ANESTHESIOLOGY

Prediction of Postoperative Deterioration in Cardiac Surgery Patients Using Electronic Health Record and Physiologic Waveform Data

Michael R. Mathis, M.D., Milo C. Engoren, M.D., Aaron M. Williams, M.D., Ben E. Biesterveld, M.D., Alfred J. Croteau, M.D., Lingrui Cai, B.S., Renaid B. Kim, B.S., Gang Liu, Ph.D., Kevin R. Ward, M.D., Kayvan Najarian, Ph.D., Jonathan Gryak, Ph.D.; BCIL Collaborators Group*

Anesthesiology 2022; 137:586–601

ABSTRACT

Background: Postoperative hemodynamic deterioration among cardiac surgical patients can indicate or lead to adverse outcomes. Whereas prediction models for such events using electronic health records or physiologic waveform data are previously described, their combined value remains incompletely defined. The authors hypothesized that models incorporating electronic health record and processed waveform signal data (electrocardiogram lead II, pulse plethysmography, arterial catheter tracing) would yield improved performance *versus* either modality alone.

Methods: Intensive care unit data were reviewed after elective adult cardiac surgical procedures at an academic center between 2013 and 2020. Model features included electronic health record features and physiologic waveforms. Tensor decomposition was used for waveform feature reduction. Machine learning–based prediction models included a 2013 to 2017 training set and a 2017 to 2020 temporal holdout test set. The primary outcome was a postoperative deterioration event, defined as a composite of low cardiac index of less than 2.0 ml min^{-1} m^{-2}, mean arterial pressure of less than 55 mmHg sustained for 120 min or longer, new or escalated inotrope/vasopressor infusion, epinephrine bolus of 1 mg or more, or intensive care unit mortality. Prediction models analyzed data 8 h before events.

Results: Among 1,555 cases, 185 (12%) experienced 276 deterioration events, most commonly including low cardiac index (7.0% of patients), new inotrope (1.9%), and sustained hypotension (1.4%). The best performing model on the 2013 to 2017 training set yielded a C-statistic of 0.803 (95% CI, 0.799 to 0.807), although performance was substantially lower in the 2017 to 2020 test set (0.709, 0.705 to 0.712). Test set performance of the combined model was greater than corresponding models limited to solely electronic health record features (0.641; 95% CI, 0.637 to 0.646) or waveform features (0.697; 95% CI, 0.693 to 0.701).

Conclusions: Clinical deterioration prediction models combining electronic health record data and waveform data were superior to either modality alone, and performance of combined models was primarily driven by waveform data. Decreased performance of prediction models during temporal validation may be explained by data set shift, a core challenge of healthcare prediction modeling.

(Anesthesiology 2022; 137:586–601)

EDITOR'S PERSPECTIVE

What We Already Know about This Topic

- Hemodynamic deterioration after cardiac surgery can range from easily reversable to severe sustained events and may lead to clinically relevant adverse outcomes
- Clinicians currently rely on close clinical observation and experiential judgment to anticipate and treat such events
- Little is known regarding machine learning approaches to real-time prediction after cardiac surgery based on data available at the bedside in the electronic health record or features extracted from commonly used physiologic monitoring devices
- The authors have previously developed advanced signal processing techniques for feature extraction from lead II of the electrocardiogram, the invasive arterial waveform, and peripheral plethysmography

What This Article Tells Us That Is New

- In this single-center, retrospective cohort study, the authors studied machine learning–based prediction models for postoperative hemodynamic deterioration using discrete electronic health record and continuous physiologic waveform data, alone or in combination, for 1,555 patients after cardiac surgery
- All patients had pulmonary artery catheters placed during surgery per institutional protocol, allowing the thermodilution-derived cardiac index to be included as a key component of the composite hemodynamic endpoint
- The best performing model in the training data set (2013 to 2017) used both data sources (area under the curve, 0.803) but was primarily driven by waveform data, suggesting that a black box waveform approach alone may have clinical utility in this setting
- However, validation of these approaches in a later data set (2017 to 2020) showed substantially decreased performance (area under the curve, 0.709), most likely consistent with the phenomena of data set shift

This article is featured in "This Month in Anesthesiology," page A1. Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are available in both the HTML and PDF versions of this article. Links to the digital files are provided in the HTML text of this article on the Journal's Web site (www.anesthesiology.org). This article has a visual abstract available in the online version. This work was presented at the International Anesthesia Research Society 2021 Annual Meeting on May 15, 2021 (virtual).

Copyright © 2022, the American Society of Anesthesiologists. All Rights Reserved. Anesthesiology 2022; 137:586–601. DOI: 10.1097/ALN.0000000000004345

Approximately 300,000 patients across the United States undergo cardiac surgeries annually, and nearly all receive postoperative intensive care unit (ICU) care.[1] In the ICU, up to 20% of cardiac surgical patients incur complications,[2,3] associated with $36,000 increased cost per case and up to eightfold increased adjusted odds of mortality.[4–6] Postoperative hemodynamic deterioration, either secondary to structural heart complications from the surgery or as a manifestation of underlying pathophysiologic processes exacerbated by the surgical insult, may lead to inadequate end-organ perfusion and precipitate life-threatening adverse events.[7,8] However, if identified early, appropriate treatments—including fluid resuscitation, pharmacologic therapies, mechanical ventilation management, and procedural interventions—may prevent or diminish postoperative adverse outcomes.[9] Through potentially reducing rates of failure to rescue, early detection and management of hemodynamic deterioration represent opportunities for significant reductions in healthcare costs and improved outcomes.[10,11]

In the ICU, a wealth of electronic health record data is collected postoperatively. In addition to electronic health record data, high-fidelity physiologic waveform data containing valuable diagnostic information are increasingly collected in perioperative and critical care settings.[12] With such data sources growing in size and complexity, emerging demands may be placed upon skilled ICU teams to synthesize, interpret, and act upon acute patient conditions conveyed through the data. In many cases, hemodynamic deterioration is recognized in a timely fashion, and life-saving treatments are administered. However, occasionally, recognizable features of hemodynamic deterioration are elicited, but due to cognitive overload or limited ICU clinical team resources, synthesis and interpretation of such features are delayed, and opportunities for early interventions are missed.[13] Still, in other cases, subclinical features of deterioration may elude human clinician detection before clinically overt patient compromise, and opportunities for early interventions are also missed.[14,15] Data science approaches may overcome such issues through improved synthesis of diverse, complex health data for detecting digital signatures of early-stage clinical deterioration.[16–20]

Through this observational study of high-fidelity electronic health record and physiologic waveform ICU data from an academic quaternary care hospital, we leveraged machine learning techniques for early detection of postoperative deterioration among patients undergoing cardiac surgical procedures. We hypothesized that patterns exist within both electronic health record data and physiologic waveform data predictive of hemodynamic deterioration and that the performance of models using both electronic health record and physiologic waveform data to predict postoperative deterioration is superior to models using either modality alone.

Materials and Methods

Study Design

We followed multidisciplinary guidelines for developing and reporting of machine learning predictive models[21] and strengthening the reporting of observational studies in epidemiology[22] throughout conducting this study (Supplemental Digital Content 1, http://links.lww.com/ALN/C892). We obtained institutional review board approval (HUM00092309) for this observational study, and patient consent was waived. We established an *a priori* study protocol and registered our observational study within a peer-review forum before data access.[23]

Submitted for publication June 28, 2021. Accepted for publication July 20, 2022. Published online first on August 11, 2022.

Michael R. Mathis, M.D.: Department of Anesthesiology, University of Michigan Health System, Ann Arbor, Michigan; Department of Computational Medicine and Bioinformatics, University of Michigan Health System, Ann Arbor, Michigan; Michigan Integrated Center for Health Analytics and Medical Prediction, Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor, Michigan; and Michigan Center for Integrative Research in Critical Care, University of Michigan, Ann Arbor, Michigan.

Milo C. Engoren, M.D.: Department of Anesthesiology, University of Michigan Health System, Ann Arbor, Michigan.

Aaron M. Williams, M.D.: Department of General Surgery, University of Michigan Health System, Ann Arbor, Michigan.

Ben E. Biesterveld, M.D.: Department of General Surgery, University of Michigan Health System, Ann Arbor, Michigan.

Alfred J. Croteau, M.D.: Department of General Surgery, Hartford HealthCare Medical Group, Hartford, Connecticut.

Lingrui Cai, B.S.: Department of Computational Medicine and Bioinformatics, University of Michigan Health System, Ann Arbor, Michigan.

Renaid B. Kim, B.S.: Department of Computational Medicine and Bioinformatics, University of Michigan Health System, Ann Arbor, Michigan.

Gang Liu, Ph.D.: Department of Computational Medicine and Bioinformatics, University of Michigan Health System, Ann Arbor, Michigan.

Kevin R. Ward, M.D.: Michigan Integrated Center for Health Analytics and Medical Prediction, Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor, Michigan; Michigan Center for Integrative Research in Critical Care, University of Michigan, Ann Arbor, Michigan; and Department of Emergency Medicine, University of Michigan Health System, Ann Arbor, Michigan.

Kayvan Najarian, Ph.D.: Department of Computational Medicine and Bioinformatics, University of Michigan Health System, Ann Arbor, Michigan; Michigan Integrated Center for Health Analytics and Medical Prediction, Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor, Michigan; and Michigan Center for Integrative Research in Critical Care, University of Michigan, Ann Arbor, Michigan.

Jonathan Gryak, Ph.D.: Department of Computational Medicine and Bioinformatics, University of Michigan Health System, Ann Arbor, Michigan; and Michigan Center for Integrative Research in Critical Care, University of Michigan, Ann Arbor, Michigan.

*Members of the BCIL Collaborators Group are listed in Appendix 1.

Study Population

We studied elective cardiac surgical procedures with full cardiopulmonary bypass performed on adult patients from February 1, 2013, to January 31, 2020, at our quaternary care center. For purposes of developing a cohort reflective of typical cardiac surgical procedures, we restricted our cohort to patients more than 40 yr old undergoing coronary artery bypass grafting, valve, and thoracic aortic procedures performed in isolation or combination. Procedures were limited to open cardiac procedures; transcatheter or robotic procedures were excluded. Cases were restricted to those with valid postoperative ICU physiologic waveform data available within the waveform repository (Michigan Anesthesiology Informatics and Systems Improvement Exchange, Ann Arbor, Michigan). To enable ascertainment of all possible deterioration events including pulmonary artery catheter–derived cardiac output (described in Appendix 2), we additionally restricted cases to those performed on patients with pulmonary artery catheter monitoring, as was our institutional practice pattern for cardiac surgical patients during the study period.

Data Source and Model Features

The data were extracted from the electronic health record (Epic Systems Corporation, Verona, Wisconsin) within the cardiovascular ICU of our institution. Prediction model features were selected on the basis of electronic health record and waveform data potentially available in near-real time at the point of care, among which included those in previously described cardiac surgical risk models[24] and other commonly used critical illness scoring systems.[25,26] These included patient demographics and comorbidities, lab values, surgical details, left ventricular ejection fraction, and nurse-validated vital signs, as well as features extracted from three widely available physiologic waveforms available during postoperative ICU care: an electrocardiogram (ECG) lead II (240 Hz sampling rate), an invasive arterial line (120 Hz), and a pulse plethysmographic waveform generated from a pulse oximeter (60 Hz). Additional physiologic waveforms available in a subset of patients (central venous pressure, pulmonary artery pressure) were not used in this study, due to (1) incomplete data, (2) nonrandom missingness, (3) frequent episodes with unusable data due to waveform artifact (*e.g.*, clamped line, wedged catheter, fluid/infusion-induced alterations), and (4) an inability to generalize the models developed to other ICU populations for which such waveforms were not commonly available. Waveform data were collected with a physiologic data integration system (Capsule, Capsule Technologies, USA).

Primary Outcome: Postoperative Deterioration Events

Postoperative hemodynamic deterioration events were determined by an *a priori* consensus agreement among three cardiac anesthesiologist and critical care physicians (M.R.M., M.C.E., K.J.G.) familiar with the cardiovascular ICU processes of care, documentation patterns, and quality of the ICU electronic health record data available for review. In contrast to traditional cardiac surgical risk models (*e.g.*, Society for Thoracic Surgeons risk calculator, EuroSCORE-II)[2,24] predicting broad complications over the entirety of the postoperative hospital stay and requiring abstraction *via* manual chart review, our approach favored dynamic, time-limited deterioration events that could potentially be detected *via* automated means in real time to enable ease of future model deployment at the point of care, with the potential for continuous retraining and tuning. To this end, we selected hemodynamic deterioration events that were (1) self-included within structured electronic health record data, (2) specific to discrete time intervals during postoperative recovery, (3) nurse-validated at the time of data entry, and (4) feasible to be retrospectively adjudicated *via* physician manual chart review.

Events were defined as a composite outcome in which any of the following occurred: new low cardiac index (less than 2.0 l min^{-1} m^{-2} assessed *via* pulmonary arterial catheter thermodilution), new sustained hypotension (mean arterial pressure of less than 55 mmHg for 120 continuous minutes or longer; *i.e.*, three consecutive nurse-validated hourly blood pressure values), epinephrine bolus of 1 mg or more (*e.g.*, advanced cardiac life support for cardiac arrest), new inotrope initiated, new vasopressor initiated, inotrope infusion dose rate escalation of 100% or more, vasopressor infusion dose rate escalation of 100% or more, and in-ICU all-cause 90-day mortality. Full specifications of postoperative deterioration event definitions are provided in Appendix 2. To focus on events of potential high relevance to ICU care teams and mitigate against potential alert fatigue, patients incurring ongoing low cardiac indices or ongoing mean arterial pressure of less than 55 mmHg for periods for periods far beyond 120 continuous minutes were not considered to have new deterioration events, unless triggering a different deterioration event (*e.g.*, new vasopressor/inotrope infusion or dose escalation) or returning to a nondeteriorated state for more than 48 h. Pulmonary arterial catheter thermodilution–based cardiac index measurements were obtained by ICU nurses educated and tested on such measurements; measurements were obtained in triplicate, and the averages were recorded. As a standard practice within our institution's cardiac surgical ICU, cardiac indices were assessed every 4 h; less frequently if instructed by the physician-directed ICU team for patients judged to be hemodynamically stable; and more frequently if therapies (*e.g.*, fluids, medications) were being actively titrated based upon the cardiac index. For the sustained hypotension outcome, in cases of multiple concurrent blood pressure data sources (*e.g.*, noninvasive blood pressure and invasive arterial line), the highest nurse-validated blood pressure was considered as the gold standard. For inotrope and vasopressor dose rate escalations, the events were restricted to those in which the previous infusion rate was above prespecified infusion-specific thresholds (*e.g.*, for a norepinephrine infusion, a threshold rate was selected as 0.10 µg kg^{-1} min^{-1}; a dose rate escalation of 100% or more was only considered if the previous rate was at or above the 0.10 µg kg^{-1} min^{-1} threshold).

Importantly, for purposes of enabling before-event prediction window ICU data to be available and improving the clinical utility of the postoperative deterioration events studied, primary outcomes were restricted to late deterioration events only. Postoperative deterioration events were defined as events occurring after 24 h postoperatively, such that at least 24 h of ICU data were available upon which to base deterioration predictions. Furthermore, patients with events occurring during the initial 24 h of ICU admission were judged by ICU physician reviewers to commonly receive continuous active resuscitation initiated at the time of patient handover from the intraoperative team, by an ICU team already cognizant of current or impending deterioration, limiting the utility of the prediction algorithm in such cases.

All postoperative deterioration events were adjudicated by manual electronic health record review with a structured questionnaire (Supplemental Digital Content 2, http://links.lww.com/ALN/C893) by four trained physician investigators familiar with local practice patterns and electronic health record documentation patterns within the ICU (M.R.M, A.J.C., A.M.W., B.E.B.) with removal of false positives. Importantly, for postoperative deterioration events occurring after a determination of comfort or end-of-life care was made by the ICU team as determined by manual review of clinical notes and inspection of nursing flowsheet data for patient care indicators (*e.g.*, extubation, disconnection from monitors, morphine boluses), events were excluded, and all ICU data generated after the change in goals of care were right-censored.

All patients with at least one postoperative deterioration event were compared to a random sample of control patients meeting study inclusion criteria and spending greater than 24 h within the ICU yet not experiencing postoperative deterioration events. Nonevent control patients were similarly adjudicated by the ICU physician panel with removal of false negatives.

Feature Analysis Time Window and Temporal Gap Before Deterioration Event

For analyzed physiologic waveforms, the data were extracted within a 15-min time window, segmented into five 3-min subwindows, and separated by an 8-h temporal gap between the window and the postoperative deterioration event (fig. 1). The data generated during the temporal gap were not used for prediction. Among nonevent control patients, fiducial time points (reference points representing nondeterioration events) were randomly selected (using a uniform probability distribution) from the duration of their ICU stay. Multiple fiducial time points were extracted from nonevent controls such that analysis windows from nonevent controls constituted 65% of the total sample size.

Physiologic Waveforms: Data Conditioning and Feature Extraction

Methods used to process and extract features from each set of physiologic waveforms have been previously described in detail[27] and are summarized in figure 2. These methods were independent of heart rhythm (*e.g.*, normal sinus, atrial fibrillation, etc.) and included waveform-specific bandpass

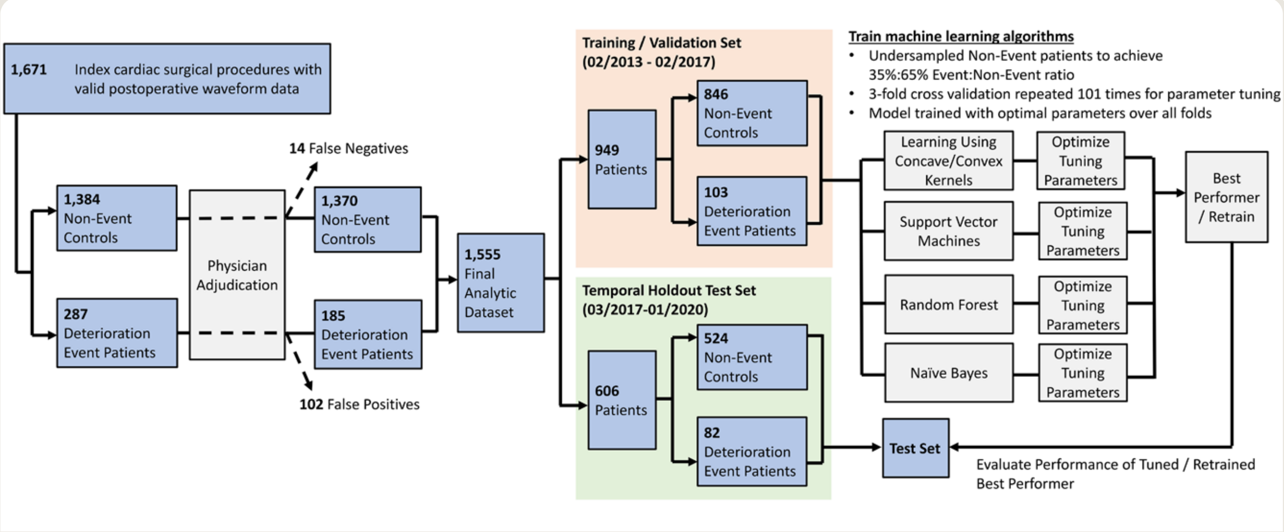
1,671 Index cardiac surgical procedures with valid postoperative waveform data
1,384 Non-Event Controls
287 Deterioration Event Patients
Physician Adjudication
14 False Negatives
102 False Positives
1,370 Non-Event Controls
185 Deterioration Event Patients
1,555 Final Analytic Dataset
Training / Validation Set (02/2013 - 02/2017)
949 Patients
846 Non-Event Controls
103 Deterioration Event Patients
Temporal Holdout Test Set (03/2017-01/2020)
606 Patients
524 Non-Event Controls
82 Deterioration Event Patients
Train machine learning algorithms
• Undersampled Non-Event patients to achieve 35%:65% Event:Non-Event ratio
• 3-fold cross validation repeated 101 times for parameter tuning
• Model trained with optimal parameters over all folds
Learning Using Concave/Convex Kernels
Support Vector Machines
Random Forest
Naïve Bayes
Optimize Tuning Parameters
Optimize Tuning Parameters
Optimize Tuning Parameters
Optimize Tuning Parameters
Best Performer / Retrain
Test Set
Evaluate Performance of Tuned / Retrained Best Performer

Fig. 1. Case selection, physician adjudication, development of feature analysis time windows, data partitioning, and machine learning model development.

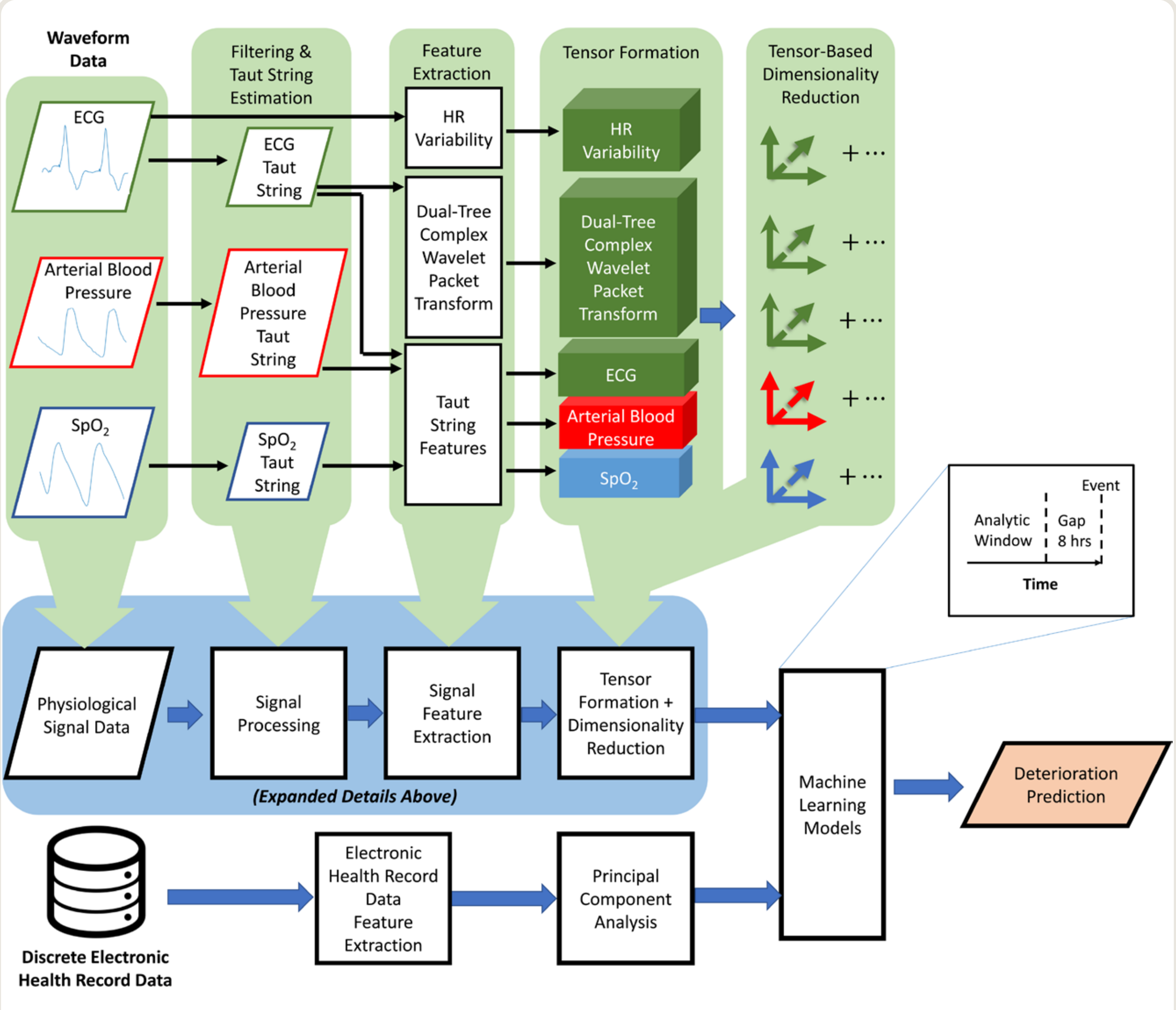
Waveform Data
Filtering & Taut String Estimation
Feature Extraction
Tensor Formation
Tensor-Based Dimensionality Reduction
ECG
Arterial Blood Pressure
SpO_2
ECG Taut String
Arterial Blood Pressure Taut String
SpO_2 Taut String
HR Variability
Dual-Tree Complex Wavelet Packet Transform
Taut String Features
HR Variability
Dual-Tree Complex Wavelet Packet Transform
ECG
Arterial Blood Pressure
SpO_2
+ ⋯
Event
Analytic Window
Gap 8 hrs
Time
Physiological Signal Data
Signal Processing
Signal Feature Extraction
Tensor Formation + Dimensionality Reduction
(Expanded Details Above)
Machine Learning Models
Deterioration Prediction
Discrete Electronic Health Record Data
Electronic Health Record Data Feature Extraction
Principal Component Analysis

Fig. 2. Overview of waveform and discrete electronic health record data conditioning, featurization, and tensor decomposition. ECG, electrocardiogram; HR, heart rate.

filtering (0.5 to 120 Hz for ECG; 1.25 to 25 Hz for arterial line; 1.75 to 10 Hz for pulse plethysmography) to remove commonly arising artifacts, as well as peak detection (systolic and diastolic peaks for arterial line and pulse plethysmography waveforms, R peaks for ECG waveform).[28] Taut string estimation[29] was applied to the filtered arterial line and pulse plethysmography signals, after which the sequence of estimated systolic and diastolic peaks were used to calculate beat-to-beat metrics (Supplemental Digital Content 3, http://links.lww.com/ALN/C894). The sequence of R peaks from the ECG signal was used to generate a heart-rate variability sequence for feature extraction. Separately, taut string estimation was applied to the filtered ECG with additional morphological and statistical predictive features extracted. Taut string estimation was also applied to the filtered ECG signal with similar predictive features extracted. Finally, a dual-tree complex wavelet packet transform[30] was applied to the taut string estimate of the filtered ECG signal with the decomposition level set to 2, with additional features extracted.

Discrete Electronic Health Record Data: Feature Extraction

Patient comorbidities were classified using the Enhanced Elixhauser Comorbidity Index derived from International Classification of Diseases, Ninth Revision (ICD-9) and ICD-10 codes.[31] Lab values were transformed into ordinal variables using commonly accepted reference ranges (Supplemental Digital Content 4, http://links.lww.com/ALN/C895). Building from previous work,[27] features newly included in prediction modeling included surgical details, left ventricular ejection fraction, additional comorbidities, nurse-validated vital signs, and urine output electronic health record data. For missing vital signs data, previous vital signs were carried forward until new data were available. Urine output was calculated on a per-hour basis with values evenly distributed across up to 8 h previous, for hours with no urine output documented.

To capture longer-term trends in patient electronic health record data, additional periods of analysis leading up to the prediction window gap before the deterioration event were extracted. Building from previous work,[27] newly studied retrospective periods of electronic health record features from the aforementioned categories of lab values and vital signs were extracted. For each retrospective period, the median of the values for each feature during that period was chosen as the feature value; four periods were extracted in total. For prediction window gaps of 8 h (primary analysis) and 12 h (sensitivity analysis), the periods were 8 h in duration; for prediction window gaps of less than 8 h (additional sensitivity analyses), the retrospective periods were 4 h in duration. All electronic health record data variables were standardized using their respective mean and standard deviations calculated from the training data set, with these values being used to standardize corresponding features in the validation and test sets.

Feature Reduction *via* Tensor Decomposition and Principal Component Analysis

Using the above methods and permuting waveform features across analytic subwindows and ϵ values as described in previous work,[27] a total of 5,150 waveform features and 486 discrete electronic health record features were extracted. To develop a reduced feature set for analysis, tensor decomposition techniques[27] were used, consisting of three steps: tensor formation, tensor structure analysis, and reduced feature extraction.

For each waveform and its attendant features, a third-order tensor (taut string ϵ value x waveform feature x subwindow) was formed for each patient in the training set. The third-order tensors were then stacked, creating a fourth-order tensor composed of all extracted waveform features from all patients in the training data set. The tensor structure was then analyzed using canonical polyadic decomposition,[32] which produces factor matrices for each mode. Higher-order singular value decompositions[33] were employed as a preprocessing step for the dual-complex wavelet packet transform, arterial line, and pulse plethysmography feature tensors to reduce computational cost. The factor matrices corresponding to the ϵ value and subwindow modes were then used to extract a reduced set of features from the tensors formed using the test data set. Through this process, the number of waveform features was reduced on average to 430.

Separately from the waveform features, the 486 discrete electronic health record features were reduced *via* a principal component analysis using a 90% explained variance threshold. The reduced waveform and discrete electronic health record features were then analyzed in a combined prediction model.

Machine Learning Models

Machine learning analyses and tensor decomposition[34] were performed using MATLAB (version 2020b, The Mathworks Inc., USA). Four machine learning models—naïve Bayes, support vector machine, random forest, and learning using concave and convex kernels—were used to predict the postoperative deterioration event target output. Whereas naïve Bayes, support vector machine, and random forest models are well established machine learning techniques, the Learning Using Concave and Convex Kernels modeling method is a novel machine learning technique used in previously published health applications[35] with notable advantageous properties including the ability to be trained using relatively small data sets and to handle outliers on a per-feature basis.

As a baseline comparison, the naïve Bayes model was trained using a normal distribution with no additional hyperparameter optimization. Configurable hyperparameters for the remaining machine learning models are described in Supplemental Digital Content 5 (http://links.lww.com/ALN/C896). The optimal combination of parameters was determined *via* a grid search approach for all models, with those corresponding to the highest receiver operating characteristic area under the curve (AUC) chosen.

To assess robustness of machine learning models to time-varying trends in clinical care and documentation patterns, cases were partitioned into a training/validation set (February 1, 2013, to February 28, 2017) and a temporal holdout test set (March 1, 2017, to January 31, 2020). Within the training/validation set, the modeling process used three-fold cross-validation repeated 101 times. Each fold contained a random selection of the total cohort; 67% of the patients with deterioration events were used to train the model, and 33% were used to perform validation (hyperparameter optimization). Within each fold for model training and validation, 35% were patients with postoperative deterioration events, and 65% were nonevent patients. No patient observations occurred in more than one fold (fig. 1). The performance of each tuned model on the training/validation set and temporal holdout test set was evaluated using AUC, positive predictive value, sensitivity, specificity, and F1 score. Model calibration was evaluated using Brier score and expected calibration error.

Relative Importance of Waveform *versus* Electronic Health Record Features

To evaluate the extent to which the prediction models were driven by the physiologic waveform *versus* discrete electronic health record features, we performed secondary analyses using waveform features only and discrete electronic health record features only. Differences in AUC between models were compared using Welch's *t* test.

Sensitivity Analyses

In addition to the primary analysis which used an 8-h prediction window temporal gap, preplanned sensitivity analyses were performed using temporal gaps of 0.5, 1, 2, 4, and 12h. In response to peer review, additional *post hoc* sensitivity analyses were performed in which (1) individual waveform features were separately evaluated, (2) the deterioration event outcome definition was revised to exclude inotrope and vasopressor infusion escalations, and (3) the deterioration event outcome definition revised to exclude *all* inotrope and vasopressor-based outcomes.

Results

Patient Population: Baseline Characteristics

Of the 1,671 cardiac surgery index postoperative ICU admissions with valid waveform data reviewed, 1,555 met study inclusion criteria after physician adjudication (fig. 1). We summarize perioperative characteristics for the entire cohort, patients with deterioration events, and nonevent controls in table 1 (extended details in Supplemental Digital Content 6, http://links.lww.com/ALN/C897). Our study population had a median age of 67 yr (interquartile range, 58 to 74), and 64% were men. Cardiac surgical procedures included valve (78% of cases; 36% of cases isolated to valve), coronary artery bypass grafting (26%; 10% isolated), and thoracic aortic (24%; 10% isolated). The most common medical comorbidities included valvular disease (87%), cardiac arrhythmias (58%), and peripheral vascular disorders (52%). The median ICU length of stay was 55h (interquartile range, 36 to 100). Time to last cardiac index before pulmonary artery catheter removal, last arterial line waveform before removal, and ICU length of stay are shown for the training/validation and temporal holdout cohorts in Supplemental Digital Content 7 (http://links.lww.com/ALN/C898). Compared to the training/validation data set, patients in the temporal holdout test set more commonly had cardiovascular comorbidities, less frequently underwent aortic or mitral valve surgeries, and had higher postoperative urine output.

Postoperative Deterioration Events

Among the 1,555 cases meeting inclusion criteria, 185 (12%) patients experienced 276 physician-adjudicated postoperative deterioration events between 24h after postoperative ICU admission and discharge from the ICU. Deterioration events most commonly included new low cardiac index less than 2.0 l min^{-1} m^{-2} (108 patients, 7.0% of total cohort; 150 events, 54% of all events), new inotrope infusion initiated (29 patients, 1.9% of total cohort; 31 events, 11% of all events), and sustained hypotension (22 patients, 1.4% of total cohort; 25 events, 9% of all events). Deterioration event types and hours after postoperative admission, split between the training/validation cohort and the temporal holdout cohort, are summarized across events in table 2 and patients in Supplemental Digital Content 8 (http://links.lww.com/ALN/C899). Compared to patients not developing deterioration events, patients with deterioration events were more commonly female, had multiple comorbidities, underwent combinations of procedures, and had postoperative lab values indicative of coagulopathy and renal dysfunction.

Machine Learning Model Performance

After feature extraction and dimensionality reduction, we describe 8-h predictive performance metrics of the training/validation set and the temporal holdout test set for machine learning models in table 3. Among the machine learning models used, the random forest model yielded the best performance with AUCs and positive predictive values of 0.803 (95% CI, 0.799 to 0.807) and 63.6% (95% CI, 62.8 to 64.4%), respectively, in the training/validation set and 0.709 (95% CI, 0.705 to 0.712) and 33.9% (95% CI, 33.1 to 34.6%), respectively, in the temporal holdout test set. This corresponded to three patients in the test set predicted to have deterioration events needed to identify a single patient with a true deterioration event (number needed to identify). We describe random forest model calibration in the temporal holdout test set (Brier score = 0.154; expected calibration error = 6.6%) compared to the training/validation model (Brier score = 0.190; expected calibration error = 15.7%) in Supplemental Digital Content 9 (http://links.lww.com/ALN/C900).

Waveform *versus* Electronic Health Record Features

We describe performance of the best performing model (random forest) on the training/validation and temporal holdout test sets, using solely waveform features or discrete electronic health record features *versus* both in figure 3. Models using both waveform and discrete electronic health record features consistently outperformed models using solely waveform features ($P < 0.001$) and solely discrete electronic health record features ($P < 0.001$). Furthermore, the model using solely waveform features (AUC, 0.697; 95% CI, 0.693 to 0.701; positive predictive value, 32.9%; 95% CI, 32.0 to 33.8%) significantly outperformed the model using solely discrete electronic health record features (AUC, 0.641; 95% CI, 0.637 to 0.646; positive predictive value, 32.2%; 95% CI, 31.3 to 33.2%; $P < 0.001$).

Sensitivity Analyses

In preplanned sensitivity analyses varying the temporal gap windows for the deterioration prediction model, statistically significant increases in model AUCs were observed for gap lengths less than 4h before deterioration events, although no statistically significant change in model AUCs were observed

Table 1. Summary Characteristics for Entire Cohort, Nonevent Controls, and Deterioration Event Patients

Category	Feature	Entire Cohort N = 1,555, n (%) or Median (Interquartile Range)	Nonevent Control Patients, N = 1,370, n (%) or Median (Interquartile Range)	Deterioration Event Patients, N = 185 n (%) or Median (Interquartile Range)	*P* Value
Demographic/anthropometric data	Age, yr	67 (58 to 74)	66 (57 to 74)	70 (63 to 77)	0.457
	Male sex	997 (64.1%)	915 (66.8%)	82 (44.3%)	< 0.001
	Race				
	Caucasian	1,370 (88.1%)	1,210 (88.3%)	160 (86.5%)	0.005
	Other	28 (1.8%)	27 (2.0%)	1 (0.5%)	
	Unknown	21 (1.4%)	19 (1.4%)	2 (1.1%)	
	African American	99 (6.4%)	83 (6.1%)	16 (8.6%)	
	Asian	26 (1.7%)	22 (1.6%)	4 (2.2%)	
	Patient Refused	8 (0.5%)	8 (0.6%)	0 (0.0%)	
	American Indian or Alaskan Native	1 (0.1%)	1 (0.1%)	0 (0.0%)	
	Native Hawaiian and other Pacific Islander	2 (0.1%)	0 (0.0%)	2 (1.1%)	
Patient medical history	Congestive heart failure	659 (42.4%)	532 (38.8%)	127 (68.6%)	< 0.001
	Cardiac arrhythmias	896 (57.6%)	762 (55.6%)	134 (72.4%)	< 0.001
	Valvular disease	1,357 (87.3%)	1,192 (87.0%)	165 (89.2%)	0.403
	Pulmonary circulation disorders	320 (20.6%)	244 (17.8%)	76 (41.1%)	< 0.001
	Peripheral vascular disorders	803 (51.6%)	713 (52.0%)	90 (48.6%)	0.386
	Hypertension, complicated	322 (20.7%)	254 (18.5%)	68 (36.8%)	< 0.001
	Hypertension, uncomplicated	1,125 (72.3%)	990 (72.3%)	135 (73.0%)	0.839
	Paralysis	28 (1.8%)	20 (1.5%)	8 (4.3%)	0.006
	Other neurologic disorders	86 (5.5%)	65 (4.7%)	21 (11.4%)	< 0.001
	Chronic pulmonary disease	509 (32.7%)	425 (31.0%)	84 (45.4%)	< 0.001
	Diabetes, complicated	173 (11.1%)	143 (10.4%)	30 (16.2%)	0.019
	Diabetes, uncomplicated	377 (24.2%)	322 (23.5%)	55 (29.7%)	0.064
	Liver disease	155 (10.0%)	125 (9.1%)	30 (16.2%)	0.003
	Coagulopathy	312 (20.1%)	252 (18.4%)	60 (32.4%)	< 0.001
	Coronary artery disease	722 (46.4%)	618 (45.1%)	104 (56.2%)	0.004
	Recent myocardial infarction	28 (1.8%)	21 (1.5%)	7 (3.8%)	0.031
	Previous open cardiac surgery	102 (6.6%)	95 (6.9%)	7 (3.8%)	0.104
	Active endocarditis	109 (7.0%)	87 (6.4%)	22 (11.9%)	0.006
Preoperative labs/studies and status	Left ventricular ejection fraction, %	55.0 (55.0 to 65.0)	56.5 (55.0 to 65.0)	55.0 (45.0 to 60.0)	0.916
	Estimated glomerular filtration rate, ml min^{-1} 1.73 m^{-2}	75.0 (58.8 to 88.5)	76.5 (61.4 to 89.4)	59.7 (41.7 to 78.9)	0.201
	Preoperative inotrope infusion	6 (0.4%)	3 (0.2%)	3 (1.6%)	0.004
	ASA physical status classification				0.067
	2	2 (0.1%)	2 (0.1%)	0 (0.0%)	
	3	308 (19.8%)	283 (20.7%)	25 (13.5%)	
	4	1,244 (80.0%)	1,085 (79.2%)	159 (85.9%)	
Surgery type	Isolated CABG	160 (10.3%)	147 (10.7%)	13 (7.0%)	0.007
	Isolated non-CABG	829 (53.3%)	737 (53.8%)	92 (49.7%)	0.007
	Two procedures	437 (28.1%)	375 (27.4%)	62 (33.5%)	0.007
	Three or more procedures	38 (2.4%)	28 (2.0%)	10 (5.4%)	0.007
Postoperative nonwaveform-derived vital signs	Spo$_2$, %	97.0 (95.0 to 99.0)	97.0 (95.0 to 99.0)	96.0 (94.8 to 98.0)	0.884
	Temperature, ºC	36.9 (36.6 to 37.2)	36.9 (36.6 to 37.2)	36.8 (36.5 to 37.3)	0.030
Postoperative outputs	Median hourly urine output, ml	35.0 (0.0 to 71.0)	38.2 (0.3 to 75.0)	25.0 (0.0 to 50.0)	0.067
Postoperative laboratory values v1.3 to 2.0	Creatinine range, mg/dl				< 0.001
	Unknown	0 (0.0%)	0 (0.0%)	0 (0.0%)	
	Less than 0.5 (females) or less than 0.7 (males)	95 (6.1%)	92 (6.7%)	3 (1.6%)	
	0.5 to 1.0 (females) or 0.7 to 1.3 (males)	1,072 (68.9%)	999 (72.9%)	73 (39.5%)	
	1.1 to 2.0 (females) or 1.3 to 2.0 (males)	282 (18.1%)	212 (15.5%)	70 (37.8%)	
	Greater than 2.0	106 (6.8%)	67 (4.9%)	39 (21.1%)	
	Glucose range, mg/dl				0.943
	Unknown	0 (0.0%)	0 (0.0%)	0 (0.0%)	
	Less than 40	0 (0.0%)	0 (0.0%)	0 (0.0%)	
	40 to 69	4 (0.3%)	3 (0.2%)	1 (0.5%)	

(Continued)

Table 1. Continued

Category	Feature	Entire Cohort N = 1,555, n (%) or Median (Interquartile Range)	Nonevent Control Patients, N = 1,370, n (%) or Median (Interquartile Range)	Deterioration Event Patients, N = 185 n (%) or Median (Interquartile Range)	*P* Value
	70 to 180	1,442 (92.7%)	1,272 (92.8%)	170 (91.9%)	
	Greater than 180	109 (7.0%)	95 (6.9%)	14 (7.6%)	
	Hemoglobin range, g/dl				0.003
	Unknown	0 (0.0%)	0 (0.0%)	0 (0.0%)	
	Less than 7.0	25 (1.6%)	16 (1.2%)	9 (4.9%)	
	7.0 to 11.9 (females) or 7.0 to 13.4 (males)	1,453 (93.4%)	1,282 (93.6%)	171 (92.4%)	
	12.0 to 16.0 (females) or 13.5 to 17.0 (males)	76 (4.9%)	71 (5.2%)	5 (2.7%)	
	Greater than 16.0 (females) or greater than 17.0 (males)	1 (0.1%)	1 (0.1%)	0 (0.0%)	
	INR range				< 0.001
	Unknown	2 (0.1%)	2 (0.1%)	0 (0.0%)	
	Less than 0.9	0 (0.0%)	0 (0.0%)	0 (0.0%)	
	0.9 to 1.2	1,350 (86.8%)	1,220 (89.1%)	130 (70.3%)	
	1.3 to 2.0	177 (11.4%)	131 (9.6%)	46 (24.9%)	
	Greater than 2.0	26 (1.7%)	17 (1.2%)	9 (4.9%)	
	Lactate range, mmol/l				0.013
	Unknown	13 (0.8%)	13 (0.9%)	0 (0.0%)	
	Less than 0.5	1 (0.1%)	1 (0.1%)	0 (0.0%)	
	0.5 to 1.6 (arterial) or 0.5 to 2.2 (venous)	867 (55.8%)	763 (55.7%)	104 (56.2%)	
	1.7 to 4.0 (arterial) or 2.3 to 4.0 (venous)	620 (39.9%)	553 (40.4%)	67 (36.2%)	
	Greater than 4.0	54 (3.5%)	40 (2.9%)	14 (7.6%)	
	Platelet count range, 10^9/l				< 0.001
	Unknown	1 (0.1%)	1 (0.1%)	0 (0.0%)	
	Less than 50	9 (0.6%)	3 (0.2%)	6 (3.2%)	
	50 to 149	939 (60.4%)	816 (59.6%)	123 (66.5%)	
	150 to 400	596 (38.3%)	541 (39.5%)	55 (29.7%)	
	Greater than 400	10 (0.6%)	9 (0.7%)	1 (0.5%)	
	Potassium range, mmol/l				0.118
	Unknown	0 (0.0%)	0 (0.0%)	0 (0.0%)	
	Less than 3.5	35 (2.3%)	32 (2.3%)	3 (1.6%)	
	3.5 to 5.0	1,422 (91.4%)	1,259 (91.9%)	163 (88.1%)	
	5.1 to 6.0	94 (6.0%)	75 (5.5%)	19 (10.3%)	
	Greater than 6.0	4 (0.3%)	4 (0.3%)	0 (0.0%)	
	Sodium range, mmol/l				< 0.001
	Unknown	0 (0.0%)	0 (0.0%)	0 (0.0%)	
	Less than 136	97 (6.2%)	83 (6.1%)	14 (7.6%)	
	136 to 146	1,398 (89.9%)	1,250 (91.2%)	148 (80.0%)	
	147 to 155	56 (3.6%)	37 (2.7%)	19 (10.3%)	
	Greater than 155	4 (0.3%)	0 (0.0%)	4 (2.2%)	
	Leukocyte count range, 10^9/l				0.096
	Unknown	0 (0.0%)	0 (0.0%)	0 (0.0%)	
	Less than 4	10 (0.6%)	9 (0.7%)	1 (0.5%)	
	4 to 10	467 (30.0%)	416 (30.4%)	51 (27.6%)	
	11 to 20	963 (61.9%)	853 (62.3%)	110 (59.5%)	
	Greater than 20	115 (7.4%)	92 (6.7%)	23 (12.4%)	

Additional characteristics are available in the extended table in Supplemental Digital Content 6 (http://links.lww.com/ALN/C897).

CABG, coronary artery bypass graft; INR, international normalized ratio; Spo$_2$, oxygen saturation measured by pulse oximetry.

between 4 and 12h (Supplemental Digital Content 10, http://links.lww.com/ALN/C901). Across all gap lengths, the random forest model consistently outperformed other machine learning models (Supplemental Digital Content 11, http://links.lww.com/ALN/C902). In *post hoc* sensitivity analyses performed in response to peer review, prediction

Table 2. Postoperative Deterioration Event Summary: Event Level

		Deterioration Events							
		Postoperative Timing				Surgical Subgroups			
All Deterioration Events	**n (%)**	**24 to 48 h Postoperative, n (%)**	**48 to 96 h Postoperative, n (%)**	**96 h to 7 Days Postoperative, n (%)**	**More than 7 Days Postoperative, n (%)**	**Isolated CABG, n (%)**	**Isolated Non-CABG, n (%)**	**Two Procedures, n (%)**	**Three or More Procedures, n (%)**
Training/validation cohort	153 (100%)	82 (54%)	34 (22%)	16 (10%)	21 (14%)	6 (4%)	70 (46%)	55 (36%)	11 (7%)
Mortality	4 (3%)	1 (1%)	0 (0%)	1 (6%)	2 (10%)	0 (0%)	2 (50%)	2 (50%)	0 (0%)
Cardiac index of less than 2.0 l min^{-1} m^{-2}	79 (52%)	58 (71%)	15 (44%)	6 (38%)	0 (0%)	1 (1%)	42 (53%)	26 (33%)	5 (6%)
Mean arterial pressure of less than 55 mmHg for more than 120 min	18 (12%)	5 (6%)	6 (18%)	2 (13%)	5 (24%)	1 (6%)	3 (17%)	9 (50%)	3 (17%)
Epinephrine bolus of 1 mg or more	5 (3%)	2 (2%)	3 (9%)	0 (0%)	0 (0%)	0 (0%)	2 (40%)	3 (60%)	0 (0%)
Inotrope infusion initiated	14 (9%)	7 (9%)	3 (9%)	1 (6%)	3 (14%)	1 (7%)	7 (50%)	6 (43%)	0 (0%)
Inotrope infusion escalated by 100% or more	9 (6%)	3 (4%)	1 (3%)	1 (6%)	4 (19%)	1 (11%)	3 (33%)	2 (22%)	0 (0%)
Vasopressor infusion initiated	9 (6%)	4 (5%)	2 (6%)	1 (6%)	2 (10%)	1 (11%)	3 (33%)	3 (33%)	2 (22%)
Vasopressor infusion escalated by 100% or more	15 (10%)	2 (2%)	4 (12%)	4 (25%)	5 (24%)	1 (7%)	8 (53%)	4 (27%)	1 (7%)
Temporal holdout cohort	123 (100%)	45 (37%)	18 (15%)	21 (17%)	39 (32%)	15 (12%)	59 (48%)	46 (37%)	1 (1%)
Mortality	10 (8%)	1 (2%)	2 (11%)	0 (0%)	7 (18%)	1 (10%)	6 (60%)	3 (30%)	0 (0%)
Cardiac index of less than 2.0 l min^{-1} m^{-2}	71 (58%)	33 (73%)	12 (67%)	11 (52%)	15 (38%)	9 (13%)	29 (41%)	33 (46%)	0 (0%)
Mean arterial pressure of less than 55 mmHg for more than 120 min	7 (6%)	3 (7%)	1 (6%)	0 (0%)	3 (8%)	0 (0%)	6 (86%)	0 (0%)	1 (14%)
Epinephrine bolus of 1 mg or more	3 (2%)	0 (0%)	0 (0%)	2 (10%)	1 (3%)	0 (0%)	1 (33%)	1 (33%)	0 (0%)
Inotrope infusion initiated	17 (14%)	4 (9%)	1 (6%)	4 (19%)	8 (21%)	1 (6%)	10 (59%)	6 (35%)	0 (0%)
Inotrope infusion escalated by 100% or more	8 (7%)	1 (2%)	2 (11%)	4 (19%)	1 (3%)	1 (13%)	5 (63%)	1 (13%)	0 (0%)
Vasopressor infusion initiated	6 (5%)	3 (7%)	0 (0%)	0 (0%)	3 (8%)	2 (33%)	2 (33%)	2 (33%)	0 (0%)
Vasopressor infusion escalated by 100% or more	1 (1%)	0 (0%)	0 (0%)	0 (0%)	1 (3%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)

CABG, coronary artery bypass graft.

models using solely arterial line or pulse plethysmography waveforms outperformed a model using solely ECG waveforms (Supplemental Digital Content 12, http://links.lww.com/ALN/C903), and deterioration prediction models using a composite outcome definition excluding inotrope or vasopressor infusion dose rate escalations of 100% or more (Supplemental Digital Content 13, http://links.lww.com/ALN/C904) and excluding all inotrope or vasopressor infusion outcome components (Supplemental Digital Content 14, http://links.lww.com/ALN/C905) yielded similar performance to the primary model.

Discussion

In this study of hemodynamic deterioration after cardiac surgery, we report a patient-level deterioration event incidence of 12% at a quaternary center. Through waveform processing, tensor decomposition, and machine learning, we developed models combining electronic health record and ICU physiologic waveform data to predict postoperative deterioration 8 h before the event. The best performing model demonstrated high performance in the training/validation set (AUC, 0.803) yet substantially decreased performance in the temporal holdout test set (0.709). Model performance was consistently greater than models limited to solely electronic health record or waveform data. Our study serves as a proof-of-concept that patterns within ICU waveform data can be combined with electronic health record data for improved early detection of postoperative deterioration events. However, the substantially decreased model performance in the temporal holdout test set highlights challenges inherent to implementation of

Table 3. Performance of Optimized Postoperative Deterioration Prediction Models with 8-h Prediction Window, Test Set

Prediction Model	AUC (95% CI)	Positive Predictive Value (95% CI)	F1 Score (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Training/validation set					
Naive Bayes	0.648 (0.639 to 0.657)	0.485 (0.476 to 0.495)	0.587 (0.582 to 0.592)	0.801 (0.788 to 0.813)	0.483 (0.457 to 0.509)
Learning using concave and convex kernels	0.783 (0.777 to 0.788)	0.620 (0.611 to 0.629)	0.670 (0.665 to 0.675)	0.751 (0.742 to 0.760)	0.730 (0.718 to 0.742)
Random forest	0.803 (0.799 to 0.807)	0.636 (0.628 to 0.644)	0.684 (0.680 to 0.688)	0.761 (0.752 to 0.769)	0.748 (0.738 to 0.758)
Support vector machines	0.787 (0.783 to 0.790)	0.640 (0.632 to 0.649)	0.681 (0.677 to 0.685)	0.745 (0.737 to 0.752)	0.757 (0.746 to 0.768)
Temporal holdout test set					
Naive Bayes	0.557 (0.549 to 0.565)	0.255 (0.247 to 0.263)	0.354 (0.349 to 0.359)	0.673 (0.639 to 0.706)	0.482 (0.434 to 0.529)
Learning using concave and convex kernels	0.708 (0.705 to 0.711)	0.336 (0.329 to 0.342)	0.428 (0.425 to 0.431)	0.617 (0.601 to 0.634)	0.703 (0.687 to 0.718)
Random forest	0.709 (0.705 to 0.712)	0.339 (0.331 to 0.346)	0.434 (0.43 to 0.438)	0.631 (0.614 to 0.647)	0.698 (0.682 to 0.714)
Support vector machines	0.677 (0.671 to 0.682)	0.346 (0.339 to 0.353)	0.429 (0.425 to 0.433)	0.584 (0.570 to 0.598)	0.730 (0.714 to 0.747)

AUC, area under the curve.

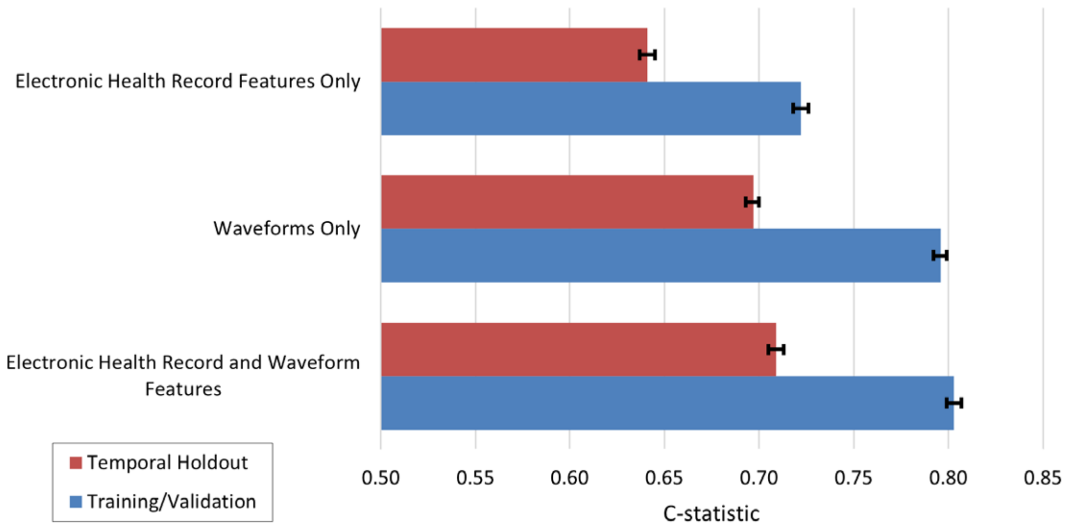

Fig. 3. Comparative performance (C-statistics) of best-performing prediction models (random forest) on test set: solely discrete electronic health record features *versus* solely physiologic waveform features *versus* both. The 95% CI values are shown in brackets.

machine learning–based prediction models in clinical settings with continuously evolving clinical documentation and practice patterns.

The performance of prediction algorithms developed in this study, although modest, were substantially below other widely used postoperative or critical illness prediction models using classical statistical approaches and developed in other ICU populations, such as EuroSCORE-II (AUC, 0.810; 95% CI, 0.782 to 0.836)[24] and SOFA (0.88; 95% CI, 0.82 to 0.94),[26] as well as meta-analyses of machine learning–based postcardiac surgery prediction models (0.88; 95% CI, 0.83 to 0.93).[36] This was likely due to the different nature of our primary outcome, which focused on dynamic and potentially unexpected deterioration events occurring more than 24 h postoperatively yet before end-of-life care rather than all deterioration events over the entirety of the postoperative period. Additionally, compared to other widely used cardiac surgery postoperative prediction models (*e.g.*, STS, Euroscore-II),[2,24] our prediction model focused on events that could be potentially detected *via* automated means in real time rather than retrospective chart review. This approach offers the advantage of being able to be deployed and continuously retrained at the point of care.

Compared to traditional cardiac surgical risk models limited to electronic health record data and other machine learning–based prediction models developed for anesthesiology and critical care settings,[17–20,37,38] our study uniquely combines both discrete electronic health record features and waveform features. Our study also leverages tensor decomposition, a method of parsing multidimensional arrays of waveform data in a computationally efficient manner, for feature reduction. Through combining such features and handling *via* a means robust to high dimensionality, our analytic techniques are an improvement to commonly used prediction and alerting systems using a single mode of physiologic waveform data or solely discrete electronic health record data lacking clinical context and leading to "alarm fatigue" from high false-positive rates.[39–41] To this end, the techniques used in our study offer incremental progress toward improved prediction model performance, although with clinical utility challenged by temporal changes in electronic health record documentation and clinical practice patterns.

The machine learning–based prediction models developed in this study are capable of drawing complex inferences from rich, complementary electronic health record and waveform data sources. For such capabilities of machine learning models to be fully realized, implementation of clinical decision support systems embedded into the electronic health record at the point of care—in a way which remains robust to changes in practice over time—must be further developed. Within anesthesiology and critical care, a regulatory landscape for such tools has been developed,[42–44] including methods to address data set shift,[45,46] although currently demonstrating limited success in improving outcomes.

Of critical concern to clinical decision support systems leveraging machine learning is the notion of relative harms of different outcomes posing competing risks, which are often implicitly understood by clinicians based upon clinical judgment.[47] Whereas the performance of healthcare prediction models may be improved *via* incorporation of physiologic waveform data and machine learning techniques, the relative harms of particular outcomes (*e.g.*, low cardiac index, sustained hypotension, mortality) have yet to be fully quantified by clinicians. Such relative valuation of harms requires consideration of nuanced clinical contexts and furthermore varies by patient preferences, clinical providers, and geographic region with varying patient care priorities or standards. In this study, for simplicity, we assume component postoperative deterioration events to have equal value, yet it is clear that some events (*e.g.*, mortality) have greater health implications than others (*e.g.*, new sustained hypotension). As techniques to improve the performance of prediction models continue to be refined, so must the methods to quantify relative harms based upon clinical judgment. One such technique incorporating relative harms based upon clinician judgment includes reinforcement learning, in which a prediction model seeks to learn an optimal individualized treatment plan; currently, early-stage applications of reinforcement learning have been deployed into anesthesiology and critical care settings.[48,49]

Study Limitations

Our study has multiple important limitations. First, we observed substantial decreases in model performance and calibration during temporal validation. Such decreases may be explained by (1) time-varied changes in electronic health record documentation and clinical practice patterns and (2) increased sampling of deterioration event patients to improve class balance during model training, leading to a generally greater number of false-positive predictions during temporal validation, offering caution to the generalizability of the model developed. These findings potentially represented data set shift, a major source of model underperformance among machine learning–based electronic health record prediction models, particularly in the setting of an information technology software/infrastructure update, changes to clinical practice, or shifts in patient demographics.[45] Next, our study used a convenience sample of ICU data among primarily Caucasian patients with pulmonary artery catheter monitoring (required for availability of thermodilution cardiac index measurements) at an academic medical center, potentially skewing toward more complex cardiac surgical procedures, limiting generalizability. Although a larger sample size may have allowed for a more diverse data set, including more cases over a longer time span would be unlikely to increase prediction model performance, given the potential data set shift requiring continuous retraining of prediction models.[50] Additionally, although we performed a temporal validation through the use of a test set with future cardiac surgeries relative to the training set, our study lacked external validation on a population within a separate ICU, postoperative surgical cohort, or institution.

Given the low prevalence of component postoperative deterioration events, a composite endpoint was used as the target output for this study to improve the class balance of the machine learning algorithms developed. Although mutually agreed upon *via* a consensus of ICU physicians, component postoperative deterioration events defining the target output of this study had subjective thresholds and are likely to have (1) varying clinical importance, (2) incomplete capture of all events potentially representing hemodynamic deterioration, (3) different early warning signs benefiting from separate prediction algorithms, and (4) different treatment implications. Additionally, several components of the composite deterioration event were determined by clinician actions (*e.g.*, new inotrope or vasopressor infusion or escalation) with varying levels of proactivity or reactivity to clinical events anticipated or having occurred (*e.g.*, low cardiac index or sustained hypotension), potentially leading to bias in the outcome measure. Furthermore, during deescalation of care, patients incurring occult low cardiac indices after pulmonary artery catheter removal not accompanied by another component of the composite outcome (*e.g.*, sustained hypotension or vasopressor infusion) potentially led to underestimation of hemodynamic deterioration outcomes.

Next, other health data sources available to the ICU team yet not reliably available for this study (*e.g.*, central venous oxygen saturation, information within operative notes or nuances discussed during postoperative handover, central venous pressure or pulmonary artery catheter waveforms with high rates of artifact or inadequate capture within the waveform repository) may have limited or biased the performance of the prediction model and decreased the likelihood of ICU teams being truly unaware of impending deterioration. Although this issue was partially mitigated through the exclusion of deterioration events occurring within 24h of transport from the operating room and ICU handover, this limited algorithm generalizability, which was unable to be explored further due to the need for at least 24h of ICU data to compute predictions. Future studies considering additional data sources, robust to nonrandom missingness, resolving artifacts manifesting within ICU waveforms and the electronic health record, and capable of handling complex unstructured data are necessary to improve predictive performance. Additionally, although the tensor decomposition technique used in this study represents a resourceful method for handling high-dimensional physiologic waveform data, the modeling of complex interactions between waveform features performed by the technique—improving model performance while mitigating potential overfitting—was at the expense of a nearly total lack of explainability to clinicians observing such waveforms.

Finally, although incorporation of physiologic waveform data led to improved prediction model performance, it remains unclear (1) how such data should be presented to an ICU team burdened with competing priorities, (2) how an ICU team should act on an algorithm prediction, and (3) whether a clinical treatment may lead to improved outcomes.

Conclusions

We report a 12% incidence of postoperative hemodynamic deterioration for patients receiving postcardiac surgery ICU care and demonstrate the value of physiologic waveforms routinely available in the ICU setting for improving clinical prediction model performance. For the clinical utility of such models to be fully realized, future studies are needed to improve model robustness to data set shift, to externally validate single-center findings, and to assess the feasibility and utility of deployment in real time.

Acknowledgments

The authors gratefully acknowledge Larry Hernandez, Ph.D., and Neriman Tokcan, Ph.D., Department of Computational Medicine and Bioinformatics, University of Michigan Medicine, Ann Arbor, Michigan, for their previous contributions to the codebase used in the analyses contained in this article.

Research Support

Supported by National Heart, Lung, and Blood Institute grant No. K01HL141701 (Bethesda, Maryland; to Dr. Mathis) and Department of Defense grant No. W81XWH-17-2-0012 (Washington, D.C.; to Drs. Gryak, Najarian, Mathis, Derksen, and Ward), The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or Department of Defense.

Competing Interests

Drs. Jonathan Gryak, Kayvan Najarian, and Harm Derksen have submitted a U.S. patent application related to this work ("Tensor amplification-based data processing"). Dr. Kevin Ward has a patent for a waveform analytic technology developed and assigned to the University of Michigan and licensed to Fiftheye, Inc. (Ann Arbor, Michigan), for which he has equity interest; is founder and equity holder in New Vital Signs, LLC (Ann Arbor, Michigan); is founder and equity holder in Precision Trauma (Ann Arbor, Michigan); is founder and equity holder in Prevada Medical (Ann Arbor, Michigan); and holds patents for predictive analytics in critical care submitted through and assigned to the University of Michigan and Virginia Commonwealth University (Richmond, Virginia). Dr. Milo Engoren received consulting fees from Aerogen (Chicago, Illinois) for nitric-oxide technology and Masimo (Irvine, California) for monitor technologies unrelated to this work and serves on a data safety monitoring board for use of extracorporeal membrane oxygenation in patients with out-of-hospital cardiac arrest, unrelated to this work. The other authors declare no competing interests.

Correspondence

Address correspondence to Dr. Mathis: University of Michigan, 1H247 UH, SPC 5048, 1500 East Medical Center Drive, Ann Arbor, Michigan 48109-5048. mathism@med.umich.edu. This article may be accessed for personal use at no charge through the Journal Web site, www.anesthesiology.org.

Supplemental Digital Content

Supplemental Digital Content 1: STROBE Checklist, http://links.lww.com/ALN/C892
Supplemental Digital Content 2: Physician Adjudicator Questionnaire, http://links.lww.com/ALN/C893
Supplemental Digital Content 3: Physiologic Waveform Feature Extraction, http://links.lww.com/ALN/C894
Supplemental Digital Content 4: Laboratory Value Reference Ranges, http://links.lww.com/ALN/C895
Supplemental Digital Content 5: Machine Learning Model Hyperparameters, http://links.lww.com/ALN/C896
Supplemental Digital Content 6: Extended Characteristics for Study Population, http://links.lww.com/ALN/C897
Supplemental Digital Content 7: ICU, Pulmonary Artery Catheter, and Arterial Line Duration, http://links.lww.com/ALN/C898
Supplemental Digital Content 8: Deterioration Event Summary – Patient Level, http://links.lww.com/ALN/C899

Supplemental Digital Content 9: Calibration Plots, http://links.lww.com/ALN/C900

Supplemental Digital Content 10: Model Performance *versus* Gap Window Length – Figure, http://links.lww.com/ALN/C901

Supplemental Digital Content 11: Model Performance *versus* Gap Window Length – Table, http://links.lww.com/ALN/C902

Supplemental Digital Content 12: Model Performance – EHR *versus* Waveform Data, http://links.lww.com/ALN/C903

Supplemental Digital Content 13: Model Performance – No Infusion Dose Escalations, http://links.lww.com/ALN/C904

Supplemental Digital Content 14: Model Performance – No Infusions, http://links.lww.com/ALN/C905

References

1. D'Agostino RS, Jacobs JP, Badhwar V, Fernandez FG, Paone G, Wormuth DW, Shahian DM: The Society of Thoracic Surgeons Adult Cardiac Surgery Database: 2018 update on outcomes and quality. Ann Thorac Surg 2018; 105:15–23
2. O'Brien SM, Feng L, He X, Xian Y, Jacobs JP, Badhwar V, Kurlansky PA, Furnary AP, Cleveland JC Jr, Lobdell KW, Vassileva C, Wyler von Ballmoos MC, Thourani VH, Rankin JS, Edgerton JR, D'Agostino RS, Desai ND, Edwards FH, Shahian DM: The Society of Thoracic Surgeons 2018 adult cardiac surgery risk models: Part 2. Statistical methods and results. Ann Thorac Surg 2018; 105:1419–28
3. Cornwell LD, Omer S, Rosengart T, Holman WL, Bakaeen FG: Changes over time in risk profiles of patients who undergo coronary artery bypass graft surgery: the Veterans Affairs Surgical Quality Improvement Program (VASQIP). JAMA Surg 2015; 150:308–15
4. Mehaffey JH, Hawkins RB, Byler M, Charles EJ, Fonner C, Kron I, Quader M, Speir A, Rich J, Ailawadi G, Virginia Cardiac Services Quality Initiative: Cost of individual complications following coronary artery bypass grafting. J Thorac Cardiovasc Surg 2018; 155:875–82.e1
5. LaPar DJ, Crosby IK, Rich JB, Fonner E Jr, Kron IL, Ailawadi G, Speir AM; Investigators for Virginia Cardiac Surgery Quality Initiative: A contemporary cost analysis of postoperative morbidity after coronary artery bypass grafting with and without concomitant aortic valve replacement to improve patient quality and cost-effective care. Ann Thorac Surg 2013; 96:1621–7
6. Glance LG, Osler TM, Mukamel DB, Dick AW: Effect of complications on mortality after coronary artery bypass grafting surgery: evidence from New York State. J Thorac Cardiovasc Surg 2007; 134:53–8
7. Lomivorotov VV, Efremov SM, Kirov MY, Fominskiy EV, Karaskov AM: Low-cardiac-output syndrome after cardiac surgery. J Cardiothorac Vasc Anesth 2017; 31:291–308
8. Gorman JH 3rd, Gorman RC, Milas BL, Acker MA: Circulatory management of the unstable cardiac patient. Semin Thorac Cardiovasc Surg 2000; 12:316–25
9. Society of Thoracic Surgeons Task Force on Resuscitation after Cardiac Surgery: The Society of Thoracic Surgeons expert consensus for the resuscitation of patients who arrest after cardiac surgery. Ann Thorac Surg 2017; 103:1005–20
10. Crawford TC, Magruder JT, Grimm JC, Suarez-Pierre A, Sciortino CM, Mandal K, Zehr KJ, Conte JV, Higgins RS, Cameron DE, Whitman GJ: Complications after cardiac operations: All are not created equal. Ann Thorac Surg 2017; 103:32–40
11. Edwards FH, Ferraris VA, Kurlansky PA, Lobdell KW, He X, O'Brien SM, Furnary AP, Rankin JS, Vassileva CM, Fazzalari FL, Magee MJ, Badhwar V, Xian Y, Jacobs JP, Wyler von Ballmoos MC, Shahian DM: Failure to rescue rates after coronary artery bypass grafting: An analysis from the Society of Thoracic Surgeons Adult Cardiac Surgery Database. Ann Thorac Surg 2016; 102:458–64
12. Vandendriessche B, Abas M, Dick TE, Loparo KA, Jacono FJ: A framework for patient state tracking by classifying multiscalar physiologic waveform features. IEEE Trans Biomed Eng 2017; 64:2890–900
13. Patel VL, Kannampallil TG, Shortliffe EH: Role of cognition in generating and mitigating clinical errors. BMJ Qual Saf 2015; 24:468–74
14. Moorman JR, Rusin CE, Lee H, Guin LE, Clark MT, Delos JB, Kattwinkel J, Lake DE: Predictive monitoring for early detection of subacute potentially catastrophic illnesses in critical care. Annu Int Conf IEEE Eng Med Biol Soc 2011; 2011:5515–8
15. Pinsky MR: Complexity modeling: Identify instability early. Crit Care Med 2010; 38:S649–55
16. Moss TJ, Lake DE, Calland JF, Enfield KB, Delos JB, Fairchild KD, Moorman JR: Signatures of subacute potentially catastrophic illness in the ICU: Model development and validation. Crit Care Med 2016; 44:1639–48
17. Tseng PY, Chen YT, Wang CH, Chiu KM, Peng YS, Hsu SP, Chen KL, Yang CY, Lee OK: Prediction of the development of acute kidney injury following cardiac surgery by machine learning. Crit Care 2020; 24:478
18. Fernandes MPB, Armengol de la Hoz M, Rangasamy V, Subramaniam B: Machine learning models with preoperative risk factors and intraoperative hypotension parameters predict mortality after cardiac surgery. J Cardiothorac Vasc Anesth 2021; 35:857–65
19. Kilic A, Goyal A, Miller JK, Gjekmarkaj E, Tam WL, Gleason TG, Sultan I, Dubrawksi A: Predictive utility

of a machine learning algorithm in estimating mortality risk in cardiac surgery. Ann Thorac Surg 2020; 109:1811–9

20. Molina RS, Molina-Rodríguez MA, Rincón FM, Maldonado JD: Cardiac operative risk in Latin America: A comparison of machine learning models *vs.* EuroSCORE-II. Ann Thorac Surg 2022; 113:92–9
21. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, Venkatesh S, Berk M: Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. J Med Internet Res 2016; 18:e323
22. Elm E von, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative: The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. Ann Intern Med 2007; 147:573–7
23. Anesthesia Clinical Research Committee. Michigan medicine - Anesthesiology clinical research. Available at: https://medicine.umich.edu/dept/anesthesiology/research/clinical-research. Accessed September 2, 2022.
24. Nashef SA, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, Lockowandt U: EuroSCORE II. Eur J Cardiothorac Surg 2012; 41:734–45
25. Le Gall JR, Lemeshow S, Saulnier F: A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA 1993; 270:2957–63
26. Ferreira FL, Bota DP, Bross A, Mélot C, Vincent JL: Serial evaluation of the SOFA score to predict outcome in critically ill patients. JAMA 2001; 286:1754–8
27. Hernandez L, Kim R, Tokcan N, Derksen H, Biesterveld BE, Croteau A, Williams AM, Mathis M, Najarian K, Gryak J: Multimodal tensor-based method for integrative and continuous patient monitoring during postoperative cardiac care. Artif Intell Med 2021; 113:102032
28. Mollakazemi MJ, Atyabi SA, Ghaffari A: Heart beat detection using a multimodal data coupling method. Physiol Meas 2015; 36:1729–42
29. Belle A, Ansari S, Spadafore M, Convertino VA, Ward KR, Derksen H, Najarian K: A Signal processing approach for detection of hemodynamic instability before decompensation. PLoS One 2016; 11:e0148544
30. Serbes G, Gulcur HO, Aydin N: Directional dual-tree complex wavelet packet transforms for processing quadrature signals. Med Biol Eng Comput 2016; 54:295–313
31. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, Saunders LD, Beck CA, Feasby TE, Ghali WA: Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care 2005; 43:1130–9
32. Kolda TG, Bader BW: Tensor decompositions and applications. SIAM Rev 2009; 51:455–500
33. De Lathauwer L, De Moor B, Vandewalle J: A multilinear singular value decomposition. SIAM J Matrix Anal Appl 2000; 21:1253–78
34. Kolda TG, Bader BW: MATLAB Tensor Toolbox. Sandia National Laboratories, 2006. Available at: https://www.osti.gov/biblio/1230898. Accessed September 2, 2022.
35. Sabeti E, Gryak J, Derksen H, Biwer C, Ansari S, Isenstein H, Kratz A, Najarian K: Learning using concave and convex kernels: Applications in predicting quality of sleep and level of fatigue in fibromyalgia. Entropy (Basel) 2019; 21:E442
36. Benedetto U, Dimagli A, Sinha S, Cocomello L, Gibbison B, Caputo M, Gaunt T, Lyon M, Holmes C, Angelini GD: Machine learning improves mortality risk prediction after cardiac surgery: Systematic review and meta-analysis. J Thorac Cardiovasc Surg 2022; 163:2075–2087.e9
37. Kendale S, Kulkarni P, Rosenberg AD, Wang J: Supervised Machine-learning Predictive Analytics for Prediction of Postinduction Hypotension. Anesthesiology 2018; 129:675–88
38. Hatib F, Jian Z, Buddi S, Lee C, Settels J, Sibert K, Rinehart J, Cannesson M: Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis. Anesthesiology 2018; 129:663–74
39. Drew BJ, Harris P, Zègre-Hemsey JK, Mammone T, Schindler D, Salas-Boni R, Bai Y, Tinoco A, Ding Q, Hu X: Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. PLoS One 2014; 9:e110274
40. Cvach M: Monitor alarm fatigue: An integrative review. Biomed Instrum Technol 2012; 46:268–77
41. Belle A, Thiagarajan R, Soroushmehr SM, Navidi F, Beard DA, Najarian K: Big data analytics in healthcare. Biomed Res Int 2015; 2015:370194
42. Kheterpal S, Shanks A, Tremper KK: Impact of a novel multiparameter decision support system on intraoperative processes of care and postoperative outcomes. Anesthesiology 2018; 128:272–82
43. Javitt GH: Regulatory Landscape for Clinical Decision Support Technology 2018; 128:247–9
44. Belard A, Buchman T, Forsberg J, Potter BK, Dente CJ, Kirk A, Elster E: Precision diagnosis: A view of the clinical decision support systems (CDSS) landscape through the lens of critical care. J Clin Monit Comput 2017; 31:261–71
45. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, Kohane IS, Saria S: The clinician and Dataset shift in artificial intelligence. N Engl J Med 2021; 385:283–6
46. Otles E, Oh J, Li B, Bochinski M, Joo H, Ortwine J, Shenoy E, Washer L, Young VB, Rao K, Wiens J: Mind the performance gap: Examining dataset shift during

prospective validation, Proceedings of the 6th Machine Learning for Healthcare Conference. Edited by Jung K, Yeung S, Sendak M, Sjoding M, Ranganath R. PMLR, 2021, pp 506–34

47. Agrawal A, Gans J, Goldfarb A: Prediction Machines: The Simple Economics of Artificial Intelligence. Brighton, MA, Harvard Business Press, 2018
48. Sandu C, Popescu D, Popescu C: Postcardiac surgery recovery process with reinforcement learning, 2015 19th International Conference on System Theory, Control and Computing. 2015, pp 658–61
49. Yu C, Liu J, Zhao H: Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. BMC Med Inform Decis Mak 2019; 19:57
50. Chen JH, Alagappan M, Goldstein MK, Asch SM, Altman RB: Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. Int J Med Inform 2017; 102:71–9

Appendix 1: Biomedical and Clinical Informatics Laboratory Collaborators

The additional biomedical and clinical informatics laboratory collaborators for this study are as follows:

Harm Derksen, Ph.D.: Department of Mathematics, Northeastern University, Boston, Massachusetts.

Kyle J. Gunnerson, M.D., F.C.C.M.: Department of Anesthesiology, University of Michigan Health System, Ann Arbor, Michigan; Department of Computational Medicine and Bioinformatics, University of Michigan Health System, Ann Arbor, Michigan; Michigan Center for Integrative Research in Critical Care, University of Michigan, Ann Arbor, Michigan; and Department of Emergency Medicine, University of Michigan Health System, 1500 East Medical Center Drive, Ann Arbor, Michigan.

Hasan B. Alam, M.B.B.S.: Department of General Surgery, Northwestern Medicine, Chicago, Illinois.

Collaborator contributions:

Harm Derksen, Ph.D., was responsible for the conception and design of the work and revising it critically for important intellectual content.

Kyle J. Gunnerson, M.D., F.C.C.M., was responsible for the conception and design of the work; the interpretation of data for the work; and revising it critically for important intellectual content.

Hasan B. Alam, M.B.B.S., was responsible for the conception and design of the work; the interpretation of data for the work; and revising it critically for important intellectual content.

Appendix 2: Postoperative Deterioration Event Definitions

Deterioration Event Type	Description
Low cardiac index	New-onset decrease in cardiac index of less than 2.0 l min^{-1} m^{-2} with either no previous cardiac index measurement within 48 h or previous cardiac index measurement within 48 h of 2.0 l min^{-1} m^{-2} or more
Sustained hypotension	New-onset decrease in mean arterial pressure less than 55 mmHg for 120 min or longer; for cases in which multiple sources of mean arterial pressure were monitored (*e.g.*, invasive arterial pressure monitors, noninvasive blood pressure measurements), the highest mean arterial pressure measurement available within a 60-min period was used
Epinephrine bolus	Intravenous administration of 1 mg or more of epinephrine
New inotrope infusion	New inotrope infusion (epinephrine, milrinone, dobutamine, or dopamine) initiated
New vasopressor infusion	New centrally administered vasopressor infusion (norepinephrine, vasopressin) initiated
Inotrope infusion rate escalation	Existing inotrope infusion rate previously above prespecified initial infusion-specific threshold (see below) and subsequently increased by 100% or more Infusion-specific minimum initial thresholds: •Epinephrine: 0.02 μg kg^{-1} min^{-1} •Milrinone: 0.250 μg kg^{-1} min^{-1} •Dobutamine: 2.0 μg kg^{-1} min^{-1} •Dopamine: 2.5 μg kg^{-1} min^{-1}
Vasopressor infusion rate escalation	Existing vasopressor infusion rate previously above prespecified initial infusion-specific threshold (see below) and subsequently increased by 100% or more Infusion-specific minimum initial thresholds: •Vasopressin: 2 units/h •Norepinephrine: 0.10 μg kg^{-1} min^{-1}
Mortality	Postoperative death of patient not currently receiving comfort care or end-of-life care measures

Events occurring after a decision to pursue patient comfort care or end-of-life care goals, as adjudicated by chart review by intensive care unit physicians, were excluded from analysis.