that it is no longer considered deviant. Despite evidence that independent double checks can detect up to 95% of potential medication errors and eliminate 58% of those identified [14], the culture of autonomy, rejection of “cookbook” medicine, and resistance to standardization hinder the reporting of minor or near-miss events and impede efforts to improve medication safety in anesthesia. Grigg and Roesler [3] summarized this culture as allowing providers to “hand scrawl on poorly labeled syringes drawn up from nonstandard, look-alike vials in a distracting environment and organize them in an arbitrary, personalized arrangement.” Organizational barriers also exist, such as fear of reprisal for errors or a blame culture, which can prevent medication errors from being reported [15].

Interventions and Recommendations

Inadequate labeling and packaging of medications is one of the primary human factors associated with a significant number of medication errors. Poor labeling can result in the administration of the wrong medication, incorrect dosage, and an incorrect route of administration. Common labeling errors include syringes that are unlabeled or inaccurately labeled and illegible, handwritten labels [16]. To address the labeling problem, various solutions have been proposed. These include using standardized color-coded labels for similar drug classes, implementing barcode-assisted labeling systems to generate medication labels, and using commercially prepared prefilled medication syringes. In a systematic review by Maximous et al [17], improved labeling led to a reduction of 37% in medication errors. Recent data have questioned the effectiveness of color-coding in preventing medication errors [18], but research on human factors engineering (HFE) highlights the importance of pattern recognition when performing pressured, high-stress tasks, such as administering high-risk medication in the operating room environment [8]. Although support for the color-coded labeling system has somewhat decreased, HFE acknowledges that identifying objects in high-stress situations relies on multiple cues [3], and the historical use of color-coding remains a crucial cue for anesthesiologists.

Digital health technologies such as computerized physician order entry (CPOE) systems and electronic medication administration records (eMAR) have been shown to reduce medication errors and improve patient safety. CPOE systems allow health care professionals to electronically enter medication orders, reducing the likelihood of errors due to illegible handwriting or transcription errors. The use of CPOE systems can ensure that medication orders are accurately and efficiently transmitted to the pharmacy and the surgical team. The eMAR systems can provide real-time information about medication...
administration during surgery, allowing the surgical team to make informed decisions about the patient’s care [19]. eMAR reduces the risk of medication errors due to incorrect dosages or administration times. In the perioperative setting, these digital health technologies can be particularly beneficial.

Barcode medication administration (BCMA) technology has been successful on nursing floors, but is not yet widely implemented in anesthesia due to the cost of the systems and a lack of a universal electronic health record (EHR) system capable of scanning barcodes and incorporating the information into the operative record [20]. BCMA systems use barcodes to verify patient and medication information, thus reducing the risk of errors due to incorrect medication selection or administration. BCMA systems can help ensure that the correct medications are administered to the correct patient at the correct time, reducing the risk of medication errors and adverse events [21]. Point-of-care barcode scanning has the potential to eliminate 17% of medication errors and 25.5% of potential adverse drug events [1]. In a study by Merry et al [22], a 21% reduction in perioperative medication errors was demonstrated when syringe labels were scanned immediately before administration. When automated drug-specific decision support and alerts were added, an additional 29% and 59% of medication errors could be eliminated, respectively. Despite the availability of barcode-assisted labeling systems in their study environment, the authors found that up to 24% of medication errors still involved a labeling error. These errors occurred when healthcare professionals bypassed the system or found a workaround. To minimize these risks, prefilled medication syringes prepared at standard concentrations and provided by the pharmacy may be the best risk reduction strategy. A failure modes and effects analysis of the use of prefilled syringes has the potential to eliminate 16 medication preparation steps and 19 potential failure modes [23]. For organizations with barcode scanning abilities, these prefilled syringes could also use barcode technology, which would eliminate compounding of medications by the anesthesia clinicians and the associated risks of this practice. Other process-based interventions, such as facilitating timing of documentation, reducing workarounds, and standardizing connections of IV drug infusions to the most proximal port, could further reduce medication errors. These process interventions have the potential to eliminate 35%, 24%, and 1.3% of medication errors, respectively [1]. Ultimately, multimodal strategies are needed, which include all potential human factor system changes and process interventions discussed above [24]. Multimodal interventions, including barcode readers with automatic auditory and visual verification of the drug, prefilled color-coded syringes, and workspace improvements including standardized stocking of anesthesia carts, have the greatest potential to reduce errors [25]. In combination, these interventions could reduce error rates by 21%-35% per administration and 37%-41% per anesthetic [17].

Artificial intelligence (AI) has the potential to revolutionize medication safety by providing real-time decision support, reducing medication errors, and improving communication among healthcare professionals [26]. AI can analyze large data sets to identify patterns and predict adverse events, allowing healthcare professionals to intervene before harm occurs. AI can also provide decision support, suggesting the most appropriate medication and dosage for a particular patient based on their medical history and other factors. Finally, AI can facilitate communication among healthcare professionals, ensuring that critical information is shared in a timely and efficient manner. In the perioperative setting, AI can be particularly valuable in predicting and preventing adverse events. For example, AI algorithms can analyze vital signs and other patient data to identify patients at risk for postoperative complications such as sepsis or acute kidney injury [27]. AI can also provide decision support to help healthcare professionals select the most appropriate medication and dosage for a particular patient, taking into account their medical history, allergies, and other factors. Finally, AI can facilitate communication among healthcare professionals, ensuring that critical information is shared in a timely and efficient manner.

AI algorithms can analyze large amounts of data to identify patterns and predict medication errors. For example, AI can analyze medication orders and patient data to identify patients at high risk for medication errors. This can help clinicians to proactively intervene to prevent errors before they occur. Natural language processing algorithms can analyze free-text notes in the EHR to identify potential medication errors. For example, natural language processing can identify notes that mention medication errors or adverse drug events. This can help clinicians to identify and address medication errors that may have been missed through other means. AI can be used to provide real-time decision support to clinicians. For example, AI algorithms can analyze medication orders and provide alerts to clinicians about potential drug interactions, dosing errors, or other safety concerns. This can help clinicians to make informed decisions about medication orders and reduce the risk of errors. Table 1 outlines specific ways that AI can be used to improve perioperative medication errors.
Table 1. Application of artificial intelligence (AI) in the patient safety of perioperative medication.

<table>
<thead>
<tr>
<th>AI technology</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive analytics</td>
<td>AI algorithms have the potential to analyze vast amounts of data, identify patterns, and predict medication errors [28]. For example, clinicians can proactively intervene to prevent errors before they occur by analyzing medication orders and patient data to identify patients at high risk for medication errors, using AI.</td>
</tr>
<tr>
<td>Natural language processing</td>
<td>Natural language processing algorithms can analyze free-text notes in the electronic health record to identify potential medication errors that may have been missed through other means, such as notes that mention medication errors or adverse drug events [29]. This approach can assist clinicians in identifying and addressing potential errors before they cause harm.</td>
</tr>
<tr>
<td>Clinical decision support</td>
<td>AI can provide real-time decision support to clinicians by analyzing medication orders and providing alerts for potential drug interactions, dosing errors, or other safety concerns [30]. This feature can assist clinicians in making informed decisions about medication orders and reducing the risk of errors.</td>
</tr>
<tr>
<td>Machine learning</td>
<td>Machine learning algorithms can be used to identify patterns and predict medication errors by analyzing medication orders and patient data. These algorithms can also be used to develop personalized medication regimens for individual patients based on their unique characteristics, which can improve medication safety and reduce the risk of adverse drug events [31].</td>
</tr>
<tr>
<td>Computer vision</td>
<td>Computer vision algorithms can be used with barcoding systems to verify medication administration [22]. For instance, computer vision can analyze barcode scans to verify that the medication matches the order and the patient's information in the electronic health record. This feature can help reduce the risk of errors due to incorrect medication administration.</td>
</tr>
</tbody>
</table>

Discussion

Ensuring patient safety is a paramount concern, especially when it comes to administering medications in the perioperative setting. Medication reconciliation is a crucial process that involves comparing a patient's medication orders with their current medication regimen to identify any discrepancies and prevent medication errors caused by incomplete or inaccurate medication histories. To ensure medication safety, it is important to perform preoperative medication reconciliation and document it accurately [32]. Standardized protocols for medication administration, such as those recommended by professional societies or institutions, can help reduce the risk of errors during drug preparation, dosage calculation, and administration [33]. Education and training are essential for improving the safety of perioperative medication. Clinicians should be trained on medication safety best practices, including the use of decision support tools, the importance of medication reconciliation, and the use of standardized protocols [34]. Effective communication among health care professionals, especially during handovers, is critical to reducing medication errors. The use of standardized communication tools and training can help improve communication among health care professionals [35]. Continuous electronic monitoring of vital signs, particularly during surgery, can help identify and manage medication-related adverse events promptly. Emerging evidence suggests that incorporating HFET principles into practice may improve patient safety by reducing cognitive workload and simplifying medication administration processes [36]. By optimizing the design of medication administration processes, we can reduce the risk of errors and improve patient safety [37].

Digital health technologies and AI can be used to enhance perioperative medication safety by detecting potential medication errors and providing decision support tools to clinicians in real time. Clinical decision support tools, such as alerts for potential drug interactions or incorrect doses, can help to prevent medication errors [38]. These tools can be integrated into EHR systems, providing clinicians with accurate and up-to-date information about medications and their effects. CPOE systems allow clinicians to enter medication orders directly into an EHR system, reducing errors caused by illegible handwriting or transcription errors [39]. These systems can also provide decision support tools, such as alerts for potential drug interactions or incorrect doses. Additionally, CPOE systems can help to standardize medication orders, reducing the risk of errors due to miscommunication or confusion. The eMAR systems allow health care professionals to electronically record medication administration, reducing errors due to incorrect dosages or administration times [40]. These systems can also be integrated with barcode scanning technology to ensure that the correct medication is administered to the correct patient at the correct time. Telepharmacy services can be used to provide medication-related support to health care professionals in remote or underserved areas, ensuring that medication orders are accurate and complete and that medications are administered safely [41]. Patient portals can be used to provide patients with information about their medications, including dosages, side effects, and potential interactions [42]. These portals can also be used to remind patients to take their medications and to provide instructions on how to properly administer medications. By incorporating these digital health technologies and AI, health care professionals can reduce the risk of medication errors and improve patient outcomes.

Improving patient safety during perioperative medication administration requires a multifaceted approach, incorporating strategies that address medication reconciliation, standardized protocols, effective communication, continuous monitoring, and HFET principles.

Conclusions

Perioperative medication safety has been largely overlooked in terms of rigorous assessment of medication events and the implementation of safety measures. Unlike other high-risk
industries, there are few safety protocols to prevent simple but
dangerous medication errors, such as those related to labeling.
Addressing systemic weaknesses that contribute to medication
errors requires HFE and cultural reforms, and a shift away from
focusing on individual blame and failure of truth-telling and
transparency to enable real reform. Simply improving vigilance
is insufficient since it does not address human factors and
systemic issues that contribute to errors. Digital health
technologies and AI offer significant promise in enhancing
perioperative medication safety. Systems such as CPOE, eMAR,
and BCMA can reduce medication errors and improve
communication among health care professionals. AI can provide
real-time decision support, predict adverse events, and facilitate
communication. However, it is necessary to develop effective
ways to measure medication errors and capture data to identify
the true scope of the problem and develop solutions for
mitigation. Standardization, medication reconciliation, education
and training, clinical decision support, barcoding and electronic
medication administration, and effective team communication
are all crucial to improving perioperative medication safety. By
implementing these strategies, health care professionals can
reduce the risk of medication errors and improve patient safety.

Authors’ Contributions
JY was responsible for the study design, data analysis, interpretation of the results, and writing of the manuscript.

Conflicts of Interest
None declared.

References
2. Cooper L, DiGiovanni N, Schultz L, Taylor AM, Nossaman B. Influences observed on incidence and reporting of medication
10.1213/ane.0000000000002521]
7. Institute of Medicine (US) Committee on Quality of Health Care in America. In: Kohn LT, Corrigan JM, Donaldson MS, 
[Medline: 31577240]
10. Abeysekera A, Bergman JJ, Kluger MT, Short TG. Drug error in anaesthetic practice: a review of 896 reports from the
Australian Incident Monitoring Study database. Anaesthesia 2005 Mar;60(3):220-227 [FREE Full text] [doi: 
10.1111/j.1365-2044.2004.03670.x] [Medline: 15710005]
10.1023/b:joms.00000021518.60670.10] [Medline: 15171066]
12. Institute for Safe Medication Practices. ISMP survey on drug shortages for hospital pharmacy directors or their designees
only. ISMP Medication Safety Alert! 2017;22(17):5-6.
13. Prielipp R, Magro M, Morell RC, Brull SJ. The normalization of deviance: do we (un)knowingly accept doing the wrong
14. Jensen LS, Merry AF, Webster CS, Weller J, Larsson L. Evidence-based strategies for preventing drug administration
[Medline: 15096243]
15. Leahy IC, Lavoie M, Zurakowski D, Baier AW, Brustowicz RM. Medication errors in a pediatric anesthesia setting:
10.1016/j.clinane.2018.05.011] [Medline: 29913393]
[Medline: 29757795]


Abbreviations

AI: artificial intelligence  
BCMA: barcode medication administration  
CPOE: computerized physician order entry  
EHR: electronic health record  
eMAR: electronic medication administration record  
HFE: human factors engineering  
IV: intravenous

©Jiancheng Ye. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 31.05.2023. 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.
The Role of Wearable Technology in Measuring and Supporting Patient Outcomes Following Total Joint Replacement: Review of the Literature

Gregory Iovanel, BSc; David Ayers, MD; Hua Zheng, PhD
Department of Orthopedics and Physical Rehabilitation, UMass Chan Medical School, Worcester, MA, United States

Corresponding Author:
Hua Zheng, PhD
Department of Orthopedics and Physical Rehabilitation
UMass Chan Medical School
55 Lake Ave N
Worcester, MA, 01655
United States
Phone: 1 774 455 4473
Email: Hua.Zheng@umassmed.edu

Abstract

Background: The incidence rate of total joint replacement (TJR) continues to increase due to the aging population and the surgery that is very successful in providing pain relief to and improving function among patients with advanced knee or hip arthritis. Improving patient outcomes and patient satisfaction after TJR remain important goals. Wearable technologies provide a novel way to capture patient function and activity data and supplement clinical measures and patient-reported outcome measures in order to better understand patient outcomes after TJR.

Objective: We examined the current literature to evaluate the potential role of wearable devices and compare them with existing methods for monitoring and improving patient rehabilitation and outcomes following TJR.

Methods: We performed a literature search by using the research databases supported by the University of Massachusetts Chan Medical School’s Lamar Soutter Library, including PubMed and Scopus, supplemented with the Google Scholar search engine. A specific search strategy was used to identify articles discussing the use of wearable devices in measuring and affecting postoperative outcomes of patients who have undergone TJR. Selected papers were organized into a spreadsheet and categorized for our qualitative literature review to assess how wearable data correlated with clinical measures and patient-reported outcome measures.

Results: A total of 9 papers were selected. The literature showed the impact of wearable devices on evaluating and improving postoperative functional outcomes. Wearable-collected data could be used to predict postoperative clinical measures, such as range of motion and Timed Up and Go times. When predicting patient-reported outcomes, specifically Hip Disability and Osteoarthritis Outcome Scores/Knee Injury and Osteoarthritis Outcome Scores and Veterans RAND 12-Item Health Survey scores, strong associations were found between changes in sensor-collected data and changes in patient-reported outcomes over time. Further, the step counts of patients who received feedback from a wearable improved over time when compared to those of patients who did not receive feedback.

Conclusions: These findings suggest that wearable technology has the potential to remotely measure and improve postoperative orthopedic patient outcomes. We anticipate that this review will facilitate further investigation into whether wearable devices are viable tools for guiding the clinical management of TJR rehabilitation.

(JMIR Perioper Med 2023;6:e39396) doi:10.2196/39396

KEYWORDS

total joint replacement; wearables; osteoarthritis; rehabilitation; mobility
Introduction

Total joint replacement (TJR) has proven to be highly effective in relieving joint pain and improving physical function for millions of patients with advanced knee or hip osteoarthritis and continues to be one of the most commonly performed surgical procedures in the United States [1-3]. As this trend persists, increased attention must be paid toward effectively monitoring and coaching patients following surgery to ensure successful rehabilitation. Traditional assessments of postoperative recovery, such as the Timed Up and Go (TUG) and 6-minute walk tests, are considered gold standards for measuring mobility, balance, and walking ability [4]. However, these assessments require in-person monitoring by health care providers and do not replicate activities of daily living. Patient-reported outcomes (PROs) have been widely used to evaluate joint pain and physical function through standardized patient questionnaires. Patients report on how they perceive their health status without the interpretation of a medical professional. Although the assessment of PROs has become part of the standard of care in many orthopedic practices, the implementation of PRO capture, the maintenance of data integrity, data interpretation, and cost management are still challenging for many practices [5-9]. The internet-based remote monitoring of patient mobility data is an alternative method of collecting patient data following surgery that has recently been introduced and warrants further evaluation.

Wearable technologies provide a novel way to capture patient function and activity data and supplement clinical measures and PRO measures (PROMs) to better understand patient recovery after TJR. Wearable technologies, in the context of health care, refer to devices that can record real-time data from an individual while worn. These devices include accelerometers, which capture the acceleration of a limb or the entire body; gyroscopes, which measure orientation and angular velocity; and inertial measurement units—a more sophisticated technology that combines an accelerometer, gyroscope, and magnetometer and is capable of reporting the movement, orientation, and position in space of a person or object [10]. Many companies manufacture such devices that can be synced to a smartphone, computer, or tablet to transmit patient mobility data securely and instantly to health care providers via an internet-based application. Medical professionals are then able to track patients’ progress in real time and tailor rehabilitation regimens for patients to follow, based on the data obtained [11,12]. Such wearable technologies could offer the possibility of capturing real-time function data on the rehabilitation and recovery of patients who have undergone TJR and eliminating the need for direct supervision. In addition, a connected mobile app can be developed to collect PROMs, thereby minimizing the need for additional PROM capture tools [13]. Current research has shown the feasibility of wearable devices and their capability for motion and activity tracking [14]. However, it is not clear whether the activity data collected by wearable devices can serve as outcome measures or as adjuncts to support outcome monitoring. There is a dearth of consensus on whether wearables can be used as effective tools, can be aligned with standard clinical measures and PROMs, or can even improve outcomes.

To promote wearable use as part of rehabilitation programs following TJR, their impact on postoperative patient outcomes, as well as their accuracy in measuring these outcomes, must be further investigated. This paper seeks to review the current landscape of orthopedic wearables literature and assess the effectiveness of available devices with respect to evaluating and improving postoperative outcomes.

Methods

A literature search was conducted by using the research databases supported by the University of Massachusetts Chan Medical School’s Lamar Soutter Library, including PubMed and Scopus, supplemented with the Google Scholar search engine. Articles published in English from 2004 to 2021 were reviewed. The search terms used to identify these articles are defined in Textbox 1.

Textbox 1. Literature search strategy (search terms used in the literature search strategy).

<table>
<thead>
<tr>
<th>Term groupings and search terms</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Wearable devices</strong></td>
</tr>
<tr>
<td>(“wearable”) AND (“devices” OR “technology”)</td>
</tr>
<tr>
<td><strong>Patient</strong></td>
</tr>
<tr>
<td>(“total joint replacement” OR “total knee replacement” OR “total hip replacement”) AND “outcomes”</td>
</tr>
<tr>
<td><strong>Rehabilitation</strong></td>
</tr>
<tr>
<td>“rehabilitation” OR “recovery”</td>
</tr>
</tbody>
</table>

The inclusion criteria included English-language articles, research studies, and studies with wearable technology that focused on comparing wearable-collected data with clinical measures or PROMs or affecting patient outcomes. The exclusion criteria were articles focusing on wearable device design, study protocols, theoretical articles, books, or book chapters. Titles and abstracts of identified articles were screened to determine eligibility based on the inclusion and exclusion criteria. Since only a limited number of papers met the inclusion criteria, a full reading was conducted for all of the eligible papers. The information was tabulated via a standardized Excel (Microsoft Corporation) form that was developed for this review.
which included the first author’s name, year of publication, name and type of the wearable device, location where the device was worn, number of patients in the study, outcome measures, and study findings (Table 1). A narrative literature review of the selected articles was conducted by 2 reviewers, providing a qualitative overview of outcome measures, data collection methods, and main findings.
<table>
<thead>
<tr>
<th>Authors, year</th>
<th>Wearable device</th>
<th>Device type</th>
<th>Device location</th>
<th>Patients, n</th>
<th>Outcome measure</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kwasnicki et al [15], 2015</td>
<td>e-AR (Imperial College London)</td>
<td>Accelerometer</td>
<td>Ear</td>
<td>14</td>
<td>TUG&lt;sup&gt;a&lt;/sup&gt; time and ROM&lt;sup&gt;b&lt;/sup&gt;</td>
<td>The classification of patients into preoperative, normal, and 24-week postoperative groups based on outcomes was 89% accurate, while classification for all time intervals was 69% accurate.</td>
</tr>
<tr>
<td>Chiang et al [16], 2017</td>
<td>APDM OPAL (APDM Wearable Technologies)</td>
<td>Accelerometer, gyroscope, magnetometer, and barometer</td>
<td>Thigh and calf (2 sensors)</td>
<td>18</td>
<td>Satisfaction</td>
<td>Only 17% of patients felt uncomfortable with the sensor belt.</td>
</tr>
<tr>
<td>Bendich et al [17], 2019</td>
<td>Fitbit Flex (Fitbit LLC)</td>
<td>Accelerometer</td>
<td>Wrist</td>
<td>22</td>
<td>Daily step count, daily minutes active, HOOS/KOOS&lt;sup&gt;c&lt;/sup&gt;, and VR-12&lt;sup&gt;d&lt;/sup&gt; score</td>
<td>Changes from preoperative levels to 6-week postoperative levels in “daily step count” and “daily minutes active” (collected with a wearable sensor) were strongly associated with improvements in HOOSs/KOOSs and VR-12 physical component scores (collected over the same period).</td>
</tr>
<tr>
<td>Chen et al [18], 2015</td>
<td>APDM OPAL</td>
<td>Accelerometer, gyroscope, and magnetometer</td>
<td>Chest, thigh, and calf (3 sensors)</td>
<td>10</td>
<td>ROM</td>
<td>The device was able to identify proper exercise posture 88.26% of the time.</td>
</tr>
<tr>
<td>Battenberg et al [19], 2017</td>
<td>Fitbit One (Fitbit LLC), Omron HJ-321 (Omron Corporation), Sportline 340 Strider (Sportline Inc), Fitbit Force (Fitbit LLC), Nike+ Fuelband SE (Nike Inc), and StepWatch Activity Monitor (OrthoCare Innovations)</td>
<td>Fitbit One (accelerometer), Omron HJ-321 (pedometer and accelerometer), Sportline 340 Strider (pedometer), Fitbit Force (accelerometer), Nike+ Fuelband SE (accelerometer), and StepWatch Activity Monitor (accelerometer)</td>
<td>Fitbit One (waist), Omron HJ-321 (waist), Sportline 340 Strider (waist), Fitbit Force (waist), Nike+ Fuelband SE (waist), and StepWatch Activity Monitor (ankle)</td>
<td>30</td>
<td>Step count</td>
<td>The waist-based devices—Fitbit One and Omron HJ-321—were &gt;90% accurate in counting steps for all activities, the wristband devices were &lt;90% accurate for most activities, and the StepWatch Activity Monitor (ankle) was &gt;95% accurate for lower cadence activities but undercounted running by 25%.</td>
</tr>
<tr>
<td>Toogood et al [20], 2016</td>
<td>Fitbit (Fitbit LLC)</td>
<td>Accelerometer</td>
<td>Ankle</td>
<td>33</td>
<td>Compliance</td>
<td>The mean compliance over 30 days was 26.7 days (89%).</td>
</tr>
<tr>
<td>Saporito et al [21], 2019</td>
<td>Custom</td>
<td>Accelerometer and barometer</td>
<td>Neck (pendant)</td>
<td>15</td>
<td>TUG time</td>
<td>A strong correlation ($\rho=0.70$) was observed between remote TUG times and standardized TUG times.</td>
</tr>
<tr>
<td>Van der Walt et al [22], 2018</td>
<td>Garmin Vivofit 2 (Garmin Ltd)</td>
<td>Accelerometer</td>
<td>Wrist</td>
<td>163</td>
<td>Step count</td>
<td>Participants receiving feedback on step goals from the device had significantly higher ($P&lt;.03$) mean daily step counts than those of participants who did not receive any feedback from the device.</td>
</tr>
<tr>
<td>Kuiken et al [23], 2004</td>
<td>Custom</td>
<td>Goniometer</td>
<td>Knee</td>
<td>11</td>
<td>ROM and mean activity rate</td>
<td>After total knee arthroplasty, patients wearing a device providing feedback had higher mean total activity rates—a measure of ROM—on days when they did not receive feedback from the device (mean 22.5, SD 11.1 activity counts per hour) than on days when they did receive feedback (mean 15.1, SD 10.9 activity counts per hour), but this was not statistically significant ($P=.11$).</td>
</tr>
</tbody>
</table>
The standard postoperative TJR outcome measures in this literature review included (1) assessments typically conducted in clinical settings, such as range of motion (ROM) assessments and the TUG test, and (2) PROMs, such as joint-specific outcome measures (Hip Disability and Osteoarthritis Outcome Score/Knee Injury and Osteoarthritis Outcome Score [HOOS/KOOS]), global health measures (Veterans RAND 12-Item Health Survey [VR-12]), patient satisfaction, and activity adherence.

**Results**

A total of 9 articles that met the inclusion criteria were identified. The articles evaluated the mobility and activity data collected through the wearable devices and compared them with standard clinical outcome measures and PROMs.

**Correlation of Wearables and Clinical Measures**

In evaluating ROM and TUG time, the wearables varied in accuracy. Kwansnicki et al. [15] observed 14 patients who underwent total knee replacement and wore the e-AR accelerometer (Imperial College London) on the ear to conduct home-based mobility assessments. The authors compared a generated sensor score, which was based on sensor data, with the results of other assessment techniques (TUG test and knee ROM). They calculated Spearman ρ correlation coefficients between sensor scores and TUG and ROM measurements to assess the strength of association. They found that perioperative sensor scores correlated, albeit not significantly for all activities, with TUG time and ROM improvements. In another study that focused on TUG measurements, Saporito et al. [21] collected standardized TUG data from 239 community-living older adults in a laboratory and sensor-based data on participants’ activities of daily living through a wearable pendant device for at least 3 days and developed a regularized linear model for estimating remote TUG times. Based on the device data of 15 patients who underwent total hip replacement, a strong correlation was observed between estimated remote TUG times and standardized TUG times via leave-one-out cross-validation.

**Correlation of Wearables and PROMs**

Data from wearable devices may correlate with PROMs. Bendich et al. [17] aimed to determine whether sensor-collected data could be used as predictors of PROMs. In their study, 22 patients who underwent TJR wore a Fitbit Flex (Fitbit LLC) device on the wrist, which allowed for the observation of potential associations between “daily step count” and “daily minutes active” data collected by the wearable and PROMs, specifically the HOOS/KOOS and VR-12, over time. The researchers found that changes observed in “daily step count” from before the operation to postoperative week 6 were strongly associated with changes in VR-12 scores, while changes observed in “daily minutes active” from before the operation to postoperative week 6 were strongly associated with changes in HOOS/KOOSs.

**Impact of the Use of Wearables on Patient Outcomes**

The authors of 2 articles discussed the impact of the use of wearable devices on postoperative TJR patient outcomes. Specifically, the researchers investigated how the ability of devices to offer feedback on exercise and rehabilitation to patients may impact patient outcomes. Van der Walt et al. [22] randomized 163 patients who underwent TJR into 2 groups; one received feedback for their rehabilitation via the Garmin VivoFit 2 (Garmin Ltd) accelerometer, and the other did not receive any feedback. They found that the mean daily step counts of the group that received feedback were significantly higher than those of the group that did not receive feedback (43% higher in postoperative week 1, 33% higher in postoperative week 2, 21% higher in postoperative week 6, and 17% higher at postoperative month 6). Surprisingly, in a study with 11 patients who underwent total knee arthroplasty, Kuiken et al. [23] found that patients who wore a device that provided feedback had a slightly higher mean total activity rate on days when they did not receive feedback from the device compared to those days when they did receive feedback from the device, although this difference was not statistically significant.

Patients reported high satisfaction with and adherence for the use of wearable devices. A study by Chiang et al. [16] found that in a group of 18 patients who underwent total knee replacement and wore a thigh- and calf-worn wearable, 83% reported no discomfort when wearing the device. In a study by Toogood et al. [20] on device adherence, the mean compliance rate for wearing an ankle-based Fitbit accelerometer (Fitbit LLC) among 33 patients who underwent total hip replacement was 89% (26.7/30 days). Although this study noted that devices were worn for 24 hours per day, apart from during washing, the daily duration of use was not specifically mentioned in the other selected studies.

**Device Data Accuracy Evaluation**

Several devices were found to be generally accurate in counting steps. Battenberg et al. [19] tested the accuracy of several widely used wearable devices in a convenience sample of 30 healthy participants. They found that the wrist-worn Fitbit One (Fitbit LLC) and Omron HJ-321 (Omron Corporation) had greater than 90% accuracy in step counting during all activities; the wristband devices, such as the Fitbit Force (Fitbit LLC) and Nike+ Fuelband SE (Nike Inc), had less than 90% accuracy for most activities; and the ankle-worn StepWatch Activity Monitor (Orthocare Innovations) was greater than 95% accurate when counting steps during lower cadence activities but undercounted steps during running by 25%. In a study by Chen et al. [18], 10 healthy participants, while wearing 3 APDM OPAL (APDM Wearable Technologies) sensors on the chest, thigh, and calf, performed 3 different rehabilitation exercises that were designed for patients with knee osteoarthritis to manage rehabilitation progress at home. The device was found to have an overall recognition accuracy of 97% for exercise type classification.
and an overall recognition accuracy of 88% for proper exercise posture.

Discussion

Wearable Data Can Be Used as Alternative Outcome Measures

Postoperative TJR recovery remains a black box to health care providers until patients report to a clinic or respond to a survey. With adequate implementation and the ability to collect data continuously, even from a remote setting, wearable devices can help health care providers to monitor progress consistently and detect early problems in rehabilitation [24]. The literature shows that function and activity data obtained from wearables, including step count and exercise tracking data, correlate with both clinical outcomes and PROMs [15,17,21]. Such wearable data are able to provide measures of patients’ objective functional outcomes that are comparable with standard clinical metrics and patient surveys. In addition, the opportunity to regularly monitor patients in real time and allow for direct feedback from and communication with health care providers can alleviate the inconveniences of unnecessary office visits and costs; patients with good progress can continue at-home rehabilitation, while patients with poor progress can be alerted to proactively visit a clinic before permanent complications occur. Further research is however needed to evaluate device bias and data accuracy to make sure that wearable results are reliable.

There has also been some support in the literature for the use of monitoring insoles, particularly for the purpose of load and gait analysis. Although preliminary findings suggest that monitoring insoles have good accuracy in measuring foot load distribution and natural gait, the few studies that have been performed are limited by small sample sizes [25,26]. Additional investigations with larger data sets will be needed.

Wearables Can Be Used to Improve Outcomes

In addition to generating data that correlate with established outcomes, wearables can also be used to improve outcomes overall by more actively engaging patients in exercise and activity [24,27]. Indeed, devices connected to mobile apps can provide feedback to patients regarding their rehabilitation routines, and the mobility metrics, such as daily step count, of patients who received such feedback significantly improved when compared to those of patients who did not receive feedback [22]. Additionally, the ability of these wearables to provide daily exercise reminders to patients and plot their progress over time sustained patients’ motivation and further contributed to outcome improvement [28,29].

Wearables and Apps Can Be Included in Future Health IT Infrastructure

As orthopedic clinical research has progressed, more data sources have emerged from which to monitor and guide patient rehabilitation and care following TJR. Whereas most patient data previously originated from electronic health records, direct patient-generated data in the form of PROMs or outcomes tracked and collected by wearables aptly supplement clinically collected data. Particularly, the ability of wearables to generate objective, continuous data showing trends in patient progress is unique in comparison to PROMs, which provide subjective data from predetermined time points, and electronic health record data, which are only collected during patients’ point-of-care visits and require medical professionals’ involvement. Moreover, with the increased emphasis on telemedicine, particularly since the COVID-19 pandemic, the remote monitoring of patient recovery via wearables represents a potential new path toward collecting patient data and guiding clinical decision-making [30,31]. These novel applications emphasize the role of wearables in the future of health IT infrastructure.

Challenges

There are still challenges to the implementation of wearable technology. Technical support will be needed for device calibration and data collection. Some research teams have assisted in the use of wearables during appointments scheduled at patients’ homes [15], hospital wards, or outpatient clinics [16]. Patients also need to be provided with training and guidance before and during the study period to ensure proper device mounting and use. Additionally, standardization must be established across different devices and across data collection in different settings to ensure that data are comparable and meaningful.

Conclusion

This review discusses the current state of the literature regarding the effectiveness of wearable devices in measuring and improving TJR outcomes, as well as the future directions of wearable device use. Wearable technologies have great potential for assessing and enhancing patients’ postoperative physical function. Wearables can be effective, alternative tools for evaluating TJR outcomes, as early findings have shown correlations among wearable-recorded data, PROMs, and clinical outcomes. The implementation and standardization of wearables should be addressed in future research.

Acknowledgments

The authors thank Sylvie Puig, PhD, for her editorial assistance.

Conflicts of Interest

DA is a scientific advisor to Exactech.

References

https://periop.jmir.org/2023/1/e39396

JMIR Perioper Med 2023 | vol. 6 | e39396 | p.58
(page number not for citation purposes)


Abbreviations

HOOS/KOOS: Hip Disability and Osteoarthritis Outcome Score/Knee Injury and Osteoarthritis Outcome Score
PRO: patient-reported outcome
PROM: patient-reported outcome measure
ROM: range of motion
TJR: total joint replacement
TUG: Timed Up and Go
VR-12: Veterans RAND 12-Item Health Survey

©Gregory Iovanel, David Ayers, Hua Zheng. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 12.01.2023. 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.
The Accuracy of Wrist-Worn Photoplethysmogram–Measured Heart and Respiratory Rates in Abdominal Surgery Patients: Observational Prospective Clinical Validation Study

Jonna A van der Stam1,2,3, MSc; Eveline H J Mestrom4, MD; Jai Scheerhoorn5, MD; Fleur E N B Jacobs6, MSc; Simon Nienhuijs5, MD, PhD; Arjen-Kars Boer2,3, PhD; Natal A W van Riel1,3,7, PhD; Helma M de Morree8, PhD; Alberto G Bonomi8, PhD; Volkher Scharnhorst1,2,3, PhD; R Arthur Bouwman4,9, MD, PhD

1Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, Netherlands
2Clinical Laboratory, Catharina Hospital, Eindhoven, Netherlands
3Expert Center Clinical Chemistry Eindhoven, Eindhoven, Netherlands
4Department of Anesthesiology, Intensive Care & Pain Medicine, Catharina Hospital, Eindhoven, Netherlands
5Department of Surgery, Catharina Hospital, Eindhoven, Netherlands
6Department of Medical Physics, Catharina Hospital, Eindhoven, Netherlands
7Department of Vascular Medicine, Amsterdam University Medical Centers, Amsterdam, Netherlands
8Patient Care & Monitoring Department, Philips Research, Eindhoven, Netherlands
9Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, Netherlands

Corresponding Author:
Jonna A van der Stam, MSc
Department of Biomedical Engineering
Eindhoven University of Technology
de Zaale
Eindhoven, 5612 AZ
Netherlands
Phone: 31 40 2398675
Email: jonna.vd.stam@catharinaziekenhuis.nl

Abstract

Background: Postoperative deterioration is often preceded by abnormal vital parameters. Therefore, vital parameters of postoperative patients are routinely measured by nursing staff. Wrist-worn sensors could potentially provide an alternative tool for the measurement of vital parameters in low-acuity settings. These devices would allow more frequent or even continuous measurements of vital parameters without relying on time-consuming manual measurements, provided their accuracy in this clinical population is established.

Objective: This study aimed to assess the accuracy of heart rate (HR) and respiratory rate (RR) measures obtained via a wearable photoplethysmography (PPG) wristband in a cohort of postoperative patients.

Methods: The accuracy of the wrist-worn PPG sensor was assessed in 62 post–abdominal surgery patients (mean age 55, SD 15 years; median BMI 34, IQR 25-40 kg/m²). The wearable obtained HR and RR measurements were compared to those of the reference monitor in the postanesthesia or intensive care unit. Bland-Altman and Clarke error grid analyses were performed to determine agreement and clinical accuracy.

Results: Data were collected for a median of 1.2 hours per patient. With a coverage of 94% for HR and 34% for RR, the device was able to provide accurate measurements for the large majority of the measurements as 98% and 93% of the measurements were within 5 bpm or 3 rpm of the reference signal. Additionally, 100% of the HR and 98% of the RR measurements were clinically acceptable on Clarke error grid analysis.

Conclusions: The wrist-worn PPG device is able to provide measurements of HR and RR that can be seen as sufficiently accurate for clinical applications. Considering the coverage, the device was able to continuously monitor HR and report RR when measurements of sufficient quality were obtained.

Trial Registration: ClinicalTrials.gov NCT03923127; https://www.clinicaltrials.gov/ct2/show/NCT03923127

https://periop.jmir.org/2023/1/e40474
KEYWORDS
Telemetry; monitoring; photoplethysmography; PPG; photoplethysmogram; wearable monitoring; vital parameter; wearable sensor; sensor; heart rate; respiratory Rate; respiration; respiratory; breathing; monitoring; wearable; postoperative; post-operative; vital sign

Introduction
Alterations in vital parameters can often be found hours before a life-threatening event occurs [1-7]. In current clinical practice, postoperative monitoring often consists of a period of continuous monitoring in an intensive care or postanesthesia care unit, followed by an admission to a general ward. Since continuous monitoring of vital parameters is not present in the general ward, nursing staff performs the so-called spot checks to monitor the patient’s vital parameters. During these spot checks, the nursing staff measures several vital parameters, often followed by manual entry or calculation of an early warning score such as the Modified Early Warning Score, to identify patients at risk of deterioration [8]. In clinical practice, these spot checks form a considerable workload, and vital parameters, especially respiratory rate (RR), are often poorly registered [9,10]. Additionally, as the name implies, these spot checks capture only vital parameters at a specific moment in time, and vital parameters during the rest of the day remain unknown.
Alternatively, wearable sensors could be used to unobtrusively and continuously measure vital parameters in postoperative patients. However, their accuracy in postoperative patients should be established prior to introduction in clinical practice.

One type of wearable sensor that can monitor a patient’s vital parameters is a photoplethysmography (PPG) wristband. This type of sensor measures the intensity of the light reflected from the skin, which indicates changes in the blood volume in peripheral circulation, to determine both heart rate (HR) and RR [11]. Wrist-worn PPG sensors have potential for use as a continuous, unobtrusive monitoring system in low-acuity settings such as the general ward.

A few studies have reported the accuracy of other PPG-based wearables in hospitalized patients; however, these trials only studied the measurement of HR [12-15]. Additionally, the accuracy of wrist plethysmography devices for HR measurements in a perioperative cohort was previously investigated and found to be clinically acceptable [15]. However, as both HR and RR have been identified as important parameters for the prediction of clinical deterioration, accuracy for both vital parameters should be established [16]. Therefore, this study aims to assess the accuracy of a wrist-worn PPG device for measuring both RR and HR in postoperative patients.

Methods

Study Population
These analyses were performed with a subpopulation of Transitional Care Study 3 (TRICA; ClinicalTrials.gov NCT03923127)—a single-center study on wearable monitoring in postoperative patients in a tertiary hospital [17,18]. All adult patients scheduled for major abdominal oncological or bariatric surgery from April 2019 to August 2020 who were willing and able to sign informed consent were eligible for participation. Patients who met any of the following criteria were not included: being pregnant or breastfeeding, having an allergy to tissue adhesives, having an antibiotic-resistant skin infection, having an active implantable device, or having any skin condition at the area of application of the devices. This subanalysis describes 68 postoperative patients, and inclusion into this subanalysis for accuracy of the wearable sensor was based on the availability of research personnel and real-time data logging equipment.

Ethics Approval
The trial was approved by the medical ethical committee METC Máxima MC, Veldhoven, The Netherlands (W19.001).

Data Collection
The wearable PPG wristband device, ELAN, was equipped with a Philips Cardio and Motion Monitoring Module (CM3, Philips Electronic Nederland BV), which contains a PPG and 3-axial accelerometer sensor. The PPG sensor measures the intensity of the green light scatter-reflected from the skin to determine changes in blood volume in the peripheral circulation with a sampling frequency of 32 Hz [19]. From the obtained PPG signal, HR and RR were determined using previously published algorithms, the RR measurements are derived from interbeat interval variability and PPG amplitude [20]. Additionally, the device reports a quality index with each measured vital value, which mostly captures the signal-to-noise ratio [15]. Only vitals with a quality index of 4 (range 0–4), are considered to be of high quality and can be included in further analysis.

Shortly after surgery, the PPG wristband was applied to the patient’s wrist in the postanesthesia care unit (PACU) or intensive care unit (ICU), depending on where the patient was recovering immediately after surgery. The wristband then continuously collected both HR and RR.

As a ground truth, the electrocardiogram (ECG)-based HR and capnography-based RR signals of 68 patients were extracted from the bedside monitor in the PACU or ICU. These signals were saved in real time for offline processing, allowing comparison between the HR and RR measured by the PPG wristband and the reference monitor. In the PACU, vital parameters from the CAR-ESCAPE monitor B650 (GE Healthcare) were extracted using iCollect software (GE Healthcare) with a sampling frequency of 250 Hz for ECG and 1 Hz for RR. In the ICU, vital parameters were extracted from the Philips IntelliVue MP70 monitor using IntelliVue software (Philips) with a sampling frequency of 100 Hz for ECG and 0.1 Hz for RR. HR was derived from the ECG on second-to-second bases using QRS detection algorithms, RR was obtained using the patient monitors’ algorithms.
Data Analysis

The obtained vital parameters from the PPG wristband and the reference monitors were synchronized using a means of cross-correlation on the HR signals, and synchronized signals were visually inspected and corrected if necessary. Patients with a reference recording length shorter than 15 minutes were excluded from the analysis.

Low-quality measurements were excluded from both the PPG and monitor data. For the PPG wristband vitals, a low quality index can originate from motion artefacts or a low signal-to-noise ratio. For HR and RR, detection of arrhythmia using an arrhythmia detection algorithm would also lead to a low quality score [21]. For the reference monitor, the logged ECG and capnography signals were visually inspected to identify low-quality measurements, based on assessment of the temporal sequence.

Baseline characteristics are expressed as mean (SD) or, in case of nonnormally distributed values, as median (IQR) values. Agreement between the PPG wristband and reference monitor measurements on a second-to-second basis was visualized using Bland-Altman plots [22]. As multiple observations from the same patients were analyzed, the bias and limits of agreement were calculated using the method for repeated measures of Zou et al [23]. Additionally, the 95% CIs around the limits of agreement were assessed using MOVER [23].

According to the American National Standards Institute consensus standard, the error for HR measurements should be ≤10% or ≤5 bpm. In this analysis, an error of ≤5 bpm for HR and ≤3 rpm for RR was considered clinically acceptable. Additionally, Clarke error grid analysis was performed to quantify the implications of the difference between the vitals measured by the reference monitor and the PPG wristband. Clarke error grid analysis was originally developed for blood glucose measurements, and the boundaries of the different zones were adapted on the basis of the Modified Early Warning Score protocol used in our hospital [8,17,24,25].

Results

In total, 68 postsurgical patients were enrolled, of whom 6 were excluded from HR analysis due to either unavailable ECG reference (n=2) or a recording length of less than 15 minutes (n=4). For RR analysis, 14 patients were excluded from further analysis due to either lack of sufficient quality capnography reference data (n=9) or insufficient recording length (n=5) (Figure 1). The characteristics of the included population are shown in (Table 1).

Figure 1. Flowchart of patient inclusion for heart rate (HR) and respiratory rate (RR) analysis. A total of 62 patients were included in the data analysis, of whom 8 were only included in the HR analysis and 54 were included in both analyses. ECG: electrocardiography.

<table>
<thead>
<tr>
<th>Enrolled patients</th>
<th>Exclusion:</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR</td>
<td>ECG not available (n=2)</td>
</tr>
<tr>
<td></td>
<td>Recording length &lt;15 minutes (n=4)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RR</th>
<th>Exclusion:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Insufficient quality capnography (n=9)</td>
</tr>
<tr>
<td></td>
<td>Recording length &lt;15 minutes (n=5)</td>
</tr>
</tbody>
</table>

Included for HR analysis n=62

Included for RR analysis n=54

Table 1. Baseline characteristics of the included population.
Table 1. Population demographics (N=62).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female, n (%)</td>
<td>33 (53)</td>
</tr>
<tr>
<td>Age (Years), mean (SD)</td>
<td>55 (15)</td>
</tr>
<tr>
<td>BMI (kg/m$^2$), median (IQR)</td>
<td>34 (25-40)</td>
</tr>
<tr>
<td>Surgery type, n (%)</td>
<td></td>
</tr>
<tr>
<td>Gastric bypass</td>
<td>21 (34)</td>
</tr>
<tr>
<td>Gastric sleeve</td>
<td>9 (15)</td>
</tr>
<tr>
<td>Esophagectomy</td>
<td>7 (11)</td>
</tr>
<tr>
<td>Hyperthermic intraperitoneal chemotherapy</td>
<td>9 (15)</td>
</tr>
<tr>
<td>Pancreatectomy</td>
<td>4 (6)</td>
</tr>
<tr>
<td>Low anterior resection or abdominoperineal resection with intraoperative radiation therapy</td>
<td>6 (10)</td>
</tr>
<tr>
<td>Low anterior resection or abdominoperineal resection without intraoperative radiation therapy</td>
<td>4 (6)</td>
</tr>
<tr>
<td>Debulking</td>
<td>2 (3)</td>
</tr>
<tr>
<td>Duration of surgery (minutes), median (IQR)</td>
<td>144 (76-342)</td>
</tr>
<tr>
<td>Postoperative admission to the intensive care unit, n (%)</td>
<td>27 (43)</td>
</tr>
</tbody>
</table>

HR Assessment

For HR assessment, a total of 146 hours of data, from both the PACU or ICU patient monitor and the wearable sensor, were collected in 62 patients. Per patient, a median of 1.2 hours of data (range 16 minutes to 10 hours) were collected. Overall, 492,987 (94%) of the PPG wristband data points were of sufficient quality to be included in the analysis. As shown in Figure 2, the percentage of sufficient-quality HR data per patient varied among patients, and a median of 96% (IQR 92%-99%) of high-quality HR data were obtained. The gaps without high-quality HR data ranged from a length of 1 second to 7.2 minutes, and 96% of the gaps were of <60 seconds.

Figure 2. Availability of photoplethysmography wristband data of high-quality data for heart rate (left) and respiratory rate (right) expressed as the percentage of seconds with high- and low-quality data.
Figure 3. Bland-Altman (top) and Clarke error grid (bottom) plots of the vital parameters obtained from the photoplethysmography (PPG) wristband and reference monitor, each data point represents 1 second. The upper figures depict Bland-Altman analysis for heart rate (HR; left) and respiratory rate (RR; right). Limits of agreement are indicated by the black lines, dashed lines represent the 95% CIs of the limits of agreement. The bottom figures depict the Clarke error grid analysis for HR (left) and RR (right) comparing the measurements of the reference monitor (x-axis) and the PPG wristband (y-axis). Zone A represents data points that differ less than 20% from the reference or are correctly identified as bradycardia or bradypnea. Zone B represents data points that differ by more than 20% but would not cause unnecessary treatment. Zone C represents points that would lead to unnecessary treatment for patients with normal vital parameters. Zone D represents failure to detect bradycardia or bradypnea, or tachycardia or tachypnea. Zone E represents data points where bradycardia or bradypnea and tachycardia or tachypnea are confused.
Table 2. Agreement and clinical accuracy of heart rate (HR) and respiratory rate (RR) measured by the photoplethysmography (PPG) wristband compared to those of the reference monitor.

<table>
<thead>
<tr>
<th>Data availability</th>
<th>HR</th>
<th>RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients, n</td>
<td>62</td>
<td>54</td>
</tr>
<tr>
<td>Measurements, n</td>
<td>526,833</td>
<td>495,217</td>
</tr>
<tr>
<td>Good-quality PPG wristband, n (%)</td>
<td>492,987 (94)</td>
<td>170,383 (34)</td>
</tr>
<tr>
<td>Good-quality reference, n (%)</td>
<td>515,991 (98)</td>
<td>367,092 (74)</td>
</tr>
<tr>
<td>Both good quality, n (%)</td>
<td>484,096 (92)</td>
<td>128,816 (26)</td>
</tr>
</tbody>
</table>

Bland-Altman analysis

<table>
<thead>
<tr>
<th></th>
<th>HR</th>
<th>RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson r</td>
<td>0.99</td>
<td>0.80</td>
</tr>
<tr>
<td>Bias (rpm/rpm), mean (SD)</td>
<td>–0.15 (1.8)</td>
<td>0.17 (2.6)</td>
</tr>
<tr>
<td>Lower limit of agreement (rpm/rpm), lower limit (95% CI)</td>
<td>–3.62 (–3.7 to –3.6)</td>
<td>–4.99 (–5.8 to –4.3)</td>
</tr>
<tr>
<td>Upper limit of agreement (rpm/rpm), upper limit (95% CI)</td>
<td>3.32 (3.3 to 3.4)</td>
<td>5.33 (4.7 to 6.2)</td>
</tr>
<tr>
<td>Within 5 bpm or 3 rpm, %</td>
<td>98</td>
<td>93</td>
</tr>
</tbody>
</table>

Clarke error grid analysis, n (%)

<table>
<thead>
<tr>
<th>Zone</th>
<th>HR</th>
<th>RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>483,716 (99.9)</td>
<td>115,434 (89.6)</td>
</tr>
<tr>
<td>B</td>
<td>369 (0.1)</td>
<td>10,781 (8.4)</td>
</tr>
<tr>
<td>C</td>
<td>0 (0)</td>
<td>61 (0)</td>
</tr>
<tr>
<td>D</td>
<td>11 (0)</td>
<td>2499 (1.9)</td>
</tr>
<tr>
<td>E</td>
<td>0 (0)</td>
<td>41 (0)</td>
</tr>
<tr>
<td>A+B</td>
<td>484,085 (100)</td>
<td>126,215 (98.0)</td>
</tr>
</tbody>
</table>

RR Assessment

For RR, a total of 138 hours of data, from both the PACU or ICU patient monitor and the wearable sensor, were collected from among 54 patients. A median of 1.2 hours (range 16 minutes to 11 hours) of data were collected per patient. Overall, 170,383 (34%) of the PPG wristband RR measurements were of sufficient quality to be included in further analysis. Figure 2 shows the availability of high-quality data per patient, a median of 20% (IQR 7%-40%) of sufficient-quality RR data were obtained. The gaps without high-quality RR data ranged from a length of 1 second to 67 minutes, and 81% of the gaps were of <60 seconds.

Bland-Altman analysis of the PPG wristband–measured RR showed a bias of 0.17 rpm and limits of agreement of –4.99 to 5.33 rpm. As 93% of the RR measurements met the predefined ≤3 rpm, the limits of agreement were wider than the predefined ±3 rpm. Clarke error grid analysis showed that 98% of the data points were within the clinically acceptable zones A and B, indicating that the differences between the PPG wristband RR and reference monitor only have limited clinical implications. Most of the remaining 2% of data points lie within zone D, which indicates failure to detect impaired RR either due to failure to detect bradypnea (1.85%) or tachypnea (0.09%). Splitting the data based on the unit patients were admitted to (ICU vs PACU) showed a numerically higher percentage of available good-quality PPG measurements in the ICU and numerically wider limits of agreement in the PACU (Multimedia Appendix 1).

Discussion

Principal Findings

The use of wearable sensors to monitor hospitalized patients is rapidly attracting attention in the clinical community. However, prior to the introduction of these devices in clinical practice, their performance in the patient population of interest needs to be established. As postoperative patients are currently only monitored using spot checks for the duration for which they are in the general ward, this population could benefit from wearable monitoring. This study focused on the performance of a wearable PPG wristband for the measurement of HR and RR in postoperative patients.

For HR, the device was able to accurately measure the vital parameter as the bias and limits of agreement were within the predefined ≤5 bpm. Any differences between the PPG wristband and reference monitor were found to be clinically acceptable since 100% of the measurements were within zones A and B of the Clarke error grid. Additionally, the wearable PPG sensor would be feasible in terms of data availability for HR as the device only reported low quality for 4% of the HR measurements.

For RR, 93% of the included measurements were within the predefined ≤3 rpm and while the bias was within this threshold, the limits of agreement were wider than the predefined cutoff. However, as 98% of the included measurements lie within zones A and B of the Clarke error grid, the device does provide...
The trial was partly financed by a grant from the Rijksdienst voor Ondernemend Nederland (grant RVO ITEA 161006).

The authors would like to thank Linda Eerikainen, Koen de Groot, and Wim Verhaegh for their contribution to the data preparation.

**Acknowledgments**

This study shows that the wearable PPG wristband can measure HR accurately and with sufficient coverage in postoperative patients. For RR, the large majority of the included data were clinically acceptable; however, the coverage of sufficient-quality RR measurements was low. Therefore, the PPG wristband would be able to perform continuous monitoring of HR and also report RR when sufficient-quality measurements are obtained. Before implementing such PPG-based wearable devices in clinical practice, both accuracy and coverage should be considered.

**Comparisons to Prior Work**

The accuracy of HR measurements by the same PPG wristband was previously studied in the PACU of our hospital. In the cohort of this study, the clinical accuracy in the PACU and ICU was confirmed with comparable results [15]. The accuracy of another wrist-worn PPG personal fitness tracker sensor for the monitoring of HR in hospitalized patients was previously studied by Kroll et al [12], who reported a bias of –4.7 and lower and upper limits of agreement of –31 and 21, respectively. Additionally, 73% of their measurements met the desired ≤5 bpm. Our findings with the ELAN PPG wristband show better agreement with the reference signal than their findings using the Fitbit Charge HR. The accuracy of another wrist-worn PPG sensor, the CardiacSense, in ambulatory patients was studied by Hochstadt et al [13]. As they reported their findings regarding the length of peak intervals rather than HR, comparison of results is difficult.

Limited data on the accuracy of RR measurements using PPG in clinical settings are available. Touw et al [26] studied the accuracy of finger-cuff PPG RR measurements in patients receiving procedural sedation and analgesia and found a bias of –2.0 rpm with limits of agreement from –12.4 to 8.4 rpm. Compared to our findings, the PPG wristband used in this study can measure RR with a smaller bias and smaller limits of agreement. Haveman et al [27] compared upper-arm–worn wearable PPG measurements of HR and RR to manual those performed by nursing staff. They found a moderate relationship for HR and a poor relationship for RR. However, their results cannot be easily compared to ours as gold-standard measurements were unavailable in their cohort. Additionally, Haveman et al [28] described lower accuracy and data availability for upper-arm–measured PPG RR during activity in volunteers. Patient activity level could therefore be a potential factor that relates to the differences between the RR in the ICU and the PACU. However, as the postoperative unit a patient is admitted to is chosen on the basis of surgery type, severity, and patient characteristics, this trial does not allow drawing conclusions regarding the origin of these differences. Papini et al [29] studied respiratory activity in a sleep-disordered population using a wrist-worn PPG device. They found a median correlation of 0.62 and a median per-patient coverage of 75.3%. Comparison of the accuracy to our findings is complicated as we reported a correlation over the entire data set; however, an overall correlation of 0.80 in this study indicates a better agreement between the 2 RR measurements. However, their median per-patient coverage of 75.3% clearly outperforms the 20% found in the present population.

**Strengths and Limitations**

This analysis was performed in a real-world, clinically relevant patient population, as postoperative patients could benefit from wearable monitoring in low-acuity care settings such as the surgical ward. However, this study had some limitations. First, while capnography is the gold standard for RR monitoring, a good-quality reference RR signal could not be obtained for 9 patients, and for the patients who could be included, 26% of the capnography data had insufficient quality to be included. Second, the data for this trial were collected in the PACU and ICU rather than the general ward. However, we believe that our findings could reasonably be transferred to the general ward as patients became alert and mobile during their stay in these recovery units. Third, the analysis of trending ability of the device could be an interesting addition to the data analysis and can be included in future research if longer monitoring times of both the wearable and reference monitor are available.

**Future Directions**

Other potential future clinical applications of PPG wearables include the measurement of activity level, blood pressure, HR variability, energy expenditure, and the detection of atrial fibrillation [21,30-33]. In future clinical use, PPG wristbands thus have the potential to provide information on even more aspects of the patients’ health status. This study shows that the ELAN PPG wristband can continuously measure HR with clinically acceptable accuracy. For RR, the device can perform clinically accurate measurements, but, due to limited coverage, can only be used to perform intermittent measurements.

**Conclusions**

The wearable PPG wristband can measure HR accurately and with sufficient coverage in postoperative patients. For RR, the large majority of the included data were clinically acceptable; however, the coverage of sufficient-quality RR measurements was low. Therefore, the PPG wristband would be able to perform continuous monitoring of HR and also report RR when sufficient-quality measurements are obtained. Before implementing such PPG-based wearable devices in clinical practice, both accuracy and coverage should be considered.
Conflicts of Interest
RAB acts as a clinical consultant for Philips Research. AGB is an employee of Philips Research. HMdM was an employee of Philips Research at the time this study was conducted.

Multimedia Appendix 1
Results split based on unit the patients were admitted to postoperatively.

References


Abbreviations
- **ECG**: electrocardiography
- **HR**: heart rate
- **ICU**: intensive care unit
- **PACU**: postanesthesia care unit
- **PPG**: photoplethysmography
- **RR**: respiratory rate
Prediction of Pelvic Organ Prolapse Postsurgical Outcome Using Biomaterial-Induced Blood Cytokine Levels: Machine Learning Approach

Mihyun Lim Waugh¹, BS; Nicholas Boltin¹, PhD; Lauren Wolf¹, PhD; Jane Goodwin², BS; Patti Parker³, BSN; Ronnie Horner⁴, PhD; Matthew Hermes⁵, PhD; Thomas Wheeler⁶, MD; Richard Goodwin², PhD; Melissa Moss¹,⁷, PhD

¹Biomedical Engineering Program, University of South Carolina, Columbia, SC, United States
²Biomedical Sciences, University of South Carolina School of Medicine - Greenville, Greenville, SC, United States
³Prisma Health Greenville, Greenville, SC, United States
⁴Department of Health Services Research and Administration, University of Nebraska Medical Center, Omaha, NE, United States
⁵Department of Bioengineering, Clemson University, Clemson, SC, United States
⁶Spartanburg Regional Healthcare System, Spartanburg, SC, United States
⁷Department of Chemical Engineering, University of South Carolina, Columbia, SC, United States

Corresponding Author:
Melissa Moss, PhD
Department of Chemical Engineering, University of South Carolina
301 Main St.
Rm 2C02
Columbia, SC, 29208
United States
Phone: 1 803 777 5604
Email: mossme@cec.sc.edu

Abstract

Background: Pelvic organ prolapse (POP) refers to symptomatic descent of the vaginal wall. To reduce surgical failure rates, surgical correction can be augmented with the insertion of polypropylene mesh. This benefit is offset by the risk of mesh complication, predominantly mesh exposure through the vaginal wall. If mesh placement is under consideration as part of prolapse repair, patient selection and counseling would benefit from the prediction of mesh exposure; yet, no such reliable preoperative method currently exists. Past studies indicate that inflammation and associated cytokine release is correlated with mesh complication. While some degree of mesh-induced cytokine response accompanies implantation, excessive or persistent cytokine responses may elicit inflammation and implant rejection.

Objective: Here, we explore the levels of biomaterial-induced blood cytokines from patients who have undergone POP repair surgery to (1) identify correlations among cytokine expression and (2) predict postsurgical mesh exposure through the vaginal wall.

Methods: Blood samples from 20 female patients who previously underwent surgical intervention with transvaginal placement of polypropylene mesh to correct POP were collected for the study. These included 10 who experienced postsurgical mesh exposure through the vaginal wall and 10 who did not. Blood samples incubated with inflammatory agent lipopolysaccharide, with sterile polypropylene mesh, or alone were analyzed for plasma levels of 13 proinflammatory and anti-inflammatory cytokines using multiplex assay. Data were analyzed by principal component analysis (PCA) to uncover associations among cytokines and identify cytokine patterns that correlate with postsurgical mesh exposure through the vaginal wall. Supervised machine learning models were created to predict the presence or absence of mesh exposure and probe the number of cytokine measurements required for effective predictions.

Results: PCA revealed that proinflammatory cytokines interferon gamma, interleukin 12p70, and interleukin 2 are the largest contributors to the variance explained in PC 1, while anti-inflammatory cytokines interleukins 10, 4, and 6 are the largest contributors to the variance explained in PC 2. Additionally, PCA distinguished cytokine correlations that implicate prospective therapies to improve postsurgical outcomes. Among machine learning models trained with all 13 cytokines, the artificial neural network, the highest performing model, predicted POP surgical outcomes with 83% (15/18) accuracy; the same model predicted
Pelvic organ prolapse (POP), defined as symptomatic descent of the vagina and surrounding pelvic organs, affects approximately 50% of parous women and 6% of nonparous women between ages 20 and 59 years [1], with almost 300,000 POP surgeries performed per year [2]. To reduce anatomical recurrence, surgical treatment may include the insertion of polypropylene mesh into the vaginal wall to provide mechanical support and reinforcement of the prolapsed organs. Unfortunately, postsurgical mesh complication, predominantly mesh exposure through the vaginal wall, occurs with some frequency and results in decreased quality of life, leaving patients with costly residual symptoms and emotional distress [3]. Patients may elect for surgical reintervention to revise or remove the mesh implantation. In fact, according to Reid et al [4], 37 (8%) out of 482 patients underwent further surgery to remove the mesh, and 7 (2%) patients repeated the prolapse surgery. These complications provoked the removal of transvaginal mesh kits from the market by the Food and Drug Administration in 2019. A clinical decision support tool to better inform both patients and surgeons about the risk of complications following POP surgery may allow for the reintroduction of this advantageous surgical augmentation.

Inflammatory responses are associated with mesh exposure due to asymptomatic mesh infection that inhibits the mesh from integrating with the surrounding environment [5]. While some degree of mesh-induced cytokine response is necessary for successful implantation, excess or unattenuated cytokine response could result in chronic inflammation and implant rejection. As chronic inflammation progresses, granulation tissues formed during the foreign body reaction will evolve into mesh encapsulation by regular dense connective tissue and myofibroblast-induced contracture around the implant, which can result in mesh exposure [6,7]. The balance between proinflammatory and anti-inflammatory agents is critical in achieving successful mesh implantation, and this balance may be influenced by the individual's response to the implant material. Thus, leveraging a patient’s immune response to the biomaterial could facilitate the prediction of postsurgical outcomes.

Leveraging a patient-specific, multifaceted immune response for the prediction of postsurgical complications is an ideal problem for the application of principal component analysis (PCA) and supervised machine learning models. In fact, this approach has been used to predict complications following other surgical procedures as well as progressive disease outcomes. In a liver transplant study, Raji and Vinod Chandra [8] applied PCA to a composite medical data set comprised of donors’ medical information as well as the recipients’ medical history and implemented an artificial neural network to predict the long-term survival of liver transplant patients. In an oral cancer retrospective study by Chu et al [9], PCA along with bivariate analyses were used to highlight correlated variables from the patient data, which included patient demographics and clinicopathological tumor data (including tumor sites, disease staging, etc), and to predict oral cancer progress.

PCA and supervised machine learning have also been applied to biological measurements for predicting medical outcomes. Tseng et al [10] built a predictive model for cardiac surgery–associated acute kidney injury (AKI) using preoperative biochemistry data in combination with patient demographic characteristics and clinical condition. Incorporating a different type of biological measurement, a glioblastoma study by Akbari et al [11] used PCA and support vector machines to distinguish multiparametric magnetic resonance imaging signatures and quantify the patterns to predict regions of tumor recurrence after surgery. Chen et al [12] demonstrated that specifically including immune data in predictive models enhances predictive capacity. These researchers implemented machine learning models using individual patient immune data, such as blood cytokine levels, to predict severe AKI after cardiac surgery and found that this approach provided a far superior prediction tool compared to a clinical factor–based model [12].

The application of PCA and machine learning to predict postsurgical complications in women after POP surgery has also shown promising results. In the study of Jelovsek et al [13], statistical modeling uses 32 candidate risk factors (ie, age, race, smoking history, etc) identified by consensus with surgical outcomes to predict postsurgical complications. This approach of using personalized preoperative decision-making based on the individual’s medical history presents a better predictive model to postsurgical complications and offers a more effective decision support tool than the practice of counseling patients using average success rates reported from large, randomized studies [13]. However, this predictive method does not leverage the patient’s potential immune response to the surgery involving polypropylene mesh.

**Conclusions:** This preliminary study, incorporating a sample size of just 20 participants, identified correlations among cytokines and demonstrated the potential of this novel approach to predict mesh exposure through the vaginal wall following transvaginal POP repair surgery. Further study with a larger sample size will be pursued to confirm these results. If corroborated, this method could provide a personalized medicine approach to assist surgeons in their recommendation of POP repair surgeries with minimal potential for adverse outcomes.

(JMIR Perioper Med 2023;6:e40402) doi:10.2196/40402

**KEYWORDS**

pelvic organ prolapse; polypropylene mesh; inflammatory response; cytokines; principal component analysis; supervised machine learning models; surgical outcome prediction; biomaterial; repair surgery

**Introduction**

Pelvic organ prolapse (POP), defined as symptomatic descent of the vagina and surrounding pelvic organs, affects approximately 50% of parous women and 6% of nonparous women between ages 20 and 59 years [1], with almost 300,000 POP surgeries performed per year [2]. To reduce anatomical recurrence, surgical treatment may include the insertion of polypropylene mesh into the vaginal wall to provide mechanical support and reinforcement of the prolapsed organs. Unfortunately, postsurgical mesh complication, predominantly mesh exposure through the vaginal wall, occurs with some frequency and results in decreased quality of life, leaving patients with costly residual symptoms and emotional distress [3]. Patients may elect for surgical reintervention to revise or remove the mesh implantation. In fact, according to Reid et al [4], 37 (8%) out of 482 patients underwent further surgery to remove the mesh, and 7 (2%) patients repeated the prolapse surgery. These complications provoked the removal of transvaginal mesh kits from the market by the Food and Drug Administration in 2019. A clinical decision support tool to better inform both patients and surgeons about the risk of complications following POP surgery may allow for the reintroduction of this advantageous surgical augmentation.

Inflammatory responses are associated with mesh exposure due to asymptomatic mesh infection that inhibits the mesh from integrating with the surrounding environment [5]. While some degree of mesh-induced cytokine response is necessary for successful implantation, excess or unattenuated cytokine response could result in chronic inflammation and implant rejection. As chronic inflammation progresses, granulation tissues formed during the foreign body reaction will evolve into mesh encapsulation by regular dense connective tissue and myofibroblast-induced contracture around the implant, which can result in mesh exposure [6,7]. The balance between proinflammatory and anti-inflammatory agents is critical in achieving successful mesh implantation, and this balance may be influenced by the individual's response to the implant material. Thus, leveraging a patient’s immune response to the biomaterial could facilitate the prediction of postsurgical outcomes.

Leveraging a patient-specific, multifaceted immune response for the prediction of postsurgical complications is an ideal problem for the application of principal component analysis (PCA) and supervised machine learning models. In fact, this approach has been used to predict complications following other surgical procedures as well as progressive disease outcomes. In a liver transplant study, Raji and Vinod Chandra [8] applied PCA to a composite medical data set comprised of donors’ medical information as well as the recipients’ medical history and implemented an artificial neural network to predict the long-term survival of liver transplant patients. In an oral cancer retrospective study by Chu et al [9], PCA along with bivariate analyses were used to highlight correlated variables from the patient data, which included patient demographics and clinicopathological tumor data (including tumor sites, disease staging, etc), and to predict oral cancer progress.

PCA and supervised machine learning have also been applied to biological measurements for predicting medical outcomes. Tseng et al [10] built a predictive model for cardiac surgery–associated acute kidney injury (AKI) using preoperative biochemistry data in combination with patient demographic characteristics and clinical condition. Incorporating a different type of biological measurement, a glioblastoma study by Akbari et al [11] used PCA and support vector machines to distinguish multiparametric magnetic resonance imaging signatures and quantify the patterns to predict regions of tumor recurrence after surgery. Chen et al [12] demonstrated that specifically including immune data in predictive models enhances predictive capacity. These researchers implemented machine learning models using individual patient immune data, such as blood cytokine levels, to predict severe AKI after cardiac surgery and found that this approach provided a far superior prediction tool compared to a clinical factor–based model [12].

The application of PCA and machine learning to predict postsurgical complications in women after POP surgery has also shown promising results. In the study of Jelovsek et al [13], statistical modeling uses 32 candidate risk factors (ie, age, race, smoking history, etc) identified by consensus with surgical outcomes to predict postsurgical complications. This approach of using personalized preoperative decision-making based on the individual’s medical history presents a better predictive model to postsurgical complications and offers a more effective decision support tool than the practice of counseling patients using average success rates reported from large, randomized studies [13]. However, this predictive method does not leverage the patient’s potential immune response to the surgery involving polypropylene mesh.
In this preliminary study, we explored the levels of baseline and stimulus-induced cytokines in blood isolated from patients who had undergone POP repair surgery with a polypropylene mesh. Proinflammatory and anti-inflammatory cytokine levels from these data were analyzed using PCA to establish the principal components (PCs) and to identify associative or opposing trends among cytokines. In addition, supervised machine learning models were applied to demonstrate predictive capabilities when models were trained with either all 13 cytokines or a smaller group of 7 cytokines determined most effective by a random forest method. The results demonstrate that leveraging PCA and supervised machine learning models to predict outcomes of vaginal mesh implantation has the potential to benefit future patients when they are faced with this surgical decision, which carries a relatively high risk of unsuccessful surgical outcome.

**Methods**

**Study Population**

In total, 20 healthy, nonpregnant female participants aged 56-89 years at Prisma Health Greenville Memorial Hospital with a history of surgical intervention to correct POP via a procedure that used polypropylene mesh were selected for the study. The participants, who were not matched, included 10 who experienced postsurgical mesh complication in which the implanted mesh protruded through the vaginal wall (also referred to as mesh exposure) and 10 participants who did not experience this complication post surgery. This sample size was estimated as an effective cohort for the pilot study using a 1-tailed t test based on an a priori power analysis, which indicates the number of patients for a given theoretical minimum study power as a function of the expected difference between patients with and those without mesh exposure, or the Cohen d. We assumed a conservative study power of 0.80 and a 100%, or 2-fold, difference in the level of a given cytokine between individuals with and without mesh exposure, equivalent to a Cohen d of 1. Here, 20 patients with equal distribution among the 2 groups are needed to observe the difference with a probability of .1. Participants with POP recurrence or taking medications that would alter inflammatory response were excluded from this study.

**Ethics Approval**

The study protocol was approved by the institutional review board (IRB) of Prisma Health (Pro00067964). Informed consent from all study participants was obtained using an IRB-approved informed consent form. All samples collected and data analyzed were deidentified and followed IRB protocol.

**Blood Sample Collection and Processing**

Blood samples were obtained from the 20 selected participants. Approximately 12 mL of blood was drawn from the upper extremity of each participant into 3 BD Vacutainer EDTA-coated tubes. Deidentified blood samples were then transferred on ice to a laboratory facility at the University of South Carolina School of Medicine Greenville for immediate processing. Each participant’s blood sample was divided into equal aliquots for 24-hour incubation at 37 °C under 3 distinct conditions: (1) incubation with inflammatory agent lipopolysaccharide (LPS) at 20 ng/mL (positive control), (2) incubation with sterile polypropylene mesh area of 2 cm × 2 cm (experimental), and (3) incubation alone (negative control). After incubation, the plasma layer was collected following centrifugation (1500 × g, 10 min, 4 °C) and immediately stored at –80 °C.

**Measurement of Blood Cytokine Levels**

Cytokine levels in each blood sample were quantified using the bead-based MILLIPLEX Human Cytokine/Chemokine/Growth Factor Panel A—Immunology Multiplex Assay (EMD Millipore Corp), which is composed of analytes for target cytokines interleukin 1α (IL-1α), IL-1β, IL-2, IL-4, IL-6, IL-8, IL-10, IL-12p40, IL-12p70, IL-17A, interferon-gamma (IFN-γ), tumor necrosis factor-alpha (TNF-α), and granulocyte-macrophage colony-stimulating factor (GM-CSF). Frozen plasma samples were thawed at room temperature and analyzed following Milliplex protocol guidelines. Cytokine concentrations were measured using a Bio-Plex 200 (Bio-Rad) and Bio-Plex Manager software (Bio-Rad). Sample volume was doubled to ensure measurable levels of cytokines, and assay output data were adjusted to reflect concentrations in plasma samples. Each multiplex assay was performed in duplicate, and cytokine levels were evaluated in 3 independent measurements.

**Data Analysis**

**Overview**

Cytokine data gathered from the multiplex immunoassay were analyzed using data mining and predictive analytical methods. PCA was used to identify important cytokines by studying their contributions to each PC as well as to discern associations between cytokines. Supervised machine learning models were created to determine whether cytokine levels can accurately predict which patients are more likely to experience mesh exposure post surgery.

**Descriptive Analytics**

The statistical programming language R (version 4.1.2; R Foundation) was used to analyze raw cytokine data values generated from the multiplex immunoassay. The imported data structure contained 60 observations (20 participants × 3 independent measurements) and 40 total variable fields (13 cytokines × 3 blood treatments + 1 target variable). The target variable was the participant’s outcome, which indicated a postsurgical complication that participants might have experienced following POP surgery. Observations marked “presence” represent participants who experienced mesh exposure through the vaginal wall. Observations marked “absence” represent participants who did not experience any mesh exposure through the vaginal wall. Univariate and multivariate methods were used to explore the data set, including identifying missing values, analyzing outliers, and visualizing frequency distributions.

**PCA**

PCA was performed using the FactoMineR package (version 2.4; R Foundation) [14] to identify associations between cytokines [15]. Before analysis, data transformations were
performed on each variable to correct for skewness in the distribution. The amount of skewness was calculated to assess the symmetry of distribution for each variable using equation 1, where \( \bar{x} \) is the sample mean and \( x_i \) and \( n \) are the individual observations and number of observations, respectively, within the sample [16]. Each cytokine’s distribution was corrected for skewness using either a natural logarithm, square root, or inverse square root method. Using equation 2, z-score standardization was also applied to scale each cytokine variable, thus ensuring that the mean was equal to 0 and SD equal to 1. Biplots were created to visualize PCs with the highest degree of variance explained. The eigenvectors were overlayed on the biplots to visualize correlations and identify hidden patterns between cytokines.

Predictive Analytics
Supervised machine learning models were created using the caret package (version 6.0-90; R Foundation) [17] in the R programming language. The 4 models trained were decision tree, logistic regression, Naive Bayes, and artificial neural network. This approach focused on the data set from the experimental group only (cytokine expression for blood incubated with polypropylene mesh). Prior to creating the predictive models, the original data were split using an industry standard of 70% for training and 30% for testing. Each group contained an equal distribution of participants who did or did not experience postsurgical mesh exposure through the vaginal wall—the prediction target for each model. Each model was then trained using the 70% (42/60) subset and a cross-validation training control. A 10-fold cross-validation with 25% (15/60) left out replicated 3 times was used on each model to avoid bias and overfitting. From this, training accuracies are reported. Additional testing was performed for the prediction accuracy of each model using the 30% (18/60) test data. Prediction accuracies are reported along with sensitivity and specificity for the prediction of participants to experience postsurgical mesh exposure.

Additionally, this process was replicated to study the effects of reducing the number of cytokines needed to predict a postsurgical mesh exposure. A random forest algorithm was used to select important cytokines for this study. These models were trained and tested for accuracy, sensitivity, and specificity as detailed above. The results are compared to models trained with all 13 cytokines.

Results
PCA
To identify significant associations among the cytokines, PCA was used to examine a total of 60 blood samples (20 participants × 3 blood treatments). Among the 20 participants, 10 experienced postsurgical mesh exposure through the vaginal wall and 10 did not. Figure 1 depicts a biplot of each blood treatment and summarizes the intercorrelated relationships among individual inflammatory mediators. The combined variances explained for PC 1 and PC 2 in blood samples incubated with LPS (Figure 1A), polypropylene mesh (Figure 1B), or alone (Figure 1C) were approximately 64%, 73%, and 66%, respectively. In all 3 treatment groups, IL-10 and IL-4 align in the same directions, as do IL-12p70 and IFN-γ, indicating a positive correlation for both of these cytokine pairs. In contrast, IL-6 and IL-12p40 were negatively correlated when comparing stimulation of blood via LPS (Figure 1A) versus polypropylene mesh (Figure 1B). Only IL-1α displayed a negative correlation when comparing blood incubated with polypropylene mesh (Figure 1B) versus blood incubated alone (Figure 1C).

Figure 1. PCA was performed using cytokine levels in blood samples of postsurgical POP subjects; the analysis included 60 blood samples (20 subjects × 3 blood treatments), wherein each sample was evaluated in 3 independent measurements performed in duplicate. Biplots illustrating individual cytokines were constructed for blood samples incubated in the presence of LPS (A), incubated in the presence of polypropylene mesh (B), or incubated alone (C). Arrow direction indicates the cytokine correlation; arrow length indicates the magnitude of the variation. IL-1α: interleukin-1 alpha; IL-1β: interleukin-1 beta; IL-2: interleukin-2; IL-4: interleukin-4; IL-6: interleukin-6; IL-8: interleukin-8; IL-10: interleukin-10; IL-12p40: interleukin-12p40; IL-12p70: interleukin-12p70; IL-17A: interleukin-17A; IFN-γ: interferon gamma; TNF-α: tumor necrosis factor-alpha; GM-CSF: granulocyte-macrophage colony-stimulating factor; PC: principal component.

When PCA was used to examine only blood samples incubated with polypropylene mesh, PC 1 and PC 2 explained 60.1% and 13.1% of the total data variance, respectively (Figure 1B). Figure 2A displays each cytokine’s contribution to PC 1 and illustrates that IFN-γ, IL-12p70, and IL-2 are the predominant contributors to the variance explained in PC 1. In addition, IL-1α, IL-17A, IL-12p40, IL-12p70, GM-CSF, and TNF-α were loaded predominately on PC 2.
and TNF-α exhibited contributions above a level expected if the contributions were uniform. All other cytokines have contributions to PC 1 similar to or less than what would be expected if the contributions of all cytokines were uniform. Figure 2B illustrates that the predominant contributors to the variance explained in PC 2 are IL-10, IL-4, and IL-6. All other cytokines have contributions to PC 2 similar to or less than what would be expected if the contribution of all cytokines were uniform.

In order to visualize associations between the participants presenting the absence or presence of postsurgical mesh exposure through the vaginal wall, a biplot illustrating individual participants was created (Figure 3). This biplot reveals a high percentage of variability represented by the first 2 PCs (79.1%). Blood samples from participants who did not experience postsurgical mesh exposure were heavily represented by positive PC 1 values, while blood samples from participants with the presence of postsurgical mesh exposure were generally represented by positive PC 2 values.

Predictive Analysis

Four supervised machine learning models incorporating all 13 cytokines were trained using 70% (42/60) of the available 60 observations (20 participants × 3 independent measurements); the remaining 30% (18/60) was used to test the models’ accuracy when predicting the presence of mesh exposure through the vaginal wall. All 4 machine learning machines achieved at least 62% (26/42) training accuracy (Table 1). Artificial neural network achieved the highest prediction accuracy of 83% (15/18), while decision tree and Naïve Bayes both achieved a prediction accuracy of 61% (11/18). Naïve Bayes, decision tree, and artificial neural network excelled at correctly predicting patients with the presence of mesh exposure postsurgery at 89% (16/18). Artificial neural network was superior for correctly predicting patients who did not experience mesh exposure postsurgery (14/18, 78%).
Table 1. Summary of supervised learning model statistics. All 13 cytokines were used to predict the presence or absence of postsurgical mesh exposure through the vaginal wall; 70% (42/60) of observations were used for training, and 30% (18/60) of observations were used for testing.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training accuracy, n (%)</th>
<th>Prediction accuracy, n (%)</th>
<th>95% CI</th>
<th>Sensitivity, n (%)</th>
<th>Specificity, n (%)</th>
<th>Prediction, κ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial neural network</td>
<td>33 (79)</td>
<td>15 (83)</td>
<td>0.586-0.964</td>
<td>14 (78)</td>
<td>16 (89)</td>
<td>0.667</td>
</tr>
<tr>
<td>Decision tree</td>
<td>27 (64)</td>
<td>11 (61)</td>
<td>0.57-0.827</td>
<td>6 (33)</td>
<td>16 (89)</td>
<td>0.222</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>26 (62)</td>
<td>11 (61)</td>
<td>0.357-0.827</td>
<td>6 (33)</td>
<td>16 (89)</td>
<td>0.222</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>31 (73)</td>
<td>9 (50)</td>
<td>0.260-0.740</td>
<td>10 (56)</td>
<td>8 (44)</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Predictive Analysis Using Feature Selection

Additional models and predictive analyses explored whether a smaller set of cytokines could achieve similar predictive results. Feature selection using a random forest method identified a group of 7 cytokines capable of yielding effective predictive analysis: IL-1β, IL-8, IL-12p40, IL-12p70, TNF-α, IL-17A, and IL-6. Figure 4 illustrates that models exhibited variation among the importance of cytokines when implementing this more targeted group of cytokines. IL-1 and IL-8 are strongly represented in all models, while IL-6 is important in only Naïve Bayes.

Table 2 illustrates that all models achieved at least 64% (27/42) training accuracy. The logistic regression model that used the 7 selected cytokines achieved a training accuracy of 81% (34/42), a prediction accuracy of 72% (13/18), a sensitivity of 67% (12/18), and a specificity of 78% (14/18), and thus outperformed compared to the logistic regression model that incorporated all of the cytokine data. Moreover, decision tree models achieved the same result when using the selected cytokines or when all cytokines were included. The prediction accuracy in Naïve Bayes and artificial neural network models executed with the 7 selected cytokines decreased by only 5% each compared to the same models that used all of the cytokine data.

Table 2. Summary of supervised learning model statistics. Feature selection via random forest was used to identify a group of 7 cytokines capable of yielding effective predictive analysis. The subset of cytokines was used to predict the presence or absence of postsurgical mesh exposure through the vaginal wall; 70% (42/60) of observations were used for training, and 30% (18/60) of observations were used for testing.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training accuracy, n (%)</th>
<th>Prediction accuracy, n (%)</th>
<th>95% CI</th>
<th>Sensitivity, n (%)</th>
<th>Specificity, n (%)</th>
<th>Prediction, κ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial neural network</td>
<td>34 (81)</td>
<td>14 (78)</td>
<td>0.524-0.936</td>
<td>12 (67)</td>
<td>16 (89)</td>
<td>0.556</td>
</tr>
<tr>
<td>Decision tree</td>
<td>27 (64)</td>
<td>11 (61)</td>
<td>0.356-0.827</td>
<td>6 (33)</td>
<td>16 (89)</td>
<td>0.222</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>30 (72)</td>
<td>10 (56)</td>
<td>0.308-0.785</td>
<td>6 (33)</td>
<td>14 (78)</td>
<td>0.111</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>34 (81)</td>
<td>13 (72)</td>
<td>0.465-0.903</td>
<td>12 (67)</td>
<td>14 (78)</td>
<td>0.444</td>
</tr>
</tbody>
</table>

Discussion

Summary

Among patients with POP who undergo mesh implantation surgery, 17% of them experience mesh exposure through the vaginal wall [18]. This rate of surgical mesh complication is significant when compared to 0.035%-5.4% mesh-related erosions reported in other mesh-based surgeries [19-23], necessitating the development of a personalized decision support tool for patients with POP. This exploratory study demonstrates a novel and efficient approach to predicting postsurgical outcomes for mesh implantation using cytokine levels in patient blood following exposure to a biomaterial. Previous studies have often used patient demographic and medical data to train machine learning programs to create predictive outcomes for POP mesh surgeries [13]. In contrast, this study uses biological material to mimic an in vivo response, thus presenting a novel, noninvasive, personalized clinical decision tool. A systematic PCA approach identifies associations among cytokines that provide physiological insight. Supervised machine learning models developed in this study demonstrate that blood cytokine

https://periop.jmir.org/2023/1/e40402
measurements may be used as a predictive tool. In addition, the number of cytokine measurements needed may be reduced without compromising predictive capabilities, rendering this approach more applicable within a clinical setting.

**Principal Findings and Comparison to Prior Work**

The PCA analysis illustrated in Figure 1 reveals several significant associations among the cytokines. Several cytokines display positive correlations when comparing the 2 different stimuli: LPS and polypropylene mesh. However, IL-6 and IL-12p40 are negatively correlated between these 2 treatments. Thus, these 2 cytokines may explain the different inflammatory responses induced by LPS versus polypropylene mesh. When comparing blood samples incubated alone to those incubated in the presence of polypropylene mesh, only IL-1α exhibits a negative correlation, demonstrating that this proinflammatory mediator might be specifically affected by the mesh stimulus. Furthermore, 2 pairs of cytokines positively correlate (IL-10 and IL-4; IL-12p70 and IFN-γ), indicating that one of the cytokines in each pair could be eliminated to reduce the number of cytokines tested in a clinical setting.

The cytokines that contribute most to each PC segregate into proinflammatory and anti-inflammatory cytokines (Figure 2). When cytokine data from patient blood incubated with mesh were analyzed using PCA, cytokines IFN-γ, IL-12p70, and IL-2 were the largest contributors to the variance explained in PC 1. These markers are identified as proinflammatory agents [24-26], which suggests that proinflammatory cytokines may heavily influence PC 1. In contrast, cytokines IL-10, IL-4, and IL-6 were the largest contributors to the variance explained in PC 2. IL-4 and IL-10 are prominent anti-inflammatory cytokines [25], suggesting that anti-inflammatory cytokines heavily influence PC 2. IL-6, previously thought to have proinflammatory function only, is recently recognized as potentially having both proinflammatory and anti-inflammatory roles in COVID-19 [27] and diabetes [28].

When juxtaposing the biplot of polypropylene-stimulated cytokine observations (Figure 1B) with that of mesh exposure outcome (Figure 3), it can be extrapolated that proinflammatory cytokines IL-12p40, IL-1α, and TNF-α are positioned in the region of the biplot that uniquely corresponds to surgical outcomes involving the presence of mesh exposure through the vaginal wall. Such juxtaposition suggests that IL-12p40, IL-1α, and TNF-α may be associated with the presence of postsurgical mesh exposure. These observations may inspire potential therapeutic strategies that could improve postsurgical outcome. For example, the surgical mesh could be designed to modulate these key proinflammatory cytokines. In this way, while supporting the pelvic structure, the mesh could simultaneously function in controlling the cytokine response to minimize biomaterial rejection.

Table 1 describes the accuracy of supervised machine learning models. Chen et al [12] similarly used blood cytokine levels in a machine learning study to predict severe AKI after cardiac surgery. Their study concluded that a logistic regression model was the most effective in discovering the cytokine associations in severe AKI. In this study, the prediction accuracy for all 4 models exceeded 60%, with the artificial neural network model demonstrating the best overall performance, predicting POP surgical outcomes with 83% (15/18) accuracy when trained with all 13 cytokines. This predictive capability is similar to that reported for prediction derived from patient medical history [4], despite this study comprising a significantly smaller patient group. Considering the small population size, these results represent relatively high prediction accuracy for health care data.

When creating models trained with a subgroup of 7 cytokines (Table 2), selected using a random forest method, the artificial neural network model maintained the greatest effectiveness with respect to sensitivity, specificity, and prediction accuracy. Moreover, the group of selected cytokines outperformed the larger group of cytokines in the logistic regression model and achieved the same results in the decision tree model. The Naïve Bayes and artificial neural network prediction accuracy dropped only 5% when using the subgroup of cytokines, thus demonstrating the resiliency of these models. These results demonstrate that predictive capabilities are retained with fewer cytokines, which would enhance clinical feasibility by reducing the cost and time associated with this clinical decision tool.

**Limitations and Future Directions**

This study implemented rigorous research methods to identify physiological relationships among cytokine markers and developed robust machine learning models to predict mesh exposure; yet, some limitations should be noted. First, because this is a pilot study, the sample size is limited to 20 participants within a single hospital system. Nevertheless, this limited sample size predicted 83% (15/18) accuracy, a level that compares favorably with another predictive model study by Chu et al [9] that achieved a prediction accuracy of 71% in a study population size of 467. Thus, the results of this pilot study indicate the utility of this approach and the merit of future studies. Future study will provide validation with a larger population of participants from multiple hospitals. Additionally, the 10 participants in each group were not matched regarding variables. To minimize confounders, patients with POP recurrence or taking medication that would alter inflammatory response were excluded and the age ranges and average age at the time of surgery within each group were similar. Future studies with a larger population, however, will benefit from matching participants with respect to these and other potentially confounding variables. Nonetheless, the results of this pilot study highlight the importance of inflammatory markers in the prediction of this postsurgical condition.

**Conclusions**

While this preliminary study is limited to a sample size of just 20 participants, this novel approach to using cytokine response to predict POP surgical outcomes has successfully distinguished important cytokines and their correlations. Moreover, these relationships point toward prospective therapies that could...
promote better surgical outcomes. Supervised learning models also demonstrate a high level of accuracy, specificity, and sensitivity, even when a smaller group of cytokine data is used. This result suggests that blood cytokine analysis might be feasibly used in a clinical setting to predict POP surgical outcomes. Further study with a larger patient population will be needed to confirm the utility of this method. If successful at a larger scale, this approach has the potential to change perspectives in which surgeons would recommend and proceed with POP repair surgeries and to prevent undesired outcomes of mesh-related surgeries in patients.

Acknowledgments
This work was supported by a Transformative Research Grant from Greenville Health System Health Sciences Center to MM, TW, RH, and RG.

Data Availability
Data are openly available in the GitHub public repository [30].

Conflicts of Interest
None declared.

References
29. BiomaterialStudyCytokine. GitHib. URL: https://github.com/tomhjlim/BiomaterialStudyCytokine.git [accessed 2023-05-12]

Abbreviations

AKI: acute kidney injury
GM-CSF: granulocyte-macrophage colony-stimulating factor
IFN: interferon
IL: interleukin
IRB: institutional review board
LPS: lipopolysaccharide
PC: principal component
PCA: principal component analysis
POP: pelvic organ prolapse
TNF: tumor necrosis factor
Early Warning Scores to Support Continuous Wireless Vital Sign Monitoring for Complication Prediction in Patients on Surgical Wards: Retrospective Observational Study

Mathilde C van Rossum1,2, PhD; Robin E M Bekhuis3,4, MSc; Ying Wang1,4, PhD; Johannes H Hegeman3, MD, PhD; Ellis C Folbert3, PhD; Miriam M R Vollenbroek-Hutten1, PhD; Cornelis J Kalkman5, MD, PhD; Ewout A Kouwenhoven1*, MD, PhD; Hermie J Hermens1*, PhD

1Department of Biomedical Signals and Systems, University of Twente, Enschede, Netherlands
2Department of Cardiovascular and Respiratory Physiology, University of Twente, Enschede, Netherlands
3Department of Surgery, Hospital Group Twente, Almelo, Netherlands
4Hospital Group Twente Academy, Hospital Group Twente, Almelo, Netherlands
5Department of Anesthesiology, University Medical Center Utrecht, Utrecht, Netherlands
*these authors contributed equally

Corresponding Author:
Mathilde C van Rossum, PhD
Department of Cardiovascular and Respiratory Physiology
University of Twente
PO box 217
Enschede, 7500 AE
Netherlands
Phone: 31 53 489 4152
Email: m.c.vanrossum@utwente.nl

Abstract

Background: Wireless vital sign sensors are increasingly being used to monitor patients on surgical wards. Although early warning scores (EWSs) are the current standard for the identification of patient deterioration in a ward setting, their usefulness for continuous monitoring is unknown.

Objective: This study aimed to explore the usability and predictive value of high-rate EWSs obtained from continuous vital sign recordings for early identification of postoperative complications and compares the performance of a sensor-based EWS alarm system with manual intermittent EWS measurements and threshold alarms applied to individual vital sign recordings (single-parameter alarms).

Methods: Continuous vital sign measurements (heart rate, respiratory rate, blood oxygen saturation, and axillary temperature) collected with wireless sensors in patients on surgical wards were used for retrospective simulation of EWSs (sensor EWSs) for different time windows (1-240 min), adopting criteria similar to EWSs based on manual vital signs measurements (nurse EWSs). Hourly sensor EWS measurements were compared between patients with (event group: 14/46, 30%) and without (control group: 32/46, 70%) postoperative complications. In addition, alarms were simulated for the sensor EWSs using a range of alarm thresholds (1-9) and compared with alarms based on nurse EWSs and single-parameter alarms. Alarm performance was evaluated using the sensitivity to predict complications within 24 hours, daily alarm rate, and false discovery rate (FDR).

Results: The hourly sensor EWSs of the event group (median 3.4, IQR 3.1-4.1) was significantly higher (P<.004) compared with the control group (median 2.8, IQR 2.4-3.2). The alarm sensitivity of the hourly sensor EWSs was the highest (80%-67%) for thresholds of 3 to 5, which was associated with alarm rates of 2 (FDR=85%) to 1.2 (FDR=83%) alarms per patient per day respectively. The sensitivity of sensor EWS–based alarms was higher than that of nurse EWS–based alarms (maximum=40%) but lower than that of single-parameter alarms (87%) for all thresholds. In contrast, the (false) alarm rates of sensor EWS–based alarms were higher than that of nurse EWS–based alarms (maximum=0.6 alarm/patient/d; FDR=80%) but lower than that of single-parameter alarms (2 alarms/patient/d; FDR=84%) for most thresholds. Alarm rates for sensor EWSs increased for shorter time windows, reaching 70 alarms per patient per day when calculated every minute.
Conclusions: EWSs obtained using wireless vital sign sensors may contribute to the early recognition of postoperative complications in a ward setting, with higher alarm sensitivity compared with manual EWS measurements. Although hourly sensor EWSs provide fewer alarms compared with single-parameter alarms, high false alarm rates can be expected when calculated over shorter time spans. Further studies are recommended to optimize care escalation criteria for continuous monitoring of vital signs in a ward setting and to evaluate the effects on patient outcomes.

(JMIR Perioper Med 2023;6:e44483) doi:10.2196/44483

KEYWORDS
early warning scores; vital signs; telemedicine; physiological monitoring; clinical alarms; postoperative complications; perioperative nursing

Introduction

Background
Surgical patients are at risk of developing postoperative complications, which may progress to life-threatening illnesses and seriously affect patient outcomes if not promptly detected and correctly treated [1]. Most postoperative complications occur in the first week after surgery and are typically present in a ward setting [2-5]. Therefore, adequate patient monitoring in surgical wards is crucial for identifying the early signs of complications [6].

In hospital wards, patient monitoring typically consists of routine vital sign checks performed by nurses every 6 to 8 hours, complemented by subjective evaluation of the patient status during nursing activities [7,8]. In addition, early warning scores (EWSs) are widely used to evaluate the risk of patient deterioration. EWSs are typically calculated by assigning points to a measured set of vital signs, where the sum of the points reflects the EWSs. EWSs are often implemented as part of rapid response systems, where they are used to trigger clinical actions or escalation of care when exceeding a predefined threshold [5,9]. However, vital sign checks and corresponding EWSs are often incomplete or not performed on time, particularly during the night or when the protocol mandates more frequent measurements [10,11]. Together with the intermittent measurement frequency, this may lead to unnoticed or delayed detection of patient deterioration.

In recent years, wireless sensors that enable mobile vital sign monitoring have been introduced. These sensors facilitate automated, less obtrusive, and continuous patient monitoring in a ward setting [12,13]. Although there is still little evidence regarding the clinical effects of continuous monitoring in this setting, various studies have suggested that continuous monitoring can aid early identification of clinical deterioration and may provide opportunities to improve outcomes in patients with complications [4,14]. However, the interpretation of the large amount of data that are generated by the sensors is still a major challenge because vital sign measurements fluctuate largely during the day and are influenced by movement and many patient-related or environmental factors [15]. Moreover, continuous manual data observation is hampered by restricted staffing levels in a ward setting and inconsistent assessment of abnormalities by caregivers [16]. Therefore, to promote an adequate and timely response to patient deterioration, automatic methods to support and identify vital sign abnormalities related to potential complications are desired.

Currently, wireless monitoring is often implemented in combination with traditional alarm strategies, where an alert is sent automatically as soon as the measurement of one of the vital signs exceeds a preset upper or lower threshold [12,17]. Although this single-parameter alarm strategy is standard in high-care units such as intensive care units, these alarms are sensitive to various disturbances and easily lead to excessive false alarm rates. As the nurse:patient ratio is lower in a ward setting, this alarm burden is a serious concern for nurse workload and could lead to alarm fatigue, thereby potentially threatening patient safety [7]. Furthermore, single-parameter abnormality detection does not align with the current use of EWSs for risk prediction in patients on surgical wards, which relies on multiple parameters.

Objective
Accordingly, it may be of interest to use EWSs instead of single-parameter alarms to detect potential abnormalities in continuously monitored patients on surgical wards, as supported by recently developed remote monitoring systems [12]. However, there is still little evidence regarding whether EWSs derived from mobile vital sign measurements can actually support the identification of deterioration in patients on surgical wards and be useful as alarm systems for continuous remote monitoring. Therefore, this exploratory study aimed to gain insight into the potential sensitivity of a sensor-based, high-rate EWS for predicting postoperative complications and to evaluate the expected daily alarm rate in comparison with single-parameter alarm criteria as well as manual EWS measurements for different alarm settings. Furthermore, the potential predictive value of “nurse worry” and patient-reported deterioration was explored to investigate the possible benefits of systematically collecting subjective data in monitoring routines.

Methods

Study Design
The study had an observational retrospective design.

Ethics Approval
This study was approved by The Medical Research Ethics Committee Twente (MoViSign study; NL65885.044.18).

https://periop.jmir.org/2023/1/e44483
Informed Consent
All included participants provided written informed consent to participate in the study and to use their data for research purposes.

Population
The study included patients (aged ≥18 years) undergoing elective esophageal or gastric resection (ie, upper gastrointestinal [GI] group) and older patients (aged ≥70 years) undergoing acute surgery for a hip fracture (ie, hip fracture group), as these groups are known to have relatively high rates of complications during ward stay. Patients were recruited preoperatively or after postoperative admission to the surgical ward (GI unit or center of geriatric traumatology). Patients were excluded if they had implanted electronic medical devices, had known allergies to materials used in wearable sensors, had suspected delirium or cognitive impairment, or were unable to decide upon study participation. Recruitment was performed during office hours only, and patients who were expected to leave the hospital within 24 hours after possible recruitment were not eligible for study participation.

Measurements
The patients received standard ward care including routine nursing observations. Nurses were instructed to perform manual measurements of vital signs and calculate a corresponding Modified Early Warning Score (MEWS) [18] at least 3 times a day. The MEWS was used to indicate the risk of patient deterioration (≤2=low risk, 3-4=intermediate risk, and ≥5=high risk), where cutoff values of 3 and 5 were used as response triggers to call a physician and the rapid response team for further patient investigation, respectively. In addition to routine care, vital signs were recorded every minute using a wearable patient monitoring system. Heart rate (HR) and respiratory rate (RR) were measured using the chest-worn LifeTouch sensor (Isansys Lifecare Ltd). Axillary temperature (AT) was recorded with the LifeTemp sensor (Isansys Lifecare Ltd), and blood oxygen saturation (SpO₂) was measured with the 3150 WristOx finger pulse oximeter (Nonin Medical Inc). All the measurements were sent to a bedside patient gateway using an encrypted Bluetooth connection. Sensor recordings were started after the patient was admitted to the surgical ward and informed consent was obtained and continued until hospital discharge (or premature patient dropout). During the recording period, the researchers checked and maintained the technical functionality of the sensor recordings at least once a day on weekdays. Medical professionals and patients were blinded to the sensor measurements to prevent bias, but nurses were instructed to temporarily detach the sensors for showering, diagnostic imaging, or surgical interventions or reinterventions.

To explore the added value of routine collection of subjective information from nurses and patients, nurses were instructed to fill in a paper checklist during every nurse shift to register the possible presence and reasons for nurse worry using an adapted version of the Dutch Early Nurse Worry Indicator Score, as specified in Multimedia Appendix 1 [19]. Furthermore, patients were asked to fill in a daily diary during their ward stay to indicate the presence of symptoms, how they were currently feeling (0 points=very poor; 5 points=very good), and how they felt compared with the previous day (0 points=much impaired; 5 points=much improved) on a 5-point Likert scale. If needed, researchers helped or encouraged the patients to fill in the diary when they visited the patients. The nurse worry checklist and patient diary were used only for research purposes. All collected measurements were deidentified for further analysis.

Events
An event was defined as a postoperative complication that was diagnosed according to local standards and that was treated during ward stay or within 7 days after hospital discharge. Events were identified retrospectively from patient records and reviewed by a surgeon (EAK or HJH) to ensure correct interpretation. As there is no way to ascertain exactly when a complication started, we recorded the start of targeted treatment as the onset time of the complication. Similarly, the end time of the complication was defined as the timing of the last therapeutic action, for example, the last medication gift. If no information regarding the last therapeutic action was available or if therapy was continued after hospital discharge (or premature patient dropout) or in case of palliative treatment, it was assumed that the complication lasted until the end of the sensor recording period. Complications classified as Clavien-Dindo class [20] of II or higher, for which treatment was started within the measurement period (ie, the period of sensor measurements during ward stay), were eligible and included for further analysis. In patients with multiple complications, eligible events were only included if treatment was started at least 24 hours after any previously included complication. Furthermore, complications were only included if at least 4 hours of HR and RR sensor measurements were available in the 24 hours window before the onset time of the event after data preprocessing (see the Data Preprocessing section). Patients with eligible complications were enrolled in the event group, whereas patients with an uncomplicated postoperative trajectory served as the control group. Patients with only ineligible complications were excluded from further analyses. To evaluate the risk of selection bias, the baseline characteristics were compared between the included and excluded patient groups.

EWS and Alarm Simulation
Overview
Sensor and nurse measurements were used to simulate corresponding EWSs. In addition, alarms were respectively simulated using the EWSs based on sensor vital signs measurements (sensor EWSs), EWSs based on manual vital signs measurements (nurse EWSs), and single vital sign recordings, with the aim of evaluating and comparing the effects of using these measurements as response triggers or active alarms. Figure 1 provides an overview of the alarm simulation and evaluation, which are further explained in the following subsections.
**Figure 1.** Overview of the calculation of the early warning score (EWS), alarm simulation, and alarm evaluation in sensor and nurse measurements.

**Sensor measurements**
- HR, RR, SpO$_2$, and AT measurements obtained by wireless sensors

**Preprocessing 1-to-240-min windows**
- Outlier removal
- Resampling to 1-, 5-, 10-, 20-, 30-, 60-, 120- and 240-min windows
- Removal of windows with <50% data availability or with <2 vital parameters

**Sensor EWS calculation**
\[
\text{Sensor EWS} = \text{HR points} + \text{RR points} + \text{SpO}_2 \text{ points} + \text{AT points}
\]

**Sensor EWS alarm criteria**
- \(\text{Sensor EWS} \geq \text{EWS threshold (set between 1-9)}\)

**Alarm classification and evaluation**
- Classification of each alarm based on the period in which the alarm is generated:
  - Sensitivity for early detection: number of complications associated with early TP alarms
  - Total sensitivity: number of complications associated with early or late TP alarms
  - Total alarm rate: average number of alarms/patient/day
  - False discovery rate: percentage of alarms classified as FP

**Nurse measurements**
- HR, RR, SpO$_2$, systolic BP, and TT measurements obtained by nurses

**Preprocessing 60-min windows**
- Outlier removal
- Resampling to 60-min window
- Removal of windows with <50% data availability

**Preprocessing**
- Removal of measurements with <2 vital parameters

**Nurse EWS calculation**
\[
\text{Nurse EWS} = \text{HR points} + \text{RR points} + \text{SpO}_2 \text{ points} + \text{systolic BP points} + \text{TT points}
\]

**Nurse EWS alarm criteria**
- \(\text{Nurse EWS} \geq \text{EWS threshold (set between 1-9)}\)

**EWS calculation criteria are described in Table 1.**
Table 1. Criteria used to calculate early warning scores (EWSs) based on sensor vital signs measurements (sensor EWSs) and early warning scores based on manual vital signs measurements (nurse EWSs).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Inclusion of vital sign</th>
<th>Points</th>
<th>Sensor EWSs</th>
<th>Nurse EWSs</th>
</tr>
</thead>
<tbody>
<tr>
<td>RR\textsuperscript{a} (brpm\textsuperscript{b})</td>
<td>✓</td>
<td>≤8</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>HR\textsuperscript{d} (brpm\textsuperscript{e})</td>
<td>✓</td>
<td>≤40</td>
<td>4-15</td>
<td>6-10</td>
</tr>
<tr>
<td>SpO\textsubscript{2}\textsuperscript{f} (%)</td>
<td>✓ ✓</td>
<td>≤91</td>
<td>80-90</td>
<td>75-85</td>
</tr>
<tr>
<td>Systolic BP\textsuperscript{h} (mm Hg)</td>
<td>✓</td>
<td>≤70</td>
<td>70-80</td>
<td>60-70</td>
</tr>
<tr>
<td>AT\textsuperscript{i} or TT\textsuperscript{j} (°C)</td>
<td>✓ ✓</td>
<td>≤34.9</td>
<td>35-38.4</td>
<td>38.5</td>
</tr>
</tbody>
</table>

\textsuperscript{a}RR: respiratory rate.
\textsuperscript{b}brpm: breaths per minute.
\textsuperscript{c}Variables included in the sensor EWSs and nurse EWSs.
\textsuperscript{d}HR: heart rate.
\textsuperscript{e}brpm: beats per minute.
\textsuperscript{f}SpO\textsubscript{2}: blood oxygen saturation.
\textsuperscript{g}No standard protocol; therefore, only included if available.
\textsuperscript{h}BP: blood pressure.
\textsuperscript{i}AT: axillary temperature.
\textsuperscript{j}TT: tympanic temperature.
\textsuperscript{k}AT used for sensor EWSs and TT used for nurse EWSs.

Data Preprocessing

Data preprocessing and analysis were performed using MATLAB (version R2021b; The Math Works Inc). Sensor recordings were preprocessed by removing implausible extreme values (HR >200 or <30 beats/min [brpm], RR <5 or >50 breaths/min [brpm], AT <30 °C or >50 °C, and SpO\textsubscript{2} <70% or >100%), most likely caused by artifacts. As no standard monitoring frequency has currently been established for monitoring patients on surgical wards, sensor measurements were analyzed repeatedly for different time windows. Accordingly, the minute-sampled vital sign recordings were resampled by averaging the signal values in successive windows of 5, 10, 20, 30, 60, 120, and 240 minutes. For each vital parameter, windows wherein data were missing for >50% of the time were disregarded.

EWS Simulation

The original and down-sampled sensor measurements as well as all vital sign measurements obtained by nurses during the sensor recording period were used for the retrospective simulation of the corresponding EWSs. For this purpose, points were assigned to each vital parameter according to the MEWS criteria [18] (Table 1). As SpO\textsubscript{2} is not included in the MEWS, the SpO\textsubscript{2} criteria were obtained from the UK National Early Warning Score criteria [21].

The sensor EWSs were calculated as the sum of the points assigned to each of the vital parameters measured by the sensor system, resulting in sensor EWSs between 0 and 11. Similarly, the vital sign measurements obtained by nurses were used to calculate the nurse EWSs, ranging between 0 and 14. The sensor EWSs and nurse EWSs were only calculated for measurements in which at least 2 vital sign values were available.

Alarm Simulation

To evaluate and compare the effect of using the sensor or nurse measurements as an active notification system, “alarms” were simulated based on the sensor EWS measurements (“sensor EWS–based alarms”) and based on all available nurse EWS measurements (“nurse EWS–based alarms”). An alarm was defined as an occurrence where the EWS increased to a higher level compared with a previous measurement, thereby exceeding a preset threshold (“EWS threshold”). When the measurements exceeded the EWS threshold from the beginning of the measurement period, the first sample was considered as the first alarm. EWS alarms were simulated repeatedly for EWS thresholds of 1 to 9, aiming to evaluate the alarms for a wide range of thresholds. However, as MEWS cutoff values of 3 and 5 were used in this study’s hospital as response triggers, these thresholds were selected as EWS threshold in the primary analysis. For the sensor EWSs, the primary investigation was based on 1-hour time windows, but alarms were also simulated for all other time windows (1-240 min).

In addition to the alarms based on EWSs, sensor measurements were used to simulate single-parameter alarms, similar to traditional physiological alarm systems. A single-parameter alarm was simulated when the vital sign measurement exceeded the predefined normal range. Upper and lower thresholds were defined by the outer limits of the EWS criteria, that is, HR ≥40 or ≤130 bpm, RR ≤8 or ≥30 brpm, SpO\textsubscript{2} ≤91%, and AT ≤34.9 °C or ≥38.5 °C. Single-parameter alarms were simulated for 1-hour time windows, and situations where multiple single-parameter alarms were present in the same window were counted as 1 alarm.
**Alarm Classification**

All the simulated alarms were classified based on the timing of the alarms, as illustrated in Figure 1. Alarms that were simulated ≤24 hours before the onset of the included events were classified as “early true-positive” (TP) alarms, thereby reflecting alarms that could theoretically promote early identification of events. Alarms that presented during the treatment period of the included events were classified as “late true-positive” alarms. Alarms that presented >24 hours before onset or after the end of the treatment period of included events or that presented in the control group were classified as false-positive (FP) alarms, that is, false alarms.

**Analysis**

**Availability and Agreement of Vital Signs Measurements**

The availability of sensor measurements, that is, data completeness, was calculated as the percentage of the total recording time where vital sign measurements were available after preprocessing in the total study population. The availability of nurse measurements was verified by the number of measurements available per 24 hours. The agreement between nurse and sensor measurements was explored for each vital parameter using Bland-Altman analysis for repeated measures [22] by comparing each available nurse measurement with the corresponding average value of the available preprocessed sensor measurements in the 5-minute window [23] before the nurse measurement.

**Group Comparison**

All the measurements (sensor recordings, sensor EWSs, nurse EWSs, nurse worry checklist, and patient diary) were compared between the event and control groups to explore their relationships with the development of complications. For all available hourly sensor EWS and nurse EWS measurements obtained during the sensor recording period, the average EWSs and the percentage of measurements where the EWSs were ≥3 (intermediate to high risk) and ≥5 (high risk) was investigated for each group. For the sensor EWSs, the number of EWS points assigned to the hourly sensor measurement windows was evaluated to investigate the contribution of the different vital parameters to the sensor EWSs. Furthermore, the percentage of hourly sensor measurements, that is, single-parameter recordings, that exceeded the upper or lower thresholds defined for the single-parameter alarms were assessed. Finally, the prevalence of worry expressed by nurses and the prevalence of (very) poor current status and (severely) impaired status indicated by patients during the study period were evaluated to explore the potential diagnostic value of this subjective information. Group differences were statistically compared using the Mann-Whitney U test. In addition, a sensitivity analysis was performed to verify the impact of patient exclusion on group differences by separately evaluating group differences after adding excluded patients to the control group.

**Alarm Evaluation**

The performance of the sensor EWS–based, nurse EWS–based, and single-parameter alarms was investigated using 2 sensitivity rates: the total alarm rate (TAR) and the false discovery rate (FDR) [24]. The sensitivity for early detection of adverse events was calculated as the percentage of events associated with early TP alarms. Similarly, the total sensitivity for event detection was defined as the percentage of events for which early TP and late TP alarms were observed. The TAR was defined as the average daily number of alarms per patient and the FDR as the percentage of alarms classified as FP. The sensitivity for early detection, total sensitivity, TAR, and FDR were compared between the sensor EWS–based, nurse EWS–based, and single-parameter alarms.

**Results**

**Population**

A total of 60 patients were included in the study, of whom 33 (55%) were patients undergoing elective esophageal or gastric resection and 27 (45%) were older patients with a hip fracture. The baseline characteristics of the participants are presented in Table 2. On average, sensor recordings were started 48 hours after the onset of surgery. Out of 60 patients, the period of vital sign recording was completed by 46 (77%) patients, whereas measurements were stopped before hospital discharge in 14 (23%) patients due to patient refusal. No temperature sensor was placed in 2 patients, and pulse oximeter recordings were not started in 2 other patients due to unavailable sensors or patient refusal.
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Upper GI group (n=33)</th>
<th>Hip fracture group (n=27)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>64 (11)</td>
<td>82 (7)</td>
</tr>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>24 (73)</td>
<td>4 (15)</td>
</tr>
<tr>
<td>Female</td>
<td>9 (27)</td>
<td>23 (85)</td>
</tr>
<tr>
<td><strong>American Society of Anesthesiologists classification, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>20 (61)</td>
<td>12 (44)</td>
</tr>
<tr>
<td>III</td>
<td>12 (36)</td>
<td>11 (41)</td>
</tr>
<tr>
<td>IV</td>
<td>1 (3)</td>
<td>3 (11)</td>
</tr>
<tr>
<td>Unknown</td>
<td>0 (0)</td>
<td>1 (4)</td>
</tr>
<tr>
<td><strong>Comorbidities, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiovascular</td>
<td>19 (58)</td>
<td>19 (70)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>3 (9)</td>
<td>7 (26)</td>
</tr>
<tr>
<td>GI</td>
<td>12 (36)</td>
<td>8 (30)</td>
</tr>
<tr>
<td>Neuropsychiatric</td>
<td>6 (18)</td>
<td>10 (37)</td>
</tr>
<tr>
<td>Pulmonary</td>
<td>6 (18)</td>
<td>8 (30)</td>
</tr>
<tr>
<td>Urogenital</td>
<td>9 (27)</td>
<td>11 (41)</td>
</tr>
<tr>
<td>Other categories(^b)</td>
<td>6 (18)</td>
<td>11 (41)</td>
</tr>
<tr>
<td>ICU(^c) readmission, n (%)</td>
<td>3 (9)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>In-hospital mortality, n (%)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Out-of-hospital mortality(^d), n (%)</td>
<td>0 (0)</td>
<td>1 (4)</td>
</tr>
<tr>
<td>Hospital readmission(^d), n (%)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

\(^a\) GI: gastrointestinal.

\(^b\) Includes endocrine, infective, neuromuscular, or thrombosis-related comorbidities.

\(^c\) ICU: intensive care unit.

\(^d\) Within 7 days of hospital discharge.

**Events**

During the measurement period or the 7-day follow-up period, a total of 40 complications were observed in 28 patients. Of all observed complications (n=40), 25 (63%) events were excluded (Clavien-Dindo class I: 10/25, 40%; within 24 hours from the previous event: 2/25, 8%; no or insufficient sensor data availability: 13/25, 52%), resulting in the exclusion of 14 patients from the analysis. Out of 40 events, 15 (38%) were included in the analysis (Clavien-Dindo class II: 10/15, 67%; III: 2/15, 13%; IV: 1/15, 7%; V: 1/15, 7%). The included events were present in a group of 14 patients (upper GI group: n=9, 64%; hip fracture group: n=5, 36%), who were selected as the event group. The control group consisted of 32 patients (upper GI group: n=17, 53%; hip fracture group, n=15, 47%) in whom no events were observed. We found no statistical differences in the baseline characteristics between the event and control groups or between the included and excluded patient groups.

**Availability and Agreement of Vital Signs Measurements**

Sensor recordings in the total study population had a median duration of 120 (IQR 93-163) hours. After preprocessing, the median data availability of HR and RR was 84% (IQR 74%-94%) of the measurement time and that of AT was 97% (IQR 84%-100%). SpO\(_2\) values were missing more frequently, with a median availability of 46% (IQR 40%-60%). The median number of vital sign observations registered by nurses was 3 (IQR 2-4) per day. On average, HR was available in 96%, RR in 48%, SpO\(_2\) in 92%, systolic blood pressure in 95%, and tympanic temperature in 93% of nurse observations.

Bland-Altman plots for all available nurse and sensor data pairs are presented in Multimedia Appendix 2. The sensor measurements of HR and RR were higher than the nurse measurements, with a mean difference of 3 bpm and 9.4 brpm, respectively, where the largest differences between measurements were particularly seen for higher measurement values. In contrast, the SpO\(_2\) (mean absolute difference 2.1%) and temperature measurements (mean difference 0.9 °C)
recorded by the sensor were lower than the nurse measurements, where the largest deviations were observed in the lower ranges. For all parameters, the limits of agreement were relatively wide compared with the observed range of values.

**Group Comparison**

Figure 2 shows the hourly sensor EWSs over time for the event and control groups, visualized using EWS thresholds of 3 and 5. High sensor EWSs were often scattered over time, and no uniform sensor EWS pattern was observed in the days or hours preceding the complication treatment. According to the average number of points that were assigned to the hourly vital sign measurements (Figure 3), RR and SpO\textsubscript{2} received 2 or 3 points in at least half of the available measurements for both the event and control groups, thereby contributing most to the sensor EWSs.

Table 3 describes the differences in the measurements between the event and control groups. The average hourly sensor EWSs of the included patients with complications were significantly higher \((P=.004)\) than those of the patients with uncomplicated postoperative trajectories. In addition, the sensor EWSs in the event group reached scores of \(\geq 5\) more often compared with the control group \((P<.001)\). For nurse EWSs, the number of measurements, as well as the average nurse EWSs and the percentage of observations where sensor EWSs of \(\geq 3\) or \(\geq 5\), were significantly higher for the event group \((P=.02, P=.009, P<.001, \text{and } P<.001, \text{respectively})\). Furthermore, nurses expressed worry more often for the patients in the event group \((P=.02)\). Indicators for nurse worry that were reported in both groups are specified in Multimedia Appendix 1. In the event group, most reported indicators of worry included “Change in breathing,” “Subjective nurse observation,” and “Patient indicates.” The patient diary results did not differ significantly between the groups \((P=.35 \text{ and } P=.37)\). However, patients in the event group seemed to report a (very) poor and a (much) impaired status more often, although these patients filled in the daily diary less frequently. According to the sensitivity analysis, adding the excluded patient group to the control group did not change any of the results, as all \(P\) values remained within the same levels of significance (ie, \(P<.001\), \(P<.05\), or \(P>.05\)).

In the 24 hours before the onset of the included events, the average sensor EWSs had a median value of 3.4 (IQR 2.5-4.0), which was significantly different from the median value of 2.0 (IQR 1.1-2.7) of the average nurse EWSs in this period \((P=.006)\) and from the median value of 2.8 (IQR 2.4-3.2) of the average sensor EWSs observed across the total measurement period in the control group \((P=.04)\).

Figure 4 shows a case example of a patient undergoing esophagectomy diagnosed on postoperative day 8 with anastomotic leakage (Clavien-Dindo class IV) using a computed tomography scan. Treatment included the placement of an esophageal stent on day 9 (defined as event onset) and antibiotic therapy, followed by intensive care unit readmission on day 11 due to progressive hemodynamic instability. The sensor recordings contained missing data periods for all the vital parameters. Furthermore, the sensor measurements differed from the nurse measurements, particularly in terms of RR and temperature. The sensor and nurse measurements showed a similar gradually increasing pattern in HR in the days before event diagnosis and treatment. However, the sensor measurements indicated an increasing, although extreme, trend in RR before diagnosis, whereas nurse RR measurements increased only after the stent placement. Correspondingly, the sensor EWSs reached values \(\geq 5\) more frequently from day 7, whereas the nurse EWSs reached values of \(\geq 5\) only on day 10. In addition to abnormal vital sign measurements, nurse worry as well as impaired or poor self-reported patient status were observed not only in the 24-hour period before stent placement but also earlier in the postoperative trajectory.
Figure 2. Early warning score based on sensor vital signs measurements (sensor EWS) of all patients enrolled in the event and control group, calculated using 1-hour windows and presented according to cutoff values of ≥3 (orange) and ≥5 (red). Complication timing (onset of treatment) is annotated by white diamond markers.

Figure 3. Number of points that were assigned to the hourly sensor vital sign measurements in patients with and without postoperative complications following the criteria used to calculate early warning scores (specified in Table 1). The bar stacks reflect the average percentage of 1-hour windows in which 0, 1, 2, or 3 points were assigned to the corresponding vital parameter. AT: axillary temperature; HR: heart rate; RR: respiratory rate; SpO\textsubscript{2}: blood oxygen saturation.
<table>
<thead>
<tr>
<th>Result</th>
<th>Event group (n=14), median (IQR)</th>
<th>Control group (n=32), mean (IQR)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sensor measurements</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measurement period(^a) (h)</td>
<td>186 (92-260)</td>
<td>119 (106-142)</td>
<td>.10</td>
</tr>
<tr>
<td><strong>Hourly sensor EWSs(^b)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensor EWS availability (% of the measurement period)</td>
<td>82 (66-98)</td>
<td>85 (68-93)</td>
<td>.94</td>
</tr>
<tr>
<td>Average sensor EWSs</td>
<td>3.5 (3.1-4.1)</td>
<td>2.8 (2.4-3.2)</td>
<td>.004</td>
</tr>
<tr>
<td>Sensor EWSs ≥3 (% of available hourly sensor measurements)</td>
<td>62 (44-83)</td>
<td>49 (41-66)</td>
<td>.08</td>
</tr>
<tr>
<td>Sensor EWSs ≥5 (% of available hourly sensor measurements)</td>
<td>25 (18-40)</td>
<td>8.1 (2.4-17)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Nurse EWSs(^c)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nurse EWSs availability (number of measurements/24 h)</td>
<td>3.4 (2.3-4.2)</td>
<td>2.7 (1.9-3.1)</td>
<td>.02</td>
</tr>
<tr>
<td>Average nurse EWSs</td>
<td>1.7 (0.9-2.3)</td>
<td>0.9 (0.4-1.3)</td>
<td>.009</td>
</tr>
<tr>
<td>Nurse EWSs ≥3 (% of available measurements)</td>
<td>22 (7.7-34)</td>
<td>0 (0-12)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Nurse EWSs ≥5 (% of available measurements)</td>
<td>4.3 (0-6.9)</td>
<td>0 (0-0)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Hourly single-parameter recordings</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 parameter out of normal range(^d) (% measurement period)</td>
<td>26 (19-41)</td>
<td>18 (13-25)</td>
<td>.03</td>
</tr>
<tr>
<td>Multiple parameters out of normal range(^d) (% measurement period)</td>
<td>3.0 (0.0-13)</td>
<td>0 (0-0.8)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Nurse worry checklist</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nurse worry checklist availability (number of checklists filled/24 h)</td>
<td>1.2 (0.8-1.9)</td>
<td>1.2 (0.9-1.6)</td>
<td>.93</td>
</tr>
<tr>
<td>Nurse worry expressed (% of available nurse worry checklists)</td>
<td>25 (0-33)</td>
<td>0 (0-17)</td>
<td>.02</td>
</tr>
<tr>
<td><strong>Patient diary</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient diary availability (number of diary forms filled/24 h)</td>
<td>0.5 (0-1)</td>
<td>0.8 (0-1)</td>
<td>.37</td>
</tr>
<tr>
<td>Patient expressed (very) poor and (much) impaired status (% of available diary forms)</td>
<td>14 (0-36)</td>
<td>0 (0-33)</td>
<td>.35</td>
</tr>
</tbody>
</table>

\(^a\)Period of sensor measurements performed during ward stay.

\(^b\)Sensor EWS: early warning score based on sensor vital sign measurements.

\(^c\)Nurse EWS: early warning score based on manual vital sign measurements.

\(^d\)Out-of-normal range conforms to single-parameter alarm criteria.
Figure 4. Case example of a patient with anastomotic leakage. AT: axillary temperature; bpm: beats per minute; brpm: breaths per minute; CT: computed tomography; EWS: early warning score; HR: heart rate; nurse EWS: EWS based on manual vital signs measurements; RR: respiratory rate; SpO2: blood oxygen saturation; sensor EWS: EWS based on sensor vital signs measurements; TT: tympanic temperature. The sensor vital sign measurements and sensor EWS measurements are calculated using 1-hour windows.

Alarm Evaluation

Sensor EWS-Based Alarms

Figure 5 displays the performance metrics of sensor EWS alarms simulated in 1-hour windows using EWS thresholds of 1 to 9. The sensitivity rates were the highest for an EWS threshold of 3, where alarms were observed 24 hours before the onset of included events in 12 out of 15 events (sensitivity for early detection=80%) and during the treatment period for all events (total sensitivity=100%). TAR followed a similar pattern as the sensitivity rates and reached the highest level for a threshold of 3, resulting in a maximum of 2 alarms per patient per day, of which 85% was classified as FP alarms. An EWS threshold of 5 was associated with a TAR of 1.2 alarms per patient per day, FDR of 83%, sensitivity for early detection of 67%, and total sensitivity of 93%.

Figure 6 shows the performance of sensor EWS alarms for window lengths of 1 to 240 minutes, illustrated for an EWS threshold of 5 as an example. The sensitivity rates and FDR decreased for longer windows. TAR strongly declined with increasing window length, ranging from 31 alarms per patient per day for window lengths of 1 minute to 0.8 alarms per patient per day for window lengths of 120 minutes. Similar trends were observed for all other thresholds, although the sensitivity decreased more rapidly with increasing window length for the other thresholds. In addition, FDR increased with the window length for thresholds of <3. The highest TAR levels were observed for a threshold of 3, reaching up to 70 alarms per patient per day for a 1-minute window length.
**Nurse EWS-Based Alarms**

Figure 7 displays the performance metrics of nurse EWS–based alarms. The sensitivity for early detection and total sensitivity were the same for most threshold values, reaching a maximum of 40%. The TAR and FDR were the highest (0.6 alarms/patient/day and 80%, respectively) for a threshold of 1 but decreased for higher threshold values. An EWS threshold of 5 was associated with a sensitivity for early detection and total sensitivity of 27%, TAR of 0.1 alarms per patient per day, and FDR of 54%. Compared with the sensor EWSs, not only the sensitivity rates but also the TAR and FDR of the nurse EWSs were lower for all thresholds >1.
Figure 7. Sensitivity, total alarm rate (TAR), and alarm classification of early warning score based on manual vital signs measurements (nurse EWS)–based alarms, calculated using early warning score (EWS) thresholds of 1 to 9. FP: false-positive; TP: true-positive.

Single-Parameter Alarms

Figure 8 shows the TAR and classification of the single-parameter alarms. In total, the single-parameter alarms resulted in a TAR of 2 alarms per patient per day. The FDR of all single-parameter alarms was 83.6%, and 4.7% of the alarms were classified as early TP and 11.7% as late TP. In 13 of the 15 events, at least 1 single-parameter alarm was observed in the 24-hour period before the onset of the event, resulting in a sensitivity for early detection of 87%. Total sensitivity was 100%. Most alarms (>0.5 alarms/patient/d) were observed for low \( \text{SpO}_2 \), low AT, and high RR, of which a large part (>80%) was classified as FP alarms. In contrast, high HR and high AT alarms were associated with lower alarm rates and lower FDR (≤30%). No single-parameter alarms were observed for low HR or low RR.

Discussion

Principal Findings

This study investigated the potential usability of high-rate EWSs obtained from wireless vital sign data for early identification of complications in patients on surgical wards and explored the performance of a sensor EWS–based alarm system in comparison with single-parameter alarms and manual EWS measurements by nurses for various alarm settings. The EWSs based on hourly sensor recordings were significantly higher in patients with complications than in patients with an uncomplicated postoperative trajectory. Furthermore, the sensitivity to predict adverse events within 24 hours was higher for sensor EWS–based alarms compared with nurse EWS–based alarms. Therefore, EWSs obtained from sensor measurements might contribute to early awareness or confirmation of patient deterioration in current ward routines. However, high EWSs...
were often recurrent in patients with and without events and could therefore lead to high false alarm rates when used as real-time alarm systems. Although hourly EWS measurements resulted in fewer alarms compared with single-parameter alarms, the number of false EWS alarms increased when using shorter time windows and varied between alarm thresholds, highlighting the importance of careful selection of alarm settings.

**Comparison With Prior Work**

**Prediction of Complications**

Traditionally, vital signs are used to monitor patients after surgery, as vital signs change 4 to 24 hours before adverse events [25]. The additional use of EWSs supports systematic assessment and responses to patient deterioration by nurses [5,26]. Coherently, it can be expected that frequent monitoring of EWSs can contribute to improved and early recognition of complications. This expectation is supported by the current case example and by other studies where abnormal vital signs and corresponding (partial) EWSs were observed more evidently, more often, or earlier in sensor recordings than in nurse measurements [23,27]. Similarly, sensor EWS–based alarms and single-parameter alarms based on hourly sensor measurements provide a higher sensitivity for predicting events within 24 hours compared with nurse EWS–based alarms. RR contributed the most to high sensor EWSs, which agrees with previous studies reporting RR as an important predictor of deterioration [5,7,28]. Furthermore, SpO₂ abnormalities were often observed, which is in line with the high prevalence of hypoxemic (micro) events found in similar observational studies [27,29].

Compared with hourly sensor EWS–based alarms, we found higher sensitivity rates for single-parameter alarms, although the differences were small for some EWS thresholds. This observation is in line with a study by Rothman et al [30] who reported that only 32% of life-threatening adverse events were preceded by abnormalities in multiple components of the EWS, whereas 27% showed only a single abnormal criterion. In contrast, Rothman et al [30] also reported that 41% of the adverse events were not preceded by abnormal vital signs at all. This raises the question of the relatively high sensitivity of both the sensor EWS–based and single-parameter alarms. Accordingly, it is plausible that abnormal vital sign measurements were not caused by the complication development but were related to normal variations within patients or the postoperative status, as supported by the observation that high sensor EWSs were also present in the control group. Similar findings were reported by Itelman et al [31], who reported that national EWSs based on 5-minute wireless vital sign recordings obtained in patients on general wards provided warnings for deterioration events in 67% of the cases on average 29 hours before detection by caregivers but also led to warnings in 78% of all patients who did not experience deterioration. Similarly, Haahr - Raunkjæer et al [29] reported that episodes of abnormal vital sign measurements were observed more often during the 24-hour period before events in patients admitted to the ward; however, the duration of abnormalities in continuous recordings did not differ between patients with and without serious adverse events.

Another possible explanation for the low specificity is that sensor-based respiration rates were typically much higher than nurse-based measurements. Whether this reflects the inaccuracy of nurse-based RR measurements [10] or RR overestimation by the sensor remains to be determined. In this context, one should be aware that the overestimation of RR falsely increases EWSs, which reduces the discriminative ability of systems. Together, these findings highlight the risk of event overdetection and highlight the need for further optimization of the combined warning criteria for wireless monitoring.

**Alarms**

The use of automatic alarm systems in monitoring routines can contribute to the early awareness and active responses of care professionals to potential deterioration [32]. However, excessive false alarm rates can lead to alarm fatigue, posing a serious risk to patient safety and a barrier to the adoption of continuous monitoring [33]. Therefore, a balanced ratio between alarm sensitivity, specificity, and alarm rate is crucial, which requires careful selection of alarm thresholds. In current practice, a variety of EWS thresholds are used, depending on the EWS variant and clinical settings. The calculation of the sensor EWSs was based on the MEWS criteria, for which studies recommended thresholds of 3 or 4 for monitoring patients on wards [34,35]. However, the results of this study indicate that the sensor EWSs resulted in the highest false alarm rates for EWS thresholds of 3 and 4. In addition, we found that sensor EWSs ranged between 2 and 4 and rarely dropped to values <2, which also explains the low alarm rates for thresholds of <3. Therefore, using a threshold of 5 may be considered to reduce the alarm burden in sensor-based alarm systems.

Alarm rates based on the hourly sensor EWSs were lower than those based on the hourly single-parameter alarms, which underlines the benefits of a multiple-parameter approach to improving specificity. However, it should be noted that the EWS alarm rates increased rapidly when using shorter time windows. Therefore, it might be appropriate to restrict automatic alarm generation based on EWSs to intervals of 1 hour or even longer in a ward setting and additionally present historical data trends for a more detailed evaluation of patient deterioration during nurse rounds. The use of summary measures for continuously measured vital signs is supported by van Goor et al [36], who reported that mean values calculated over multiple hour time frames can be helpful in supporting the identification of respiratory insufficiency. Although assessing data over longer time windows instead of every few minutes will limit alarm sensitivity, this approach would still be an improvement compared with intermittent nurse observations by allowing more frequent and less-static EWS measurements. Moreover, such an approach might be well accepted, as it protects nurses from alarm overload and promotes the use of the system as a support system without compromising the nurses’ leading role in the decision-making process. For safe and effective implementation, it is crucial that caregivers are trained in interpreting the data and alarms and that high-risk patients requiring continuous surveillance are admitted to a high-care unit with a nurse:patient ratio that allows immediate response in case of acute deterioration [7,37].
Sensor Versus Manual Vital Signs Measurements

The use of wireless sensors provides new opportunities for monitoring vital signs in ward settings. By enabling continuous recording, wireless monitoring supports the comprehensive investigation of data abnormalities and trends, reducing the likelihood that signs of deterioration are missed due to intermittent or incomplete nurse measurements [10,27]. In addition, reducing the time needed for routine measurements can reduce nurses’ workloads, leaving more time for patient observation and support. However, remote vital sign monitoring technology is still immature, and standards for the implementation of remote patient monitoring in ward routines are lacking. Furthermore, sensor measurements during patient mobilization still present practical challenges; for example, wireless connection issues and limited patient adherence [27,33] can lead to data loss. In our study, we encountered most measurement issues for SpO₂, where we found a relatively high rate of missing data related to detachment of the finger probe by patients, loss of Bluetooth connection, and short battery life, as also reported in another study using the same device [29]. Technological improvements can be achieved by developing new sensors that can derive accurate SpO₂ values from locations on the body that are less prone to (movement) artifacts and sensor dislodgment. Furthermore, adequate methods for correcting missing data periods in postanalysis and real-time alarm systems are desired.

Second, wearable sensor measurements are prone to artifacts caused by motion and poor skin contact, and clinical validation studies have shown variable accuracy of remote systems [38-40]. These sensor inaccuracies could explain the observed low agreement between the sensor and nurse data pairs, where the average bias was particularly large for RR and SpO₂ as also observed in other studies [23]. However, it must be noted that nurse observations also have variable accuracy, especially for RR [10]. In addition, discrepancies between the sensor and nurse measurements can be expected when measurements are performed using different sensing techniques, sensor locations, and under different circumstances; for example, in this study, sensor measurements recorded AT instead of tympanic temperature and were performed during daily activities, where vital sign measurements can be influenced by physical activity, whereas nurse measurements were performed during rest. Together, these factors underline that sensor measurements and corresponding EWSs are not a one-on-one replacement of measurements performed by nurses, where each approach has its benefits and limitations. These differences highlight the need for appropriate response trigger criteria for continuous monitoring, perhaps requiring different classifications of EWS points or the implementation of alternative detection algorithms.

Finally, when seeking continuous risk prediction methods or fully automatic EWS assessments, it should be considered that currently available wireless monitoring systems are restricted to measuring only a selection of vital parameters. For example, the level of consciousness is implemented in standard EWS or patient monitoring routines [9] but cannot be measured automatically using current sensor systems. Until reliable sensing and preprocessing techniques are available for all relevant parameters, it would be of interest to develop ways to optimally combine sensor and nurse measurements in care routines. Various commercially available systems for wireless or automatic patient monitoring already support or are used for the calculation of complete EWS every few hours [12,41], often after verification of valid measurement values and manual insertion of parameters that cannot yet be measured by the system, such as blood pressure or consciousness. Following this approach, Weenk et al [23] performed a study that included patients on wards, where an EWS was calculated every 30 minutes using 2 different wireless vital sign sensor systems and complementary nurse measurements. Higher EWSs were seen earlier in combined recordings than in nurse measurements alone, suggesting that integration of manual and automatic monitoring could contribute to better continuous patient monitoring and improve early identification of deterioration.

Nurse Worry and Patient-Reported Indicators

Subjective evaluation of patient status is an essential element of patient monitoring and clinical decision-making [7,37]. The presence of nurse worry has been described as a predictor of deterioration and is increasingly used as an individual calling criterion or as part of EWS variations [19,42]. The importance of nurse worry was confirmed by this study, where we found that a nurse worry checklist could support the identification of patients with events and may reveal early signs of deterioration, as illustrated by the case example. Accordingly, wireless monitoring should be accompanied by routine nurse evaluations, which could mitigate the risk of overreliance on technology [7]. To ensure consistent collection and valuation of subjective information that cannot be retrieved with sensors, it can be beneficial to embed structured bedside observation scores such as the Dutch Early Nurse Worry Indicator Score [19] in electronic patient files and warning criteria.

In addition to nurse worry, well-being indicators reported by patients or by their relatives provide important signs of deterioration [8], and the presence of patient concerns has been suggested as an additional care escalation criterion [43]. Correspondingly, we found that patient-reported deterioration was often reported as a reason for nurse worry in patients in the event group. However, signs of deterioration are not always recognized or expressed by patients or their relatives [43]. In this study, exploring a patient diary that aimed for the independent and structured assessment of patient indicators, we noticed that patients with complications seemed to report a poor or impaired status more frequently. Despite regular encouragement from researchers to fill in the diary, the responsiveness to it was limited. Interestingly, patients with events seemed less responsive to the diary compared with the control group, which might be a sign of deterioration by itself. Further investigation of patient-reported indicators for use in monitoring routines, perhaps using patient-tailored questionnaires or 2-way communication systems that promote patient response, is recommended.

Strengths and Limitations

This exploratory study included 2 surgical populations with a relatively high risk of developing postoperative complications and, therefore, may benefit from improved monitoring. The
heterogeneity of our study population and the variety of included events allowed the exploration of the predictive value of continuous EWSs for different algorithm settings. However, the total number of included patients and events eligible for analysis were limited, and it should be noted that clinical deterioration may present differently across other settings. Therefore, validation of this study’s results in larger cohorts and other specific or general populations is needed once remote monitoring systems are mature and implemented more widely. Similarly, verification of the performance of high-rate EWSs for monitoring systems using other sensors or different parameter sets is desired. Finally, prospective studies assessing the clinical effects of using high-rate EWSs are needed.

The criteria used for the simulation of the sensor EWSs were based on clinical standards to allow comparison with current practice; therefore, the optimization of alarm criteria was limited to exploring different cutoff thresholds and sample frequencies. Furthermore, the EWSs were obtained using average vital sign values, whereas the use of other summary measures can result in different EWSs [44]. Because the results of this study revealed that sensor EWS–based alarms could lead to overdetection and high false alarm rates, further investigation of the best-suited preprocessing methods as well as alternative alarm criteria or methods is highly recommended; for example, it might be beneficial to personalize the allocation of EWS points by adapting normal ranges based on previous recordings of individuals or patient groups [24] or to correct for the influence of physical activity using embedded motion sensors. Moreover, because patterns of deterioration depend on the underlying cause [2] and patient characteristics, it could be useful to implement algorithms tailored to detect specific complications or patient groups [45]. Finally, trend-based metrics, machine learning techniques, or dynamic models may provide better performance for identifying abnormal patterns in continuous vital sign data compared with traditional EWSs, and event prediction may be improved further by integrating additional patient information extracted from the electronic patient record [16,28,45-51].

Although many studies have focused on the prediction of life-threatening events such as cardiac arrest, this study focused on all minor and major complications that required additional care in a ward setting. Patients with Clavien-Dindo class I complications that, by definition, did not require pharmacological or surgical treatment were excluded from the analysis. The sensitivity analysis indicated that the inclusion of these patients in the control group did not affect the results, indicating that the predictive value of vital signs or EWSs is limited for these events. Following the concept that early warning systems should alert only when action is required, this study used the onset of therapeutic actions as a surrogate marker of complication onset. However, the timing of therapeutic onset depends highly on the diagnostic definitions and is prone to delays in diagnostic or therapeutic activities. Furthermore, delays in clinical reporting hamper accurate retrospective determination of therapy onset. Therefore, prospective clinical trials are needed to investigate the effects of using a high-rate sensor-based EWSs as a diagnostic tool or alarm system on time to treatment and the corresponding patient outcomes.

Conclusions

Hourly EWSs obtained using wireless vital sign sensors resulted in higher sensitivity for predicting postoperative complications compared with the currently used 8-hour, nurse-based EWSs. Therefore, sensor EWSs may contribute to the improved detection of clinical deterioration in patients on surgical wards. In addition, using the EWS as an active alarm trigger results in fewer alarms compared with single-parameter alarms in sensor measurements. However, false alarm rates can increase rapidly when calculating EWSs in short time spans, risking alarm fatigue and overdetection. To prevent alarm overload in a ward setting, we suggest restricting EWS-based alarm generation to windows of at least 1 hour or longer. In addition, investigation of alternative care escalation criteria for wireless monitoring is warranted because automatic vital sign recordings differ from nurse measurements and are currently only available for a selected set of vital parameters. Finally, it is recommended to embed a structured collection of subjective nurse and patient information into monitoring routines, as this might provide complementary signs of deterioration. Future clinical studies are needed to evaluate the clinical effects of using the EWSs and the corresponding alarm system for continuous monitoring of patients on wards.

Acknowledgments

The authors thank all the patients who participated in the observational study and are grateful to the nurses and hospital staff who supported the study and data collection.

This study was supported by the Pioneers in Health Care Innovation fund (PIHC 2017).

Data Availability

The deidentified vital sign measurements that were analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Nurse worry indicators.

References


Abbreviations

AT: axillary temperature
bpm: beats per minute
brpm: breaths per minute
EWS: early warning score
FDR: false discovery rate
FP: false-positive
GI: gastrointestinal
HR: heart rate
MEWS: Modified Early Warning Score
nurse EWS: early warning score based on manual vital signs measurements
RR: respiratory rate
sensor EWS: early warning score based on sensor vital signs measurements
SpO2: blood oxygen saturation
TAR: total alarm rate
TP: true-positive
A Real-Time Mobile Intervention to Reduce Sedentary Behavior Before and After Cancer Surgery: Pilot Randomized Controlled Trial

Carissa A Low¹, PhD; Michaela Danko¹, BSN; Krina C Durica¹, MA; Julio Vega¹, PhD; Meng Li¹, MS; Abhineeth Reddy Kunta¹, MS; Raghu Mulukutla², MS; Yiyi Ren², MS; Susan M Sereika¹, PhD; David L Bartlett³, MD; Dana H Bovbjerg¹, PhD; Anind K Dey², PhD; John M Jakicic⁴, PhD

¹University of Pittsburgh, Pittsburgh, PA, United States
²University of Washington, Seattle, WA, United States
³Allegheny Health Network Cancer Institute, Pittsburgh, PA, United States
⁴University of Kansas Medical Center, Kansas City, KS, United States

Corresponding Author:
Carissa A Low, PhD
University of Pittsburgh
3347 Forbes Avenue
Suite 200
Pittsburgh, PA, 15213
United States
Phone: 1 4126235973
Email: lowca@upmc.edu

Abstract

Background: Sedentary behavior (SB) is prevalent after abdominal cancer surgery, and interventions targeting perioperative SB could improve postoperative recovery and outcomes. We conducted a pilot study to evaluate the feasibility and preliminary effects of a real-time mobile intervention that detects and disrupts prolonged SB before and after cancer surgery, relative to a monitoring-only control condition.

Objective: Our aim was to evaluate the feasibility and preliminary effects of a perioperative SB intervention on objective activity behavior, patient-reported quality of life and symptoms, and 30-day readmissions.

Methods: Patients scheduled for surgery for metastatic gastrointestinal cancer (n=26) were enrolled and randomized to receive either the SB intervention or activity monitoring only. Both groups used a Fitbit smartwatch and companion smartphone app to rate daily symptoms and collect continuous objective activity behavior data starting from at least 10 days before surgery through 30 days post discharge. Participants in the intervention group also received prompts to walk after any SB bout that exceeded a prespecified threshold, with less frequent prompts on days that patients reported more severe symptoms. Participants completed end-of-study ratings of acceptability, and we also examined adherence to assessments and to walking prompts. In addition, we examined effects of the intervention on objective SB and step counts, patient-reported quality of life and depressive and physical symptoms, as well as readmissions.

Results: Accrual (74%), retention (88%), and acceptability ratings (mean overall satisfaction 88.5/100, SD 9.1) were relatively high. However, adherence to assessments and engagement with the SB intervention decreased significantly after surgery and did not recover to preoperative levels after postoperative discharge. All participants exhibited significant increases in SB and symptoms and decreases in steps and quality of life after surgery, and participants randomized to the SB intervention unexpectedly had longer maximum SB bouts relative to the control group. No significant benefits of the intervention with regard to activity, quality of life, symptoms, or readmission were observed.

Conclusions: Perioperative patients with metastatic gastrointestinal cancer were interested in a real-time SB intervention and rated the intervention as highly acceptable, but engagement with the intervention and with daily symptom and activity monitoring decreased significantly after surgery. There were no significant effects of the intervention on step counts, patient-reported quality of life or symptoms, and postoperative readmissions, and there was an apparent adverse effect on maximum SB. Results highlight...
the need for additional work to modify the intervention to make reducing SB and engaging with mobile health technology after abdominal cancer surgery more feasible and beneficial.

**Trial Registration:** ClinicalTrials.gov NCT03211806; https://tinyurl.com/3apwkk

*(JMIR Perioper Med 2023;6:e41425)* doi:10.2196/41425

**KEYWORDS**
sedentary behavior; mobile health; smartphone; mobile phone; wearable device; surgical oncology; physical activity; perioperative cancer patients; surgical recovery; abdominal cancer surgery; perioperative intervention; activity monitoring

**Introduction**

Surgical treatment is a critical component of curative therapy for most cancers, but risks for postoperative complications, unplanned readmissions, and persistent functional impairments are common, especially for abdominal cancers, where rates of complications and readmissions can range from 25%-50% [1-3]. These high rates of adverse postoperative outcomes place patients at risk for functional limitations and impaired quality of life as well as high health care costs and utilization. Supportive interventions aimed at optimizing perioperative health and functioning are needed for this high-risk surgical oncology population.

Physical activity is one modifiable behavior that holds promise for affecting postoperative recovery and outcomes [4-6]. Indeed, prehabilitation programs that promote physical activity before surgery have been linked to improved preoperative functional capacity [7] and shorter length of stay after cancer surgery [8]. Similarly, early mobilization after surgery, generally defined as out of bed activity by perioperative day one, is recommended as part of Enhanced Recovery after Surgery pathways, although evidence of benefit is mixed [9]. Because both prehabilitation and Enhanced Recovery after Surgery often include nutritional interventions and other components, it is difficult to determine whether and to what extent increased physical activity alone can reduce postoperative risks. Moreover, postoperative symptoms such as pain and fatigue make increasing physical activity after cancer surgery challenging and may compromise adherence to exercise interventions [10,11]. In the perioperative surgical oncology context, interventions aimed at disrupting prolonged sedentary behavior (SB) with brief walking breaks may be more attainable than more structured exercise interventions, especially if the intervention can adapt to changing symptom burden over the perioperative course. To date, no studies have tested the impact of perioperative SB disruption on surgical oncology outcomes [12].

The goal of this study was to pilot-test a personalized mobile technology–supported intervention to reduce SB before and after gastrointestinal cancer surgery. This intervention uses a smartwatch and smartphone to collect daily symptom ratings that are used to tailor the frequency of prompts to disrupt prolonged SB in real time, which we hypothesized would increase the feasibility of the intervention. We previously described the development and usability of this intervention in a single-arm pilot trial [13]. In this pilot randomized controlled trial, we compared patients randomized to receive the SB intervention to those whose activity and symptoms were monitored only. The primary outcome of this initial pilot trial was feasibility, defined as accrual and retention, end-of-study acceptability ratings, and adherence to intervention assessments and activity prompts. Secondary outcomes included objective activity and SB, patient-reported quality of life and symptoms, as well as postoperative readmissions.

**Methods**

**Participants**

Participants were recruited from the outpatient clinics of 6 surgeons specializing in abdominal cancer surgery at a National Cancer Institute–designated comprehensive cancer center. Participants were recruited between June 2019 and March 2021 at their preoperative surgical oncology clinic visit. Study accrual was paused from March to December of 2020 due to the COVID-19 pandemic. Research staff provided study information and email reminders to the 6 surgical oncology care teams and asked the nurse or physician assistant to identify potential patients at the time of their consent to surgery, confirm their eligibility, and to either consent them directly or connect them with the research team for consent and onboarding. The study was open to English-speaking adults scheduled for surgical treatment of metastatic gastrointestinal or peritoneal cancer and able to stand and walk unassisted. Exclusion criteria included having less than 10 days to scheduled surgery date, to provide adequate time for participants to become familiar with study technology and activity prompts prior to surgery. No participants had sensory or motor impairments that interfered with use of the study apps.

**Study Procedures**

Following completion of written informed consent, participants were randomized via random number generator to either the SB intervention (which included activity monitoring) or activity monitoring only. They were provided with a Fitbit Versa smartwatch (first generation) paired with a Google Pixel 2 smartphone on which Detecting Activity to Support Healing (DASH) study apps (Intervention or Monitoring-only) as well as the Fitbit app had been installed. From the time of consent to 30 days after hospital discharge following their surgery, participants were asked to keep the devices charged, to wear the smartwatch as much as possible, to rate their daily experience of symptom severity once each morning, and for intervention participants only, to respond to activity prompts. Participants completed a questionnaire at study entry to collect information about demographic variables, health behaviors, and experience with mobile technology. Before surgery, during inpatient recovery, and approximately 30 days after postoperative discharge, participants also completed...
standardized measures of depressive and physical symptoms and quality of life. At the end of the study, all participants completed a semistructured interview about their experiences with the devices and a questionnaire about the acceptability and usability of the apps.

As previously described [13], all participants used the DASH Android smartphone study app to rate the daily severity of 10 symptoms (ie, pain, fatigue or tiredness, sleep disturbance, trouble concentrating or remembering things, feeling sad or down, feeling anxious or worried, shortness of breath, numbness or tingling, nausea, and diarrhea or constipation), using a scale from 0 (ie, symptom not present) to 10 (ie, symptom as bad as you can imagine). Participants were randomized to either the DASH intervention or monitoring-only control condition. Participants randomized to the DASH intervention received a Fitbit smartwatch app that used the most recent symptom rating to set a threshold for SB bouts and used real-time step count data to trigger activity prompt notifications when prespecified SB thresholds were exceeded (60 consecutive minutes of SB when all symptoms were rated less than 7 out of 10 or 120 consecutive minutes of SB if any symptom was rated as 7 or higher). For the purposes of this study, SB was operationalized as a minute with fewer than 10 steps logged by the Fitbit, to allow for incidental stepping and arm movements that might be misclassified as steps, while also classifying very slow walking as activity, given the perioperative context and likely diminished gait cadence during early postoperative recovery [14]. When SB thresholds were exceeded, an activity prompt (“Ready for a short walk?”) was sent to the smartwatch. If 30 or more steps were logged within 15 minutes of an activity prompt, participants received a positive feedback message (“Great job being active!”). Prompts were sent only between each participant’s waking time and bedtime, which were set by participants in the Android app and could be adjusted during the study. Participants randomized to monitoring-only received a Fitbit smartwatch app that measured steps but did not send activity prompts.

**Ethics Approval**

All procedures were approved by the University of Pittsburgh Institutional Review Board (STUDY 19030389) and registered on ClinicalTrials.gov (NCT03211806).

**Measures**

Primary outcome measures assessing feasibility were (1) accrual and retention rates (ie, percentage of participants approached who enrolled in the research and percentage of participants enrolled who completed the study); (2) acceptability, based on end-of-study responses to the System Usability Scale [15] and the following questions: “On a scale of 0-100, how easy was it to use the smartphone/watch?” “On a scale of 0-100, how pleasant was the smartphone/watch interface (appearance, design, usability?)” and “On a scale of 0-100, how satisfied were you with the overall system (including the smartphone and watch and all notifications?)”; as well as (3) adherence (percentage of days symptom ratings were completed and at least 8 hours of Fitbit data were logged, and for intervention participants only, percentage of activity prompts after which steps were detected).

Secondary outcome measures included (1) objective SB (maximum and mean SB bout duration per day based on Fitbit minutes with less than 10 steps logged); (2) objective physical activity (Fitbit step count per day); (3) patient-reported symptoms (depressive symptoms via Center for Epidemiological Studies-Depression [16]) and physical symptoms via questions adapted from the MD Anderson Symptom Inventory and based on the National Cancer Institute’s Symptom Management and Health-Related Quality of Life Steering Committee recommendations [17,18]; (4) patient-reported quality of life via the Functional Assessment of Cancer Therapy [19]; and (5) readmissions within 30 days after index hospital discharge, extracted from electronic medical records.

**Analytic Approach**

Group differences in baseline participant characteristics and end-of-study acceptance and usability ratings were examined using independent sample two-tailed $t$ tests and chi-squared tests. Linear mixed modeling assuming the best-fitting variance-covariance structure for the repeated assessments was used to explore the effect of the intervention over the 3 study time points (ie, preoperative, inpatient, and after discharge) for the outcomes of adherence, SB, physical activity, psychological and physical symptoms, and quality of life. The models included a fixed, between-subjects effect for randomized group assignment as well as a fixed, within-subject effect for time and group interaction by time. In addition, to test statistics ($F$ test values) and corresponding $P$ values from the model, least square means with standard errors are presented. Two outcomes, average SB and steps, were square-root transformed due to positively skewed residual distributions when modeling the outcome in its original metric. Data for Fitbit step counts and SB bout duration were only included from days that the Fitbit was worn at least 8 hours, and sleep episodes as identified by the Fitbit were excluded from SB bouts.

**Results**

**Participant Characteristics**

As shown in Table 1, the sample was primarily White and predominantly male, with most patients undergoing cytoreductive surgery with hyperthermic intraperitoneal chemotherapy. Participants randomized to the intervention arm had significantly higher BMI and were less likely to be a former smoker compared to those randomized to the control arm. Participants started using the DASH apps a mean of 19.6 (range 8-47) days prior to surgery, throughout their inpatient stay as feasible (which lasted an average of 10.9, range 5-24 days), and for 30 days post discharge, for an average of 57.2 total days (range 44-92 days) of study participation.
Table 1. Participant characteristics.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>All participants (n=26)</th>
<th>Intervention arm (n=13)</th>
<th>Monitoring-only arm (n=13)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>56.2 (10.5)</td>
<td>54.9 (6.5)</td>
<td>57.5 (13.5)</td>
</tr>
<tr>
<td>Sex, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>11 (42.3)</td>
<td>7 (53.8)</td>
<td>4 (30.8)</td>
</tr>
<tr>
<td>Male</td>
<td>15 (57.7)</td>
<td>6 (46.2)</td>
<td>9 (69.2)</td>
</tr>
<tr>
<td>Race, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>24 (92.3)</td>
<td>13 (100)</td>
<td>11 (84.6)</td>
</tr>
<tr>
<td>Black</td>
<td>1 (3.8)</td>
<td>0 (0)</td>
<td>1 (7.7)</td>
</tr>
<tr>
<td>More than one</td>
<td>1 (3.8)</td>
<td>0 (0)</td>
<td>1 (7.7)</td>
</tr>
<tr>
<td>BMI, mean (SD)</td>
<td>27.4 (5.3)</td>
<td>30.2 (5.8)</td>
<td>24.8 (3.3)</td>
</tr>
<tr>
<td>Smoking history, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current smoker</td>
<td>1 (4)</td>
<td>1 (8.3)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Former smoker</td>
<td>5 (20)</td>
<td>0 (0)</td>
<td>5 (38.5)</td>
</tr>
<tr>
<td>Never a smoker</td>
<td>19 (76)</td>
<td>11 (91.7)</td>
<td>8 (61.5)</td>
</tr>
<tr>
<td>Exercise frequency, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seldom or never</td>
<td>6 (24)</td>
<td>4 (33.3)</td>
<td>2 (15.4)</td>
</tr>
<tr>
<td>1-2 times per week</td>
<td>7 (28)</td>
<td>5 (41.7)</td>
<td>2 (15.4)</td>
</tr>
<tr>
<td>3-4 times per week</td>
<td>10 (40)</td>
<td>3 (25)</td>
<td>7 (53.8)</td>
</tr>
<tr>
<td>&gt;5 times per week</td>
<td>2 (8)</td>
<td>0 (0)</td>
<td>2 (15.4)</td>
</tr>
<tr>
<td>Has Wi-Fi at home, n (%)</td>
<td>22 (88)</td>
<td>11 (91.7)</td>
<td>11 (84.6)</td>
</tr>
<tr>
<td>Owns a smartphone, n (%)</td>
<td>26 (100)</td>
<td>12 (100)</td>
<td>13 (100)</td>
</tr>
<tr>
<td>Owns an activity tracker, n (%)</td>
<td>4 (16)</td>
<td>2 (16.7)</td>
<td>2 (15.4)</td>
</tr>
<tr>
<td>CS+HIPEC surgery, n (%)</td>
<td>17 (65.4)</td>
<td>10 (76.9)</td>
<td>7 (53.8)</td>
</tr>
</tbody>
</table>

*CS+HIPEC: cytoreductive surgery with hyperthermic intraperitoneal chemotherapy.

Primary Outcomes

Accrual and Retention

Of the 35 eligible patients approached, 26 consented to the study (74% accrual rate). Reasons for not participating were “too busy/overwhelmed” (n=2), “not good with technology” (n=3), “already wear a smartwatch/activity monitor and did not want to wear two” (n=2), and “had to leave clinic so did not have time to discuss the study” (n=2). The retention rate for the study was 88%, with 3 participants withdrawing (2 participants before starting to use the devices—one in intervention and one in monitoring-only condition—and 1 participant in the intervention condition 18 days after surgery due to poor health and readmission).

Acceptance

A total of 20 participants completed the end-of-study interview, and those in both the intervention and monitoring-only conditions rated the phone and watch interfaces as pleasant and easy to use and the overall system as satisfactory and usable (Table 2).

Table 2. Mean (SD) participant ratings of interface and system usability.

<table>
<thead>
<tr>
<th>Variable (range 0-100)</th>
<th>All (n=20)</th>
<th>Intervention arm (n=9)</th>
<th>Monitoring-only arm (n=11)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone—ease of use</td>
<td>93.1 (7.2)</td>
<td>91.1 (8.3)</td>
<td>94.6 (6.1)</td>
<td>.29</td>
</tr>
<tr>
<td>Watch—ease of use</td>
<td>91.6 (12.6)</td>
<td>94.4 (5.1)</td>
<td>89.3 (16.3)</td>
<td>.38</td>
</tr>
<tr>
<td>Phone—pleasantness</td>
<td>87.7 (13.7)</td>
<td>90.0 (13.2)</td>
<td>85.8 (14.5)</td>
<td>.51</td>
</tr>
<tr>
<td>Watch—pleasantness</td>
<td>87.9 (14.9)</td>
<td>91.7 (7.9)</td>
<td>84.8 (18.7)</td>
<td>.32</td>
</tr>
<tr>
<td>Overall satisfaction</td>
<td>88.5 (9.1)</td>
<td>91.3 (5.5)</td>
<td>86.2 (11.0)</td>
<td>.22</td>
</tr>
<tr>
<td>System Usability Scale</td>
<td>85.1 (11.5)</td>
<td>83.6 (9.2)</td>
<td>86.4 (13.5)</td>
<td>.61</td>
</tr>
</tbody>
</table>

When asked what they thought of the study, participants in the intervention condition reported that the activity prompts were motivating, especially prior to surgery. For example, one participant (P13) noted that “it got me moving more than I
normally would have.” However, participants also noted that the prompts were not as motivating or as easy to respond to after surgery, especially in the hospital. One participant (P6) said the following:

At the beginning, I thought it was awesome and was very excited about the step counting and found it motivational; after surgery, I had a lot of trouble getting interest back, and the watch wasn’t enough to be motivating; I lost interest because I had other priorities health-wise.

Another participant (P7) said the following:

It’s so much easier to get up presurgery. Maybe a hierarchy of prompts tapping different motivations [would be better] because it takes so much more to get up post-surgery.

Across both conditions, participants mentioned that they enjoyed tracking their step or sleep data in the Fitbit app. One participant (P16, in the monitoring-only group) noted the following:

Part of my recovery was setting step goals for myself and increasing that goal.

Some participants, like P2 (in the monitoring-only group), also noticed associations between activity and how they felt, noting “days with higher steps always felt better symptom-wise, looking back.” Many participants felt that physical activity was beneficial for their physical and psychological recovery, as P13 (in the intervention group) said the following:

Moving and walking helped prevent scar tissue development. If I had stayed sedentary, I think I would have been in much worse shape.

Adherence

Over the course of the study, daily symptom ratings were completed on 62% (874/1416) of days, ranging from 14% (9/65) to 95% (55/58) of days across individual participants. Fitbits were worn on 77% (1091/1416) of study days, and 91% (990/1091) of these days had at least 8 hours of Fitbit data available. On average, 69% (977/1416) of days were included in Fitbit analyses; across individual participants, the percentage of days with at least 8 hours of Fitbit data ranged from 17% (9/52 days with ≥8 hours of data) to 100% (58/58 days with ≥8 hours of data). As shown in Figure 1, participants became less adherent with both symptom reporting and wearing the Fitbit after surgery (symptom reporting: $F_{\text{time}}=22.9; P<.001$; and Fitbit: $F_{\text{time}}=9.2; P=.001$), but there were no significant differences in adherence between the two study groups (symptom reporting: $F_{\text{group}}=0.0; P=.95$; $F_{\text{group}\times\text{time}}=0.3; P=.78$; and Fitbit: $F_{\text{group}}=0.2; P=.663$; $F_{\text{group}\times\text{time}}=0.4; P=.68$).

For participants in the intervention group, an average of 5.8 activity prompts were sent per day, and participants took steps and received positive feedback after 22% (418/1925) of activity prompts. This varied substantially from before surgery (mean 3.3, SD 1.8 prompts per day; 200/407, 49% of prompts resulted in walking) to after surgery in the hospital (mean 7.8, SD 2.6 prompts per day; 29/462, 6% of prompts resulted in walking) to postdischarge recovery (mean 6.2, SD 2.6 prompts per day; 189/1056, 18% of prompts resulted in walking).

Secondary Outcomes

**Fitbit-Measured Sedentary Behavior Bouts and Steps**

On average, participants logged 3642 (SD 3365) steps per day with a mean SB bout duration of 61 (SD 80) minutes and a maximum SB bout duration of 248 (SD 155) minutes. For all participants, step counts decreased significantly, and mean and maximum SB bout duration increased significantly from before surgery to during inpatient recovery (Figure 2; step count: $F_{\text{time}}=60.5; P<.001$; maximum SB bout: $F_{\text{time}}=10.1; P<.001$; and mean SB bout: $F_{\text{time}}=28.1; P<.001$). The intervention had no significant effect on step count ($F_{\text{group}}=2.3; P=0.15$; and $F_{\text{group}\times\text{time}}=1.2; P=.32$) or mean SB bout duration ($F_{\text{group}}=1.5; P=.24$; and $F_{\text{group}\times\text{time}}=0.5; P=.24$). Unexpectedly, participants randomized to the intervention had longer maximum SB bouts overall ($F_{\text{group}}=6.16; P=.02$; and $F_{\text{group}\times\text{time}}=1.48; P=.24$). One
important limitation to note is that these mean step count and SB bout values are based on the subset of participants who were compliant with wearing the smartwatch. Although there were no significant group differences in Fitbit compliance, intervention participants tended to wear the watch for fewer hours per day, and some mentioned removing the watch when they knew they would not be able to get up and walk or when trying to nap or rest. Because we included all days with at least 8 hours of total but not necessarily consecutive wear time, we may have misclassified some episodes during which participants were not wearing the watch as sedentary bouts. When hours of wear time per day was included in models, the group effect on maximum SB bout duration was no longer significant (data not shown).

Figure 2. (A) Fitbit daily mean step count, (B) maximum sedentary bout duration, and (C) mean sedentary bout duration.

**Patient-Reported Measures**

Similar to the other outcomes, we observed a significant time effect for quality of life ($F_{time}=21.4; P<.001$; Figure 3), depressive symptoms ($F_{time}=10.9; P<.001$), and physical symptoms ($F_{time}=24.0; P<.001$), but no significant group (quality of life: $F_{group}=0.3; P=.60$; depressive symptoms: $F_{group}=1.6; P=.22$; and physical symptoms: $F_{group}=0.1; P=.76$) or group $\times$ time effect (quality of life: $F_{group\times time}=0.0; P=.97$; depressive symptoms: $F_{group\times time}=0.3; P=.78$; and physical symptoms: $F_{group\times time}=0.9; P=.41$). All participants regardless of condition reported worsening quality of life and symptoms after surgery. In total, 5 of 12 participants who started in the intervention condition were readmitted within 30 days, compared to 4 of 12 participants who started in the monitoring-only conditions, and there was no significant group difference in readmission rate ($\chi^2_1=0.2; P=.67$).

Figure 3. (A) Patient-reported quality of life, (B) depressive symptoms, and (C) physical symptoms. CES-D: Center for Epidemiological Studies-Depression; FACT: Functional Assessment of Cancer Therapy; MDASI: MD Anderson Symptom Inventory.
**Discussion**

**Principal Findings**

In this paper, we described results from a pilot randomized trial testing a SB intervention in patients undergoing surgery for metastatic gastrointestinal cancer. To our knowledge, this is the first SB intervention developed specifically for patients undergoing cancer surgery at high risk for adverse postoperative outcomes [12]. Although patients were willing to participate and remain in the trial and rated the intervention as highly acceptable, engagement with the intervention and with daily symptom and activity monitoring decreased significantly after surgery. There were no significant effects of the intervention on step counts, patient-reported quality of life or symptoms, and postoperative readmissions. Contrary to hypotheses, participants randomized to the intervention group exhibited longer maximum SB bouts relative to the monitoring-only condition.

The SB intervention tested in this trial was designed to make replacing prolonged SB with short walking breaks more feasible in the perioperative context by reducing the frequency of SB prompts on days that patients reported high symptom burden. Given the significant drops in adherence and engagement that occurred after surgery and particularly during inpatient recovery as well as the fact that participants may have been less likely to complete symptom ratings on days they felt particularly unwell, additional modifications to the intervention are needed to make postoperative activity more feasible. A number of participants in the intervention condition noted that it was very difficult to get out of bed to walk when prompted, especially without assistance; this challenge and the associated frustration could have led to decreased self-efficacy to adhere to the intervention that carried into the postdischarge period, leading to increased SB and decreased adherence. Pausing the intervention until patients were recovering at home, adapting the intervention to involve caregivers or hospital staff and timing prompts around their availability to assist with postoperative ambulation, or replacing walking with light stretches or activities that could be done in bed while seated could all be options for future interventions targeting activity among postoperative inpatients. Adherence to assessments also decreased in the monitoring-only group after surgery, suggesting that either reduced frequency of assessments or additional support and reminders may be needed to make collection of continuous activity and daily symptom ratings feasible postoperatively.

The lack of observed benefits with regard to SB and activity limitations after surgery, could result in higher adherence over time if participants perceive the SB intervention to be meaningfully increasing activity and to have potential health benefits. In the future, providing education about the risks of perioperative SB, personalized goal setting to inform prompts, and coaching and problem-solving to overcome barriers to disrupting SB could be considered as additional interventional components to enhance a perioperative SB intervention [12,20]. Involving patients in the co-design of perioperative SB interventions could also result in enhanced engagement and benefit [21].

**Comparison to Prior Work**

Although adherence to reporting symptoms and wearing the Fitbit declined significantly after surgery, rates were consistent with other work on wearables [22] and symptom reporting [23] during cancer treatment and with an earlier study of activity monitoring after cancer surgery [24]. As in our earlier field trial [13], adherence also varied substantially between participants, ranging from approximately 15% to 100% for both symptom reporting and Fitbit wearing throughout the perioperative period. Another approach to consider in future work is a more stepped-care approach, with more frequent contact or high-touch support for patients with poor adherence. Of note, because we used the same wearable device to both deliver the intervention and measure objective activity, poor adherence and engagement resulted in less accurate assessments of activity and SB, which may also have affected results, particularly if patients in the intervention group began wearing the device for fewer hours each day after surgery due to inability to respond to the walking prompts or to minimize disruptions caused by the prompts.

Although the intervention yielded no significant benefits for patients, this study highlights the continued need for interventions to improve postoperative recovery following surgery for metastatic abdominal cancer. Consistent with other studies, nearly 40% of patients in our sample experienced an unplanned hospital readmission, and participants remained significantly less active 30 days after postoperative discharge, relative to their preoperative activity levels. In addition to improved interventions targeting SB and activity, interventions aimed at remotely monitoring and addressing worsening pain or other symptoms as well as other causes of readmission (eg, dehydration) hold promise for their ability to support this high-risk population.

**Strengths and Limitations**

Strengths of this study include the randomized design and the use of real-time symptom ratings and step data to trigger personalized just-in-time activity prompts. Focusing on disrupting SB rather than increasing physical activity is novel in the context of cancer surgery. Starting the intervention prior to scheduled surgery allowed participants to become familiar with the devices and begin increasing activity prior to surgery and hospitalization, while there may also be clinical value in continuing an intervention shortly after surgery when SB is very prevalent. The use of off-the-shelf consumer devices is also highly scalable.
This study had a number of important limitations. First, the sample size was smaller than originally intended due to COVID-19 pandemic–related disruptions to accrual and is an important limitation of this study. Second, developing a system capable of remotely detecting real-time step counts using a consumer wearable device proved challenging and required us to provide study Android phones to participants to use for collecting symptom ratings and synchronizing the Fitbit smartwatch. All enrolled participants already owned a personal smartphone; requiring them to also carry and charge an additional study smartphone across perioperative transitions of care may have contributed to adherence challenges; future interventions in this area should be deployed on participants’ existing phones to potentially improve feasibility. Third, we elected to use a monitoring-only control so that the only difference between the two study groups was the activity prompts in recognition of the fact that merely using an activity monitor can promote physical activity among cancer patients [25]; alternative control conditions could have yielded different results. Finally, all participants were undergoing surgery for metastatic peritoneal or gastrointestinal cancer, and results may not generalize to other perioperative groups or contexts.

Future Directions
As described above, additional intervention refinement and testing is needed to make real-time SB disruption more feasible and engaging for an abdominal cancer surgery population, particularly during the postoperative period. Given the small sample in this work, larger trials of activity modification are necessary once the intervention has been improved to be more feasible and potentially more robust. There have been significant advances in consumer wearable technology since the DASH apps were developed in 2018, and future work should consider interventions that leverage Apple HealthKit or GoogleFit and work across different activity monitoring devices.

Conclusions
In conclusion, although patients undergoing abdominal cancer surgery were interested in a real-time SB intervention and rated the intervention as highly acceptable, adherence and engagement decreased significantly after surgery, and there were no observed benefits of the intervention on objective activity, quality of life, symptoms, or readmissions. Further research may be needed to understand factors that influence SB following surgery and to make reducing SB and engaging with mobile health technology after surgery more feasible and beneficial for these high-risk patients.

Acknowledgments
The authors gratefully acknowledge the assistance of Michelle Liebdzinski and Heather Jones in recruiting patients as well as the patient participants who made this research possible. This research was funded by the National Cancer Institute (K07CA204380 and P30CA047904).

Data Availability
Data generated and analyzed during this study are not publicly available to protect patient health information but are available as deidentified data from the corresponding author on reasonable request.

Conflicts of Interest
JMJ is part of the Scientific Advisory Board for Wondr Health, Inc, is a Consultant for Educational Initiatives, Inc, and is the Principal Investigator on a research grant awarded to the University of Kansas Medical Center by Epitomee Medical, Inc.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.2).
[PDF File (Adobe PDF File), 95 KB - periop_v6i1e41425_app1.pdf ]

References


Abbreviations

DASH: Detecting Activity to Support Healing
SB: sedentary behavior
A Mobile App for Postoperative Pain Management Among Older Veterans Undergoing Total Knee Arthroplasty: Mixed Methods Feasibility and Acceptability Pilot Study

Jessica Kelley Morgan¹, PhD; Caitlin R Rawlins², RN, MSN; Steven K Walther¹, MBA; Andrew Harvey¹, BS; Annmarie O'Donnell¹, BS; Marla Greene², RN, BSN; Troy G Schmidt², MD

¹Continuous Precision Medicine, Research Triangle Park, NC, United States
²Western North Carolina Veterans Affairs Health Care System, Asheville, NC, United States

Corresponding Author:
Jessica Kelley Morgan, PhD
Continuous Precision Medicine
800 Park Offices Drive
Suite 3718
Research Triangle Park, NC, 27709-3547
United States
Phone: 1 9196185405
Email: jmorgan@cpmed.io

Abstract

Background: Prescription opioid misuse risk is disproportionate among veterans; military veterans wounded in combat misuse prescription opioids at an even higher rate (46.2%). Opioid misuse is costly in terms of morbidity, mortality, and humanitarian and economic burden and costs the Civilian Health and Medical Program of the Department of Veterans Affairs more than US $1.13 billion annually. Preventing opioid misuse at the time of prescription is a critical component in the response to the opioid crisis. The CPMRx mobile app has been shown to decrease the odds of opioid misuse during the postoperative period.

Objective: The overarching purpose of this feasibility pilot study was to explore whether deploying a mobile app (CPMRx) to track postoperative pain and medication use is feasible in a Department of Veterans Affairs medical center. In support of this goal, we had four complementary specific aims: (1) determine the technological and logistical feasibility of the mobile app, (2) assess the acceptability of the mobile app to participants, (3) measure demand for and engagement with the mobile app, and (4) explore the potential use of the mobile app to patients and providers.

Methods: Participants (N=10) were veterans undergoing total knee arthroplasty within the Veterans Health Administration provided with the CPMRx app to self-manage their pain during their 7-day at-home recovery following surgery. CPMRx uses scientifically validated tools to help clinicians understand how a patient can use the least amount of medication while getting the most benefit. The suite of software includes a mobile app for patients that includes a behavioral health intervention and a clinical decision support tool for health care providers that provides feedback about pain and medication use trends. Patients filled out paper questionnaires regarding acceptability at their postoperative follow-up appointment.

Results: Overall, quantitative measures of acceptability were high. The average rating for the amount of time required to use the app was 4.9 of 5 (5=“very little”), and the average rating for ease of use was 4.4 of 5 (5=“very easy”). Open-ended questions also revealed that most participants found ease of use to be high. Demand and engagement were high as well with a mean number of mobile app entries of 34.1 (SD 20.1) during the postoperative period. There were no reported technological or logistical issues with the mobile app. Participants took an average of 25.13 (SD 14.37) opioid tablets to manage their postoperative pain.

Conclusions: Results of this study revealed that the use of a mobile app for pain and medication management during postoperative recovery was both feasible and acceptable in older veterans undergoing total knee arthroplasty within the Veterans Health Administration. The wide variation in opioid consumption across participants revealed the potential use of the mobile app to provide actionable insights to clinicians if adopted more widely.

(JMIR Perioper Med 2023;6:e50116) doi:10.2196/50116
Introduction

Background

The opioid crisis has resulted in a significant increase in opioid misuse and opioid use disorder; as a result, opioid-related overdose deaths have climbed to staggeringly high levels. Opioid misuse, use disorders, and overdoses are costly in terms of morbidity, mortality, and humanitarian and economic burden. It is estimated that in 2019 alone, more than 9.7 million adults living in the United States (3.5%) misused prescription opioids [1], and drug overdose deaths in the United States rose 29.4% in 2020 to an estimated 93,331, including 69,710 (74.7%) involving opioids [2-3]. Prescription opioid misuse risk is disproportionate among veterans; military veterans wounded in combat misuse prescription opioids at an even higher rate (46.2%) [4].

The cost of opioid overdose, abuse, and dependence in the United States is estimated to be US $78.5 billion annually [5]. If reduced quality of life from opioid use disorder and the value of life lost due to fatal opioid overdose are included, the estimate increases to more than US $1 trillion [6]. Opioid misuse costs the Civilian Health and Medical Program of the Department of Veterans Affairs more than US $1.13 billion annually [6]. Adjusted annual health care costs for diagnosed opioid misuse patients are higher than those for patients without diagnosed misuse, and the prevalence of diagnosed opioid misuse is almost 7 times higher for those in the Veterans Health Administration than in commercial health plans, translating to a significant economic burden for this population [7].

One known risk factor for opioid misuse and dependence is opioid prescription following surgical procedures, both minor and major [8]. Extant data suggest that nonmedical prescription opioid use is a strong risk factor for heroin use initiation among both civilian [9] and veteran populations [10]. For the patient being prescribed opioids, there is a need to ensure appropriate use to prevent habit-forming behaviors during postoperative recovery, as it has been estimated that 12.5% of people who are prescribed opioids misuse them [11] and, as noted earlier, this is even more pronounced among some veterans (46.2%) [4]. For members of the patient’s family and others in the community, it is important that excess prescription opioids are not made available for misuse; indeed, about half (50.8%) of people who misused prescription pain relievers reported obtaining them from a friend or relative, either for free, by purchasing them, or by taking them without asking, and an additional 35.7% got them through a prescription by a single doctor [1]. Several recent studies have also found wide variation in postoperative prescribing practices and noted systemic overprescription [12-14], which may inadvertently promote continued opioid use after their indication is no longer warranted. Taken together, these findings underscore the criticality of managing the use of opioids during the postoperative period in combating the opioid crisis and highlight this time as an essential point of intervention.

Mobile Health

Over 97% of Americans own cell phones, 85% of which are smartphones. Smartphone ownership is even more common among younger adults; 96% of persons aged between 18 and 29 years, 95% of persons aged between 30 and 49 years, and 83% of persons aged between 50 and 64 years own smartphones [15]. As a result, technology-based health interventions are becoming increasingly common. Among older adults, however, smartphone ownership is less common; among adults over the age of 65 years, smartphone ownership drops to 61% [15]. Mobile health (mHealth) includes everything from the use of mobile technology, such as smartphones, to software, including mobile apps, to facilitate or enhance health care [16]. Mobile technologies are used in multiple ways, including the use of smart apps, web-based software, and SMS text messaging [17]. Smart apps are being used to augment treatment for substance abuse disorders, promote self-management of chronic conditions, increase engagement with health research, assess or measure symptoms, and foster adherence to both treatments and appointments [18-21]. Evidence suggests that symptom reporting with smart apps is well tolerated by patients and has better validity and reliability than hard copies that often rely on recall [22]. One pilot study found that 78% of patients with chronic pain who downloaded an app to track their pain symptom ratings used it with an average of 16.4 daily assessments submitted by patients within the first month. Patients reported that the app was easy to use and were willing to continue to use the app even after the study was complete [22]. Additionally, electronic monitoring (EM) of prescription medication use is a budding technology with a variety of applications and can be integrated into most drug packaging, including pill dispensers and blister packs, to help measure adherence [23]. Prior research has shown that, among veterans receiving outpatient care for posttraumatic stress disorder, age significantly predicted ownership of mHealth devices but not use or interest in mHealth apps among device owners [16]. Therefore, although older veterans may be less likely to own a personal smart device, access should be adequate, and interest in using mHealth apps is high [16].

CPMRx App

Continuous Precision Medicine has developed a suite of software, CPMRx, that uses scientifically validated tools to help clinicians understand how a patient can use the least amount of medication while getting the most benefit. The suite of software includes a mobile app for patients that includes a behavioral health intervention and a clinical decision support tool for health care providers that provides feedback about pain and medication use trends.

We have used a public health research pipeline approach for designing, evaluating, and implementing behavioral health interventions, and the CPMRx mobile app has been shown to be feasible and acceptable for supporting postoperative pain management across several patient populations and surgery types, including young adults undergoing third molar extraction.
(dental clinic, Womack Army Medical Center) [24,25], adults undergoing tonsillectomy (ear, nose, and throat, WakeMed Hospital) [26], and women undergoing cesarean section (obstetrics and gynecology, Temple University Health) [27,28]. The effectiveness of the mobile app and clinical decision support tool has also been validated. The implementation of CPMRx provides actionable insights that can be used to establish more precise prescribing guidelines. At Womack Army Medical Center, for example, clinicians were able to reduce the overprescription of an opioid by 10,000 pills annually for a single surgery type (ie, third molar extraction) [24]. In a recently conducted randomized controlled trial at Temple University (N=100), the CPMRx app was shown to reduce the odds of prescription opioid misuse during the postoperative period by 92% [28].

Specific Aims
The overarching purpose of this single-arm prospective pilot study was to explore whether deploying a mobile app to track postoperative pain and medication use is feasible in a Department of Veterans Affairs Medical Center. In support of this goal, we had four complementary specific aims: (1) determine the technological and logistical feasibility of the mobile app, (2) assess the acceptability of the mobile app to participants, (3) measure demand for and engagement with the mobile app, and (4) explore the potential use of the mobile app to patients and providers.

Methods
Participants and Procedures
Participants (N=10) were veterans undergoing total knee arthroplasty at the Charles George VA Medical Center (CGVAMC). Veterans used the CPMRx app to self-manage their pain and track medication use during their 7-day at-home recovery following surgery.

A convenience sample of participants was recruited during presurgical appointments at the CGVAMC. Presurgical appointments are conducted by orthopedic nurse case managers for patients undergoing total knee arthroplasty performed by CGVAMC orthopedic surgeons. During these appointments, patients are educated about the procedure and scheduled for a physical, history, and anesthesia evaluation to ensure they are eligible for the procedure. For this study, in addition to scheduling the physical, history, and anesthesia evaluation, the orthopedic nurse case manager assisted in recruitment by briefing the patient on the research study. The orthopedic nurse coordinator made note of all patients who expressed interest in participating in the research study. During the preoperative appointment, patients provided informed consent and HIPAA (Health Insurance Portability and Accountability Act) authorization.

On the day of discharge from the postsurgical ward, patients typically receive patient education from the registered nurse assigned to them. The discharging provider ensures an order is placed for standard postoperative medications that include opioid pain medication (oxycodone 5 mg every 6 hours as needed or hydromorphone 2 mg every 6 hours as needed, 42 in total) and acetaminophen (650 mg every 6 hours as needed, 84 tablets of 325 mg in total), as well as the other medications typically prescribed following a patient’s respective procedure. For this study, all pain medications provided to study participants were placed into a smart EM blister pack to record time stamps whenever a medication was removed by the patient. All participants received discharge instructions and education as part of the standard of care and additionally received instruction on the mobile app, tablet, and EM blister packs.

Participants self-managed their pain at home and used the mobile app and EM blister packs for the first week (7 days) following discharge to record pain score and medication use, as necessary. Patients met with the orthopedic nurse case manager for a typically scheduled postoperative appointment between 7 and 10 days after surgery. At this appointment, they returned the tablet and EM blister packs and filled out an acceptability survey. All unused medications were disposed of by the study team.

Software and Hardware
CPMRx software delivers a user-friendly platform that (1) allows users to report dose-by-dose pain scores, (2) helps users consider whether a dose is needed, and (3) creates use traceability. The software allows a user to report their pain score using a modified Visual Analog Scale when a dose is taken and includes a user-directed “gamification” component. This component delivers positive reinforcement cues to the user for managing their pain within recommended treatment protocols with the goal of providing education and incentivizing patients to make smarter and more informed decisions about dose frequency and amount. The software collects and organizes data that can be accessed by clinicians to view trending analysis for pain scores, adherence to treatment plans, and time between doses.

For this study, the mobile app was installed on study-provided smart tablets that were locked into “kiosk mode” (ie, altered to only provide access to the CPMRx mobile app and necessary system functions). The EM pill blister packs used in this study were manufactured by Information Mediary Corporation. Compliance data were extracted from blister packs using a desktop radio frequency identification reader and were then transferred into an electronic record for research purposes.

Measures
Sociodemographic variables were extracted from patients’ electronic medical records. These included age (in years), sex (male or female), race (White or European American, Black or African American, Asian American, American Indian or Alaska Native, Native Hawaiian or Pacific Islander, 2 or more races, or unknown), current smoking status (smoker or nonsmoker), and history of opioid prescription (dichotomized as yes or no).

Patients self-reported pain within the mobile app on a modified Visual Analog Scale with tick marks at 0-10. The screen reads “How much pain are you in?” and there is a face that changes colors as the user chooses the pain rating on the slider. For those with dexterity issues, the tick marks can simply be tapped.
Acceptability assessments included 2 quantitative measures and 7 open-ended questions. Quantitative measures were ease of use (rated on a 5-point scale from 1="very hard" to 5="very easy") and amount of time required (rated on a 5-point scale from 1="too much" to 5="very little"). Open-ended questions sought to better understand the overall user experience and possibilities for improvement (eg, “What did you like most about your experience with the app?” and “How would you improve this app?”). Demand and engagement were measured as the actual use of the mobile app during the postoperative period, and logistical and technological feasibility were assessed by reports of issues with app use. The total number of opioids used was also measured to determine their potential use in informing clinical decision-making.

Data Analysis
All quantitative analyses were run using SAS (version 9.4; SAS Institute). Univariate statistics were used to summarize both continuous (eg, means, SDs, and medians) and categorical (eg, proportions and total numbers) variables for sample characteristics. Descriptive statistics were run to determine means and SDs of quantitative measures of acceptability.

Qualitative data were coded through thematic analysis using NVivo (version 14; Lumivero). Due to the simplicity of the text, a single researcher coded all open-ended responses and generated themes using an inductive semantic approach.

Ethical Considerations
This study was approved by the institutional review board at the CGVAMC and was conducted in the performance of a cooperative research and development agreement between Continuous Precision Medicine and the Veterans Health Administration Innovation Ecosystem. All participants provided informed consent and HIPAA authorization. All study data were deidentified. Participants were provided a US $50 gift card for their participation.

Results

Sample Characteristics
The sample for this pilot was quite homogeneous; 9 (90%) of the veterans were male, and all 10 (100%) were White nonsmokers with no history of prior opioid prescription. The mean age of the participants was 68.8 (SD 9.7) years with a range of 54 to 81 years.

Feasibility Outcomes
Overall, quantitative measures of acceptability were high. The average rating for the amount of time required to use the app was 4.88 (SD 0.35; range 4.0-5.0), and the average rating for ease of use was 4.38 (SD 1.06). One participant (age 77 years, male) rated ease of use of the mobile app as hard (2 of 5). Thematic analysis of open-ended questions also revealed ease of use as a central theme, and that most participants found ease of use to be high. For instance, in response to the question, “What did you like most about your experience with the app?” participants cited “ease of documentation,” “very straightforward,” “easy to use,” “easy to understand,” and “increases awareness of medication use and very easy to use.” Demand and engagement were high as well, with a mean number of mobile app entries of 34.1 (SD 20.1) during the postoperative period. There were no reported technological or logistical issues with the mobile app. Participants took an average of 25.13 (SD 14.37) opioid tablets to manage their postoperative pain.

Discussion
The results of this study revealed that the use of a mobile app for pain and medication management during postoperative recovery was both feasible and acceptable in older veterans undergoing total knee arthroplasty within the Veterans Health Administration. Overall, participants rated ease of use and amount of time required as highly acceptable, with no logistical or technological issues being reported. This is congruent with our prior findings across several studies, including those involving young soldiers undergoing third molar extraction [24,25], adults undergoing tonsillectomy [26], and women undergoing cesarean section [27,28]. This study was the first to include a report by a patient of difficulty using the mobile app, which does warrant discussion. As mentioned previously, a male participant aged 77 years reported that he found the mobile app hard to use and noted in his comments that it was “not intuitive.” Although there is a tutorial embedded within the app, it may be that an additional point of socialization to the software in this population may be helpful. It is our recommendation that the clinical point of contact (eg, the orthopedic nurse case manager) demonstrate how the patient should record information within the app prior to discharge, particularly if the patient reports unfamiliarity with mobile app technologies. This in-person tutorial should mitigate any acceptability concerns for the patient and would, in our estimation, require less than 2 additional minutes of the nurse’s time.

Demand for and engagement with the mobile app were high, which supports prior research among veterans suggesting an interest in adopting mHealth technologies [16]. The number of entries by participants showed that repeated inputs (at each time of medication use) were not too burdensome over the course of the 7-day prescription. In fact, more entries were logged over this 7-day postoperative period than in prior pain studies [22]. Finally, the wide variation in opioid consumption across participants revealed the potential use of the mobile app to provide actionable insights to clinicians if adopted more widely. Taken together, these results suggest that scaling use of the CPMRx mobile app in this population is both feasible and acceptable. Given the small sample size, generalizability of findings is limited. Future research should examine whether these patterns hold across surgical services within the Veterans Health Administration and for longer lengths of care as would be the case for patients dealing with chronic pain. Implementing the CPMRx suite of software may improve clinical care by reducing opioid misuse and preventing habit-forming behaviors at the individual level as well as decreasing excess prescription opioids available for diversion and thereby reducing risk at the community level.
Acknowledgments

This material is based upon work supported in part by the North Carolina Biotechnology Center and conducted in the performance of a cooperative research and development agreement between Continuous Precision Medicine, the Veterans Health Administration Innovation Ecosystem, and the Western North Carolina Veterans Affairs Health Care System. This material is the result of work supported with resources and the use of facilities at the Charles George VA Medical Center, Asheville, North Carolina. We would like to thank Matthew D Hollifield, PharmD, BCPS, clinical pharmacist, Charles George VA Medical Center, for his assistance in data collection preparation and logistics.

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Disclaimer

The contents of this publication do not represent the views of the US Department of Veterans Affairs or the US government.

Conflicts of Interest

JKM, SKW, AH, and AO are employees of Continuous Precision Medicine. To mitigate the potential for perceived conflict, all data collection was conducted by members of the Charles George VA Medical Center study team.

References


24. Rethman M, Morgan JK, Moorhead-Beardsley B, Walther SK, Arnold JP. Opioid use and misuse among active-duty soldiers following postoperative prescription: using predictors to move toward precision prescribing practices. 2022 Presented at: American Psychological Association Annual Convention; August 2022; Minneapolis, MN. URL: https://www.researchgate.net/publication/362407942_Predictors_of_Opioid_Use_and_Misuse_among_Active-Duty_Soldiers_Following_Postoperative_Prescription#:~:text=4%3a%20x%20liter_Overall%20%3bmodel%20predicting%20total%20use%20was%20significant%20(rho%3a%200.3%20significant%20predictors)


Abbreviations

CGVAMC: Charles George VA Medical Center
EM: electronic monitoring
HIPAA: Health Insurance Portability and Accountability Act
mHealth: mobile health

https://periop.jmir.org/2023/1/e50116
Original Paper

Perioperative Risk Assessment of Patients Using the MyRISK Digital Score Completed Before the Preanesthetic Consultation: Prospective Observational Study

Fabrice Ferré1*, MD, PhD; Rodolphe Laurent1*, MD; Philippine Furelau1, MD; Emmanuel Doumard2, PhD; Anne Ferrier1, MD; Laetitia Bosch1, MD; Cyndie Ba1, MD; Rémi Menut1, MD; Matt Kurrek3, MD; Thomas Geeraerts1, MD, PhD; Antoine Piau4, MD, PhD; Vincent Minville1, MD, PhD

1Département d’Anesthésie-Réanimation, Hôpital Pierre-Paul Riquet, Centre Hospitalier Universitaire Purpan, Toulouse, France
2Institut de Recherche en Informatique de Toulouse, Université Toulouse III Paul Sabatier, Toulouse, France
3Department of Anesthesia, University of Toronto, Toronto, ON, Canada
4Département de Gériatrie, Centre Hospitalier Universitaire Rangueil, Toulouse, France
*these authors contributed equally

Corresponding Author:
Fabrice Ferré, MD, PhD
Département d’Anesthésie-Réanimation
Hôpital Pierre-Paul Riquet
Centre Hospitalier Universitaire Purpan
Place du Dr Baylac
Toulouse, 31300
France
Phone: 33 0561779988
Fax: 33 0561772170
Email: fabriceferre31@gmail.com

Abstract

Background: The ongoing COVID-19 pandemic has highlighted the potential of digital health solutions to adapt the organization of care in a crisis context.

Objective: Our aim was to describe the relationship between the MyRISK score, derived from self-reported data collected by a chatbot before the preanesthetic consultation, and the occurrence of postoperative complications.

Methods: This was a single-center prospective observational study that included 401 patients. The 16 items composing the MyRISK score were selected using the Delphi method. An algorithm was used to stratify patients with low (green), intermediate (orange), and high (red) risk. The primary end point concerned postoperative complications occurring in the first 6 months after surgery (composite criterion), collected by telephone and by consulting the electronic medical database. A logistic regression analysis was carried out to identify the explanatory variables associated with the complications. A machine learning model was trained to predict the MyRISK score using a larger data set of 1823 patients classified as green or red to reclassify individuals classified as orange as either modified green or modified red. User satisfaction and usability were assessed.

Results: Of the 389 patients analyzed for the primary end point, 16 (4.1%) experienced a postoperative complication. A red score was independently associated with postoperative complications (odds ratio 5.9, 95% CI 1.5-22.3; P=.009). A modified red score was strongly correlated with postoperative complications (odds ratio 21.8, 95% CI 2.8-171.5; P=.003) and predicted postoperative complications with high sensitivity (94%) and high negative predictive value (99%) but with low specificity (49%) and very low positive predictive value (7%; area under the receiver operating characteristic curve=0.71). Patient satisfaction numeric rating scale and system usability scale median scores were 8.0 (IQR 7.0-9.0) out of 10 and 90.0 (IQR 82.5-95.0) out of 100, respectively.

Conclusions: The MyRISK digital perioperative risk score established before the preanesthetic consultation was independently associated with the occurrence of postoperative complications. Its negative predictive strength was increased using a machine learning model to reclassify patients identified as being at intermediate risk. This reliable numerical categorization could be used to objectively refer patients with low risk to teleconsultation.
Introduction

Background

In France, the process of a patient undergoing elective surgery includes several essential steps such as the surgical consultation, preanesthetic consultation (PAC), and preanesthetic visit [1]. The decree of December 5, 1994, explicitly states that an anesthetist should carry out the PAC [2]. This consultation contributes to the preanesthetic evaluation of the patient’s health status, justifying the prescription of complementary examinations (e.g., laboratory tests) and any specialized consultations that allow a perioperative risk assessment formalized by the American Society of Anesthesiologists (ASA) score [3]. This perioperative risk evaluation is used, for example, to determine a patient’s eligibility for an ambulatory care pathway, an enhanced recovery after surgery program, or, conversely, a postoperative stay in the intensive care unit [4,5]. As it stands, this state-of-the-art evaluation requires medical expertise.

Beyond the low reproducibility of the ASA score [6,7], the perioperative risk global assessment during the PAC is not well standardized and may be incomplete, especially when consultation time is limited. This is why in some anesthesia teams, patients are asked to complete a questionnaire in paper form before their PAC, allowing them to specify, for example, their past medical and surgical history or their usual treatments. The patient is then asked to hand the completed questionnaire to the anesthetist during the PAC [8]. However, electronic patient-reported outcome measures offer many advantages over paper-based collection [9-13]: preferred modality; (directly) visualized results; higher data quality and response rate; decreased completion time; facilitates patient-clinician communication, improving the decision-making process; and so on. With regard to the preanesthetic questionnaire, the digital version is considered more efficient than the paper form [14]. Moreover, it has been shown that the quality of perioperative care can be improved by a digitalized preoperative information and assessment program [14-16], particularly through automatic reminders or clinical decision support. Health digital tools are therefore definitely of major interest in the perioperative setting.

Objectives

In 2020, the COVID-19–related restrictions accelerated the implementation of organizational digital health innovations. In the context of the COVID-19 crisis, the anesthesia department at the Toulouse University Hospital in Purpan, Toulouse, France, decided to digitalize the PAC by implementing teleconsultations (as much as possible to reduce interpersonal contact) and a digital conversational agent (aka chatbot) that allowed collection of medical data before the PAC. An approach assessing the relevance of the data collected as well as user satisfaction seemed essential to validate the sustainable use of this digital tool. The main objective of this study was to demonstrate that our chatbot was able to stratify patients according to their perioperative risk level. We hypothesized that the MyRISK perioperative risk score, established before the PAC according to a predefined algorithm based on data collected digitally, was correlated with the occurrence of postoperative complications at 6 months. Our secondary objectives were to improve the prognostic predictive value of this score using a machine learning model to reclassify patients classified as intermediate risk and assess patients’ and physicians’ satisfaction when using this digital health tool.

Methods

Experimental Design

This single-center prospective observational study was conducted in the anesthesia department of the Toulouse University Hospital. To our knowledge, the correlation with postoperative complications of a digital score based on self-reported medical data has never been described. Thus, no assumptions could be made regarding the relative risk and the positive and negative predictive values of being classified as high perioperative risk by the MyRISK score. Given the relatively low postoperative complication rate after scheduled orthopedic surgery (almost 5% [17]), we estimated that approximately 500 patients should be included to meet our objectives (based on expert opinion).

Population

All patients aged ≥18 years who were scheduled for orthopedic surgery at the Toulouse University Hospital between June 1, 2020, and October 31, 2020, were eligible. The exclusion criteria were the presence of a protection regime for adults (guardianship, curatorship, or safeguard of justice), patients who did not speak French, the presence of a major sensory handicap (blindness or deafness) compromising the comprehension of the information, patients who did not complete the digital questionnaire through the chatbot (this criterion was considered a refusal to participate), and patients who expressed their opposition to participating in this study.

MyRISK Score

The preanesthetic digital conversational agent, Medical Assistant Experience (MAX), was developed by 2 anesthetists of the Toulouse University Hospital (FF and VM) in collaboration with a company that creates secure health companions (BOTdesign, Toulouse, France; Figure 1: Multimedia Appendices 1 and 2). Its content was based on the preexisting paper form questionnaire with the addition of anesthetic items (Appendices 1 and 2). Its content was based on the preexisting paper form questionnaire with the addition of anesthetic items considered relevant, such as those allowing the calculation of perioperative scores published in the literature. As an example, we can cite the calculation of the Amsterdam Preoperative Anxiety and Information Scale score [18]; the Lee cardiovascular complication risk score [19]; or the snoring, tiredness, observed apnea, blood pressure, BMI, age, neck...
circumference, and gender (STOP-BANG) obstructive sleep apnea screening score [20] in which 7 of the 8 items are collected from the patient’s responses to the conversational agent.

Access to the chatbot was made possible once the surgical decision was made, after which the patient received an email inviting them to create their personal account using a smartphone, tablet device, or computer. The data collected were editable at any time by the patient.

The MyRISK score was developed using the Delphi method [21]. The first step was to identify among all the items of data collected by the chatbot those that were relevant, that is, considered to have weight in the development of a predictive risk score. After 2 rounds of discussion, a panel of 6 experts reached a consensus on 16 items, which were then retained (Table 1). The second step involved defining the independent risk level (1, 2, or 3) of each of the 16 items (Table 1). The third step concerned developing an algorithm based on these 16 items to stratify the global perioperative risk level into 3 categories corresponding to a presumed low, intermediate, or high global perioperative risk. Briefly, patients were classified as low risk when all 16 criteria were level 1, as intermediate risk when ≥1 of the 16 criteria were level 2, and as high risk when ≥3 level 2 criteria or ≥1 level 3 criterion were present.

Finally, to make the MyRISK score a visual tool, a green, orange, or red dot was assigned to the low, intermediate, and high global perioperative risk levels, respectively. This color coding was easily accessible and visible on the digital dashboard of patients enrolled in the MAX program. Patients were then considered to have a green, orange, or red MyRISK score.

Figure 1. Screenshots of the digital conversational agent Medical Assistant Experience (BOTdesign, Toulouse, France). The patient is asked to complete a self-assessment of its predictive criteria for difficult intubation (Mallampati and upper lip bite tests).
Table 1. MyRISK score criteria.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years) [22]</td>
<td>&lt;65</td>
<td>65 to 80</td>
<td>&gt;80</td>
</tr>
<tr>
<td>BMI (kg/m²) [23]</td>
<td>&lt;30</td>
<td>30 to 40</td>
<td>&gt;40</td>
</tr>
<tr>
<td>Drug allergies</td>
<td>No</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Hemostasis disorders</td>
<td>No</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Number of medications</td>
<td>0</td>
<td>1 to 5</td>
<td>&gt;5</td>
</tr>
<tr>
<td>Active smoking [24]</td>
<td>No</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Asthma</td>
<td>No</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Sleep apnea syndrome [25]</td>
<td>No</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Maximum level of activity (METb) [26]</td>
<td>Walking up 2 flights of stairs without stopping; walking in the street (5-7 km/h); important domestic activities (washing the floor); and sports activities</td>
<td>N/A</td>
<td>Activities of daily living (meals and toileting); walking in the house; and walking in the street (3-5 km/h)</td>
</tr>
<tr>
<td>Cardiac symptoms during exercise</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
</tr>
<tr>
<td>Hypertension</td>
<td>No</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Cardiac disease [27]</td>
<td>No</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Respiratory disease [24]</td>
<td>No</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Renal disease [28]</td>
<td>No</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Neurological disorders [29]</td>
<td>No</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Diabetes</td>
<td>No</td>
<td>Yes</td>
<td>N/A</td>
</tr>
</tbody>
</table>

aN/A: not applicable.
bMET: metabolic equivalent.

Postoperative Complications

Each patient was interviewed by telephone 6 months after surgery by one of the physicians involved in the study. After information was provided on study objectives and oral consent, the telephone survey was used to ask patients about the potential occurrence of postoperative complications. The survey was guided by a computerized structured questionnaire, allowing the secure collection of pseudonymized data. After oral consent, the computerized postoperative patient record form was consulted to ensure that there were no missing data concerning the occurrence of postoperative complications. The average duration of the interview was 9.8 (SD 9) minutes.

We considered that patients had a postoperative complication if they had experienced at least one adverse event among those listed in Textbox 1.
**Potential postoperative complications**

- Acute renal failure (creatinine increase $\geq 0.3$ mg/dL $\geq 26.5$ μmol/L in 48 hours or creatinine increase $\geq 1.5 \times$ baseline creatinine in <7 days or diuresis $<0.5$ mL/kg/h for 6 hours or hospitalization for acute kidney injury) [30]
- Myocardial infarction (such as an increase in troponin associated with at least one of the following: signs of ischemia, ST-segment change, development of left branch block on electrocardiogram, or hospitalization for angina or myocardial infarction) [31-33]
- Acute heart failure (such as the presence of clinical, radiological, or echocardiographic signs; N-terminal pro brain natriuretic peptide elevation $\geq 900$ pg/mL; or hospitalization for cardiac heart failure or cardiogenic pulmonary edema) [31,32]
- De novo atrial fibrillation (confirmed on electrocardiogram) [31,34]
- Transient ischemic attack or stroke (confirmed by computed tomography or magnetic resonance imaging or any hospitalization for stroke or transient ischemic attack) [35,36]
- Infection (such as fever requiring antibiotic therapy, hospitalization for fever, or suspected or documented infection)
- Respiratory complication (such as lung disease or respiratory compromise requiring oxygen or noninvasive or invasive ventilation or any hospitalization for respiratory compromise or lung disease) [37]
- Thromboembolic event (confirmed on Doppler ultrasound or computed tomography) [38]
- Hemorrhage (requiring transfusion)
- Rehospitalization
- Death

**User Satisfaction**

Patient satisfaction with the quality and usability of the chatbot was assessed by the system usability scale (SUS). The SUS is a validated standardized tool for collecting users’ opinions on the perceived ease of use of a digitalized system [39,40]. Briefly, the SUS assesses user experience and acceptability (Table 2). Ten statements (5 positive and 5 negative) are listed. Users respond to each statement using a Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The overall score is calculated to consider items with reversed valences.

<table>
<thead>
<tr>
<th>Statements</th>
<th>Scoringa</th>
</tr>
</thead>
<tbody>
<tr>
<td>I think that I would like to use this system</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>I found the system unnecessarily complex</td>
<td></td>
</tr>
<tr>
<td>I thought the system was easy to use</td>
<td></td>
</tr>
<tr>
<td>I think that I would need the support of a technical person to be able to use this system</td>
<td></td>
</tr>
<tr>
<td>I found the various functions in the system were well integrated</td>
<td></td>
</tr>
<tr>
<td>I thought there was too much inconsistency in this system</td>
<td></td>
</tr>
<tr>
<td>I would imagine that most people would learn to use this system very quickly</td>
<td></td>
</tr>
<tr>
<td>I found the system very cumbersome to use</td>
<td></td>
</tr>
<tr>
<td>I felt very confident using the system</td>
<td></td>
</tr>
<tr>
<td>I needed to learn a lot of things before I could get going with this system</td>
<td></td>
</tr>
</tbody>
</table>

*aInstruction on using the System Usability Scale: Please, circle the appropriate score for each statement from 1 (strongly disagree) to 5 (strongly agree).*

**Machine Learning Model**

Our goal was to train a machine learning model to predict the MyRISK score of patients having either a green or red score. This trained model was then asked to predict the MyRISK score of individuals classified as orange and reclassify them as either modified green or modified red.
**Data Preprocessing**

The data set used to train the model was extracted from a larger database of patients. We filtered out duplicates of individuals and features with >70% missing values. We also transformed nominal features following the *one-hot encoding* method, creating a new binary feature for each unique value. We finally filtered out individuals classified as *orange* for the prediction task. The final processed data set was composed of 1823 individuals classified as *green* or *red* for 83 features.

**Feature Selection**

To filter out redundant features, the recursive feature elimination (RFE) method was used [41]. Briefly, the following steps were applied: data were split into training and test sets; the model was trained on the training set, and its performance was evaluated on the test set; each feature contribution was evaluated, and the least contributing feature was identified (local Shapley additive explanations method [42]) and removed before going back to the first step.

The RFE algorithm returns a list of model performances in the training and test sets for each feature. The evolution of the training and test accuracies of the model through the RFE are presented in Figure 2. In total, 25 features were finally selected in the data set for the final model training.

**Figure 2.** Evolution of the training and test accuracies of the model through the recursive feature elimination method. The test performance of the model starts to worsen significantly when <25 features are considered. The 25 most contributing features were finally selected for the final model training.

---

**Model Training and Performance**

The extreme gradient boosting classifying model was used to train the model on the 25 selected features [43]. Using a hyperparameter grid search, the training and test confusion matrices were obtained with good accuracy scores (97.5% and 96%, respectively).

All processing stages were performed using open-source software machine learning libraries *sklearn* 1.0.1 and extreme gradient boosting 1.5 (for the model) in Python programming language (version 3.9.7; Python Software Foundation).

**Statistical Analysis**

Qualitative data were expressed as numbers (%). Quantitative data were expressed as median (IQR) or mean (SD) as appropriate. Categorical variables were compared using the Fisher exact test or the chi-square test. Quantitative variables were compared using the nonparametric Mann-Whitney *U* test. Multivariate analysis (stepwise logistic regression analysis) was performed to identify explanatory variables for the occurrence of postoperative complications at 6 months. The analysis was performed on an intention-to-treat basis. Statistical analysis was performed using MedCalc (version 12.6.1; MedCalc Software Ltd). *P*<.05 was considered statistically significant.

**Ethics Approval**

This research is considered an experiment in educational sciences aiming to evaluate the participative and pedagogical quality of a new digital tool implemented in current practice. Hence, this research was deemed to fall outside the Jardé law, meaning that no formal ethics approval was required for this study.

Information about the participants’ health conditions was collected by the chatbot after they had created their account, but the company BOTdesign had access neither to the patients’ identity nor to their IP address. This strategy of data protection was decided in agreement with the eHealth committee of the University Hospital Center of Toulouse. The connection to the digital questionnaire was secure (following the *General Data Protection Regulation* guidelines). An email invitation to log in was sent to the patients after the appointment with the surgeon. Patients chose their own secret password to create their MAX account. Each patient was given oral and written information about this research before enrollment, ensuring that they did not have any objection to participating. This study did not present any risk for the participants and did not modify the usual care pathway or the time required for patient management.
**Results**

**Population**

Of the 1000 eligible patients who were scheduled for orthopedic surgery at the Toulouse University Hospital between June 1, 2020, and October 31, 2020, a total of 434 (43.4%) patients logged in to the chatbot. Of these 434 patients, 401 (92.4%) agreed to participate in this study. The characteristics of the studied population are presented in Table 3.

### Table 3. Demographic characteristics of the patients.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), median (IQR)</td>
<td>39 (27-54)</td>
</tr>
<tr>
<td>Sex (n=401), n (%)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>241 (60.1)</td>
</tr>
<tr>
<td>Female</td>
<td>160 (39.9)</td>
</tr>
<tr>
<td>BMI (kg/m^2), mean (SD)</td>
<td>25 (4)</td>
</tr>
<tr>
<td>Education (n=388), n (%)</td>
<td></td>
</tr>
<tr>
<td>Undergraduate</td>
<td>83 (21.4)</td>
</tr>
<tr>
<td>Graduate</td>
<td>135 (34.8)</td>
</tr>
<tr>
<td>Postgraduate</td>
<td>170 (43.8)</td>
</tr>
<tr>
<td>Ambulatory care pathway (n=388), n (%)</td>
<td>297 (76.5)</td>
</tr>
<tr>
<td>Surgical risk (n=389), n (%)</td>
<td></td>
</tr>
<tr>
<td>Minor</td>
<td>284 (73)</td>
</tr>
<tr>
<td>Intermediate</td>
<td>89 (22.9)</td>
</tr>
<tr>
<td>Major</td>
<td>16 (4.1)</td>
</tr>
<tr>
<td>ASA^a score (n=389), n (%)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>282 (72.5)</td>
</tr>
<tr>
<td>2</td>
<td>89 (22.9)</td>
</tr>
<tr>
<td>3</td>
<td>18 (4.6)</td>
</tr>
<tr>
<td>4</td>
<td>0 (0)</td>
</tr>
<tr>
<td>MyRISK score (n=389), n (%)</td>
<td></td>
</tr>
<tr>
<td>Green (low risk)</td>
<td>100 (25.7)</td>
</tr>
<tr>
<td>Orange (intermediate risk)</td>
<td>150 (38.6)</td>
</tr>
<tr>
<td>Red (high risk)</td>
<td>139 (35.7)</td>
</tr>
<tr>
<td>STOP-BANG^b score modified (out of 7), median (IQR)</td>
<td>1 (1-2)</td>
</tr>
<tr>
<td>Lee score modified (out of 4), median (IQR)</td>
<td>0 (0-1)</td>
</tr>
<tr>
<td>APAIS^c anesthesia score (out of 15), median (IQR)</td>
<td>5 (3-7)</td>
</tr>
</tbody>
</table>

^aASA: American Society of Anesthesiologists.
^bSTOP-BANG: snoring, tiredness, observed apnea, blood pressure, BMI, age, neck circumference, and gender.
^cAPAIS: Amsterdam Preoperative Anxiety and Information Scale.

**User Satisfaction**

The median satisfaction score of patients regarding the use of the chatbot as assessed by the NRS was 8.0 (IQR 7.0-9.0) out of 10. The median satisfaction score regarding the use of the digital questionnaire as assessed by the SUS was 90.0 (IQR 82.5-95.0) out of 100.

The median SUS score was higher for users who chose a tablet device or smartphone (97/391, 24.8%) than for those who chose a computer (294/391, 75.2%); 92.5 (IQR 85.0-97.5) versus 90.0 (IQR 82.0-95.0), respectively (P=.01).

A large majority of the PACs were teleconsultations (331/401, 82.5%). Regarding the patients’ wishes for a future PAC if indicated, 54.7% (181/331) of the patients who received a teleconsultation wished to keep this mode of PAC in the future, whereas 19% (13/70) of the patients who received a face-to-face consultation wished to keep the same mode of PAC (P=.08).

The mean patient satisfaction score (NRS) regarding the teleconsultation was 8.4 (SD 1.59) out of 10.
The satisfaction score of the anesthesiologists (n=18) regarding the use of the digital platform was collected. Their median satisfaction score was 7.0 (IQR 6.0-8.0) out of 10, and their median SUS usability score was 72.5 (IQR 63.1-88.1) out of 100.

**Postoperative Complications**

Of the 389 patients analyzed, 16 (4.1%) had a postoperative complication at 6 months. No deaths were reported. The results of the univariate analysis are presented in Table 4.

A dependency relationship between the ASA and MyRISK scores was found (Table 5).

Compared with ASA score=1, an ASA score of ≥3 was independently associated with the occurrence of postoperative complications at 6 months (odds ratio [OR] 5.8, 95% CI 1.7-20.2; \( P = .006 \)). In comparison with a green score, a red score was independently associated with the occurrence of postoperative complications at 6 months (OR 5.9, 95% CI 1.5-22.3; \( P = .009 \)). Age and surgical risk included in the analysis were not identified as independent variables of the occurrence of postoperative complications. The area under the receiver operating characteristic curve (AUC) of the selected model was 0.78 (95% CI 0.73-0.82).

Finally, a red score predicted postoperative complications with 75% sensitivity, 98% negative predictive value, 66% specificity, and 9% positive predictive value (AUC=0.70).
Table 4. Comparison between patients with postoperative complications and those without postoperative complications (univariate analysis; N=389).

<table>
<thead>
<tr>
<th></th>
<th>No postoperative complications (n=373)</th>
<th>Postoperative complications (n=16)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), median (IQR)</td>
<td>39.0 (27.0-53.0)</td>
<td>56.5 (44.0-68.0)</td>
<td>.007</td>
</tr>
<tr>
<td>Ambulatory care pathway, n (%)</td>
<td>292 (78.3)</td>
<td>5 (31.2)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Surgical risk, n (%)</strong></td>
<td></td>
<td></td>
<td>.06</td>
</tr>
<tr>
<td>Minor</td>
<td>276 (74)</td>
<td>8 (50)</td>
<td></td>
</tr>
<tr>
<td>Intermediate</td>
<td>83 (22.2)</td>
<td>6 (37.5)</td>
<td></td>
</tr>
<tr>
<td>Major</td>
<td>14 (3.8)</td>
<td>2 (12.5)</td>
<td></td>
</tr>
<tr>
<td><strong>ASA&lt;sup&gt;a&lt;/sup&gt; score, n (%)</strong></td>
<td></td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>1</td>
<td>277 (74.3)</td>
<td>5 (31.2)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>83 (22.2)</td>
<td>6 (37.5)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>13 (3.5)</td>
<td>5 (31.2)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td><strong>MyRISK score, n (%)</strong></td>
<td></td>
<td></td>
<td>.002</td>
</tr>
<tr>
<td>Green (low risk)</td>
<td>100 (26.8)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Orange (intermediate risk)</td>
<td>146 (39.1)</td>
<td>4 (25)</td>
<td></td>
</tr>
<tr>
<td>Red (high risk)</td>
<td>127 (34.1)</td>
<td>12 (75)</td>
<td></td>
</tr>
<tr>
<td>Number of medications, median (IQR)</td>
<td>0.0 (0.0-2.0)</td>
<td>0.5 (0.0-1.0)</td>
<td>.65</td>
</tr>
<tr>
<td>Active smoking, n (%)</td>
<td>172 (46)</td>
<td>8 (50)</td>
<td>.80</td>
</tr>
<tr>
<td>Asthma, n (%)</td>
<td>43 (11.5)</td>
<td>2 (12.5)</td>
<td>.99</td>
</tr>
<tr>
<td>Sleep apnea syndrome, n (%)</td>
<td>14 (3.7)</td>
<td>2 (12.5)</td>
<td>.14</td>
</tr>
<tr>
<td>Cardiovascular disease, n (%)</td>
<td>24 (6.4)</td>
<td>3 (18.7)</td>
<td>.09</td>
</tr>
<tr>
<td>Renal disease, n (%)</td>
<td>7 (1.8)</td>
<td>0 (0)</td>
<td>.99</td>
</tr>
<tr>
<td>Neurological disease, n (%)</td>
<td>26 (6.9)</td>
<td>2 (12.5)</td>
<td>.33</td>
</tr>
<tr>
<td>Digestive disease, n (%)</td>
<td>55 (14.7)</td>
<td>3 (18.7)</td>
<td>.72</td>
</tr>
<tr>
<td>Diabetes, n (%)</td>
<td>19 (5)</td>
<td>1 (6.2)</td>
<td>.58</td>
</tr>
<tr>
<td>APAIS&lt;sup&gt;b&lt;/sup&gt; anesthesia score (out of 15), median (IQR)</td>
<td>5.0 (3.0-7.0)</td>
<td>6.0 (4.0-7.5)</td>
<td>.61</td>
</tr>
<tr>
<td>Lee score modified (out of 4), median (IQR)</td>
<td>0.0 (0.0-0.0)</td>
<td>0.0 (0.0-0.0)</td>
<td>.52</td>
</tr>
<tr>
<td>Apfel score modified (out of 3), median (IQR)</td>
<td>1.0 (0.0-1.0)</td>
<td>1.0 (0.0-2.0)</td>
<td>.48</td>
</tr>
<tr>
<td>STOP-BANG&lt;sup&gt;c&lt;/sup&gt; score modified (out of 7), median (IQR)</td>
<td>1.0 (1.0-2.0)</td>
<td>2.0 (1.0-2.5)</td>
<td>.46</td>
</tr>
</tbody>
</table>

<sup>a</sup>ASA: American Society of Anesthesiologists.

<sup>b</sup>APAIS: Amsterdam Preoperative Anxiety and Information Scale.

<sup>c</sup>STOP-BANG: snoring, tiredness, observed apnea, blood pressure, BMI, age, neck circumference, and gender.

Table 5. Correlation between American Society of Anesthesiologists (ASA) and MyRISK scores (N=400).

<table>
<thead>
<tr>
<th>MyRISK score</th>
<th>ASA score, n (%)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Green (low risk)</td>
<td>98 (24.4)</td>
<td>7 (1.7)</td>
</tr>
<tr>
<td>Orange (intermediate risk)</td>
<td>124 (30.9)</td>
<td>28 (7)</td>
</tr>
<tr>
<td>Red (high risk)</td>
<td>66 (16.5)</td>
<td>58 (14.5)</td>
</tr>
</tbody>
</table>
Recalculation of the Predictive Value of the MyRISK Score Using a Machine Learning Model

Reclassification of Patients With an Orange MyRISK Score

Among the 389 patients analyzed for the primary end point, 150 (38.6%) were initially classified as orange. Of these 150 patients, 4 (2.7%) experienced postoperative complications. Of the 146 patients classified as orange with no postoperative complications, 65 (44.5%) were reclassified as modified red and 81 (55.5%) as modified green using the trained model. Similarly, of the 4 patients with postoperative complications, 3 (75%) were finally reclassified as modified red and 1 (25%) as modified green.

Concerning these 4 patients, the contribution of each feature was computed with the local Shapley additive explanations method (refer to the Methods section for details). The results are presented in Figure 3.
**Predictive Value of the Modified MyRISK Score**

Once the 4 patients classified as *orange* were reclassified, a *modified red* MyRISK score was identified as strongly associated with postoperative complications at 6 months (OR 21.8, 95% CI 2.8-171.5; *P*=.003). An ASA score of ≥3 was also associated with postoperative complications (OR 4.7, 95% CI 1.4-16; *P*=.01).

Finally, a *modified red* score predicted postoperative complications with high sensitivity (94%) and high negative predictive value (99%) but with low specificity (49%) and very low positive predictive value (7%; AUC=0.71).
Discussion

Principal Findings

Through this study, we were able to validate the prognostic predictive value of a perioperative risk score established from data collected by a digital conversational agent implemented to assist the anesthetist during the PAC. In this context, the MyRISK score was correlated with the incidence of complications occurring in the first 6 months postoperatively. The strength of its predictive value was increased using a machine learning model to reclassify patients classified as intermediate risk. The use of an objective method allowing perioperative risk stratification according to a color code (ie, visual tool) and available before the PAC could be relevant for physicians. Finally, the use of this innovative digital tool seems to fully satisfy users.

We were able to identify a dependency between the level of perioperative risk at the end of the medical clinical evaluation (ie, ASA score) and the one calculated digitally before the PAC (ie, MyRISK score). This correlation had already been found by Zuidema et al [44] in a 2011 study using a 22-item numerical questionnaire administered to 14,349 patients, the authors highlighted that a computerized risk assessment could perform well and correlate with the clinical assessment of the ASA score (AUC=0.953), while limiting the interindividual variability of a clinician-assessed ASA score. In this context, it is interesting to note that the primary factor in numerical misclassification of the ASA score was an incomplete or incorrect patient response. More recently, Enneking et al [45] presented a 5-criteria composite preoperative risk score (patient-centered anesthesia triage system score) that correlated well with the ASA score (AUC=0.75, 95% CI 0.69-0.83), highlighting its usefulness for patient triage. Since 2017, the authors have been using this score in clinical practice to propose the systematic performance of a teleconsultation for patients classified as no risk.

By analyzing complications occurring in the first 6 postoperative months, we were able to validate the independent prognostic predictive value of the MyRISK score on the occurrence of serious perioperative adverse events. To our knowledge, this approach of validating a numerical risk score on objective criteria (ie, perioperative complications) has never been described in the literature. Thus, by reliably classifying patients according to their level of perioperative risk, the use of the (modified) MyRISK score could allow, before the PAC, triaging of patients by proposing the most appropriate modality of consultation (eg, teleconsultation for patients classified as [modified] green and face-to-face consultation for patients classified as [modified] red). Secondary benefits linked to this triage modality are expected: patients’ experience could be improved by reducing waiting time and optimizing consultation time. In addition, the face-to-face consultation could be dedicated to the management of patients with the most complex conditions who require, for example, specialized examinations or consultations. Moreover, as the role of anesthesiologists in the postoperative management of patients is growing, the implementation of a postanesthetic consultation could be a future trend, particularly for patients classified as high risk. In this setting, digital tools help to keep patients and caregivers connected for better follow-up, allowing an early detection or even prevention of postoperative complications.

The current health context has temporarily established the need to reduce travel throughout the country to limit interpersonal contact. Thus, the French National Authority for Health has recommended the use of telemedicine to enable remote management of patients [46]. Finally, the COVID-19 pandemic has highlighted the potential of digital health solutions in facilitating the adaptation of the organization of care in a crisis situation [47]. In this context, the development of the digital solution MAX helped us to organize the resumption of surgical activity when restrictions were lifted. A relevant distribution of patients between teleconsultation and face-to-face consultation using the (modified) MyRISK score could allow the indefinite continuance of teleconsultation after the pandemic. However, only 25.7% (100/389) of the patients were classified as low perioperative risk (ie, green). Questions still remain as to the proper organizational management of patients classified as intermediate risk (ie, orange). It is worth noting that the percentage of patients classified as low perioperative risk (theoretically eligible for teleconsultation) increases to 47% (183/389) when using the modified classification.

In our study, patient satisfaction with the use of the digital questionnaire as well as the PAC process was excellent. The developed digital conversational agent seems to be an adequate platform for the collection of patients’ medical information, as shown by the excellent usability score obtained. Moreover, the usability seems to be better when completing the digital questionnaire on a smartphone or tablet device. Our results highlight the enthusiasm of patients for using a digital health platform. These results are in agreement with those described by VanDenKerkhof et al [14], where patient comfort was increased by >70% by computerizing the preanesthetic questionnaire.

Although we did not strictly evaluate the reliability of the data collected by the digital conversational agent, Osman et al [48] demonstrated a response reliability of >90% when a computerized preanesthetic questionnaire was used. Thus, there is consensus in the literature now of the reliability of digital collection of information [49].

Several factors may have favored patient acceptance of this new digital solution. The young age of the patients enrolled (median age 39.0, IQR 27.0-54.0 years) and their level of education (305/389, 78.4%, had graduate or postgraduate degrees) probably explain the very high levels of satisfaction and usability obtained. These results are in agreement with those obtained by Kruse et al [50]. Indeed, the age, level of education, and computer skills of the patients were the 3 main barriers to telehealth adoption identified by the authors [50]. The acceptability and satisfaction of the patients obtained during the use of this digital support encourages our department to develop telemedicine solutions. This enthusiasm is reinforced by the good satisfaction ratings provided by the members of the medical team during the use of this platform, which is probably linked to the automatic integration of the data collected by the digital conversational agent into the computerized PAC file.
However, a qualitative analysis of the main difficulties encountered by the physicians highlighted the absolute necessity of good interoperability among the various software systems.

In our study, the rate of postoperative complications observed at 6 months was 4.1% (16/389). No deaths were recorded. These results are in accordance with those already published in the literature. Indeed, in 2013, Chikuda et al [51] evaluated postoperative morbidity and mortality rates in scheduled orthopedic surgery among >100,000 patients [51]. In this context, the morbidity and mortality rates were 4.2% and 0.11%, respectively.

Our study includes several limitations. First, the results obtained in preoperative scheduled surgeries cannot be extrapolated to the context of urgent surgeries where the use of a digital conversational agent to assist the PAC seems difficult to achieve. Second, 82.5% (331/401) of the PACs analyzed in the study were teleconsultations. The period when the patients were included corresponded to the end of the first lockdown, which explains this high rate of teleconsultations that is not very representative of the subsequent evolution of the practices of our unit (approximately 40% currently). Third, the potential benefits of allocating patients to teleconsultation or face-to-face consultation according to the MyRISK score deserve to be studied in more detail; for example, analysis of patient-perceived quality-of-care indicators (eg, patient-reported outcome measures and patient-reported experience measures) [52] could demonstrate that the experience of care perceived by patients classified as green receiving teleconsultation is optimal, notably by avoiding unnecessary travel that disrupts their personal and professional schedules. Fourth, collection of the overall consumption of care by consulting the national health data system could have increased the prognostic predictive value of the MyRISK score by a more global and exhaustive analysis of the postoperative evolution of patients. This type of medicoeconomic approach should be favored in the future. Fifth, the experimental design of our study did not allow us to evaluate the reasons for the nonconnection to MAX by a significant number of patients. Thus, it seems likely that the satisfaction and usability scores were overestimated because they were collected only from the user population of patients. Analysis of the overall data is fundamental to understanding the potential explanatory factors of the digital divide in this context. Fifth, further studies are needed to extend the validation of the MyRISK score to other surgical populations and thus generalize our results. Finally, there is a need for studies examining the impact that the use of these tools has on clinical decision-making and patient outcomes.

Conclusions

To conclude, we were able to demonstrate the prognostic value of a perioperative risk score established from data collected by a digital conversational agent implemented before the PAC. In this setting, the MyRISK score was associated with the occurrence of postoperative complications at 6 months after surgery. The strength of its predictive value was increased using a machine learning model to reclassify patients classified as intermediate risk. The excellent levels of satisfaction and usability obtained from patients encourage us to develop and use this digital solution in health care. Further studies evaluating the overall use of the MyRISK score are necessary before using this digital stratification method to guide patients to teleconsultation or face-to-face consultation or to provide a perioperative personalized care pathway for patients at highest risk.

Acknowledgments

The authors would like to thank Olivier Thuillart and Jean-Louis Fraysse from BOTdesign (Toulouse, France), the company that created the digital companion Medical Assistant Experience. Support was provided solely from department sources. This work should be attributed to the Département d’Anesthésie Réanimation et Médecine Périopératoire, Centre Hospitalier Universitaire de Toulouse Purpan, Toulouse, France.

Authors' Contributions

FF, RL, and PF participated in the design of the study, performed all of the statistical analyses, and wrote the manuscript. ED implemented the machine learning model. MK, TG, AP, and VM participated in the design of the study and helped to draft the manuscript. AF, LB, CB, and RM performed all of the preanesthetic consultations. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Patient responding to the preanesthetic digital conversational agent.

[ PNG File , 809 KB - periop_v61e39044_app1.png ]

Multimedia Appendix 2

Example of patient completing the digital questionnaire.

[ MOV File , 11775 KB - periop_v61e39044_app2.mov ]
References


Abbreviations

ASA: American Society of Anesthesiologists
AUC: area under the receiver operating characteristic curve
MAX: Medical Assistant Experience
NRS: numeric rating scale
OR: odds ratio
PAC: preanaesthetic consultation
RFE: recursive feature elimination
STOP-BANG: snoring, tiredness, observed apnea, blood pressure, BMI, age, neck circumference, and gender
SUS: system usability scale

©Fabrice Ferré, Rodolphe Laurent, Philippe Furelau, Emmanuel Doumand, Anne Ferrier, Laetitia Bosch, Cyndie Ba, Rémi Menut, Matt Kurrek, Thomas Geeraerts, Antoine Piau, Vincent Minville. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 16.01.2023. 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.
Remote Home Monitoring of Continuous Vital Sign Measurements by Wearables in Patients Discharged After Colorectal Surgery: Observational Feasibility Study

Jobbe P L Leenen¹,²,³, RN, MSc; Vera Ardesch¹,⁴, RN, MSc; Gijsbert Patijn¹,², MD, PhD

¹Connected Care Center, Isala, Zwolle, Netherlands
²Department of Surgery, Isala, Zwolle, Netherlands
³Isala Academy, Isala, Zwolle, Netherlands
⁴Flexpool General Wards, Department of Care Support, Isala, Zwolle, Netherlands

Corresponding Author:
Jobbe P L Leenen, RN, MSc
Connected Care Center
Isala
Dr. van Heesweg 2
Zwolle, 8025 AB
Netherlands
Phone: 31 640275833
Email: j.p.l.leenen@isala.nl

Abstract

Background: Hospital stays after colorectal surgery are increasingly being reduced by enhanced recovery and early discharge protocols. As a result, postoperative complications may frequently manifest after discharge in the home setting, potentially leading to emergency room presentations and readmissions. Virtual care interventions after hospital discharge may capture clinical deterioration at an early stage and hold promise for the prevention of readmissions and overall better outcomes. Recent technological advances have enabled continuous vital sign monitoring by wearable wireless sensor devices. However, the potential of these devices for virtual care interventions for patients discharged after colorectal surgery is currently unknown.

Objective: We aimed to determine the feasibility of a virtual care intervention consisting of continuous vital sign monitoring with wearable wireless sensors and teleconsultations for patients discharged after colorectal surgery.

Methods: In a single-center observational cohort study, patients were monitored at home for 5 consecutive days after discharge. Daily vital sign trend assessments and telephone consultations were performed by a remote patient-monitoring department. Intervention performance was evaluated by analyzing vital sign trend assessments and telephone consultation reports. Outcomes were categorized as “no concern,” “slight concern,” or “serious concern.” Serious concern prompted contact with the surgeon on call. In addition, the quality of the vital sign data was determined, and the patient experience was evaluated.

Results: Among 21 patients who participated in this study, 104 of 105 (99%) measurements of vital sign trends were successful. Of these 104 vital sign trend assessments, 68% (n=71) did not raise any concern, 16% (n=17) were unable to be assessed because of data loss, and none led to contacting the surgeon. Of 62 of 63 (98%) successfully performed telephone consultations, 53 (86%) did not raise any concerns and only 1 resulted in contacting the surgeon. A 68% agreement was found between vital sign trend assessments and telephone consultations. Overall completeness of the 2347 hours of vital sign trend data was 46.3% (range 5%-100%). Patient satisfaction score was 8 (IQR 7-9) of 10.

Conclusions: A home monitoring intervention of patients discharged after colorectal surgery was found to be feasible, given its high performance and high patient acceptability. However, the intervention design needs further optimization before the true value of remote monitoring for early discharge protocols, prevention of readmissions, and overall patient outcomes can be adequately determined.

(JMIR Perioper Med 2023;6:e45113) doi:10.2196/45113

https://periop.jmir.org/2023/1/e45113

JMIR Perioper Med 2023 | vol. 6 | e45113 | p.135
(page number not for citation purposes)
KEYWORDS
telemedicine; remote monitoring; home monitoring; virtual care intervention; colorectal surgery; continuous vital signs monitoring; wearable wireless monitoring; clinical deterioration; readmission; virtual care; online intervention; e-health; remote monitoring; telehealth; vital signs

Introduction

Colorectal surgery is known for high complication and readmission rates [1-5]. In the last decade, enhanced recovery programs, such as the Enhanced Recovery After Surgery (ERAS) program, have been adopted widely and have resulted in significantly shorter hospital lengths of stay, with discharge as early as postoperative day 1 or 2 [6-8]. Serious postoperative complications such as anastomotic leak, abscess, ileus, thrombosis, or surgical site infection may therefore manifest themselves in the home setting [5,9]. Follow-up of these patients is generally limited to outpatient clinic visits that do not take place until several weeks after discharge. Late recognition of signs and symptoms by patients may cause delayed detection and lead to inferior clinical outcomes and readmissions [1,5,10].

Virtual care interventions such as remote patient monitoring with mobile health apps have the potential to further reduce lengths of hospital stay and prevent unnecessary readmissions, reduce emergency department visits, and help mitigate increasing nursing staff shortages and overall health care costs [11-13]. Such interventions require only limited investment in time and money if deployed at sufficient scale [14]. Virtual care interventions may therefore help to remove barriers to further expand early discharge protocols.

Recent technological advances have enabled continuous vital sign monitoring by wearable wireless sensor devices [15,16]. These devices have been shown to be able to accurately detect deviating vital sign trends [15,17]. Such technology may allow patients to safely recover at home while being carefully monitored with the intention to capture possible clinical deterioration at an early stage.

Several studies have demonstrated the feasibility of continuous vital sign monitoring in the hospital with wireless wearables [18-21], but there are only a few studies on remote home monitoring after discharge, and these have mostly been limited to intermittent monitoring of vital signs [11,22]. Only one study showed that continuous vital sign monitoring was technically feasible and well accepted by patients discharged after esophageal surgery [23]. It is unknown if a comparable intervention is feasible in patients after colorectal surgery. Therefore, the objective of this study was to determine the feasibility of a virtual care intervention consisting of continuous vital sign monitoring with wearable wireless sensors and teleconsultations supported by a central, remote patient-monitoring department for patients discharged after colorectal surgery.

Methods

Study Design and Setting

A single-center observational cohort study was conducted in May and June 2022 in a 1250-bed teaching hospital in the Netherlands. This study is reported in accord with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [24].

Participants

Patients scheduled for elective colorectal resection in May and June 2022 were approached for consent for the remote monitoring intervention during the preadmission call by a nurse. Inclusion criteria were as follows: age ≥18 years, elective colorectal resection, primary anastomosis, admission to the participating surgical ward, an uncomplicated clinical course (duration of admission <7 days), presence of an adequate caregiver at home, possession of a mobile phone, and no cognitive impairments. Exclusion criteria were as follows: unable to wear a continuous monitoring device due to a pacemaker or allergy, no desire for treatment (or no desire for referral) in the event of clinical deterioration, cognitive impairment at discharge, discharge to a rehabilitation or nursing home, physical limitations that would hinder participation, and insufficient command of the Dutch language.

Remote Home Monitoring Intervention

The remote home monitoring intervention was developed based on a previous study [25] that evaluated in-hospital continuous vital sign monitoring developed in cooperation with an abdominal surgeon, ward nurses, and nurses at the remote monitoring department of the hospital. The intervention consisted of 5 days of follow-up with 3 consultations by phone (every other day) and daily evaluation of vital sign trends with the wearable sensor (Figure 1).
Figure 1. Overview of the care path for included patients.

The consultations by phone were performed by the specialized cardiac care nurses at the remote monitoring department, who were dedicated to and had expertise in virtual care and remote monitoring of patients. The department is operated every day of the week from 8 AM to 11 PM. The consultation consisted of an assessment of the patient’s condition regarding pain symptoms, wound condition, stool, and nausea. In addition, vital sign trends measured by the wearable sensor were assessed daily and at the time of the telephone consultation.

Vital signs of patients were measured by the Conformit Européenne–marked Healthdot sensor and Intellivue Guardian Solution (IGS) software system (Philips). The wireless sensor is a previously validated patch worn on the patient’s chest that continuously records heart rate (HR) in beats per minute (bpm) and respiratory rate (RR) in respirations per minute (rpm) with a battery life of 14 days (Multimedia Appendix 1). The 2 vital sign measurements are transmitted wirelessly every 5 minutes through a long range, low power Internet of Things (LoRa) connection to the IGS software. Within the IGS software, vital sign trends are visualized and, complementary to the hospital Modified Early Warning Score (MEWS) protocol, the Deel EWS (D-EWS; Deel is Dutch for “partial”), a partial MEWS score, was aggregated every hour to promote adequate interpretation (Multimedia Appendix 2). The scores are based on the thresholds for HR and RR when sufficient data were present. There were no active alarms generated to the nurses when scores deviated. If any acute help was needed, patients were informed to call emergency services.

Study Procedures

Before the start of the study, nurses at the remote patient monitoring department were educated by the project manager (JPLL) and a colorectal surgeon (GP) in a 1.5-hour information session about colorectal resections, the most common complications, and nursing care for these patients. Moreover, the trend assessment used in this study was informed by case studies. In a 2-month period (May-June 2022), patients scheduled for colorectal resection were approached by the ward nurse, received information about the home monitoring intervention, and were asked to provide verbal informed consent during a preoperative telephone consultation. Directly postoperatively, the sensor was attached to the patient at the ward and in-hospital continuous monitoring was performed. When the in-hospital postoperative course was uncomplicated, the patient was eligible for home monitoring in this study. At discharge, the ward nurse asked the remote monitoring department to start the monitoring.

Data Collection

Baseline characteristics were collected, including age, sex, height, weight, BMI, American Society of Anesthesiologists (ASA) score, Charlson Comorbidity Index [26], type of colorectal surgery and procedure (open or laparoscopic), indication for surgery (malignant or benign), length of hospital stay in days, MEWS at discharge, heart rate and respiratory rate at discharge, and readmission ≤ 30 days. The primary end point was feasibility, defined by the following 3 outcomes: intervention performance, quality of vital sign data, and patient experience.

Intervention Performance

For each patient, we assessed whether the intervention was carried out according to protocol with daily vital sign assessment and telephone consultations. Daily vital sign trend assessments were scored as 0 (no cause for concern), 1 (slight concern), and 2 (serious concern). A score of 1 resulted in a wait-and-see approach and a score of 2 resulted in contact with the surgeon on call. An X was scored if the nurse was unable to assess the vital sign trend because of insufficient vital sign data. After the telephone consultation, the same scoring was done by assessing pain (rated by the Numeric Rating Scale), wound infection (redness/pain of the wound area), and experience of nausea (yes or no) and defecation (yes or no). Additionally, discrepancies in scoring between vital sign trend assessments and telephone consultations were registered. Finally, the time spent for the intervention and for each consultation was registered in time frames of 5 minutes.
Quality of the Vital Sign Data

The quality of vital sign data was defined as the proportion of trend assessments that were impossible to perform, as well as completeness of the vital sign data, the generated D-EWS scores, heart rate and respiration measurements, and the distribution of data gaps based on their duration: <1 hour, 1 to 4 hours, 4 to 8 hours, 8 to 16 hours, and 16 to 24 hours.

Patient Experience

Patient experience was measured at the end of the remote monitoring intervention by an 11-item questionnaire. Two questions assessed the overall experience of the patient and informal caregiver (when present) with a 10-point Likert scale, 7 questions assessed specific domains with a 5-point Likert scale, and 2 questions had open answers (Multimedia Appendix 3).

Statistical Analysis

Descriptive statistics were used to evaluate patient demographics and to assess the feasibility of home monitoring. Normally distributed continuous data are presented as the mean (SD). Likewise, nonnormally distributed data are presented as the median (IQR). Normality was determined with the Kolmogorov-Smirnov test and visually by a quantile-quantile plot and histogram. Nominal data were presented as frequencies (n) and percentages (%). All data were analyzed with SPSS Statistics (version 26; IBM Corp). Formal sample size calculation was challenging given the observational feasibility study design, but a sample in the range of 20 to 25 is considered adequate for this type of study [27]. Therefore, we aimed to include 20 patients and obtain data from 100 home monitoring days.

Ethical Considerations

The Medical Ethics Review Committee of Isala waived the need for ethical approval (210414). The study was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from each participant in the study.

Results

Study Characteristics

A total of 25 patients were screened, of whom 23 (92%) agreed to participate. After 1 patient dropped out and 1 patient had in-hospital complications, a total of 21 patients participated in the study. The mean age was 67 (SD 13) years, and 8 (38%) participants were male. The mean length of stay was 4.4 (SD 2.2) days, and hemicolectomies were mostly right sided (n=8, 38%). Three patients (14%) were readmitted to the hospital within 30 days, but there were no alterations in clinical decision-making or the clinical course based on the intervention (Multimedia Appendix 4). A full description of the participants is given in Table 1.
Table 1. Patient characteristics (n=21).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>67 (13)</td>
</tr>
<tr>
<td>Sex (male), n (%</td>
<td>8 (38)</td>
</tr>
<tr>
<td>Height (cm), mean (SD)</td>
<td>173 (9)</td>
</tr>
<tr>
<td>Weight (kg), mean (SD)</td>
<td>78 (15)</td>
</tr>
<tr>
<td>BMI (kg/m²), mean (SD)</td>
<td>25.9 (4.2)</td>
</tr>
<tr>
<td>American Society of Anesthesiologists class, n (%)</td>
<td></td>
</tr>
<tr>
<td>1-2</td>
<td>15 (71)</td>
</tr>
<tr>
<td>3-4</td>
<td>6 (29)</td>
</tr>
<tr>
<td>Charlson Comorbidity Index, mean (SD)</td>
<td>3.9 (2.3)</td>
</tr>
<tr>
<td>Type of surgery, n (%)</td>
<td></td>
</tr>
<tr>
<td>Right hemicolecotony</td>
<td>8 (38)</td>
</tr>
<tr>
<td>Sigmoid resection</td>
<td>6 (29)</td>
</tr>
<tr>
<td>Abdominal perianal resection</td>
<td>3 (14)</td>
</tr>
<tr>
<td>Wig resection colon</td>
<td>2 (10)</td>
</tr>
<tr>
<td>Ileocecal resection</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Low anterior resection</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Laparoscopic procedure, n (%)</td>
<td>20 (95)</td>
</tr>
<tr>
<td>Indication for surgery, n (%)</td>
<td></td>
</tr>
<tr>
<td>Malignant</td>
<td>16 (76)</td>
</tr>
<tr>
<td>Benign</td>
<td>5 (24)</td>
</tr>
<tr>
<td>Length of hospital stay (days), mean (SD)</td>
<td>4.4 (2.2)</td>
</tr>
<tr>
<td>Modified Early Warning Score at discharge, n (%)</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>8 (38)</td>
</tr>
<tr>
<td>1</td>
<td>12 (57)</td>
</tr>
<tr>
<td>2</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Heart rate at discharge (bpm), mean (SD)</td>
<td>81 (13)</td>
</tr>
<tr>
<td>Respiratory rate at discharge (rpm), mean (SD)</td>
<td>14 (1)</td>
</tr>
<tr>
<td>Readmission &lt;30 days, n (%)</td>
<td>3 (14)</td>
</tr>
</tbody>
</table>

**Intervention Performance**

Of 21 patients, 1 received only telephone consultations, because the vital sign patch was removed by accident on the ward by a ward nurse. The median time spent for the home monitoring was 40 (IQR 35-40) minutes per patient, and 49% (51/104) of the daily home monitoring checks took 5 minutes or less for the nurses to perform (Table 2). A total of 104 (99%) vital sign trend assessments were performed. The majority (71/104, 68%) of assessments did not raise concerns and 16% (17/104) could not be assessed because of excessive data loss. Nonetheless, no cases were found where the nurse was seriously concerned about a vital sign trend.

Considering the telephone consultations, 62 of 63 (98%) were performed because in one case, the patient could not be reached. The median pain score was 3 (IQR 2-4) at the first consultation and 2 (IQR 1-2) at the third consultation. In 1 of 62 consultations (1.7%) the wound assessment was painful and resulted in a score of 1 (slightly concerned) for which a wait-and-see approach was taken. In 6 of 62 consultations (10%), the patients had not yet defecated. In the majority of consultations (53/62, 85%), no concerns were raised. Only 1 of 62 consultations (1.7%) resulted in contact with the surgeon on call, due to blood loss during defecation, for which a wait-and-see policy was agreed to be followed.

Considering agreement, 42 (68%) of 62 telephone consultations were in agreement with the trend assessments, and in 8 (13%) telephone consultations, agreement was not possible because of a lack of trend data. Only once, the telephone consultation raised concern but the trend assessment did not (Figure 2).
<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total time spent by nurses on assessments per patient (minutes), median (IQR)</td>
<td>40 (35.0-40.0)</td>
</tr>
<tr>
<td>Time taken by nurses to complete daily home monitoring checks (n=104), n (%)</td>
<td></td>
</tr>
<tr>
<td>≤5 minutes</td>
<td>51 (49)</td>
</tr>
<tr>
<td>&gt;5 to ≤10 minutes</td>
<td>47 (45)</td>
</tr>
<tr>
<td>&gt;10 to ≤15 minutes</td>
<td>5 (5)</td>
</tr>
<tr>
<td>25 minutes</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Result of vital sign assessment (n=104), n (%)</td>
<td></td>
</tr>
<tr>
<td>No concern</td>
<td>71 (68)</td>
</tr>
<tr>
<td>Slight concern</td>
<td>16 (15)</td>
</tr>
<tr>
<td>Concern</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Unable to assess</td>
<td>17 (16)</td>
</tr>
<tr>
<td>Result of telephone consultations (n=62), n (%)</td>
<td></td>
</tr>
<tr>
<td>No concern</td>
<td>53 (85)</td>
</tr>
<tr>
<td>Slight concern</td>
<td>8 (13)</td>
</tr>
<tr>
<td>Concern</td>
<td>1 (2)</td>
</tr>
<tr>
<td>Numeric Rating Scale pain score (range 0-10) , median (IQR)</td>
<td></td>
</tr>
<tr>
<td>Day 1</td>
<td>3 (2-4)</td>
</tr>
<tr>
<td>Day 3</td>
<td>2 (1-4)</td>
</tr>
<tr>
<td>Day 5</td>
<td>2 (1-2)</td>
</tr>
<tr>
<td>Wound assessments of rubor/dolor (n=21), n (%)</td>
<td></td>
</tr>
<tr>
<td>Absence of defecation (n=63), n (%)</td>
<td>1 (5)</td>
</tr>
<tr>
<td></td>
<td>6 (10)</td>
</tr>
</tbody>
</table>

Figure 2. Agreement between trend assessments and telephone consultations (n=62). Values are shown as n (%). Green indicates agreement; orange indicates that there was no agreement and a wait-and-see approach was taken; red indicates that there was no agreement. X indicates that trend data were lacking. TC: telephone consultation.

Quality of the Vital Sign Data
Monitoring data were available for 20 patients for a total of 2347 hours of vital sign data with a median of 116.6 (IQR 115.5-118.7) hours per patient (Table 3). Data gaps of 30 minutes or less occurred in 95.2% (12,402/13,039) of measurements. Gaps between 30 and 60 minutes were detected in 2.6% (338/13,039) of measurements, and gaps between 1 and 16 hours were detected in 2.3% (299/13,039) of measurements. This resulted in an overall completeness of vital sign trends of 46.3% (range 5%-100%).

Of the 13,039 completed measurements, HR measurements contained a median of 15.6% (IQR 9.7%-35%) artifact data and RR measurements contained a median of 21.8% (IQR 9.3%-25.9%) artifact data. A total of 215 D-EWS scores were calculated based on the data, of which 40 (18.6%) were 3 or higher.
Table 3. Quality of vital sign data (total vital sign measurements=13,039).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monitoring time (hours), median (IQR)</td>
<td>116.6 (115.5-118.7)</td>
</tr>
<tr>
<td>Monitoring time (minutes), median (IQR)</td>
<td>6995 (6929-7124)</td>
</tr>
<tr>
<td>Heart rate (beats per minute), mean (SD)</td>
<td>74 (12)</td>
</tr>
<tr>
<td>Respiratory rate (breaths per minute), mean (SD)</td>
<td>17 (2)</td>
</tr>
<tr>
<td>D-EWS(^a) (n=1153), n (%)</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>215 (18.6)</td>
</tr>
<tr>
<td>1</td>
<td>701 (60.8)</td>
</tr>
<tr>
<td>2</td>
<td>197 (17.1)</td>
</tr>
<tr>
<td>3 or higher</td>
<td>40 (3.5)</td>
</tr>
<tr>
<td>Vital sign measurements per patient, median (IQR)</td>
<td>519 (284-1085)</td>
</tr>
<tr>
<td>Completeness of vital sign data, median (IQR)</td>
<td>36.6 (20.1-77.7)</td>
</tr>
<tr>
<td>Artifacts in heart rate measurements, median (IQR)</td>
<td>15.6 (9.7-35.0)</td>
</tr>
<tr>
<td>Artifacts in respiratory rate measurements, median (IQR)</td>
<td>21.8 (9.3-25.9)</td>
</tr>
<tr>
<td>Data gaps (n=13,039), n (%)</td>
<td></td>
</tr>
<tr>
<td>0-30 minutes</td>
<td>12,402 (95.1)</td>
</tr>
<tr>
<td>30-60 minutes</td>
<td>338 (2.6)</td>
</tr>
<tr>
<td>60-120 minutes</td>
<td>169 (1.3)</td>
</tr>
<tr>
<td>2-4 hours</td>
<td>91 (0.7)</td>
</tr>
<tr>
<td>4-8 hours</td>
<td>26 (0.2)</td>
</tr>
<tr>
<td>8-16 hours</td>
<td>13 (0.1)</td>
</tr>
</tbody>
</table>

\(^a\)D-EWS: Deel Early Warning Score (deel is Dutch for “partial”).

Patient Experience

Twenty patients returned the questionnaire (Table 4). The median satisfaction scores were 8 (IQR 7-9) of 10 for patients and 8 (IQR 5-8) of 10 for caregivers (n=7). The majority of patients (n=18, 90%) found the sensor comfortable to wear, although several remarks were made about experiences of discomfort during lateral sleep because of the rigidity of the sensor’s material (Multimedia Appendix 5). Further, the majority of patients liked to have insight in vital sign trends (n=18, 90%) and to have telephone consultations (n=17, 85%); a majority also found contact by phone adequate (n=17, 85%). Two patients (10%) felt safer with home monitoring, whereas 7 (35%) were neutral, and 11 (55%) did not feel safer. For future home monitoring interventions, 70% of patients agreed that they would want to be discharged from the hospital at an earlier stage if there were a home monitoring intervention. Possible reasons for this, based on the remarks section of the questionnaire, were that patients experienced better recovery in the home setting overall and because they experienced sleep interruptions at night while being admitted in the hospital (Multimedia Appendix 5).
Table 4. Results of the patient survey.

<table>
<thead>
<tr>
<th>Question</th>
<th>Median score (IQR)</th>
<th>Responses, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Likert-scale items (range 1 to 10)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In general, how did you experience the home monitoring period? (n=20)</td>
<td>8 (7-9)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>How did your informal caregiver experience the home monitoring period overall? (n=7)</td>
<td>8 (5-8)</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>Likert-scale items (range 1 to 5; n=20)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I found the wearable sensor comfortable.</td>
<td>5 (4-5)</td>
<td>2 (10)</td>
</tr>
<tr>
<td>I liked that healthcare professionals could see my vital signs (heart rate, breathing) on a daily basis.</td>
<td>5 (4-3-5)</td>
<td>1 (5)</td>
</tr>
<tr>
<td>I need to have insight into my vital signs measurements (heartbeat, breathing).</td>
<td>4 (3-5)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>I liked the telephone contact with the healthcare professionals.</td>
<td>5 (4-5)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>The telephone contacts were sufficient.</td>
<td>4 (4-5)</td>
<td>1 (5)</td>
</tr>
<tr>
<td>I felt safer with the home monitoring than if I had not had it.</td>
<td>4 (3-3-5)</td>
<td>11 (55)</td>
</tr>
<tr>
<td>If in the future you were allowed to go home a day earlier with home monitoring, would you want to?</td>
<td>4 (3-5)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

aRepresents a score of 1 to 4 on the 10-item Likert scale or 1 to 2 on the 5-item Likert scale.
bRepresents a score of 5 on the 10-item Likert scale or 3 on the 5-item Likert scale.
cRepresents a score of 6 to 10 on the 10-item Likert scale or 4 to 5 on the 5-item Likert scale.

Discussion

Principal Findings

In this study, we found that a remote home monitoring intervention for colorectal surgical patients consisting of vital sign trend assessments and telephone consultations performed by nurses at a remote home monitoring department was feasible. The intervention performance and patient acceptability were high, whereas the quality of vital sign data was still variable.

Intervention Performance

Our results show that 5 days of follow-up with vital sign trend assessments combined with telephone consultations could be successfully performed. The agreement between these 2 methods of patient assessment was adequate (nearly 70%). In this small cohort, the majority of assessments did not raise concerns about patient recovery, and no subsequent interventions were required, but this needs confirmation in a larger cohort. Given the limited time spent by nurses delivering the intervention, this type of postdischarge monitoring may become cost-efficient when operated at sufficient scale [14], especially if it proves to facilitate even shorter hospital stays and prevent readmissions.

To achieve this, further design optimizations are first needed. Telephone consultations may be replaced by a symptom questionnaire and self-registration in a mobile digital app by patients themselves, which promotes self-management and reduces workload for the monitoring nurse [28].

Importantly, the optimal measurement frequency for the detection of deviations in a timely manner among postdischarge surgical patients at home is still unknown [17,29]. A limited number of measurements per day rather than continuous measurements may be sufficient for the timely detection of deterioration for this patient category. Also, daily assessment of vital sign trends with the D-EWS score may not be needed for the majority of patients, as these types of early warning scores are designed to provide alerts for serious clinical deterioration, which is rare postdischarge in the home setting. Comparative studies of continuous versus intermittent vital sign measurements are needed to further investigate the optimal intervention for this patient category.

Quality of Vital Sign Data

Besides the design of the intervention, the technology of wearable wireless devices for continuous vital sign home monitoring also needs further optimization. In our study, trends were generated by performing vital sign measurement every 5 minutes. Despite the relatively low amount of measurement artifacts and data gaps, completeness of data was still low (median 37% per patient). In particular, the highly variable completeness was partly caused by data gaps longer than 4 hours, which hampered adequate trend assessment by the nurses. A possible explanation for this data loss is that the connectivity of the wireless LoRa network is currently influenced by external factors, such as indoor coverage, which can be improved by installing an amplifier for home monitoring.

Patient Experiences

Patient acceptance was very high in this study, showing that patients were content to participate in this type of care after discharge. This may be explained by the limited time investment required of patients for the intervention and the additional patient-nurse interaction provided by telephone consultations.
Furthermore, the majority of patients also indicated that they would not mind being discharged earlier if they received this remote home monitoring intervention as standard care, which shows this type of home monitoring intervention could be incorporated into early discharge protocols. In addition, when these vital sign measurements are recorded and submitted by the patients themselves instead of by automated technology, patients may participate more actively in recovery management, which is associated with better outcomes [31-33] and may encourage a more sustainable, healthier lifestyle [11,31]. On the other hand, contextual factors, such as circadian rhythm and patient mobilization, are not taken into account with single measurements made by patients themselves. Use of a wearable sensor for continuous monitoring may provide a more holistic impression of activity and patients’ circadian rhythm, but the significant associated cost of the sensors must be considered [14].

Comparison With Other Work

Previous literature on this topic is scarce, but our results on intervention performance and compliance are in line with 2 previous home monitoring interventions that were comparable to ours [23,34]. Our findings are in contrast with a previous systematic review showing that remote monitoring interventions using mobile health were associated with improved surveillance, earlier detection of complications, and more timely interventions, preventing further clinical decline [11]. However, the designs of the postoperative home monitoring interventions that were reviewed were highly variable, and the interventions did not use continuous vital sign measurements with wearable devices. With regard to patient acceptability and satisfaction, previously reported comparable interventions found similarly high rates [23,34,35].

Strengths and Limitations

This is one of the first studies to examine the feasibility of a home monitoring intervention for colorectal surgery patients using an available certified wearable sensor, software, and infrastructure. However, when interpreting the findings of this study, some limitations should be considered. First, although our primary aim was to determine feasibility, the generalizability of our results is limited because of the study design and relatively small sample size. Research on adequate and timely detection of clinical deterioration with remote home monitoring interventions (potentially preventing readmissions) requires large cohorts to assess the true benefits. Second, assessment of continuous vital sign trend data specifically for perioperative patients was a novel method for the monitoring care professionals, which may have influenced the adequacy of the assessments, even though the care professionals were specialized and trained in remote patient monitoring. Third, we used a self-developed questionnaire to assess patient experiences, since a validated questionnaire suitable for this type of intervention was not available. Finally, we did not include cost as an end point in this feasibility study. An early assessment of costs may be relevant for future optimization of such interventions, especially as wearable sensor technology is still associated with significant up-front investments and device cost.

Future Directions

Although these are promising initial results, further studies on the design of a remote home monitoring intervention for this patient category are needed before the true value for early discharge protocols, prevention of readmissions, and patient outcomes can be adequately determined. First, future research is needed to explore the needs of patients and health care professionals for each care pathway regarding telephone consultations and vital sign measurements. Consideration should be given to replacing telephone consultations with indirect digital contact through a mobile app for monitoring and coaching to achieve even more efficient care. Subsequently, the expansion of vital sign measurements to include other parameters, such as body temperature and mobility, may be relevant to capture the full status of the patient. In addition, the true added value of continuous versus intermittent vital sign measurements and the optimal frequency and timing of teleconsultations need to be determined for each patient category. Notifications generated by a combination of vital sign measurements and symptoms could assist monitoring professionals in clinical decision-making. Ultimately, optimal home monitoring protocols may eventually be personalized depending on surgery type, postoperative course, and individual patient characteristics, whereby deviating vital sign trend detection will be automated by algorithms that will support the monitoring department staff.

Conclusion

We found that a remote home monitoring intervention for colorectal surgical patients with wearable wireless sensors and telephone consultations was feasible, considering the high intervention performance and high patient acceptability. However, the design of a remote home monitoring intervention for this patient category should be further optimized before its true value for early discharge protocols, prevention of readmissions, and patient outcomes can be adequately determined.

Acknowledgments

The authors would like to thank all patients and the nurses at the Medical Coordination Office for participation in the study and Philips Healthcare and Isala I&I for technical preparation. The study was funded by the hospital Innovation & Science Funds (grant INNO2016).

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.
Conflicts of Interest
None declared.

Multimedia Appendix 1
The Philips Healthdot wearable sensor.
[DOCX File, 64 KB - periop_v6i1e45113_app1.docx]

Multimedia Appendix 2
Thresholds of the partial Early Warning (D-EWS) scores.
[DOCX File, 14 KB - periop_v6i1e45113_app2.docx]

Multimedia Appendix 3
Patient questionnaire.
[DOCX File, 17 KB - periop_v6i1e45113_app3.docx]

Multimedia Appendix 4
Case description of readmissions.
[DOCX File, 597 KB - periop_v6i1e45113_app4.docx]

Multimedia Appendix 5
Remarks of patients at the questionnaire (translated from Dutch).
[DOCX File, 13 KB - periop_v6i1e45113_app5.docx]

References


Abbreviations

ASA: American Society of Anesthesiologists
D-EWS: Deel Early Warning Score
ERAS: Enhanced Recovery After Surgery
HR: heart rate
IGS: Intellivue Guardian Solution
LoRa: long range, low power Internet of Things
RR: respiratory rate

©Jobbe P L Leenen, Vera Ardesch, Gijsbert Patijn. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 05.05.2023. 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.
Reducing Alcohol Use Before and After Surgery: Qualitative Study of Two Treatment Approaches

Lyndsay Chapman¹, MPH; Tom Ren², BS; Jake Solka¹, MS; Angela R Bazzi³,⁴, MPH, PhD; Brian Borsari⁵,⁶, PhD; Michael J Mello⁷,⁸, MPH, MD; Anne C Fernandez¹, MA, PhD

¹Department of Psychiatry, University of Michigan Medical School, Ann Arbor, MI, United States
²College of Medicine, Central Michigan University, Mount Pleasant, MI, United States
³Herbert Wertheim School of Public Health, University of California, San Diego, San Diego, CA, United States
⁴Department of Community Health Sciences, Boston University School of Public Health, Boston, MA, United States
⁵Mental Health Service, San Francisco VA Medical Center, San Francisco, CA, United States
⁶Department of Psychiatry and Behavioral Sciences, University of California, San Francisco, San Francisco, CA, United States
⁷Department of Emergency Medicine, Rhode Island Hospital, Providence, RI, United States
⁸Department of Health Services, Policy and Practice, Brown University School of Public Health, Providence, RI, United States

Corresponding Author:
Anne C Fernandez, MA, PhD
Department of Psychiatry
University of Michigan Medical School
North Campus Research Complex
2800 Plymouth Rd, Bldg 16
Ann Arbor, MI, 48109
United States
Phone: 1 734 763 9600
Email: acfernan@umich.edu

Abstract

Background: High-risk alcohol use is a common preventable risk factor for postoperative complications, admission to intensive care, and longer hospital stays. Short-term abstinence from alcohol use (2 to 4 weeks) prior to surgery is linked to a lower likelihood of postoperative complications.

Objective: The study aimed to explore the acceptability and feasibility of 2 brief counseling approaches to reduce alcohol use in elective surgical patients with high-risk alcohol use in the perioperative period.

Methods: A semistructured interview study was conducted with a group of “high responders” (who reduced alcohol use ≥50% postbaseline) and “low responders” (who reduced alcohol use by ≤25% postbaseline) after their completion of a pilot trial to explore the acceptability and perceived impacts on drinking behaviors of the 2 counseling interventions delivered remotely by phone or video call. Interview transcripts were analyzed using thematic analysis.

Results: In total, 19 participants (10 high responders and 9 low responders) from the parent trial took part in interviews. Three main themes were identified: (1) the intervention content was novel and impactful, (2) the choice of intervention modality enhanced participant engagement in the intervention, and (3) factors external to the interventions also influenced alcohol use.

Conclusions: The findings support the acceptability of both high- and low-intensity brief counseling approaches. Elective surgical patients are interested in receiving alcohol-focused education, and further research is needed to test the effectiveness of these interventions in reducing drinking before and after surgery.

Trial Registration: ClinicalTrials.gov NCT03929562; https://clinicaltrials.gov/ct2/show/NCT03929562

(JMIR Perioper Med 2023;6:e42532) doi:10.2196/42532

KEYWORDS
alcohol use; brief intervention; surgery; preoperative; alcohol use disorder; alcohol; substance use; substance abuse; postoperative; perioperative; counseling; surgical
Introduction

High-risk alcohol use (often defined as 2 or more drinks per day [1], or a score of 5 or more on the Alcohol Use Disorders Identification Test Consumption [AUDIT-C] [2,3] screening tool) is a common preventable risk factor for postoperative complications, admission to intensive care, and longer hospital stays [2-6]. Importantly, alcohol-associated adverse surgical outcomes are not specific to certain surgeries or subpopulations; instead, they are evident across heterogeneous patients and surgery types [3,5]. Short-term abstinence from alcohol use (2 to 4 weeks) prior to surgery is linked to a lower likelihood of postoperative complications [7-13]. Likewise, abstinence of 5 to 6 weeks after surgery is recommended to reduce one’s risk of experiencing complications such as delayed wound healing, infection, and impaired cardiac function [11,14]. In addition, the majority of surgical patients receive an opioid prescription after surgery [15,16], for which concurrent alcohol use is dangerous and even lethal [17,18]. Despite these recommendations and research, elective surgical patients are rarely offered alcohol-focused assessment, education, intervention, or treatment referrals prior to surgery [6,19,20]. For example, one study found only 25% of studies of alcohol use in surgical patients used standardized alcohol assessments, and a qualitative study found surgical care providers often did not have interest in screening or discussing alcohol use with patients.

Currently, the field lacks rigorous research on brief interventions to reduce high-risk alcohol use prior to surgery. Existing preoperative interventions involve pharmacotherapy and frequent in-person visits to promote abstinence, making them most appropriate for individuals at the higher end of the alcohol use and alcohol use disorder spectrum [7,9]. A broader array of treatment options are still needed to address the full spectrum of patients drinking prior to surgery. To address this gap, we developed 2 brief counseling approaches—a low-intensity “brief advice” intervention and a higher-intensity “health coaching” intervention—to reduce preoperative alcohol use. Consistent with the different preferences and implementation needs of patients and clinical staff identified in our formative research, these 2 interventions varied in content, intensity, and modality [21].

In this paper, we describe the acceptability and feasibility of the brief advice and health coaching interventions from qualitative exit interviews with treatment “high responders” (who reduced alcohol use ≥50% postbaseline) and “low responders” (who reduced alcohol use by ≤25% postbaseline). In this way, we sought to gather contextual information and potential reasons for high versus low intervention response, including participants’ perspectives on the impact of specific intervention elements, modalities, and any unanticipated influences on their alcohol use related to surgery preparation or recovery.

Methods

Study Design and Sample

This paper reports results of qualitative exit interviews conducted as part of the Alcohol Screening and Preoperative Intervention Research (ASPIRE) study, a randomized pilot trial of 2 preoperative alcohol interventions [22]. ASPIRE recruited 51 participants between July 2019 and February 2021 (with a pause from April to July 2020 due to COVID-19–related reductions in elective surgeries) from a large academic medical center in the Midwestern United States. To be eligible for the parent trial, participants had to meet the following inclusion criteria: (1) be scheduled for elective or semielective surgery in the next 35 to 120 days in select subspecialties (plastic, knee or hip arthroplasty, minimally invasive, endocrine, gynecology, urology, colorectal, or hepatobiliary); (2) receive regional or general anesthesia; (3) be aged 18 to 75 years; (4) have positive or unknown alcohol use in their social history in the electronic health record; and (5) meet criteria for high-risk alcohol use (AUDIT-C score ≥5). Selected surgical types included orthopedic, plastic, urology, gynecological, hepatobiliary, and general surgery. Individuals were excluded if they were undergoing surgeries requiring only local anesthesia, could not read or understand English, or had substantial cognitive impairment or evidence of psychotic symptoms (ie, delusions or hallucinations). Research staff contacted potential participants by phone (call or text). Eligible participants provided written informed consent for all trial procedures including qualitative exit interviews. Randomization into the brief advice or health coaching intervention conditions was stratified based on sex and alcohol screening scores.

Ethics Approval

The University of Michigan’s Institutional Review Board approved all study protocols (HUM00156743). This is a registered clinical trial (ClinicalTrials.gov NCT03929562).

Study Interventions

Health Coaching

The health coaching intervention consisted of a trained, graduate-level health coach delivering 2 telehealth intervention sessions 4 and 2 weeks prior to surgery. Participants had the choice of taking part in-person (prepandemic only) or through teleconferencing compliant with the US Health Insurance Portability and Accountability Act. Sessions lasted 45 minutes each. Intervention content was based on principles of health coaching, an evidence-based, collaborative care approach that partners patients with coaches who teach strategies to self-manage health behavior. Key features include motivational interviewing, goal setting, and discussion of strategies for alcohol use reduction or cessation [23]. Our health coaching sessions also included a 4-page visual and text-based session guide (emailed to participants in advance) that included health education, personalized feedback, and tailored surgical risk messaging based on the health belief model [24,25]. Health coaching sessions introduced the concept of “prehabilitation,” which involves improving health through behavior change before surgery in order to help improve postoperative outcomes.
This message framing emphasized two main health belief factors: (1) the desire to avoid a health problem (ie, surgical complications) and (2) the belief that a given behavior (ie, alcohol cessation) can prevent the problem. Coaching conversations were designed to increase motivation to reduce pre- and postoperative alcohol use. These sessions allowed participants to reflect on and discuss their alcohol use and produce personal perioperative health goals.

**Brief Advice**

The brief advice intervention consisted of one 10-minute phone session led by the same graduate-level health coach 4 weeks prior to surgery. The phone call was accompanied by a 2-page handout (mailed to participants prior to the session) that included information about participants’ reported alcohol use from baseline surveys, educational information about alcohol and surgical health, and advice to stop alcohol use for 4 weeks prior to surgery and for 6 weeks after surgery. Treatment resources and alcohol withdrawal information were included in the handout.

**Quantitative Assessment and Responder Classification**

The ASPIRE pilot trial involved 3 quantitative assessment time points: at baseline and 1- and 4-months postbaseline. After the final (4-month) follow-up, treatment response and low response were assessed by calculating percent change in alcohol use at follow-up time points relative to baseline.

**Alcohol Use**

Changes in alcohol use were assessed at baseline and follow-up visits using the AUDIT-C [26] and a web-based version of the Timeline Follow Back (TLFB) [27]. The 3-item AUDIT-C assesses alcohol quantity and frequency with an overall score ranging from 0 to 12 (higher scores reflect higher levels of alcohol use). The TLFB uses a calendar format on which participants self-report the number of standard drinks (approximately 14 g of alcohol) consumed each day to yield measures of average drinks per week and percent of days abstinent. Our assessment time frames included the past 3 months for baseline and final follow-up visits, and past month at the 1-month follow-up visit.

**Responder Classification**

Those who reduced alcohol use more than ≥50% relative to baseline on both the AUDIT-C and TLFB (average drinks per week) were classified as high responders. Those who reduced alcohol use by ≤25% relative to baseline on both the AUDIT-C and TLFB were classified as low responders. Percentage cutoffs were data driven, chosen based on the distribution of change in alcohol use at follow-up. We used 2 measures to classify responder status to ensure the highest level of certainty in our responder and low-responder classification. These 2 measures are highly correlated but also provide slightly different alcohol use data for decision-making.

**Qualitative Data Collection**

Following completion of the final follow-up assessments at 4 months for the ASPIRE pilot trial, we invited participants to take part in qualitative exit interviews based on study condition (brief advice and health coaching) and treatment response. Our goal was to enroll approximately at least 25% of the trial sample of N=51. Of the 26 participants invited based on these criteria, 19 agreed and completed exit interviews.

A trained graduate-level interviewer conducted qualitative exit interviews via the internet (through Health Insurance Portability and Accountability Act–compliant teleconferencing) using a semistructured interview guide containing open-ended questions and detailed probes designed to explore how the interventions (or low-intervention factors) may have influenced alcohol and other substance use during the perioperative period (see Multimedia Appendix 1 for the guide). Questions also explored participants’ perspectives on the acceptability (ie, whether the sessions and content were appropriate, appealing, and impactful) and feasibility (ie, whether participation was convenient, clear, and achievable), as well as how comfortable they felt sharing accurate and honest information about substance use with study staff and medical providers. Study staff regularly reviewed transcripts against audio recordings to confirm accuracy and deidentification.

**Qualitative Coding and Data Analysis**

We undertook thematic analysis of transcribed data that began with a collaborative codebook development process involving 4 research team members (the principal investigator, study coordinator, research assistant, and a lead qualitative investigator) [28-30]. First, key topics of interest were independently reviewed from the interview guide and several selected transcripts to develop a preliminary list of potential codes and definitions [31]. We met to discuss and refine this preliminary list, which we compiled into a draft codebook. Interviews were audio-recorded and transcribed verbatim by a professional transcription company that omitted potentially identifiable information. The draft codebook was then independently tested using another set of transcripts and the team members met again to discuss and refine the codebook.

Through 3 additional rounds of this iterative codebook development and testing process, discrepancies in code application were identified and resolved; codes were merged, added, and removed; and code definitions were revised as necessary until consensus was reached on a final codebook. To validate the final codebook and assess consistency in code application, 2 core coders (the study coordinator and research assistant) double-coded 2 transcripts using MAXQDA (VERBI Software). After assessing and determining that there was a high degree of consistency, we divided up the remaining transcripts between the 2 coders, who then independently coded the remaining transcripts under the supervision of the principal and lead qualitative investigators. We continued holding weekly meetings to review coding progress, discuss emergent themes, and identify and clarify preliminary findings [32]. To delve deeper into emergent themes that were identified during an initial round of coding (described below), midway through data collection, new questions were added on COVID-19 experiences, and intervention changes that emerged in earlier interviews were suggested. For our thematic analysis for this paper, we focused on understanding intervention acceptability and feasibility and potential reasons for intervention response. The transcripts across these domains were then reviewed to
identify key themes, which are described below and exemplified using anonymized quotes.

**Results**

**Overview**

Among 19 participants, 9 (47%) were female, 9 (47%) were from the health coaching intervention condition, 10 (53%) were from the brief advice condition, 9 (47%) were considered high responders, and 10 (53%) were considered low responders (Table 1). From our analysis of interview data, we identified the following three themes: (1) the intervention content was novel and impactful, (2) the choice of intervention modality enhanced participant engagement in the intervention, and (3) factors external to the interventions also influenced alcohol use. These themes are detailed below and summarized in Table 2.

<table>
<thead>
<tr>
<th>Female sex, n (%)</th>
<th>Brief advice (n=10)</th>
<th>Health coaching (n=9)</th>
<th>Total (n=19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifies as woman, n (%)</td>
<td>4 (40)</td>
<td>5 (56)</td>
<td>9 (47)</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>47.5 (9.75)</td>
<td>58 (14)</td>
<td>50 (17)</td>
</tr>
<tr>
<td>White, n (%)</td>
<td>10 (100)</td>
<td>8 (89)</td>
<td>18 (95)</td>
</tr>
<tr>
<td>Non-Hispanic, n (%)</td>
<td>10 (100)</td>
<td>9 (100)</td>
<td>19 (100)</td>
</tr>
<tr>
<td>Low responder, n (%)</td>
<td>5 (50)</td>
<td>5 (56)</td>
<td>10 (53)</td>
</tr>
</tbody>
</table>

**Surgical category, n (%)**

| Orthopedics | Brief advice (n=10) | 2 (20) | 5 (56) | 7 (37) |
| Plastic | 4 (40) | 3 (33) | 7 (37) |
| Minimally invasive | 2 (20) | N/Aa | 2 (11) |
| Otherb | 2 (20) | 1 (11) | 3 (16) |

**Drinks per week, mean (SD)**

<table>
<thead>
<tr>
<th>Brief advice (n=10)</th>
<th>Health coaching (n=9)</th>
<th>Total (n=19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.4 (12)</td>
<td>9.3 (14.8)</td>
<td>9.6 (14.2)</td>
</tr>
</tbody>
</table>

**Average AUDIT-C score, mean (SD)**

<table>
<thead>
<tr>
<th>Brief advice (n=10)</th>
<th>Health coaching (n=9)</th>
<th>Total (n=19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.9 (1.1)</td>
<td>6.1 (1.1)</td>
<td>6.0 (1.1)</td>
</tr>
</tbody>
</table>

**Tobacco use, n (%)**

<table>
<thead>
<tr>
<th>Brief advice (n=10)</th>
<th>Health coaching (n=9)</th>
<th>Total (n=19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (10)</td>
<td>5 (56)</td>
<td>6 (32)</td>
</tr>
</tbody>
</table>

**Marijuana use, n (%)**

<table>
<thead>
<tr>
<th>Brief advice (n=10)</th>
<th>Health coaching (n=9)</th>
<th>Total (n=19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (20)</td>
<td>4 (44)</td>
<td>6 (32)</td>
</tr>
</tbody>
</table>

**Prescription opioid use, n (%)**

<table>
<thead>
<tr>
<th>Brief advice (n=10)</th>
<th>Health coaching (n=9)</th>
<th>Total (n=19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (10)</td>
<td>3 (33)</td>
<td>4 (21)</td>
</tr>
</tbody>
</table>

**Other drug usec, n (%)**

<table>
<thead>
<tr>
<th>Brief advice (n=10)</th>
<th>Health coaching (n=9)</th>
<th>Total (n=19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (20)</td>
<td>1 (11)</td>
<td>3 (16)</td>
</tr>
</tbody>
</table>

aN/A: not applicable.
bUrology, endocrine, and gynecology.
cAUDIT-C: Alcohol Use Disorders Identification Test–Consumption.
dAmphetamines and hallucinogens.
Table 2. Summary of qualitative findings by theme and responder status.

<table>
<thead>
<tr>
<th>Theme</th>
<th>High responders (n=9)</th>
<th>Low responders (n=10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention content</td>
<td>Perceived information as novel and personally relevant for improving their surgical health.</td>
<td>Information is novel and believable but perceived as only relevant to others who have a more serious “alcohol problem.”</td>
</tr>
<tr>
<td></td>
<td>Changed risk perceptions related to alcohol use and surgery.</td>
<td>The information about alcohol and surgical risk was not shared before past surgeries and therefore inconsistent.</td>
</tr>
<tr>
<td></td>
<td>Opportunity for longer discussions with a nonjudgmental health coach was important (health coaching only).</td>
<td></td>
</tr>
<tr>
<td>Intervention modality</td>
<td>Participants reported intervention was an acceptable modality and length.</td>
<td>Phone modality limited personal connection, and length was too short to discuss a sensitive topic like alcohol use (brief advice only).</td>
</tr>
<tr>
<td>Other influential factors: study assessments</td>
<td>Completing study assessments, which included completing a daily calendar of alcohol use increased awareness and motivation to change.</td>
<td>Completing study assessments increased awareness of alcohol use and motivation but intervention did not (brief advice only).</td>
</tr>
<tr>
<td>Other influential factors: intervention language</td>
<td>No thematic findings.</td>
<td>The term intervention caused negative reactions and contributed to participants thinking their alcohol use was not heavy enough for study participation.</td>
</tr>
<tr>
<td>Other influential factors: surgery factors</td>
<td>Surgery preparation and recovery influenced choice to reduce or stop drinking for a period of time independent of the intervention content.</td>
<td>Less intensive surgery and shorter healing time linked to faster return to alcohol use.</td>
</tr>
</tbody>
</table>

**Intervention Content Was Novel and Impactful**

Overall, participants found both intervention conditions to be acceptable, of appropriate length, and impactful. Several participants stated that the intervention directly influenced their perioperative alcohol consumption: “It was very enjoyable. It was easy...there hasn’t been a thing that’s made me uncomfortable or that I haven’t agreed with [in] the study.” In general, participants reacted positively to the information presented in both intervention conditions. Participants considered the information regarding the risks of preoperative alcohol use on surgical outcomes to be novel (yet believable) and motivating. By receiving this information, participants described feeling more prepared for surgery and “in control” of their surgical outcomes. As one health coaching intervention high responder explained:

*I really did appreciate having that opportunity to say, “This is something that’s a big deal for my body.” Anything I can do to ensure the success of the surgery is something I want to do. Then, I’m talking about [the study] with a lot of people, saying, “I never realized this, but if you are going through something medically, it’s a really good idea to get yourself as ready as possible.”*[55 years old, female]

Several participants commented on what they learned about the connections between alcohol use and the body’s ability to heal, including alcohol’s impact on the immune system. In considering what “stood out” from the brief advice condition, one high responder explained:

*I’ve never had any sort of information beyond just, you know, the day of the surgery, “Don’t eat anything before or after midnight.” You know, the general pre-surgical conditions to get through the day. But this [intervention] made an attempt to look out for my bigger picture health for a longer period of time before and after the surgery.* [47 years old, male]

Other high responders generally perceived the information shared as part of the intervention as impactful, identifying it as one of the main reasons they reduced their preoperative alcohol use. One brief advice high responder (48 years old, male) stated “If I was not a part of the study or aware of the study, I probably would not have reduced [my drinking].” Health coaching participants described how conversations with a nonjudgmental health coach provided them an opportunity to carefully consider changing their alcohol use.

Both interventions presented information about alcohol use in the preoperative period that some participants found helped them make a “connection” between alcohol use and health, ultimately changing their personal risk perceptions. Participants found the concept of prehabilitation, or optimizing health prior to surgery, to be particularly impactful because the health coach encouraged them to view themselves as important members of their surgical teams who could enhance their own health, as opposed to passively receiving surgery. As such, they felt an improved sense of control over their surgical outcomes. As one health coaching high responder explained:

*Participation gave me more control over the anxiety having to do with the surgery because I’m a person who likes to control things, and it just gave me a sense of control over one more aspect of the surgery that I could do the best I could do to have success.* [55 years old, female]
Among low responders in both intervention conditions, there was a sense that the information presented was true but did not apply to them personally. This was particularly true for participants on the lower end of the AUDIT-C eligibility assessment, who did not view their drinking habits as risky. These individuals felt they had been included in the study erroneously, with one health coaching low responder (71 years old, female) saying, “it just didn’t feel like a good fit for me.” A brief advice low responder (47 years old, male) further elaborated that he thought the content was “focused on people that drink a lot all the time...while I don’t have a problem having a beer or two on the weekends, it’s really not an issue for me.”

Some participants who had undergone major surgeries in the past described never receiving information about alcohol use and related surgical risks. Accordingly, some participants mentioned feeling surprised that their doctors had not communicated this kind of information to them in the past. One brief advice low responder (22 years old, male) explained that he would have appreciated receiving better information from his surgical team given the significance of the topic, saying, “I've had two surgeries in the past on my knee...these issues literally never came up. I was never briefed on what complications might arise around alcohol or tobacco use, so that was new.”

**The Choice of Intervention Modality Enhanced Participant Engagement in the Intervention**

By design, intervention modality differed between conditions. While the brief advice condition was only delivered by phone, health coaching participants were given a choice between phone or face-to-face sessions (by videoconferencing or, prior to COVID-19, in-person). Overall, it appeared that having some choice in intervention modality increased participants’ satisfaction; while nearly all health coaching participants were satisfied with their modality of choice, brief advice participants, who had no choice, indicated mixed experiences and dissatisfaction with phone participation.

Some health coaching participants who chose face-to-face intervention delivery expressed high satisfaction with this modality, explaining that they felt more engaged with the information presented, took intervention recommendations more seriously, and were more open and honest because it felt more “personal.” As one health coaching high responder (58 years old, male) explained, “You’re talking to somebody, a human, and it just makes you want to participate and follow the rules, so [there’s] more accountability.” Most face-to-face sessions involved videoconferencing, which participants could schedule around their work and personal lives; several noted that this flexibility enabled them to save time and energy compared with traveling to in-person sessions. They were also able to participate in a setting of their choice, which in addition to being convenient, was more comfortable. Instances of dissatisfaction included one participant stating the sessions were “a bit too long” and another participant who would have preferred in-person sessions, but this was impossible due to COVID-19 pandemic restrictions at that time.

In contrast, brief advice participants’ experiences with phone sessions were more mixed, possibly because they lacked any choice in intervention modality. While some participants discussed specific advantages of phone sessions (eg, greater ease of use and convenience over videoconferencing or traveling to an in-person appointment), others felt they would have been more comfortable face-to-face, with one participant (48 years old, male) saying he would be “a lot more engaged [having] a video in front of me and [knowing] I’m being watched at the same time.” Furthermore, they perceived a poorer connection between themselves and the interventionist, viewing the phone modality as “un-intimate.” As a brief advice low responder explained:

[I was] probably not as comfortable, for whatever reason. It was a phone conversation. I was in a parking lot during work, in between meetings, in my car. I guess the setting wasn’t super great, and now I’m at home [during qualitative interview]. I’m sitting down, I’m not on my phone. Those are all factors. [48 years old, female]

Most felt that the length of health coaching sessions (~45 minutes) was appropriate and afforded them sufficient time for discussion without feeling rushed while also showing respect for their time. In contrast, while most participants in the brief advice condition felt that the session length was acceptable, some felt the length was too short to feel comfortable opening up about the sensitive topic of alcohol use.

**Factors External to the Interventions Also Influenced Alcohol Use**

There were several factors external to the intervention content that influenced participants’ alcohol consumption, including (1) self-report and tracking of alcohol use for the ASPIRE trial assessments, (2) “intervention” language used in the study, and (3) surgical experiences.

**Self-Report and Tracking**

Assessments for the ASPIRE study included self-reported alcohol consumption using a 3-month retrospective TLFB calendar procedure, which affected participants in 2 ways. First, high responders described that knowing the study team would review their data and goals promoted a sense of “accountability” and not wanting to be caught “breaking the rules” by drinking at higher levels. One health coaching participant (63 years old, male) explained “I think setting the goals and following through, knowing that someone was gonna get it at the end of the time period.... That helped a lot.” Second, self-report and tracking gave some participants an improved awareness of their drinking habits by enabling them to visually recognize behaviors they wanted to change. Some discussed how the calendar helped reveal patterns they had not noticed before, as one brief advice low responder (43 years old, female) explained: “When you look at that calendar, after you fill it out and, go, oh my God, I drank every day for two weeks...there’s no need to be doing it every day. It just seems excessive.” This was especially true for those in the brief advice condition, as study assessment activities, which totaled ~2.5 hours across the duration of the ASPIRE study, were proportionally much longer than the ~10-minute intervention session. As one brief advice high responder (43 years old, female) explained, “I don’t think the
intervention content was that impactful. I think it was more of just the [activities], doing the survey, kind of as a journal, that had more of an impact on me.”

“Intervention” Language Used in the Study
Some participants reacted negatively to language used during recruitment and study implementation. Study materials initially referred to both conditions as “interventions,” which implied to participants that they had an alcohol-related “problem” significant enough to require help. As one health coaching low responder (71 years old, female) stated, “I mean, it’s hard not to feel defensive when somebody says, ‘Oh, well, you’re being put in an intervention.’” For some participants, this term carried negative connotations leading them to question whether they were a “good fit” for the study, particularly if they did not see themselves as having a “problem.” As one health coaching low responder (71 years old, female) explained:

“I was a little taken aback by this sense of, oh my God, you know, an “intervention.” Like, where did that come from, you know? This is not something that’s ever been a particular problem or concern, so it was just a little bit of a surprise... And all of a sudden, I’m in this category like, “You need an intervention.” I was like, “What? Excuse me? Hello? No, don’t think so.” But anyway, I was mildly taken aback by all that.

Surgical Experiences
Most study participants viewed their surgeries as an important life event; some further recognized surgery as dangerous and requiring careful preparation and viewed alcohol consumption as harmful for surgical recovery even before the intervention. For these individuals, reducing alcohol use prior to surgery seemed intuitive, and they believed they would have done so even without the intervention. One brief advice high responder explained:

“I know [surgery is] dangerous, just baseline dangerous, so the better health you can be in, the better the surgery will go and the recovery. I guess I was already at that point [of wanting to be healthier for the surgery], so I was—I don’t think [the intervention] had any impact. [47 years old, male]

Postoperatively, recovery prevented many participants’ alcohol consumption for at least some period due to not feeling well enough to eat, drink, or ambulate. As one brief advice high responder explained:

“Half the time, I didn’t feel like even eatin’ or doin’ anything, so that might not be—it might not be the study that made me do what I did. It might have been the pain level and other factors. [47 years old, male]

Others intuitively recognized that drinking could negatively impact their recovery and believed they would have reduced their alcohol consumption even without the intervention. Some participants also decided to abstain from alcohol beyond 6 weeks after surgery to continue improving their health. Others recognized that mixing alcohol and opioid-based medications is dangerous and chose to abstain from alcohol in the postoperative period for such reasons. Those with easier, shorter recovery periods did not view surgery as having a large impact on their decision to drink alcohol one way or another. Low responders began to drink alcohol as soon as they felt better from surgery, even within 6 weeks, as one brief advice participant explained:

“[Alcohol] was non-existent the first few days after [the surgery], and then it picked back up shortly after...once I was off [OxyContin], then I was drinking. [43 years old, female]

Discussion
Principal Findings
This study describes the experiences of participants in a preoperative pilot trial of alcohol-focused interventions of varying intensities and modalities. Participants in our qualitative exit interviews were selected to represent high responders and low responders to help better understand potential effective and ineffective elements of the interventions to inform a future sequential multiple assignment randomized controlled trial, assigning participants to various treatment options based on initial treatment response. Overall, high responders in our pilot trial described intervention content as novel and relevant. The health coaching condition, which involved motivational interviewing and goal setting, provided an opportunity to reflect on alcohol use in an open, nonjudgmental conversation with supporting materials that framed the patients’ roles as active, thereby empowering them to take actions to improve their surgical outcomes. Surgery preparation and recovery appeared to motivate change in alcohol use for participants, with the intervention content helping some participants link their alcohol use to surgical health for the first time. While most participants found both interventions to be novel and motivating, a few believed the link between alcohol use and surgical health was intuitive and that they would have stopped or reduced alcohol use before and after surgery regardless of the intervention. Likewise, longer surgical recovery and prolonged medication use led to longer alcohol cessation for some participants.

Though all participants met “risky” alcohol use eligibility requirements for the study using a validated alcohol use screening and risk identification tool, many did not perceive themselves as drinking enough to put them at risk, especially in the low-responder group. Instead, some of these participants described dismissing the intervention content about alcohol use and surgical risk because they felt it did not apply to them. A modified approach that uses different terminology or alternative discussion points could help address this barrier to change. This could include acknowledging that some participants may be surprised by feedback suggesting their level of alcohol use could be linked to surgical risks and discussing their reactions to this information.

There were also aspects of study participation and surgery that influenced participants’ alcohol use separate from the intervention content. There was evidence of assessment reactivity, particularly for participants in the brief advice condition for whom the baseline and follow-up study assessments represented a much larger proportion of their participation time than the actual intervention itself. Being able
to visually see patterns of alcohol use on our assessment calendar was illuminating and motivating for some, which is consistent with research finding that the simple act of measuring behaviors and comparing them to external standards or goals can result in lasting behavior change [33], with larger effects on goal attainment when outcomes are physically tracked and recorded [34].

Importantly, several participants reacted negatively to the term intervention, which was used to describe study treatment conditions in several study-related documents and contact scripts. For these individuals, the term intervention triggered negative associations with confrontational Johnson-style interventions [35] during which a person with an addiction is confronted by family and friends and sent to long-term rehabilitation, sometimes against their will. After learning of this unintended association, and the negative reaction across several participants, in July 2020, we removed this term from study documents. For example, in the consent form, we changed references from study interventions to study conditions or treatments. In addition, using prehabilitation program language could also reduce stigma by indicating the program’s goals are to optimize a patient’s health for surgery and subsequent recovery. In our next study phase, we will engage community members in reviewing all our study documents prior to recruitment for clarity and interpretation.

Limitations and Future Directions
This study has several limitations. Recruitment was paused due to the COVID-19 pandemic and the resulting cancellation of elective surgeries. When surgeries and recruitment resumed at a reduced capacity, it is possible that surgeries and participants were different in significant ways from those scheduled prior to the pandemic. Interviews were conducted by a researcher on the study team, which may have increased the risk of social desirability bias; however, it was a team member who had not been in contact with participants prior to the interview (ie, not the health coach or recruiter), and we reminded participants that the purpose of the interview was to inform intervention improvement, and they were encouraged to provide honest and critical feedback. Finally, while qualitative data can help deepen understanding of participants’ intervention experience, results from this study should not be considered generalizable to broader populations. The next step in research includes a fully powered sequential multiple assignment randomized trial that will randomly assign participants to treatment at baseline and then reassign them to different treatments and or treatment intensities based on initial treatment response and low response.

Conclusions
In conclusion, through in-depth interviews with a select sample of participants in the ASPIRE pilot trial, we found that both health coaching and brief advice conditions were acceptable pending some minor improvements and modifications. Further research is needed to test the effectiveness of these interventions in reducing drinking before and after surgery, both in the short and long term. Future research and programmatic work on these topics should carefully consider the use of lay versus technical (ie, psychiatric) terminology (eg, intervention) to avoid unintended negative reactions. Future research should also investigate whether assessments and feedback alone, without being coupled with brief interventions or health coaching, may be sufficient in changing participant alcohol use behaviors before or after surgery.

Acknowledgments
This study was funded by a grant from the National Institute on Alcohol Abuse and Alcoholism (K23AA023869) to ACF as the principal investigator. ARB’s effort was supported by a career development award from the National Institute on Drug Abuse (K01DA043412).

Conflicts of Interest
None declared.

Multimedia Appendix 1
Patient qualitative interview guide.
[DOCX File, 56 KB - periop_v6i1e42532_app1.docx ]

References


Abbreviations

ASPIRE: Alcohol Screening and Preoperative Intervention Research
AUDIT-C: Alcohol Use Disorders Identification Test Consumption
TLFB: Timeline Follow Back

©Lyndsay Chapman, Tom Ren, Jake Solka, Angela R Bazzi, Brian Borsari, Michael J Mello, Anne C Fernandez. Originally published in JMIR Perioperative Medicine (http://periop.jmir.org), 26.07.2023. 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.
Background: Anesthesiologists require an understanding of their patients’ outcomes to evaluate their performance and improve their practice. Traditionally, anesthesiologists had limited information about their surgical outpatients’ outcomes due to minimal contact post discharge. Leveraging digital health innovations for analyzing personal and population outcomes may improve perioperative care. BC Children’s Hospital’s postoperative follow-up registry for outpatient surgeries collects short-term outcomes such as pain, nausea, and vomiting. Yet, these data were previously not available to anesthesiologists.

Objective: This quality improvement study aimed to visualize postoperative outcome data to allow anesthesiologists to reflect on their care and compare their performance with their peers.

Methods: The postoperative follow-up registry contains nurse-reported postoperative outcomes, including opioid and antiemetic administration in the postanesthetic care unit (PACU), and family-reported outcomes, including pain, nausea, and vomiting, within 24 hours post discharge. Dashboards were iteratively co-designed with 5 anesthesiologists, and a department-wide usability survey gathered anesthesiologists’ feedback on the dashboards, allowing further design improvements. A final dashboard version has been deployed, with data updated weekly.

Results: The dashboard contains three sections: (1) 24-hour outcomes, (2) PACU outcomes, and (3) a practice profile containing individual anesthesiologist’s case mix, grouped by age groups, sex, and surgical service. At the time of evaluation, the dashboard included 24-hour data from 7877 cases collected from September 2020 to February 2023 and PACU data from 8716 cases collected from April 2021 to February 2023. The co-design process and usability evaluation indicated that anesthesiologists preferred simpler designs for data summaries but also required the ability to explore details of specific outcomes and cases if needed. Anesthesiologists considered security and confidentiality to be key features of the design and most deemed the dashboard information useful and potentially beneficial for their practice.

Conclusions: We designed and deployed a dynamic, personalized dashboard for anesthesiologists to review their outpatients’ short-term postoperative outcomes. This dashboard facilitates personal reflection on individual practice in the context of peer and departmental performance and, hence, the opportunity to evaluate iterative practice changes. Further work is required to establish their effect on improving individual and department performance and patient outcomes.
KEYWORDS
quality improvement; feedback; anesthesiologists; patient reported outcome measures; data display; user-centered design; surgical outcome; discharge; anesthesiology; postoperative care; registry; dashboard; interactive; practice; performance; patient outcome; mobile phone

Introduction
Anesthesiologists benefit from receiving feedback on their patients’ outcomes and can use it to evaluate and improve their practice. The perioperative period is a data-rich environment with the potential for innovation through digital health tools and predictive analytics. Data-driven performance feedback can improve perioperative practice and outcomes [1,2], including antibiotic administration [3,4], drug costs [5], operating room booking efficiency [6], and temperature monitoring compliance [7]. Feedback is most effective when it is locally relevant and derived from a credible source [8]. Personalized feedback, provided in (near) real time [3,9], is more effective than retrospective and department-wide feedback [8]. However, evaluating postoperative care metrics can be challenging, particularly for day-case procedures: access to this information is often only available from fragmented data sources, data are difficult to access, or are presented in user-unfriendly formats.

In anesthesia, practitioners commonly work in isolation (1 anesthesiologist per patient) and do not often have a chance to compare variations in individual practitioners’ anesthetic techniques and outcomes unless required by a critical or near-miss event. While providing care, anesthesiologists make multiple decisions influencing pain and nausea outcomes [10]. Longitudinal follow-up of patients is often limited to inpatients; in high turnover pediatric operating rooms, anesthesiologist may not have time to check in on a recovering patient before they are discharged home, and postanesthetic care unit (PACU) nursing does not routinely inform an anesthesiologist when administering ordered doses of analgesic or antiemetic rescue medications. Hence, an anesthesiologist may not know how their patients are faring in the postoperative period. This suggests that a system to collate and visualize data for comparative feedback may allow anesthesiologists to fine-tune their decisions, optimizing pain and nausea outcomes.

As part of an organizational quality initiative, a postoperative follow-up (POFU) registry has been established at BC Children’s Hospital, a tertiary pediatric hospital in Vancouver, British Columbia, Canada. Its purpose is to understand the recovery experience of day-case surgical patients and to facilitate associated quality improvement endeavors. The POFU registry is maintained by PACU clerks and nurses, who record day-surgery patient information and short-term outcomes from PACU and then follow-up with families via telephone to gather patient-reported outcomes at 24 hours post discharge. These data are recorded using the Research Electronic Data Capture (REDCap) web application (Vanderbilt University) [11,12] hosted locally. Each patient’s contact information, demographics, clinician information, and procedure characteristics are entered using operating room scheduling system data. Validation checks are enabled for each field to ensure minimal artifacts; the data steward runs further reports and alerts to optimize data quality. Family-reported 24-hour outcomes (including postoperative pain, nausea, and vomiting) are collected by nurse telephone follow-up beginning in September 2020. PACU opioid and antiemetic administration, indicating early treatment of postoperative pain and nausea, have also been captured since April 2021.

This study’s primary aim was to turn these outcome data into accessible and actionable information by creating dashboards, which allow the anesthesiologists to evaluate their patients’ postoperative outcomes and reflect on their care. We also aimed to evaluate anesthesiologists’ perception of the dashboard. By providing anesthesiologists with single-point access to these outcome data, allowing them to reflect on their care, drill down on details, and compare their performance to their peers and to the department aggregate in a time-efficient and user-friendly way, we aim to facilitate ongoing individual and departmental practices of improvement.

Methods

Study Design
We initially conducted a literature search of previous dashboard designs and partnered with a group of anesthesiologists in our department to identify their information needs and to co-design dashboards using an iterative development process. We then designed a dashboard architecture based on the POFU registry data; this incorporated key security features required to meet institutional policies and the confidentiality requirements of our anesthesiologists. We deployed the final design to the anesthesia department and conducted a preliminary usability evaluation.

Ethical Considerations
The University of British Columbia and Children’s and Women’s Health Centre of British Columbia Research Ethics Board determined this work to be established under a quality improvement or quality assurance (QIQA) framework (reviewed June 29, 2021), for which they do not require ethical review, in accordance with Article 2.5 of the Canadian Tri-Council Policy Statement. Data used in this study were obtained from the POFU registry, also established under a QIQA framework. This paper adheres to the Standards for Quality Improvement Reporting Excellence (version 2.0) guidelines [13].

Registry Data and Dashboard Architecture

Exploratory Data Analysis, Cleaning, and Processing
We performed an exploratory analysis in Python (Python Software Foundation) to confirm that the collected registry data were clean; variables were evaluated for any out-of-range values; anesthesiologist codes (departmental QIQA identifiers) and procedure codes were verified against lists of valid entries. Age, sex, and procedure group of PACU and missing 24-hour outcome data were compared to examine any significant differences in the underlying data. Cases with missing 24-hour outcomes or unanswered phone calls were excluded from the
24-hour outcome analysis but not censored from the PACU outcome analysis.

Nausea and vomiting data were collected using a 4-point scale (none, mild, moderate, or severe). Pain scores in PACU are collected using developmentally appropriate observational or self-report tools, and 24-hour pain scores are obtained by proxy from a parent or caregiver using a 0-10 numeric rating scale [14]. We then assigned the same 4-point labels to the numeric pain ratings obtained: 0=none, 1-3=mild, 4-6=moderate, and 7-10=severe. For most dashboard purposes, we dichotomized variables into none or mild versus moderate or severe.

**Dashboard Architecture and Data Sources**

Our dashboard design aimed to provide a self-service platform for anesthesiologists to confidentially review their performance through postoperative (PACU and 24-hour) outcomes and anonymized peer comparison (Figure 1). The dashboards were developed using business analytics software Power BI Report Server (Microsoft), hosted by the BC Children’s Hospital Research Institute (BCCHR). Power BI uses Active Directory Federation Services (Microsoft) to leverage hospital and research institute single sign-on and can, in principle, load data from REDCap, network drives, or other systems via application programming interfaces. At our institution, it is maintained by the BCCHR Data Management team, who manages access to team spaces in the server and determines access and usage policies.

Figure 1. The overview of the data flow from POFU data collection for PACU and 24-hour phone call into the POFU dashboard. PACU: postanesthetic care unit; POFU: postoperative follow-up.

Deidentified POFU registry data are downloaded weekly by the data steward as a comma separated values file and stored on a BCCHR secure data drive, that is, instead of employing an application programming interface token and directly linking these systems. BCCHR data management policies imposed the need for this workaround, as the POFU REDCap registry contains personal health information, parental contact information, and anesthesiologist and surgeon identifiers. The dashboards require 3 files: the POFU weekly data export (excluding any identifying data), a Power BI username-QIQA code mapping file (see dashboard security and confidentiality), and an auxiliary file containing outcome definitions (severity level and numeric score mapping), procedure definitions (procedure code and group mapping), and age categories (National Institute of Child Health and Human Development pediatric age categories) [15].

The POFU registry and auxiliary files are set as automated data sources in Power BI, automatically refreshing the dashboard whenever files are updated. In contrast, the mapping file was added manually as a data source by the POFU data steward and cannot be downloaded by any other team member. Power BI pulls new data from these sources every Monday morning via a scheduled refresh.

The dashboard is available to anesthesiologists through a web browser and is typically accessed on a desktop or laptop computer; it can be viewed on a tablet or a smartphone, though it has not been optimized for use in this way.

**Dashboard Security and Confidentiality**

Comparative data are required to contextualize practice patterns for postoperative outcomes, but our team of co-design anesthesiologists insisted that the security and confidentiality of both patient and provider data were imperative for...
dashboard support: patient confidentiality is maintained by excluding identifiers from the data accessed by the dashboard; provider confidentiality is maintained through the use of departmental QIQA codes for anesthesiologists; see dashboard architecture and data sources. Data for the active anesthesiologist (ie, the user accessing the dashboard) and the department’s aggregate performance are presented together, but access to the underlying data is restricted, preventing access to other anesthesiologists’ data.

Power BI’s row-level security feature filters the nonaggregated case data to include only data for cases the active anesthesiologist performed; hence, each user can only see their own cases. This is achieved via the Power BI username-QIQA code mapping. Access to this file is restricted to the POFU data steward, maintaining strict confidentiality of the anesthesiologists’ identifiers.

Dashboard Design and Evaluation

Iterative Co-Design Sessions

We performed a brief literature search of papers published in 2015-2021 to understand existing anesthesia dashboards; keywords included dashboard, run charts, personalized feedback, anesthesia feedback, and surgery feedback [7,16-23]. Our preliminary visualization designs were based on the ideas drawn from this literature search. Guided by the recommendation to incorporate users’ feedback into the dashboard development [16,19,23-26], we developed our system using a participatory design approach: we discussed design ideas with 5 anesthesiologists using iterative feedback to improve visualization. We adopted a convenience approach to selecting the design team, which consisted of the POFU clinical lead and other anesthesiologists who expressed an interest in contributing to the design process. Designs were demonstrated through screenshots and dynamic working prototypes during 6 feedback sessions; the visualizations were iteratively refined based on anesthesiologists’ comments and observing their use of the prototype dashboards (Figure 2). Each iterative feedback meeting on Zoom (Zoom Video Communications Inc) lasted approximately 1 hour.

The initial dashboard prototypes showed a given anesthesiologist’s average 24-hour postoperative pain, nausea, and vomiting scores and compared them with the department’s average. Outcome scale distributions were also plotted. A second design added PACU outcomes (opioid and antiemetic administration). These were shown to the anesthesiologists, and the visualizations were iteratively refined to arrive at the final set of dashboards.

After the iterative development process, dashboards were deployed to the department in September 2021; the work was presented at departmental rounds, and anesthesiologists were emailed links to the tool with instructions and a contact email for further information. Subsequently, 2 further feedback sessions were conducted with our collaborating anesthesiologists to refine the final dashboard designs.

Usability Survey

To gather clinicians’ feedback on the initial version of the dashboard, a usability survey was distributed to all department anesthesiologists in January 2022 after this initial version of the dashboard had been available for 4 months. It consisted of...
21 required and 18 optional follow-up questions on aspects such as frequency of use (including reason, if infrequent), ease of use (including clarity of information displayed and usefulness of the instructions), content (helpfulness and suggestions for other functionality), impact on practice, and overall opinion of the dashboard (Multimedia Appendix 1).

Based on the feedback received from the usability survey, the dashboards were redesigned and redeployed to the department in June 2022; the dashboards were again presented at department rounds, and anesthesiologist superusers were identified to provide training and peer support.

Dashboard Component Design

Overview

There are three dashboard components: (1) 24-hour outcome summary, (2) PACU outcome summary, and (3) anesthesiologists’ practice profile.

24-Hour Outcome Summary

Guided by previous reports [19,24-26], 2 sections were developed initially: (1) average severity scores of patient outcomes [24,26], and (2) run charts of average severity score (Figures 3A and 3D). Inspired by Parks et al [24], we initially designed a bar chart that displayed the anesthesiologists’ average severity scores in descending order, labeled with QIQA identifiers and the active anesthesiologist (user) highlighted (Figure 3A). Following feedback, we removed their codes to add a deidentification layer and added the severity score next to the bars (Figure 3B). The usability survey respondents indicated that the “pain severity score” calculation was unclear. Subsequently, we changed the metric from “pain severity score” (average of pain score categories) to “pain incidence rate” (occurrence of moderate or severe pain), which was more readily comprehensible (Figure 3C).

Figure 3. Progression of various sections of the dashboards with average severity scores in A, B, and pain incidence rates in C; monthly run charts of the anesthesiologist, along with team’s aggregate outcomes in D, E, and F.

Based on co-design input, the active anesthesiologist’s average severity score and the department’s aggregate were initially represented in a monthly run chart with a median score superimposed (Figure 3D). The anesthesiologists subsequently suggested a run chart for the department-level score and a bar chart for the active anesthesiologist’s score (Figure 3E). Finally, the title was modified to specify that the plot now reported incidence rates, and the colors were modified for consistency with other plots.

PACU Outcome Summary

The second phase of development added PACU outcome summary to the dashboards. PACU nurse interventions requiring opioid (both ≥1 dose and ≥4 doses) or antiemetic (≥1 dose) administration were plotted. The progression in design was similar to the 24-hour outcome dashboard. Monthly average PACU opioid or antiemetic administration rates of both the active anesthesiologist and the department were included in the final dashboards.

Anesthesiologists’ Practice Profile

A practice profile page was designed based on anesthesiologists’ interest in incorporating individual case mix information (age category, sex, and surgical service) into their interpretation of their outcome data.

Data Analysis

We summarized the 24-hour outcome data and PACU outcome data in the POFU database from the initial implementation of the data collection tool to illustrate the data available to anesthesiologists for visualization. We analyzed the anesthesiologists’ usability survey responses and presented the results descriptively, along with the final dashboard designs. Finally, we conducted a preliminary analysis of the impact of the dashboards on postoperative patient outcomes by comparing

https://periop.jmir.org/2023/1/e47398

JMIR Perioper Med 2023 | vol. 6 | e47398 | p.161

(page number not for citation purposes)
the baseline preimplementation period for PACU outcomes (April 2021 to June 2021) and 24-hour outcomes (September 2020 to June 2021) with the postimplementation period (September 2021 to February 2023) by plotting the month-by-month department incidence of each outcome and the overall average incidence for the period; to reduce bias, July 2021 to August 2021 have been plotted, but not considered in the comparison, as dashboard co-design was conducted during this period and was available to some members of the department. Changes between pre- and postimplementation periods were compared using the Wilcoxon rank-sum test.

Results

Data Set Characteristics
The 24-hour outcome dashboard contains data collected from September 2020 to February 2023. Of 12,082 total cases, 7877 (65.2%) postoperative phone calls were successfully completed to collect 24-hour outcomes from the family; 316 (4.0%) had moderate or severe pain, 73 (0.9%) had moderate or severe nausea, and 84 (1.1%) had moderate or severe vomiting. The distributions of age, sex, and PACU outcomes did not differ between patients with successful calls and those without. The PACU outcome dashboard contains 8716 cases collected from April 2021 to February 2023; 634 (7.3%) of these patients were administered at least 1 opioid dose, 78 (0.9%) required repeated opioid (≥4) doses, and 93 (1.1%) required an antiemetic.

Usability Survey
The January 2022 usability survey (Multimedia Appendix 1) was completed by 17 of 29 (59%) anesthesiologists, including 1 of 17 (6%) who had been practicing 6-11 years, 5 of 17 (29%) practicing 11-15 years, 6 of 17 (35%) practicing 16-20 years, and 5 of 17 (29%) practicing ≥20 years; respondents included 4 of 17 (24%) who performed cardiac anesthesia as part of their practice, 6 of 17 (35%) who performed anesthesia for neurosurgery, and 4 of 17 (24%) who performed anesthesia for spine surgery. Among all respondents, 15 of 17 had used the dashboard during the 4 months since the initial version had been deployed, though only 2 of 17 (12%) had used it regularly. Of those who had used the dashboard, 9 of 15 (60%) reported that it was easy to navigate, and 9 of 15 (60%) thought the information was clearly presented. On the other hand, 3 of 15 (20%) users had found the navigation difficult, 2 of 15 (13%) thought the information needed to be clearer, and only 7 of 15 (47%) had found the help text and user instructions helpful.

Overall, the information provided was considered helpful by 12 of 15 (80%) of those that had used the dashboard, some of whom indicated that it had impacted their practice: for 8 of 15 (53%), this impact had been minimal, but 2 of 15 (13%) considered its impact significant. Comments indicated that the perceived benefits were primarily related to an increased awareness of the need for higher intraoperative analgesic and antiemetic dosing in some cases and an opportunity to engage with trainees on this issue. Most respondents (12/17, 71%) confirmed they were comfortable with how this information about their practice was being collected and presented to them, although 4 of 17 (24%) were concerned there may be negative consequences to having these data available. Concerns were that the quality indicators presented did not consider the multiple other influences on patient outcomes and, depending on how the indicators are subsequently used, could apportion blame inappropriately, with possible professional or legal consequences; this further highlights the need to guarantee provider confidentiality.

Concerning future use, 14 of 17 (82%) respondents recommended a regular reminder email, and 11 of 17 (64%) indicated they would refer to the dashboard at least monthly, with 10 of 17 (59%) believing that it had the potential to have a significant benefit for their practice. Additional information requirements identified for future work included: duration of PACU stay, 13 of 17 (76%); antibiotics timing, 8 of 17 (47%); perioperative hypothermia, 8 of 17 (47%); difficult intubation recorded, 6 of 17 (35%); and incidence of hypotension, 5 of 17 (29%).

Final Dashboards Deployed

Overview
The final dashboard components, deployed in June 2022, are (1) 24-hour outcome summary (Figure 4), (2) PACU outcome summary (Figure 5), and (3) anesthesiologists’ practice profile (Figure 6).
Figure 4. The final 24-hour postoperative outcome summary dashboard with notes.

Figure 5. The final PACU outcomes dashboard design. PACU: postanesthetic care unit.
24-Hour Outcome Summary
The 24-hour outcome summary displays each outcome (pain, nausea, and vomiting) and compares the active anesthesiologist’s monthly incidence rate with the department’s rate (Figure 4, left). Their peers’ incidence rates are given in descending order, with the active anesthesiologist highlighted (Figure 4, right). Guided by the usability survey, we added the active anesthesiologist’s and overall department statistics (Figure 4, bottom).

The View 24hr Outcome Report button leads to a detailed report, where the anesthesiologist can view additional outcome information via switch tabs at the top of the page (Figure S1 in Multimedia Appendix 2): this report displays the anesthesiologist’s case distribution across the 4 severity levels and the outcome incidence rate by sex, age category, and procedure group. All data are shown compared to the department’s rates.

Both the outcome summary and the detailed report, support filtering based on patient sex, age category, and surgical service using drop-down menus. The “show data” option lets users view deidentified data contained within each plot. For a patient-level or filtered view of an anesthesiologist’s caseload, the user can apply a “drill-through” feature on all the personal charts to dive deeper into the data using advanced filters.

Based on the usability survey, we reconfigured the user instruction manual as a help page (Figure S2 in Multimedia Appendix 2), with a screenshot of the outcome page and information about each section, including formulas to calculate incidence rates.

PACU Outcome Summary
The PACU outcome summary uses the same visualization techniques as the 24-hour outcome summary: (1) plot for overall change over time and (2) for peer comparison, the opioid rescue rate is further divided into pain that required at least 1 dose and pain that required ≥4 doses of rescue medication (opioids 4+).

The PACU outcome summary shows opioid and antiemetic administration rates on separate pages: PACU opioids (Figure 5) and PACU antiemetics (Figure S3 in Multimedia Appendix 2). A toggle switch between opioids and opioids 4+ allows the anesthesiologists to view their performance against their peers (Figure 5, right). This PACU summary has the same help and data management capabilities as the 24-hour outcomes dashboard.

Final Anesthesiologists’ Practice Profile
A practice profile allows the anesthesiologist to view their case mix in stacked bar plots grouped by sex, age category, and surgical service in monthly intervals (Figure 6). The practice profile supports filtering the anesthesiologist’s case mix based on PACU and 24-hour postoperative outcomes. The user can also view the average outcomes of each caseload group as a tool tip or “drill-through” to see a detailed list of cases.

Preliminary Analysis of Dashboard Impact
The department’s aggregate incidence rate for all outcomes is reasonably low but with significant month to month variability (Figure 7). Incidence rates were not different between the preimplementation and postimplementation periods: median differences were 0.7 (95% CI –2.5 to 3.4; P=.52) for PACU opioid administration, –0.0 (95% CI –0.7 to 0.5; P=.76) for PACU opioids 4+ administration, –0.5 (95% CI –1.0 to 0.2; P=.16) for PACU antiemetic administration, –0.1 (95% CI –1.2 to 0.8; P=.86) for 24-hour pain, –0.0 (95% CI –0.4 to 0.3; P=.88) for 24-hour nausea, and 0.0 (95% CI –0.5 to 0.5; P>.99) for 24-hour vomiting.
Figure 7. Variability in the department’s monthly aggregate PACU outcome rates, including (A) opioid rescue, (B) opioid rescue ≥4 doses, and (C) antiemetic administration, and 24-hour postoperative outcome rates, including (D) pain, (E) nausea, and (F) vomiting. The department’s monthly aggregate outcome rates are plotted as blue dots; the red dotted line indicates the end of the preimplementation period; the green dotted line indicates the start of the postimplementation period; the solid red and green horizontal lines indicate the medians of the pre- and postimplementation periods, respectively. PACU: postanesthetic care unit.

Discussion

Principal Results

We designed and implemented dynamic dashboards for the anesthesiologists in our department to review their outpatients’ postoperative short-term outcomes and compare these outcomes to their peers to facilitate personal performance evaluation and potential practice improvements. This project built on previous work establishing the POFU registry to collect and electronically aggregate these outcome data. The dashboard is updated weekly and presents (1) 24-hour outcomes, (2) PACU outcomes, and (3) a practice profile containing individual anesthesiologist’s case mix, grouped by age groups, sex, and surgical service. Co-design and usability evaluation indicated requirements for uncluttered summary data visualization and the ability to explore specific outcome details if needed. Anesthesiologists in our department generally found the dashboard information helpful and believed it would significantly benefit their practice. Our study confirmed that ensuring security and patient and provider confidentiality in these dashboards was crucial for successful uptake.

Comparison With Prior Work

There are significant limitations in using postoperative outcomes such as pain, nausea, and vomiting as anesthesiologist performance indicators. There are many confounding factors, including procedural details and the surgeon being responsible for prescribing analgesics. These confounders will likely increase with increasing time after the procedure. However, even for immediate postanesthesia outcomes, confounders, including the PACU nursing team assessing and managing pain, have been shown to invalidate interanesthesiologist performance comparisons [27]. It is also essential to consider the impact of the case mix: Schulz et al [28] suggest that case-mix adjustment of measures such as length of PACU stay may provide more meaningful indicators than unadjusted metrics. However, the dashboard was not founded on the idea that an anesthesiologist’s practice is the sole determinant of their patients’ outcomes; it is a tool for anesthesiologists to reflect on their practice in context. Providing the data is only the first step toward better understanding their practices, but it is necessary. Anesthesiologists know that multiple variables outside their control contribute to these data, especially for 24-hour outcomes. Understanding their patients’ variability should allow them to interpret this information appropriately.

Clinical dashboards are being widely explored in an attempt to better understand and optimize patient outcomes in a range of anesthesia settings [29], including cardiac [30] and pediatric anesthesia [18,20,31]. Data analytics focus on process, as well as outcome, metrics [32], and there is often a need to extract the required data from a range of in-hospital systems [31] or to supplement these data with patient- or family-reported outcomes [20,33]. These initiatives may or may not have a positive impact on performance. For example, regular team and individualized feedback reduced temperature monitoring delays during spine surgery among 1 group of anesthesiologists [7]. In contrast, another group found that audit and feedback did not improve the intraoperative temperature management [34]. However, such
initiatives may often be the best route to improving practice, patient experience, and outcomes, particularly in acute pain management, given the challenges faced in implementing and adopting integrated electronic medical records in this area [35].

While dashboards and follow-up phone calls may function as intended, be economically justifiable, and be well received by anesthesiologists [20,31,36], it will be more challenging to demonstrate a beneficial impact on patient-relevant outcomes rather than process outcomes. Brenn et al [20] reported an 88% response rate for postoperative phone calls to 42,688 pediatric outpatients, but were unable to link satisfaction with complication rates and suggested that reducing wait times and streamlining operations are more important to families. In contrast, Kim et al [33] found that implementing a follow-up call (even from a nonmedical professional) 48 hours after day surgery reduced family anxiety, though it did not improve family satisfaction. We have yet to examine the effect of our dashboard on improving our patients’ outcomes.

Through an iterative design process, we learned the importance of collaborating with the end users of the dashboards to maximize usability. This collaboration in the co-design process familiarized us with local clinical terminology (eg, “day-care patients” in preference to “ambulatory patients” or “outpatients”), which helped us design dashboards that are more locally relevant. These collaborative design processes have been adopted by other researcher-developers [18,19,23,31,33], although not all such dashboards have been co-designed with end users or undergone usability evaluation [17,30]. We used a usability survey to guide us toward more acceptable dynamic dashboards with various analysis options.

Limitations
Some limitations of our dashboard designs and development approach should be noted. First, there is no risk adjustment of performance scores based on case mix: an anesthesiologist with a significant proportion of patients in whom postoperative pain or nausea have a higher baseline incidence cannot be validly compared to an anesthesiologist with a different case mix. The dashboard users recognize this limitation, and we implemented the practice profile section partly in an effort to address this issue. We may explore risk adjustment strategies in future but must be cautious not to overturn our users’ requirement for concise summary performance reports.

Second, the POFU registry includes age, sex, surgical service, and procedure date but does not contain other presurgical (risk) factors such as weight, comorbidities, existing medication, allergies, number or type of previous procedures, or child and parent anxiety levels. In the future, we plan to add additional patient data, surgical and anesthetic techniques, and PACU length of stay by integrating with the hospital’s Anesthesia Information Management System (SurgiNet Anesthesia, Cerner), which may allow us to provide our anesthesiologists with further insights.

Third, our current data are limited to outcomes occurring within 24 hours of discharge: PACU data recorded according to institutional practice and the 24-hour outcome data, for which a nurse makes only a single phone call to the patient’s family. The success rate to date has only been 7800 out of 12,081 (65%) cases; a higher rate would improve validity. To alleviate the missing 24-hour outcome data, we are exploring the development of self-reporting tools for pain, nausea, and vomiting, such as Panda [37], a mobile postoperative pain management app. We also aim to determine if families are willing to provide outcomes beyond 24 hours.

Finally, our usability questionnaire was not a standardized instrument or evaluated for reliability or validity, which limits its generalizability. It aimed to evaluate specific dashboard features as a final step before department-wide deployment.

Further Evaluation
The dashboard version has been deployed and is being used by department anesthesiologists, with data updated weekly. We will perform an ongoing usability evaluation to examine the dashboards’ usefulness, determine if anesthesiologists have any issues with the visualizations, and explore suggestions for additional information or functionality, including further optimization for use on a tablet or smartphone if required. We plan to evaluate usage patterns: how many anesthesiologists use the dashboards, how frequently, for what purpose, and over what period. Finally, we aim to extend our analysis of changes in the department’s aggregate and individual outcomes post deployment. Our preliminary analysis did not demonstrate any significant changes in outcomes, which may be in part because our PACU and 24-hour follow-up outcomes were already reasonably well-optimized compared to other institutions’ [38,39]. Our future work will focus on tracking and reducing the variability in outcome rates between different practitioners.

Ultimately, evaluating the impact of this initiative on our patients’ outcomes is a significant undertaking and, hence, a long-term goal. It will involve integrating with our Anesthesia Information Management System (SurgiNet Anesthesia, Cerner), identifying other key outcomes that matter to our patients and their families, and applying the appropriate systems and resources to collect, process, and analyze this information.

Conclusions
The dynamic, personalized dashboards we designed and deployed have allowed the anesthesiologists in our department to review their outpatients’ short-term postoperative outcomes and reflect on their practice in the context of peer and departmental performance. Key lessons from this implementation include the value of adopting a participatory approach to development, with co-design workshops and usability evaluation; the importance of establishing a robust approach to the security and confidentiality of patient and provider data in gaining user trust; and the preference for presenting uncluttered summary data combined with the opportunity to drill down into specific cases if required. This dashboard solution has allowed our department’s anesthesiologists to visualize previously unavailable data collected as part of a broader quality initiative. It should provide the opportunity to evaluate iterative practice changes, although further work will be required to monitor its effect on individual and department performance and patient outcomes.
Acknowledgments

Data reported in this study were retrieved from the BC Children’s Hospital postoperative follow-up (POFU) database. The authors would like to thank the POFU registry collaborators: the Pediatric Anesthesia Research Team, BC Children’s Hospital (BCCH) Procedures and Surgical Services, and BCCH Surgery and Anesthesia, for collecting the data and granting us access to them for this work. We thank the BC Children’s Hospital Foundation for its support. This work was supported, in part, by a UBC Canada’s Digital Technology Supercluster expansion research grant. The POFU registry was supported by a Medical Allied Staff Engagement Society Engagement Fund grant (awarded to NRF) and a BCCH Surgical Program Improvement Fund grant (awarded to NRF and Elizabeth Lamb). MG holds a Michael Smith Health Research BC scholar award; RSS held a Mitacs postdoctoral fellowship while this work was conducted. This work was previously presented, in part, at the Society for Technology in Anesthesia 2022 Virtual Annual Meeting, January 2022 and the Canadian Pediatric Anesthesia Society Annual Meeting in Vancouver in September 2022.

Conflicts of Interest

None declared.

Multimedia Appendix 1

POFU Anesthesia Dashboard - usability and utility survey.  
[PDF File (Adobe PDF File), 89 KB - periop_v6i1e47398_app1.pdf ]

Multimedia Appendix 2

View 24hr Outcome Report, Help page, and PACU Antiemetics.  
[PDF File (Adobe PDF File), 530 KB - periop_v6i1e47398_app2.pdf ]

References


Abbreviations

- **BCCHR**: British Columbia Children’s Hospital Research Institute
- **PACU**: postanesthetic care unit
- **POFU**: postoperative follow-up registry
- **QIQA**: quality improvement or quality assurance
- **REDCap**: Research Electronic Data Capture
An Ensemble Learning Approach to Improving Prediction of Case Duration for Spine Surgery: Algorithm Development and Validation

Rodney Allanigue Gabriel1,2, MD, MAS; Bhavya Harjai2, BS; Sierra Simpson2, PhD; Austin Liu Du3, BS; Jeffrey Logan Tully3, MD; Olivier George4, PhD; Ruth Waterman5, MD

1Division of Biomedical Informatics, Department of Medicine, University of California, San Diego, San Diego, CA, United States
2Division of Perioperative Informatics, Department of Anesthesiology, University of California, San Diego, San Diego, CA, United States
3School of Medicine, University of California, San Diego, San Diego, CA, United States
4Department of Psychiatry, University of California, San Diego, San Diego, CA, United States
5Department of Anesthesiology, University of California, San Diego, San Diego, CA, United States

Corresponding Author:
Rodney Allanigue Gabriel, MD, MAS
Division of Biomedical Informatics
Department of Medicine
University of California, San Diego
9300 Campus Point Dr
San Diego, CA, 92037
United States
Phone: 1 858 657 7000
Email: ragabriel@health.ucsd.edu

Abstract

Background: Estimating surgical case duration accurately is an important operating room efficiency metric. Current predictive techniques in spine surgery include less sophisticated approaches such as classical multivariable statistical models. Machine learning approaches have been used to predict outcomes such as length of stay and time returning to normal work, but have not been focused on case duration.

Objective: The primary objective of this 4-year, single-academic-center, retrospective study was to use an ensemble learning approach that may improve the accuracy of scheduled case duration for spine surgery. The primary outcome measure was case duration.

Methods: We compared machine learning models using surgical and patient features to our institutional method, which used historic averages and surgeon adjustments as needed. We implemented multivariable linear regression, random forest, bagging, and XGBoost (Extreme Gradient Boosting) and calculated the average $R^2$, root-mean-square error (RMSE), explained variance, and mean absolute error (MAE) using k-fold cross-validation. We then used the SHAP (Shapley Additive Explanations) explainer model to determine feature importance.

Results: A total of 3189 patients who underwent spine surgery were included. The institution’s current method of predicting case times has a very poor coefficient of determination with actual times ($R^2=0.213$). On k-fold cross-validation, the linear regression model had an explained variance score of 0.345, an $R^2$ of 0.34, an RMSE of 162.84 minutes, and an MAE of 127.22 minutes. Among all models, the XGBoost regressor performed the best with an explained variance score of 0.778, an $R^2$ of 0.770, an RMSE of 92.95 minutes, and an MAE of 44.31 minutes. Based on SHAP analysis of the XGBoost regression, body mass index, spinal fusions, surgical procedure, and number of spine levels involved were the features with the most impact on the model.

Conclusions: Using ensemble learning-based predictive models, specifically XGBoost regression, can improve the accuracy of the estimation of spine surgery times.

(JMIR Perioper Med 2023;6:e39650) doi:10.2196/39650
KEYWORDS

ensemble learning; machine learning; spine surgery; case duration; prediction accuracy; operating room efficiency; learning; surgery; spine; operating room; case; model; patient; surgeon; linear regression; accuracy; estimation; time

Introduction

Surgeons often estimate case durations when scheduling operative time; durations may also be tied to historical averages or Current Procedural Terminology (CPT) codes, practices that are prone to substantial inaccuracies [15]. Classical statistical methods have been used to further improve the prediction of case durations [16-18]. The proliferation of electronic health records and the associated generation of vast amounts of previously uncaptured patient data have allowed for more sophisticated analytics in several clinical arenas, including the operating room [19]. With large enough data sets, specialized algorithms can develop complex predictive models after being exposed to a number of prior examples in a process known as machine learning [20].

Current predictive techniques in spine surgery include less sophisticated approaches such as classical multivariable statistical models. While a variety of features and outcomes, such as length of stay, prescription duration, and time to return to normal work, have been predicted in previous studies, there has been little focus on case duration [21-25]. To our knowledge, no other studies have focused on using machine learning models to predict the surgical case duration for the spine surgery population, but the method has been implemented in other procedures [26-28]. Spine surgery consists of heterogeneous anatomical and technical components that should theoretically be taken into account when estimating case duration. The primary objective of this study is to develop machine learning-based predictive models using patient and surgery-specific features. Specifically, we use ensemble learning, which combines multiple predictive models to determine an overall prediction of the outcome. We hypothesize that such models can outperform those that estimate case duration based on historic averages and surgeon preference (which may not be scalable or transferable outside of a given institution).

Methods

Ethics Approval

This retrospective study was approved (approval protocol 210098) by the Human Research Protections Program at the University of California, San Diego for the collection of data from our electronic medical record system. For this study, the informed consent requirement was waived. Data were collected retrospectively from the electronic medical record system of our institution’s operating room data. Data from all patients that underwent spine surgery from 3 different orthopedic spine surgeons from January 2018 to September 2021 was extracted. We excluded all patients that had missing data for actual case duration; all other features with missing values were categorized as unknown or imputed if they were continuous variables (described below). This retrospective observational study abided by the EQUATOR guidelines.

Primary Objective and Data Collection

The primary outcome measurement was a continuous value, defined as the actual operating room case duration measured in minutes (from patient wheeling into the operating room to exiting the operating room). We implemented predictive models using various machine learning algorithms to predict the actual case duration. We compared this to our current system’s practice of estimating case duration, which is equal to the mean of the last 3 times the surgical procedure was performed, with the ability of the surgeon to change times based on their preference. The models developed were multivariable linear regression, random forest regressors, bagging regressors, and XGBoost (Extreme Gradient Boosting) regressors.

The independent features in the models were (1) categorical features, which included surgical procedure (39 unique procedures), surgeon identification (3 different surgeons), American Society of Anesthesiologists Physical Status (ASA PS) score (ie, comorbidity burden), sex, specific surgical details (kyphoplasty, discectomy, fusion, and laminectomy), the anterior approach involved (ie, approach surgeon used to access the spine), and level of spine region involved (eg, cervical, thoracic, lumbar, or a combination of levels); and (2) continuous features, which included the number of spine levels involved in the surgery (from 1 to 7) and body mass index (kg/m²) (Table 1).

For missing data on the ASA PS class, the value was defined as “unknown.” For missing data on body mass index, the value was imputed by using the average BMI among all patients with known data for this feature.
### Characteristics of all the cases included in analysis (N=3315).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Instrumentation</th>
<th>Approach</th>
<th>Fusion</th>
<th>Levels</th>
<th>Other</th>
<th>Participants, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Surgical Procedure</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discectomy</td>
<td>No</td>
<td>Anterior</td>
<td>Yes</td>
<td>1</td>
<td>Fusion</td>
<td>89 (2.7)</td>
</tr>
<tr>
<td>Discectomy</td>
<td>No</td>
<td>Anterior</td>
<td>Yes</td>
<td>2</td>
<td>Fusion</td>
<td>127 (3.8)</td>
</tr>
<tr>
<td>Discectomy</td>
<td>No</td>
<td>Anterior</td>
<td>Yes</td>
<td>3+</td>
<td>Fusion</td>
<td>202 (6.1)</td>
</tr>
<tr>
<td>Deformity fusion</td>
<td>No</td>
<td>Posterior</td>
<td>Yes</td>
<td>1-6 seg.</td>
<td>For deformity</td>
<td>8 (0.2)</td>
</tr>
<tr>
<td>Deformity fusion</td>
<td>No</td>
<td>Posterior</td>
<td>Yes</td>
<td>7-12 seg.</td>
<td>For deformity</td>
<td>2 (0.1)</td>
</tr>
<tr>
<td>Lumbar fusion</td>
<td>No</td>
<td>Anterior</td>
<td>Yes</td>
<td>2</td>
<td>Lumbar</td>
<td>270 (8.1)</td>
</tr>
<tr>
<td>Lumbar fusion</td>
<td>No</td>
<td>Anterior</td>
<td>Yes</td>
<td>3</td>
<td>Lumbar</td>
<td>14 (0.4)</td>
</tr>
<tr>
<td>Oblique lumbar interbody fusion</td>
<td>No</td>
<td>Anterior</td>
<td>Yes</td>
<td>1</td>
<td>Lumbar</td>
<td>1 (0.0)</td>
</tr>
<tr>
<td>Transforaminal lumbar interbody fusion</td>
<td>No</td>
<td>Transforaminal</td>
<td>Yes</td>
<td>1</td>
<td>Lumbar</td>
<td>9 (0.3)</td>
</tr>
<tr>
<td>Extreme lateral interbody fusion</td>
<td>No</td>
<td>Lateral</td>
<td>Yes</td>
<td>1</td>
<td>Lumbar</td>
<td>251 (7.6)</td>
</tr>
<tr>
<td>Extreme lateral interbody fusion</td>
<td>No</td>
<td>Lateral</td>
<td>Yes</td>
<td>2</td>
<td>Lumbar</td>
<td>198 (6.0)</td>
</tr>
<tr>
<td>Extreme lateral interbody fusion</td>
<td>No</td>
<td>Lateral</td>
<td>Yes</td>
<td>3</td>
<td>Lumbar</td>
<td>63 (1.9)</td>
</tr>
<tr>
<td>Extreme lateral interbody fusion</td>
<td>No</td>
<td>Lateral</td>
<td>Yes</td>
<td>4</td>
<td>Lumbar</td>
<td>16 (0.5)</td>
</tr>
<tr>
<td>Thoracic fusion</td>
<td>No</td>
<td>Posterior</td>
<td>Yes</td>
<td>1</td>
<td>Thoracic</td>
<td>1 (0.0)</td>
</tr>
<tr>
<td>Thoracic fusion</td>
<td>No</td>
<td>Posterior</td>
<td>Yes</td>
<td>2</td>
<td>Thoracic</td>
<td>7 (0.2)</td>
</tr>
<tr>
<td>Thoracic fusion</td>
<td>No</td>
<td>Posterior</td>
<td>Yes</td>
<td>3</td>
<td>Thoracic</td>
<td>17 (0.5)</td>
</tr>
<tr>
<td>Thoracic fusion</td>
<td>No</td>
<td>Posterior</td>
<td>Yes</td>
<td>4</td>
<td>Thoracic</td>
<td>12 (0.4)</td>
</tr>
<tr>
<td>Thoracic fusion</td>
<td>No</td>
<td>Posterior</td>
<td>Yes</td>
<td>5+</td>
<td>Thoracic</td>
<td>35 (1.1)</td>
</tr>
<tr>
<td>Kyphoplasty or vertebroplasty</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>1</td>
<td>All</td>
<td>316 (9.5)</td>
</tr>
<tr>
<td>Kyphoplasty</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>2</td>
<td>Thoracolumbar</td>
<td>40 (1.2)</td>
</tr>
<tr>
<td>Kyphoplasty</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>3</td>
<td>Thoracolumbar</td>
<td>19 (0.6)</td>
</tr>
<tr>
<td>Kyphoplasty</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>4</td>
<td>Thoracolumbar</td>
<td>21 (0.6)</td>
</tr>
<tr>
<td>Laminectomy or decompressive laminectomy</td>
<td>No</td>
<td>Posterior</td>
<td>No</td>
<td>1</td>
<td>Lumbar</td>
<td>148 (4.5)</td>
</tr>
<tr>
<td>Laminectomy or decompressive laminectomy</td>
<td>No</td>
<td>Posterior</td>
<td>No</td>
<td>2</td>
<td>Lumbar</td>
<td>106 (3.2)</td>
</tr>
<tr>
<td>Laminectomy or decompressive laminectomy</td>
<td>No</td>
<td>Posterior</td>
<td>No</td>
<td>3</td>
<td>Lumbar</td>
<td>110 (3.3)</td>
</tr>
<tr>
<td>Laminectomy</td>
<td>No</td>
<td>Posterior</td>
<td>Yes</td>
<td>5</td>
<td>Cervical</td>
<td>109 (3.3)</td>
</tr>
<tr>
<td>Laminectomy</td>
<td>No</td>
<td>Posterior</td>
<td>Yes</td>
<td>1-4</td>
<td>Cervical</td>
<td>115 (3.5)</td>
</tr>
<tr>
<td>Laminectomy</td>
<td>No</td>
<td>Posterior</td>
<td>No</td>
<td>1-2</td>
<td>Cervical</td>
<td>3 (0.1)</td>
</tr>
<tr>
<td>Laminectomy</td>
<td>No</td>
<td>Posterior</td>
<td>No</td>
<td>2</td>
<td>Cervical</td>
<td>31 (0.9)</td>
</tr>
<tr>
<td>Laminectomy</td>
<td>Yes</td>
<td>Posterior</td>
<td>Yes</td>
<td>1</td>
<td>Lumbar</td>
<td>219 (6.6)</td>
</tr>
<tr>
<td>Laminectomy</td>
<td>Yes</td>
<td>Posterior</td>
<td>Yes</td>
<td>2</td>
<td>Lumbar</td>
<td>259 (7.8)</td>
</tr>
<tr>
<td>Laminectomy</td>
<td>Yes</td>
<td>Posterior</td>
<td>Yes</td>
<td>3</td>
<td>Lumbar</td>
<td>162 (4.9)</td>
</tr>
<tr>
<td>Laminectomy</td>
<td>Yes</td>
<td>Posterior</td>
<td>Yes</td>
<td>4+</td>
<td>Lumbar</td>
<td>259 (7.8)</td>
</tr>
<tr>
<td>Laminectomy</td>
<td>Yes</td>
<td>Posterior</td>
<td>Yes</td>
<td>1</td>
<td>Thoracic</td>
<td>9 (0.3)</td>
</tr>
<tr>
<td>Laminectomy</td>
<td>Yes</td>
<td>Posterior</td>
<td>Yes</td>
<td>2</td>
<td>Thoracic</td>
<td>9 (0.3)</td>
</tr>
<tr>
<td>Laminectomy</td>
<td>Yes</td>
<td>Posterior</td>
<td>Yes</td>
<td>3</td>
<td>Thoracic</td>
<td>7 (0.2)</td>
</tr>
<tr>
<td>Laminectomy</td>
<td>Yes</td>
<td>Posterior</td>
<td>Yes</td>
<td>4</td>
<td>Thoracic</td>
<td>18 (0.5)</td>
</tr>
<tr>
<td>Laminectomy</td>
<td>Yes</td>
<td>Posterior</td>
<td>Yes</td>
<td>5</td>
<td>Thoracic</td>
<td>30 (0.9)</td>
</tr>
<tr>
<td>Laminectomy</td>
<td>Yes</td>
<td>Posterior</td>
<td>Yes</td>
<td>6+</td>
<td>Thoracic</td>
<td>3 (0.1)</td>
</tr>
<tr>
<td>Characteristics</td>
<td>Instrumentation</td>
<td>Approach</td>
<td>Fusion</td>
<td>Levels</td>
<td>Other</td>
<td>Participants, n (%)</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-----------------</td>
<td>----------</td>
<td>--------</td>
<td>--------</td>
<td>-------</td>
<td>--------------------</td>
</tr>
<tr>
<td><strong>Specific surgical procedure included</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kyphoplasty</td>
<td>N/A(^a)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>396 (11.9)</td>
</tr>
<tr>
<td>Discectomy</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>418 (12.6)</td>
</tr>
<tr>
<td>Fusion</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>2521 (76.0)</td>
</tr>
<tr>
<td>Laminectomy</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>1597 (48.2)</td>
</tr>
<tr>
<td><strong>Anterior approach involved</strong></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>702 (21.1)</td>
</tr>
<tr>
<td><strong>Number of spine levels involved</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>1043 (31.5)</td>
</tr>
<tr>
<td>2</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>1050 (31.7)</td>
</tr>
<tr>
<td>3</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>594 (17.9)</td>
</tr>
<tr>
<td>4</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>441 (13.3)</td>
</tr>
<tr>
<td>5</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>174 (5.2)</td>
</tr>
<tr>
<td>6</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>11 (0.3)</td>
</tr>
<tr>
<td>7</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>2 (0.1)</td>
</tr>
<tr>
<td><strong>Surgeon performing procedure</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>1676 (50.6)</td>
</tr>
<tr>
<td>B</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>191 (5.8)</td>
</tr>
<tr>
<td>C</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>1448 (43.7)</td>
</tr>
<tr>
<td><strong>Level of spine involved</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cervical</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>567 (17.1)</td>
</tr>
<tr>
<td>Thoracic</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>228 (6.9)</td>
</tr>
<tr>
<td>Lumbar</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>2165 (65.3)</td>
</tr>
<tr>
<td>Male sex</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>1800 (54.3)</td>
</tr>
<tr>
<td>BMI (kg/m(^2)), mean (SD)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>29.7 (6.3)</td>
</tr>
<tr>
<td><strong>ASA PS(^b) classification score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>46 (1.4)</td>
</tr>
<tr>
<td>2</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>1140 (34.4)</td>
</tr>
<tr>
<td>3</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>2008 (60.6)</td>
</tr>
<tr>
<td>4</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>112 (3.4)</td>
</tr>
<tr>
<td>Unknown</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>9 (0.3)</td>
</tr>
</tbody>
</table>

\(^a\)N/A: not applicable.
\(^b\)ASA PS: American Society of Anesthesiologists Physical Status.

**Analysis Packages and Metrics**

Python (version 3.7.5; Python Software Foundation) was used for all statistical analyses. The code is provided in the webpage [29]. We calculated the $R^2$, root-mean-square error (RMSE), mean absolute error (MAE), explained variance, and maximum error for each iteration of k-fold cross-validation (described below) and used those scores to calculate the median scores and plot feature importance using SHAP (Shapley Additive Explanations) and prediction error plots (Figure 1).
Machine Learning Models

Overview
We compared various machine learning-based predictive models to our institution’s conventional model, which predicted case duration using average times (over the last 5 times the surgery was performed by that surgeon) based on the CPT code of the surgery plus adjustments from the surgical attending based on clinical judgment or preference. First, we developed a model using multivariable linear regression. We then evaluated the use of ensemble learning (a process in which multiple models are combined) to calculate a prediction. In this case, we used random forest, bagging, and XGBoost-based regressors (Multimedia Appendix 1). For each model, all features were included as inputs.

Multivariable Linear Regression
This is a statistical model that asserts a continuous outcome based on the weighted combination of the underlying independent variables. We tested an L2-penalty-based regression model without specifying individual class weights. This model provides a baseline score and helps make the case for improvement over the evaluation metrics.

Random Forest Regressor
Random forest is an ensemble approach (a technique that combines the predictions from multiple machine learning algorithms to make more accurate predictions than any individual model) of decision trees. It is a robust and reliable nonparametric supervised learning algorithm that acts as a means to test further improvements in metrics and determine the feature importance of a data set. The number of tree estimators was set to 1000, the criterion chosen was “squared error,” and the minimum number of samples required to split an internal node was set to 2. All other parameters were left at their default values.

Bagging Regressor
Bagging or bootstrap aggregation is another way to build ensemble models. Bagging methods build several estimators on different randomly selected subsets of data. Unlike random forest models, bagging models are not sensitive to the specific data on which they are trained. They would give a similar score even when trained on a subset of the data. Bagging regressors are also generally more immune to overfitting. We built a bagging regressor using the scikit-learn package, where replacement was allowed. The number of estimators was set to 1000 with the base of decision tree regressors, and the samples were drawn with replacement (bootstrap was set to True). All other parameters were left at their default values.

XGBoost Regressor
Boosting is another approach to ensemble learning in which decision trees are built sequentially so that each subsequent tree aims to reduce the error from the previous tree. Thus, each subsequent tree learns from previous trees and updates the residual errors. Unlike bagging, boosting uses decision trees with fewer splits; XGBoost is an implementation of a gradient-boosted tree algorithm [30]. We built an XGBoost regressor using the xgboost version 1.7.1 package (xgboost developers). The number of estimators used was 1000, the tree method was set to “auto,” and the booster was set to “gtree.” All other parameters were left at their default values, as described in the documentation of the library.
Feature Importance

An important function of a model is to uncover potential features that contribute to a given outcome. If a model can predict surgical outcomes efficiently with good specificity, then we can assume that the features of interest that are identified may be relevant and important to the actual surgical outcome. These models can often be opaque with many trees and features of interest, making interpretation of the data difficult. To aid in model interpretation, we used the SHAP model [31]. This module allows for a value to be assigned to each feature used to predict the outcome of a model. Additionally, it provides whether that feature negatively or positively impacts the outcome of that given prediction. If the score is very high or very low, that feature weighs heavily on the model. If the score is close to zero or not well separated, that feature is of lesser importance. Once features are identified and given SHAP values, interpretability is improved because features are concrete and have been assigned importance. Features can then be validated based on scientific rationale and further analysis.

K-Fold Cross-Validation

To perform a more robust evaluation of our models, we implemented k-fold cross-validation to observe the $R^2$, MAE, RMSE, explained variance, and maximum error for 10 folds after a shuffle. The data set was first shuffled to account for any sorting and then split into 10 folds, where 1 fold serves as the test set and the remaining 9 sets serve as the training set. This was repeated until all folds had the opportunity to serve as the test set. For each iteration, our performance metrics were calculated on the test set. The median of each performance metric ($R^2$, RMSE, MAE, explained variance, and maximum error) was calculated thereafter.

Results

Overview

There were 3523 spine surgeries identified during this period. After exclusion criteria were applied, 3189 surgeries were included in the final analysis. Among these, there were 39 different kinds of spine surgeries included. The majority of cases involved spinal fusion (n=2433, 76.0%) and were performed in the lumbar region (n=2082, 65.3%). The median ASA PS score was 3, and the majority of patients were male (n=1732, 54.3%; Table 1). The mean of actual surgical case duration among all surgeries was 335.5 (SD 199.9) minutes.

Performance Evaluation Using Linear Regression

Using all features (Table 1), we developed various machine learning algorithms to predict case duration. The base model, which was the conventional approach against which all machine learning models were compared, was based on our current system’s method to predict surgical times, which is based on the average of the surgical procedures’ case times over the last 5 instances with the ability for the surgeon to change times based on clinical judgment or preference. There was a poor coefficient of determination between the predicted time and actual time based on this approach ($R^2$=–0.213). We then performed multivariable linear regression trained on 80% of the data and tested on 20% of separate data, which had an $R^2$ of 0.34. Features that were statistically significant in this model included laminectomy (estimate=218.51, $P$<.001), number of levels performed, ASA PS classification score, and lumbar involvement (estimate=218.51, $P$<.001; Table 2).

Next, we implemented ensemble learning approaches to predicting case duration, in which the models were trained on 80% of the data and tested on a separate 20% of the data. The reason for the 80:20 split was to visualize the $R^2$ metric for each model (Figure 2). The $R^2$ metrics for the linear regressor, bagging regressor, random forest regressor, and XGBoost regressor, as well as the currently used method, were 0.407, 0.812, 0.812, 0.832, and 0.213, respectively.
Table 2. Results of the multivariable linear regression model predicting actual case duration. We included all features in the model. Because surgical procedure had 39 different procedures, we omitted the values from the table, however, they were included in the model.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE (minutes)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>−61.89</td>
<td>119.13</td>
<td>.60</td>
</tr>
<tr>
<td><strong>Specific surgical procedure included</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kyphoplasty</td>
<td>−225.43</td>
<td>90.5</td>
<td>.01</td>
</tr>
<tr>
<td>Discectomy</td>
<td>−33.94</td>
<td>94.7</td>
<td>.72</td>
</tr>
<tr>
<td>Fusion</td>
<td>6.17</td>
<td>75.9</td>
<td>.94</td>
</tr>
<tr>
<td>Laminctomy</td>
<td>218.51</td>
<td>41.7</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Anterior approach involved</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>102.06</td>
<td>93.6</td>
<td>.28</td>
</tr>
<tr>
<td><strong>Number of spine levels involved</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>62.57</td>
<td>77.8</td>
<td>.42</td>
</tr>
<tr>
<td>3</td>
<td>−18.15</td>
<td>80.6</td>
<td>.82</td>
</tr>
<tr>
<td>4</td>
<td>74.41</td>
<td>65.3</td>
<td>.25</td>
</tr>
<tr>
<td>5</td>
<td>246.18</td>
<td>61.7</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>6</td>
<td>212.36</td>
<td>106.7</td>
<td>.04</td>
</tr>
<tr>
<td>7</td>
<td>445.13</td>
<td>143.1</td>
<td>.002</td>
</tr>
<tr>
<td><strong>Surgeon performing procedure</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>−38.37</td>
<td>13.1</td>
<td>.003</td>
</tr>
<tr>
<td>C</td>
<td>4.32</td>
<td>6.0</td>
<td>.47</td>
</tr>
<tr>
<td><strong>Level of spine involved</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cervical</td>
<td>115.84</td>
<td>53.6</td>
<td>.03</td>
</tr>
<tr>
<td>Thoracic</td>
<td>26.56</td>
<td>34.6</td>
<td>.44</td>
</tr>
<tr>
<td>Lumbar</td>
<td>218.51</td>
<td>52.3</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Male sex</td>
<td>173.34</td>
<td>171.7</td>
<td>.31</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>0.38</td>
<td>0.52</td>
<td>.47</td>
</tr>
<tr>
<td><strong>ASA PS</strong>&lt;sup&gt;a&lt;/sup&gt; classification score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>66.03</td>
<td>30.1</td>
<td>.32</td>
</tr>
<tr>
<td>3</td>
<td>97.43</td>
<td>25.9</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>4</td>
<td>29.98</td>
<td>30.1</td>
<td>.32</td>
</tr>
<tr>
<td>Unknown</td>
<td>41.97</td>
<td>60.2</td>
<td>.49</td>
</tr>
</tbody>
</table>

<sup>a</sup>ASA PS: American Society of Anesthesiologists Physical Status.

Figure 2. Illustration of the correlation between actual times and predicted surgical times for spine surgery calculated by each model type: predicted times based on procedural averages and surgeon preference or customization, multivariable linear regression, random forest, bagging, and Extreme Gradient Boosting (XGBoost). The data set was split 80:20 (training:test), and the model was trained on the training set and validated on the test set.
Median Performance Metrics of Models Using k-Fold Cross-Validation

We calculated various performance metrics for each model by applying a k-fold cross-validation approach and calculated the median scores for each model (Table 3). The linear regression model had an explained variance score of 0.34, an $R^2$ of 0.40, an RMSE of 162.84 minutes, and an MAE of 127.22 minutes. Among all models, the XGBoost regressor performed the best with an explained variance score of 0.778, an $R^2$ of 0.77, an RMSE of 92.95 minutes, and an MAE of 44.31 minutes.

SHAP analysis was performed to describe the features of the XGBoost model with the most impact on model prediction since it was the best-performing model based on the $R^2$ (Figure 3). Figure 3A illustrates the most important features per fold, whereas Figure 3B illustrates the ranks of each feature’s importance per fold. BMI and spine fusion were consistently the top 2 most impactful features. In order of feature importance, there were then surgical procedure, number of spine levels, operating surgeon, the anatomic location being the lumbar spine, ASA PS classification score, sex, kyphoplasty, the anatomic location being the cervical spine, anterior approach, laminectomy, the anatomic location being the thoracic spine, and discectomy.

Table 3. Performance of each machine learning approach predicting case duration of spine surgery. Calculation is based on the median quantified by k-fold cross-validation for the bagging regressor, linear regression, random forest regressor, and XGBoost regressor. Current method is based on average of the last 5 instances of the surgery with surgeons input to modify time.

<table>
<thead>
<tr>
<th>Model or method</th>
<th>Explained variance</th>
<th>Max error</th>
<th>MAE $^a$ (minutes)</th>
<th>RMSE $^b$ (minutes)</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current method</td>
<td>0.012</td>
<td>847</td>
<td>180.32</td>
<td>243.30</td>
<td>−0.57</td>
</tr>
<tr>
<td>Linear regression</td>
<td>0.345</td>
<td>526.29</td>
<td>127.21</td>
<td>162.84</td>
<td>0.34</td>
</tr>
<tr>
<td>RF $^c$ regressor</td>
<td>0.768</td>
<td>454.59</td>
<td>62.82</td>
<td>96.51</td>
<td>0.76</td>
</tr>
<tr>
<td>Bagging regressor</td>
<td>0.769</td>
<td>454.90</td>
<td>62.83</td>
<td>96.51</td>
<td>0.76</td>
</tr>
<tr>
<td>XGBoost regressor</td>
<td>0.778</td>
<td>475.72</td>
<td>44.31</td>
<td>92.95</td>
<td>0.77</td>
</tr>
</tbody>
</table>

$^a$MAE: mean absolute error.

$^b$RMSE: root mean square error.

$^c$RF: random forest.
Discussion

Principal Findings

We found that the use of ensemble learning with the patient and procedural-specific features (variables that are known preoperatively and attainable from the electronic medical record system) outperformed the prediction of spine surgery case duration when compared to models that use historic averages and surgeon preferences. Unique to our approach of predicting surgical time for this heterogeneous surgical population was the granularity of features (eg, patient and surgical characteristics) combined with an ensemble learning approach. The reference model (the time estimated based on historic averages and surgeon preference) had poor performance. We then implemented machine learning-based models using features including procedural details (ie, number of spine levels, patient positioning, surgeon, level of spine region involved, etc) and patient-specific details (ie, body mass index, sex, ASA score, etc) and demonstrated improved performance. While linear regression improved $R^2$ to 0.34, the use of XGBoost, random forest, and bagging improved it further (0.77, 0.71, and 0.71, respectively). Such models could be relatively easy to integrate into a resource-capable electronic medical record system, given that the included features could be obtained automatically from the electronic record preoperatively.

The usage of historic averages or CPT code-based estimations for spine surgery scheduling may be inaccurate given that some determinants of case duration may not be accounted for in the prediction. These features include surgeon experience, level of the spine region involved, number of levels, type of surgery (ie, kyphoplasty, fusion, laminectomy, etc), need for multiple surgeries, patient positioning, and patient body mass index. The inclusion of these features into our models results in a substantial improvement in prediction accuracy. Accurate prediction of operation times has long been discussed as a means to improve operating room efficiency and patient care [14]. Recent implementations of such models have demonstrated these improvements across a variety of measures. A recent randomized
clinical trial found that a machine-learning approach increased prediction accuracy, decreased start-time delay, and decreased total patient wait time [32]. A similar randomized controlled trial demonstrated increased throughput and decreased staff burnout [32,33]. Subsequently, decreases in delays and wait times result in lower costs and increased caseloads, which can further drive cost-effectiveness [34,35]. Associations between wait times and postoperative complications provide evidence that proper identification and mitigation of delays can improve outcomes as well [36]. Overall, improvements in patient scheduling, case duration, and staffing may result in enhanced efficiency and potentially superior patient outcomes. Understanding and identifying the features that are key in lessening the burden of misused surgical time is crucial with the trending increases in caseload burden and impacted hospital resources.

Ensemble learning essentially uses an “ensemble” of predictive models and calculates the overall prediction based on the individual predictions from each model within the “ensemble.” In this case, we leveraged ensemble learning using decision tree-based machine learning algorithms: random forest, bagging, and boosting. Our results demonstrated a substantial improvement with XGBoost compared to the other ensemble approaches as well as linear regression. XGBoost often performs better than random forest because it prunes nodes if the gain of a node is minimal to the model [30]. Random forest generates the tree to a greater depth because it does not prune nodes and relies on a majority vote for the outcome. This can result in overfitting in random forest models. Random forest may also give preference to classes that are associated with categorical variables, which do not occur in XGBoost. Because XGBoost is an iterative process, it gives preference to features that enable the regressor to predict low-participation classes. Additionally, XGBoost is more efficient with the unbalanced data sets often seen in medical or biological data. Alternatives such as linear regression work well when the data is straightforward and well-distributed. The more complex the data set, the better a bagging or tree-based model will work. With ensemble approaches, nonlinear relationships between features may be captured, and a “strong” model is developed based on learning from “weaker” models, in which residual errors are improved. Thus, the use of ensemble learning in this clinical scenario—where there is a complex interplay between features—may be superior to a statistical approach that only models linear relationships (ie, linear regression). Future studies may benefit from other approaches such as support vector machines, which could be implemented to focus on accuracy, or penalized regressors, which could provide increased interpretability.

Oftentimes, machine learning approaches are described as “black boxes” because the interpretation of the importance of features to the predictive model is challenging. The implementation of an explainer model such as SHAP values is one way to elucidate the importance of features. In this study, SHAP identified that BMI is the most important feature of the model and provides weight and context to the feature about the other identified features [37,38]. BMI may be associated with increased case duration due to the additional technical and positioning challenges. Sex was also identified as an important feature. This finding is congruent with current research that demonstrates women are more likely to have bone loss earlier than men, and bone loss has been shown to affect surgical outcomes and recovery due to poor bone remodeling and healing [39,40]. Other interesting features with an important impact included the operating surgeons themselves, the ASA PS classification score, and the number of spine levels operated. It makes sense to include surgeons as a feature in predictive modeling as each physician may have different styles and comfort levels that could impact surgical time. The ASA PS classification score represents a patient’s comorbidity burden and could suggest that patients with a higher comorbidity burden would require longer anesthesia times. Finally, it makes sense that the number of spine levels contributes to case duration, as this has a potentially linear relationship to how long surgery would take. Being able to put various features into the context of the research question is essential for translating the findings into actionable metrics. Overall, the SHAP analysis identified clinically relevant features for future exploration and evaluation.

There are several limitations to the study, mainly its retrospective nature; thus, the collection and accuracy of the data are only as reliable as what is recorded in the electronic medical record system. The current institutional practice for estimating scheduled case duration was based on the historic averages of the last 5 surgeries, with the surgeon’s ability to change the times based on clinical judgment or preference. We do not have data on why and when surgeons changed the times. In addition, there were some missing data for actual case duration, but this only led to the removal of 5.9% of the initial data set. There may also be several features not included in the models that may substantially contribute to time estimates, including surgical resident involvement (and their level of training) or surgical instruments used. Furthermore, there are other machine learning algorithms that we did not test, including support vector machines and penalized regressors. Despite these limitations, we were able to develop a predictive model using XGBoost with a high $R^2$ value (>0.7). These findings would need to be validated externally and prospectively to determine their generalizability to spine surgeries.

**Conclusions**

Operating room efficiency is a key factor in maintaining and growing institutional profits. Additionally, improvements in operating room efficiency contribute to enhanced patient care and satisfaction. Given the technical and anatomical heterogeneity in spine surgeries, it has been a challenge to predict case duration using conventional methods at our institution. This method can be applied in the future to standard and heterogenous surgical procedures with or without class imbalance to identify key obstacles to future surgical efficiency; however, it is crucial to develop robust models to more accurately predict schedule case length. In our study, we demonstrated that patient and surgical features that are easy to collect from the electronic medical record can improve the estimation of surgical times using machine learning-based predictive models. Future implementation of machine learning-based models presents an alternative pathway to use
electronic medical record data to advance surgical efficiency and enrich patient outcomes.

Acknowledgments

SS and OG are supported by the National Institutes of Health Grants DA043799 and DA044451. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Authors’ Contributions

RAG, BH, and SS are responsible for the study design. RAG, BH, SS, ALD, JLT, and RW are responsible for the analysis and interpretation of data and drafting of the article. OG is responsible for supporting SS and for reviewing the article.

Conflicts of Interest

The University of California has received funding and product for other research projects from Epimed International (Farmers Branch, TX); Infutronics (Natick, MA); Precision Genetics (Greenville County, SC); and SPR Therapeutics (Cleveland, OH) for author RAG. The University of California San Diego is a consultant for Avanos (Alpharetta, GA), which RAG represents. SS is the founder of BrilliantBiome, Inc.

Multimedia Appendix 1

Illustration of how each ensemble learning algorithm computes the prediction: (A) random forest, (B) bagging, and (C) Extreme Gradient Boosting (XGBoost).

References


Abbreviations

ASA PS: American Society of Anesthesiologists Physical Status
MAE: mean absolute error
RMSE: root-mean-square error
SHAP: Shapley Additive Explanations
XGBoost: Extreme Gradient Boosting
Assessing Barriers to Implementation of Machine Learning and Artificial Intelligence–Based Tools in Critical Care: Web-Based Survey Study

Eric Mlodzinski1, MD; Gabriel Wardi1,2, MD, MPH; Clare Viglione3, MPH, RD; Shamim Nemati4, PhD; Laura Crotty Alexander1,5, MD; Atul Malhotra1, MD

1Division of Pulmonary, Critical Care, Sleep and Physiology, University of California, San Diego, CA, United States
2Department of Emergency Medicine, University of California, San Diego, CA, United States
3Dissemination and Implementation Science Center, Altman Clinical & Translational Research Institute, University of California, San Diego, CA, United States
4Department of Biomedical Informatics, University of California, San Diego, CA, United States
5Section of Pulmonary and Critical Care, Veterans Affairs San Diego Healthcare System, San Diego, CA, United States

Corresponding Author:
Eric Mlodzinski, MD
Division of Pulmonary, Critical Care, Sleep and Physiology
University of California
9300 Campus Point Drive
San Diego, CA, 92037
United States
Phone: 1 858 657 7118
Email: emlodz@gmail.com

Abstract

Background: Although there is considerable interest in machine learning (ML) and artificial intelligence (AI) in critical care, the implementation of effective algorithms into practice has been limited.

Objective: We sought to understand physician perspectives of a novel intubation prediction tool. Further, we sought to understand health care provider and nonprovider perspectives on the use of ML in health care. We aim to use the data gathered to elucidate implementation barriers and determinants of this intubation prediction tool, as well as ML/AI-based algorithms in critical care and health care in general.

Methods: We developed 2 anonymous surveys in Qualtrics, 1 single-center survey distributed to 99 critical care physicians via email, and 1 social media survey distributed via Facebook and Twitter with branching logic to tailor questions for providers and nonproviders. The surveys included a mixture of categorical, Likert scale, and free-text items. Likert scale means with SD were reported from 1 to 5. We used student t tests to examine the differences between groups. In addition, Likert scale responses were converted into 3 categories, and percentage values were reported in order to demonstrate the distribution of responses. Qualitative free-text responses were reviewed by a member of the study team to determine validity, and content analysis was performed to determine common themes in responses.

Results: Out of 99 critical care physicians, 47 (48%) completed the single-center survey. Perceived knowledge of ML was low with a mean Likert score of 2.4 out of 5 (SD 0.96), with 7.5% of respondents rating their knowledge as a 4 or 5. The willingness to use the ML-based algorithm was 3.32 out of 5 (SD 0.95), with 75% of respondents answering 3 out of 5. The social media survey had 770 total responses with 605 (79%) providers and 165 (21%) nonproviders. We found no difference in providers’ perceived knowledge based on level of experience in either survey. We found that nonproviders had significantly less perceived knowledge of ML (mean 3.04 out of 5, SD 1.52 vs mean 3.43, SD 0.94; P<.001) and comfort with ML (mean 3.28 out of 5, SD 1.02 vs mean 3.53, SD 0.93; P=.004) than providers. Free-text responses revealed multiple shared concerns, including accuracy/reliability, data bias, patient safety, and privacy/security risks.

Conclusions: These data suggest that providers and nonproviders have positive perceptions of ML-based tools, and that a tool to predict the need for intubation would be of interest to critical care providers. There were many shared concerns about ML/AI in health care elucidated by the surveys. These results provide a baseline evaluation of implementation barriers and determinants of ML-based tools in critical care and health care in general.
of ML/AI-based tools that will be important in their optimal implementation and adoption in the critical care setting and health care in general.

*(JMIR Perioper Med 2023;6:e41056)*  doi:10.2196/41056

**KEYWORDS**
surveys and questionnaires; machine learning; artificial intelligence; critical care; respiratory insufficiency; survey; Qualtrics; questionnaire; perception; trust; perspective; attitude; intubation; predict; barrier; adoption; implementation

### Introduction

Machine learning (ML) is increasingly used for the development of predictive models in health care, although implementation into clinical care has been limited [1-5]. We have recently reported a deep learning algorithm to predict the need for intubation in patients at risk of respiratory failure in the intensive care unit (ICU) [6]. This algorithm was validated on multiple data sets and was shown to outperform expert clinicians as well as an established predictive model [7]. However, this algorithm is not yet widely implemented.

ML algorithms have been published across nearly all fields of medicine, with models developed for the interpretation of clinical imaging and pathology slides, to assist in the diagnosis of skin lesions, and to predict clinical decompensation and mortality risks in specific populations [8]. In the pulmonary and critical care space, there have been prediction models developed to identify and risk stratify pulmonary nodules on computed tomography scans, and sepsis prediction algorithms to detect clinical decompensation prior to patients meeting clinical sepsis criteria [9-12]. Mechanical ventilation is another area that has seen a growing number of algorithms, with particular focus on predicting successful weaning, ventilator-associated complications, ventilator asynchrony, and timing of need for intubation [3]. However, despite the novelty and potential utility of these models, most have not been used in patient care [1,2]. Thus, further efforts to elucidate barriers and facilitators to implementation are clearly warranted.

There have been a small number of studies evaluating perceptions of ML-based tools among providers or nonproviders using surveys or semistructured interviews. One study of 12 providers in the United Kingdom assessing perception of artificial intelligence (AI) in the National Health Services found concerns over a lack of infrastructure, funding, and a common language [4]. Another study of general practitioners in the United Kingdom suggested that these clinicians felt there was only limited potential for these technologies in their practice [13]. Richardson et al [14] assessed patient perceptions of AI in health care via semistructured interviews and found an overall positive perception of advances in health care technology but also concerns over safety, costs, and patient autonomy. Another study evaluating patient perceptions of ML/AI being used in a skin cancer screening tool found a favorable reception of this technology but only if being used to assist and not replace physician judgment [15]. A similar analysis of patient perception of ML/AI in skeletal radiography revealed a strong preference toward physician interpretation over an AI tool [16]. To our knowledge, there are no studies assessing potential implementation determinants of such tools in the critical care setting.

To identify barriers and facilitators to implementation of a novel tool that predicts the need for mechanical ventilation, as well as to better understand perceptions of ML-based tools across health care, we emailed surveys to providers and shared surveys via social media for both nonproviders and providers. We gathered qualitative information from the surveys to identify potential barriers, which may need to be addressed prior to optimal implementation of these approaches. We further sought to determine whether providers’ level of confidence with and perceived knowledge of ML would be a function of their level of experience. Overall, based on our experience and prior studies, we hypothesized that senior physicians may be less comfortable with ML algorithms compared to their more junior counterparts [16]. Second, we hypothesized that nonproviders would be more skeptical of machine learning tools than the providers who may be using them.

### Methods

#### Survey Development

Our survey items were created and reviewed by a team of 4 critical care providers, 1 machine learning expert, and 1 implementation science expert to ensure completeness, functionality, and appropriate format based on published recommendations for surveys [17-21]. Survey structure and questions were not altered after survey dissemination. Respondents were provided with informed consent (Multimedia Appendices 1 and 2) and had the option to remain completely anonymous.

#### Ethical Considerations

The University of California, San Diego (UCSD) institutional review board reviewed the study and waived the need for approval (UCSD IRB Project #210349XX, “Survey of ICU Clinicians Regarding the Implementation of a Novel EMR-Based Algorithm to Predict Need for Mechanical Ventilation in ICU Patients,” with an amendment for expanded survey with social media recruitment, initial waiver of approval date March 30, 2021, amendment waiver of approval date August 12, 2021).

#### Single-Center Critical Care Physician Survey

Our single-center physician survey (Multimedia Appendix 1) was an open, voluntary, anonymous questionnaire that consisted of 8 items and was distributed to 99 critical care physician trainees and faculty at our institution via email. The survey consisted of 3 pages of content, with 6 multiple-choice and 2 free-response questions. Likert scales of 1-5 were used for...
opinion-based questions, with 1 representing the most negative and 5 the most positive outcome, and 3 representing a “moderate” response. The results are presented as means with SD. Likert scales were also converted to 3 groups with 1 and 2 representing “low,” 3 representing “moderate,” and 4 and 5 representing “high,” and the percentage of each category was reported. Respondents could go back and change answers prior to submitting the survey if desired. Data were collected over a 2-week period in May 2021.

Social Media Survey

Our social media survey (Multimedia Appendix 2) was an open, voluntary, anonymous survey distributed via Twitter and Facebook posts (Multimedia Appendix 3) by our research team. The survey contained 3 pages of content and consisted of an initial question distinguishing medical providers from nonproviders, which then branched into an 11-question survey for providers and 10-question survey for nonproviders. Professions that were under the category of providers included: physicians (practicing or in-training), advanced practice providers, nurses, and medical students. Each survey included a mixture of multiple-choice and free-response questions. Likert scales of 1-5 were used for opinion-based questions, with 1 representing the most negative and 5 the most positive outcome, with 3 representing either a “neutral” or “moderate” response depending on the question. Outcomes are presented as means (SD). Likert scales were also converted to three groups with 1 and 2 representing “low” or “negative,” 3 representing “moderate” or “neutral,” and 4 and 5 representing “high” or “positive,” and the percentage of each category was reported. One adaptive question was used. Respondents could go back and change answers prior to submitting the survey if desired. Providers were offered the chance to complete the survey as a nonprovider as well, although this was not tracked. Data were collected over a 1-month period from September to October 2021. An incentive of an Amazon gift card was offered in the social media survey for one of the respondents chosen randomly.

Survey Analyses

We used the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) to guide our survey reporting and analysis (Multimedia Appendix 4) [22]. No view rate or participation rate was known for either survey. Data were collected and stored securely in Qualtrics for both surveys. The estimated time of each survey was approximately 5 minutes.

We performed a completeness check of our survey data and removed survey responses with <25% completion rate or response times of >30 seconds. No cookies were used in tracking responses. For our social media-based survey, we performed a quality analysis of the survey data and removed responses deemed suspicious for “bot” activity. We screened for suspicious responses by flagging responses with exact matching free-text responses with timestamps within a 4-hour period, and these responses were removed. We also screened for duplicate email addresses, and responses sharing the same email addresses were removed.

Secondary analyses were completed to determine if perceived knowledge of ML/AI and comfort with using ML/AI tools differed by level of provider experience or between nonproviders and providers. We also studied whether prior understanding of ML had an association with potential barriers to implementation of ML algorithms into clinical practice.

Data were analyzed using Excel (v.18.2110.13110.0; Microsoft Corporation) and SPSS Statistics (version 28; IBM Corp.). Descriptive statistics were summarized as previously indicated. Independent t tests were used for comparison of means of each group of interest, with t statistic value, df, and the P value reported for each outcome. Cohen d (for unequal group sizes) values were calculated to estimate effect size between groups where significant differences existed. To minimize extreme responding bias and allow for binary analysis, certain Likert scale results were converted into binary format (using a score of 4-5 as a “positive” response and 1-3 as a “negative” response), which was based on similar methodology used in a previous survey-based study design [23]. Chi-square tests were used for comparing binary responses. Odds ratios with 95% confidence intervals are presented when applicable. A 2-sided α<.05 was considered significant.

Qualitative Analysis

The free-text responses were reviewed by a member of the study team. We removed responses that were deemed to be uninterpretable (due to content unrelated to the topic or nonsensical language). The percentage of respondents who provided valid responses for our qualitative questions were determined. We performed content analysis for each free-text response for both surveys regarding concerns of use of ML/AI in practice and determined shared themes across all surveys. Each response could be categorized into one or more themes. Only themes with at least 5% of responses fitting within that category were reported.

Results

Single-Center Critical Care Physician Survey

Out of 79 physicians, 47 completed this internal institutional survey. The results of the survey are displayed in Tables 1 and 2 and Figure 1. All means are presented with SD. A total of 31 (59%) respondents were attendings, 19 (36%) were fellows, and 2 (4%) were residents. Perceived knowledge of ML was low (mean 2.40, SD 0.96), with 7.5% of respondents rating their knowledge as a 4 or 5. A total of 8 (15%) respondents had knowingly used an ML-based tool in their clinical practice. Confidence in predicting the need for mechanical ventilation due to COVID-19 pneumonia (mean 3.57, SD 0.79) was lower than for respiratory failure due to all other causes (mean 3.89, SD 0.78). Overall, willingness to use an ML-based algorithm was 3.32 (SD 0.95), with 75% of respondents rating their confidence in predicting the need for mechanical ventilation as a 4 or 5. A total of 8 (15%) respondents had knowingly used an ML-based tool in their clinical practice. Confidence in predicting the need for mechanical ventilation due to COVID-19 pneumonia (mean 3.57, SD 0.79) was lower than for respiratory failure due to all other causes (mean 3.89, SD 0.78). Overall, willingness to use an ML-based algorithm was 3.32 (SD 0.95), with 75% of respondents rating their confidence in predicting the need for mechanical ventilation as a 4 or 5. A total of 8 (15%) respondents had knowingly used an ML-based tool in their clinical practice.
out of 47 valid responses, with 2 responses removed. Shared themes and responses per theme included: accuracy/reliability (n=7, 39%), workflow interruptions/alert fatigue (n=6, 33%), patient safety (n=1, 6%), and data bias (n=2, 11%). Representative examples are shown in Multimedia Appendix 5. For question 8 regarding suggestions on ways to improve implementation, 16 (34%) out of 47 participants provided valid responses, with no responses removed. Shared themes and responses per theme included: prospective data/proof of efficacy (n=10, 63%), electronic medical record integration (n=3, 19%), and data transparency (n=3, 19%).

Table 1. Characteristics of respondents to a single-center survey.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Respondents, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Response characteristics</strong></td>
<td></td>
</tr>
<tr>
<td>Response rate</td>
<td>53/99 (53)</td>
</tr>
<tr>
<td>Completion rate</td>
<td>47/53 (89)</td>
</tr>
<tr>
<td><strong>Level of experience (n=53)</strong></td>
<td></td>
</tr>
<tr>
<td>Attending</td>
<td>31 (58)</td>
</tr>
<tr>
<td>Fellow</td>
<td>19 (36)</td>
</tr>
<tr>
<td>Resident</td>
<td>2 (3)</td>
</tr>
<tr>
<td>Other</td>
<td>1 (2)</td>
</tr>
<tr>
<td><strong>Prior use of machine learning (n=53)</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>8 (15)</td>
</tr>
<tr>
<td>No</td>
<td>35 (66)</td>
</tr>
<tr>
<td>Unsure</td>
<td>10 (18)</td>
</tr>
</tbody>
</table>

Table 2. Mean scores of Likert scale questions for a single-center survey.

<table>
<thead>
<tr>
<th>Survey question</th>
<th>Score (1-5), mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q2. Level of knowledge of ML&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.40 (0.968)</td>
</tr>
<tr>
<td>Q4a. Confidence in the ability to predict the need for mechanical ventilation in COVID-19</td>
<td>3.57 (0.801)</td>
</tr>
<tr>
<td>Q4b. Confidence in the ability to predict the need for mechanical ventilation for all other causes</td>
<td>3.89 (0.787)</td>
</tr>
<tr>
<td>Q5. Willingness to use ML-based tools to predict respiratory failure</td>
<td>3.32 (0.958)</td>
</tr>
<tr>
<td><strong>Factors impacting the likelihood of using the tool</strong></td>
<td></td>
</tr>
<tr>
<td>Q6a. High-quality evidence available</td>
<td>4.28 (0.772)</td>
</tr>
<tr>
<td>Q6b. Limited workflow interruption</td>
<td>4.09 (0.974)</td>
</tr>
<tr>
<td>Q6c. Transparency of the data</td>
<td>4.13 (0.797)</td>
</tr>
<tr>
<td>Q6d. Real-time probability data of likelihood of need for mechanical ventilation</td>
<td>3.91 (0.974)</td>
</tr>
<tr>
<td>Q6e. Support from other intensive care unit clinicians and hospital leadership</td>
<td>3.57 (1.12)</td>
</tr>
<tr>
<td>Q6f. Standardized education on ML</td>
<td>3.34 (1.07)</td>
</tr>
</tbody>
</table>

<sup>a</sup>ML: machine learning.
Social Media Provider and Nonprovider Survey

We received 1196 responses, with 914 provider and 282 nonprovider responses. We excluded a total of 426 (35.6%) responses, 309 (33.8%) provider responses and 117 (41.5%) nonprovider responses. The reasons for exclusion included duplicate open-ended responses (n=324, 76.1%), duplicate email addresses (n=30, 7%), and <25% completion rate or <30-second time to completion (n=72, 16.9%). Our final analysis included a total of 770 total responses made up of 605 (78.6%) providers and 165 (21.4%) nonprovider responses. Descriptive results are displayed in Tables 3 and 4, Figures 2 and 3, and Multimedia Appendix 6. Physicians made up most of the respondents of the provider survey (n=372, 61.5%), with more attendings than medical trainees. A total of 21% (n=127) of respondents reported working in a critical care setting. Mean baseline understanding of ML/AI was 3.43 (SD 0.97), with 49% of respondents reporting a “high” level of knowledge. A total of 74% of respondents reported having used ML. Overall comfort with using an ML tool in patient care was 3.53 (SD 0.967), with 51.7% of respondents reporting a “high” level of comfort.

For the free-text question, providers and nonproviders were asked to share any concerns they had regarding ML or AI in health care. For the providers, 312 (52%) participants of 605 provided valid responses to the free-text question, with 28 responses removed. Of the 312 total responses, 56 (16.5%) reported no concerns and 256 (75.3%) responses included a concern. Shared themes and responses per theme included the following: accuracy/reliability (n=58, 22.7%), data bias (n=35, 13.7%), patient safety/outcomes (n=34, 13.3%), doctor-patient relationship (n=28, 10.9%), privacy/security (n=22, 8.6%), workflow (n=19, 7.4%), and costs (n=14, 5.5%). Representative examples are shown in Multimedia Appendix 5. For nonproviders, 109 (66%) participants provided valid free response, with 7 responses excluded. Of those 109 valid responses, 6 (5.5%) reported no concerns and 103 (94.5%) provided concerns. Shared themes and responses per theme included the following: accuracy/reliability (n=22, 21.4%), data bias (n=22, 21.4%), privacy/security (n=16, 15.5%), patient safety/outcomes (n=11, 10.7%), lack of knowledge of ML/AI (n=11, 10.7%), and doctor-patient relationship (n=10, 10.3%). Representative examples are shown in Multimedia Appendix 5.
Table 3. Likert scale responses of social media survey health care provider subgroup.

<table>
<thead>
<tr>
<th>Survey question</th>
<th>Score (1-5), mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q2. How would you rate your current understanding of ML(^a)/AI(^b) as they apply to health care?</td>
<td>3.43 (0.970)</td>
</tr>
<tr>
<td>Q3a. How useful was this tool?</td>
<td>3.73 (0.917)</td>
</tr>
<tr>
<td>Q4. How comfortable would you feel using an ML or AI-based tool to make a clinical decision regarding your patients?</td>
<td>3.53 (0.967)</td>
</tr>
<tr>
<td>Q5. How concerned are you that ML/AI will make some health care jobs/specialties obsolete?</td>
<td>3.41 (1.13)</td>
</tr>
</tbody>
</table>

Please choose the option which best describes your opinion on how the implementation of ML/AI-based tools into routine clinical practice would impact each of the following

- Q6a. Patient care
- Q6b. Efficiency in your daily practice
- Q6c. Patient-provider relationship

<table>
<thead>
<tr>
<th>Survey question</th>
<th>Score (1-5), mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please rate the extent to which each of the following factors would increase your likelihood of using an ML or AI-based tool in your clinical practice</td>
<td></td>
</tr>
<tr>
<td>Q7a. High-quality evidence of tool’s efficacy</td>
<td>3.66 (0.995)</td>
</tr>
<tr>
<td>Q7b. Transparency of data</td>
<td>3.67 (1.02)</td>
</tr>
<tr>
<td>Q7c. Workflow interruptions</td>
<td>3.56 (1.06)</td>
</tr>
<tr>
<td>Q7d. Standardized education on ML/AI tools</td>
<td>3.63 (1.001)</td>
</tr>
<tr>
<td>Q7e. Support from administration</td>
<td>3.65 (1.025)</td>
</tr>
</tbody>
</table>

\(^a\)ML: machine learning.  
\(^b\)AI: artificial intelligence.

Table 4. Likert scale responses of social media survey nonprovider subgroup.

<table>
<thead>
<tr>
<th>Survey question</th>
<th>Score (1-5), mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q2. How much confidence do you have in medical professionals' ability to make the correct decision for your medical care?</td>
<td>3.66 (0.959)</td>
</tr>
<tr>
<td>Q3. How would you rate your current understanding of ML(^a) and AI(^b) as they apply to health care?</td>
<td>3.03 (1.233)</td>
</tr>
<tr>
<td>Q4. How comfortable would you be with having a computer algorithm using ML/AI assisting in making decisions about your medical care?</td>
<td>3.27 (1.013)</td>
</tr>
<tr>
<td>Q5. How do you think the implementation of more ML/AI-based algorithms into the medical system will impact your medical care?</td>
<td>3.40 (1.014)</td>
</tr>
<tr>
<td>Q6. How do you think the implementation of more ML/AI-based algorithms into the medical system will impact your relationship with your medical team</td>
<td>3.09 (0.931)</td>
</tr>
</tbody>
</table>

Please rate the extent to which each of the following factors would increase your comfort level with an ML or AI-based tool being used in your medical care

- Q8a. High-quality evidence that it is as good or better than trained clinicians
- Q8b. High-quality evidence that it can improve patient outcomes
- Q8c. Knowing how the tool was developed
- Q8d. Knowing that the tool would improve efficiency

<table>
<thead>
<tr>
<th>Survey question</th>
<th>Score (1-5), mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please rate the extent to which each of the following factors would increase your comfort level with an ML or AI-based tool being used in your medical care</td>
<td></td>
</tr>
<tr>
<td>Q8a. High-quality evidence that it is as good or better than trained clinicians</td>
<td>3.65 (1.127)</td>
</tr>
<tr>
<td>Q8b. High-quality evidence that it can improve patient outcomes</td>
<td>3.77 (1.214)</td>
</tr>
<tr>
<td>Q8c. Knowing how the tool was developed</td>
<td>3.62 (1.082)</td>
</tr>
<tr>
<td>Q8d. Knowing that the tool would improve efficiency</td>
<td>3.56 (1.141)</td>
</tr>
</tbody>
</table>

\(^a\)ML: machine learning.  
\(^b\)AI: artificial intelligence.
Figure 2. Provider survey Likert scale results. Responses were separated into 3 categories; “low,” “moderate,” or “high,” depicted in the top graph, and “negative,” “neutral,” or “positive,” depicted in the bottom graph. Results are reported as percentage of valid responses out of 100%. Question content can be found in Table 3 and Multimedia Appendix 2.

<table>
<thead>
<tr>
<th>Question Label</th>
<th>Response (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q2</td>
<td>16.3 34.7 49</td>
</tr>
<tr>
<td>Q3a</td>
<td>3.6 30.9 60.5</td>
</tr>
<tr>
<td>Q4</td>
<td>13.3 35.1 51.7</td>
</tr>
<tr>
<td>Q5</td>
<td>10.2 28.1 52.7</td>
</tr>
<tr>
<td>Q7a</td>
<td>12.1 28.6 59.2</td>
</tr>
<tr>
<td>Q7b</td>
<td>11.9 22.4 55.7</td>
</tr>
<tr>
<td>Q7c</td>
<td>15.3 25.5 59.2</td>
</tr>
<tr>
<td>Q7d</td>
<td>12.5 29.2 58.3</td>
</tr>
<tr>
<td>Q7e</td>
<td>12.9 27.7 59.4</td>
</tr>
</tbody>
</table>

![Low, Moderate, High]

<table>
<thead>
<tr>
<th>Question Label</th>
<th>Response (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q6a</td>
<td>15.5 27.5 57.1</td>
</tr>
<tr>
<td>Q6b</td>
<td>12.3 28.1 59.6</td>
</tr>
<tr>
<td>Q6c</td>
<td>21.7 33.4 44.9</td>
</tr>
</tbody>
</table>

![Negative, Neutral, Positive]
Secondary Analyses
In the single-center survey, there was no significant difference between critical care trainees and attendings in terms of overall perceived knowledge of ML (mean 2.19, SD 0.991 vs mean 2.50, SD 0.862; \( t_{50} = 1.19; P = .24 \)) or willingness to use an ML prediction tool (mean 3.38, SD 1.05 vs 3.30, SD 0.907; \( t_{44} = 0.248; P = .80 \)). For the social media survey, there was no significant difference between trainees and attending physicians in perceived knowledge (mean 3.53, SD 0.840 vs mean 3.42, SD 1.03; \( t_{495} = 1.20; P = .23 \)) or comfort with ML tools (mean 3.60, SD 0.850 vs mean 3.51, SD 1.04; \( t_{492} = 0.943; P = .35 \)). There was a significant difference between physician and nonprovider knowledge of ML in health care (mean 3.43, SD 0.941 vs mean 3.04, SD 1.53; \( t_{352} = 4.15; P < .001 \)) and with comfort in using these tools (mean 3.53, SD 0.935 vs mean 3.28, SD 1.02; \( t_{46} = 2.90; P = .004 \)). Cohen \( d \) values were 0.33 and 0.28, respectively, suggesting a low effect size. Comparison of critical care physicians between the 2 surveys regarding their perceived knowledge of ML revealed a significantly lower perceived knowledge among the single-center survey respondents (mean 2.40, SD 0.936 vs mean 3.27 SD 1.01; \( t_{141} = 5.08; P < .001 \)). Cohen \( d \) value was 0.91, suggesting a large effect size. In a binary analysis of providers’ baseline knowledge (high vs low), there was not a significant association between baseline knowledge and willingness to use ML in patient care (OR 2.270, 95% CI 0.694-7.424; \( P = .17 \)). In a binary analysis of nonproviders’ perceived knowledge of ML (high vs low), there was a significant association between higher knowledge of ML and more comfort with ML being used in patient care (OR 6.25, 95% CI 3.05-12.84; \( P < .001 \)). The results are displayed in Table 5.
Table 5. Secondary analysis.

<table>
<thead>
<tr>
<th>Secondary analysis subgroups</th>
<th>Score (1-5), mean (SD)</th>
<th>t score</th>
<th>P valuea</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single-center survey</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline knowledge of MLb</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attending</td>
<td>2.52 (0.991)</td>
<td>-1.19</td>
<td>.24</td>
</tr>
<tr>
<td>Trainee</td>
<td>2.19 (0.862)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Willingness to use ML tool</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attending</td>
<td>3.30 (0.907)</td>
<td>0.248</td>
<td>.81</td>
</tr>
<tr>
<td>Trainee</td>
<td>3.38 (1.05)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Social media survey</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline knowledge of ML</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attending</td>
<td>3.42 (1.03)</td>
<td>-1.20</td>
<td>.23</td>
</tr>
<tr>
<td>Trainee</td>
<td>3.53 (0.840)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comfort with using ML</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attending</td>
<td>3.51 (1.04)</td>
<td>0.943</td>
<td>.35</td>
</tr>
<tr>
<td>Trainee</td>
<td>3.60 (0.850)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline knowledge of ML</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Provider</td>
<td>3.43 (0.941)</td>
<td>4.15</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Nonprovider</td>
<td>3.04 (1.53)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comfort with using ML tool</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Provider</td>
<td>3.53 (0.935)</td>
<td>2.90</td>
<td>.004</td>
</tr>
<tr>
<td>Nonprovider</td>
<td>3.28 (1.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cross-survey analysis</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline knowledge of ML</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical care providers single center</td>
<td>2.4 (0.936)</td>
<td>5.08</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Critical care providers social media</td>
<td>3.27 (1.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square analysis social media surveyc</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Association of nonprovider comfort with ML and baseline knowledge of ML</td>
<td>6.25 (3.05-12.84)</td>
<td>N/Ad</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Association of provider comfort with ML and baseline knowledge of ML</td>
<td>2.270 (0.694-7.424)</td>
<td>N/A</td>
<td>.167</td>
</tr>
</tbody>
</table>

a Independent student t tests were used for comparison of means; chi-square tests were used for comparing binary responses. P<.05 was considered significant.
bML: machine learning.
cOdds ratios and 95% CI are provided for this category.
dN/A: not applicable.

Discussion

Principal Findings

To our knowledge, this is the first study to explore nonprovider and provider perspectives of novel ML-based tools in critical care as well as potential implementation determinants of these tools. We found that both providers and nonproviders have favorable attitudes toward the use of ML in health care, although there remained a small but significant difference between these 2 groups with providers having more comfort overall. Nonproviders with more perceived knowledge of the concept of ML/AI were more likely to feel favorable toward its use in patient care. This finding suggests that efforts to implement ML tools may require increased focus on nonprovider education and buy-in as skepticism may be more pronounced in this group [14,24]. Second, we observed no major difference in the level of knowledge or comfort among providers regardless of their level of experience in either survey, which contradicts our preconceived notions of older providers being less comfortable with technological advancements in medicine. Third, we identified nonprovider and provider concerns about potential systemic bias in data used in ML tools, patient safety, negative effects on the doctor-patient relationship, and data privacy/security. Among providers, we also identified workflow...
interruptions as a major concern, and among nonproviders, limited knowledge of ML/AI was a major concern. These are critical factors that will need to be addressed to ensure user confidence in the data and algorithms. We also saw a large difference in comfort with ML among our own institution’s critical care physicians compared to the more generalized critical care physicians, suggesting that institutional differences are likely to exist and that implementation methods may need to be tailored for each institution.

Our single-center survey provided important information regarding physician acceptance of a novel algorithm for predicting the onset of mechanical ventilation in patients at risk of respiratory failure. One of the goals of this survey was to serve as a needs assessment for this tool, and based on our results, it appears that providers at our institution feel that this tool would be beneficial in the clinical context of early prediction of the need for intubation among patients with respiratory failure due to COVID-19 and all other causes [6]. These results support our team’s efforts in moving forward with the next steps of implementation, which will involve optimizing the interface for provider ease of use, preliminary prospective studies of its efficacy, and improving the sensitivity and specificity. Future steps include co-creating implementation strategies with a multidisciplinary team of clinicians, patients, implementation scientists, and medical informaticists to address identified determinants to improve the uptake and implementation of this algorithm. This process will also include additional surveys and structured interviews to assess ongoing effectiveness and to iteratively refine the algorithm to optimize its utility and improve clinical care.

**Strengths and Limitations**

One of the study’s main strengths was the use of multiple platforms including social media for dissemination of our surveys, improving the generalizability of our results and allowing us to reach a large sample size. In addition, our surveys were unique in that we were able to gather both nonprovider and provider perspectives simultaneously. We also screened for suspicious responses in the social media surveys and removed these to increase the reliability of our survey results. Despite our study’s strengths, we acknowledge the following limitations. First, due to privacy issues, we did not collect demographic or other personal data regarding the respondents. Thus, we are unable to draw conclusions regarding whether certain members of nonprovider and provider communities may be more amenable to ML methods (eg, based on gender or race). Second, our conclusions are limited to the population studied as our surveys were in English and only reached those with electronic access. Third, as with any survey, there are risks of both selection bias as well as participation bias. For selection bias in the first survey, we emailed ICU providers but did not gather any systematic data from ICU nurses or pharmacists or others who may be impacted by these tools. Regarding participation bias, it is likely that the individuals responding to social media survey would be those with an interest in this topic and thus may be more comfortable with these methods than others. Fourth, the truly open nature of the social media survey led to unanticipated issues with bot responses, and while steps were taken to remove suspicious responses, to our knowledge, there is no validated means of screening for bot activity. Fifth, there was no specific implementation conceptual framework used in the development of questions addressing implementation barriers and facilitators. Despite these limitations, we view our findings as an important step toward the successful implementation of ML/AI methods to improve patient care.

**Conclusions**

Both providers and nonproviders have overall positive perspectives on the use of ML-based tools in health care, although nonproviders remain more skeptical. In addition, it appears that a tool to help predict onset of the need for intubation would be both useful and acceptable among critical care providers. Our study revealed shared concerns regarding accuracy and reliability, data bias, privacy/security, patient safety, the doctor-patient relationship, and workflow interruptions. These data provide a baseline assessment of health care provider and nonprovider perceptions of ML/AI-based tools that will be crucial in optimizing their clinical utility.
Authors' Contributions

All authors were involved in the conception and design of the study. EM was involved in data acquisition. EM, GW, and SN were involved in the data analysis. EM, GW, and AM were involved in manuscript writing. All authors were involved in manuscript review and final approval.

Conflicts of Interest

SN is a cofounder, advisor, and holds equity in Healcisio Inc, a startup company which is developing products related to the research described in this paper. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

Multimedia Appendix 1
Single-center critical care physician survey.
[DOCX File, 19 KB - periop_v6i1e41056_app1.docx]

Multimedia Appendix 2
Social media survey content.
[DOCX File, 22 KB - periop_v6i1e41056_app2.docx]

Multimedia Appendix 3
Social media recruitment post examples.
[DOCX File, 14 KB - periop_v6i1e41056_app3.docx]

Multimedia Appendix 4
Checklist for Reporting Results of Internet E-Surveys (CHERRIES).
[DOCX File, 19 KB - periop_v6i1e41056_app4.docx]

Multimedia Appendix 5
Qualitative evaluation of provider and patient concerns regarding ML/AI in healthcare.
[DOCX File, 23 KB - periop_v6i1e41056_app5.docx]

Multimedia Appendix 6
Demographics and categorical responses to social media survey.
[DOCX File, 25 KB - periop_v6i1e41056_app6.docx]

References


Abbreviations
AI: artificial intelligence
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
ICU: intensive care unit
ML: machine learning
UCSD: University of California, San Diego