Interpretable Multi-Task Deep Neural Networks for Dynamic Predictions of Postoperative Complications

Benjamin Shickel\textsuperscript{a, e}, Tyler J. Loftus\textsuperscript{b, e}, Shounak Datta\textsuperscript{c, e}, Tezcan Ozrazgat-Baslanti\textsuperscript{c, e}, Azra Bihorac\textsuperscript{c, e}, Parisa Rashidi\textsuperscript{d, e}\textsuperscript{*}

\textsuperscript{a} Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL USA
\textsuperscript{b} Department of Surgery, University of Florida, Gainesville, FL USA
\textsuperscript{c} Department of Medicine, University of Florida, Gainesville, FL USA
\textsuperscript{d} Department of Biomedical Engineering, University of Florida, Gainesville, FL USA
\textsuperscript{e} Precision and Intelligent Systems in Medicine (PrismaP), University of Florida, Gainesville, FL USA
\textsuperscript{*} These authors contributed equally.

ABSTRACT

Accurate prediction of postoperative complications can inform shared decisions between patients and surgeons regarding the appropriateness of surgery, preoperative risk-reduction strategies, and postoperative resource use. Traditional predictive analytic tools are hindered by suboptimal performance and usability. We hypothesized that novel deep learning techniques would outperform logistic regression models in predicting postoperative complications. In a single-center longitudinal cohort of 43,943 adult patients undergoing 52,529 major inpatient surgeries, deep learning yielded greater discrimination than logistic regression for all nine complications. Predictive performance was strongest when leveraging the full spectrum of preoperative and intraoperative physiologic time-series electronic health record data. A single multi-task deep learning model yielded greater performance than separate models trained on individual complications. Integrated gradients interpretability mechanisms demonstrated the substantial importance of missing data. Interpretable, multi-task deep neural networks made accurate, patient-level predictions that harbor the potential to augment surgical decision-making.

INTRODUCTION
In the United States, more than 15 million inpatient surgeries are performed annually.\textsuperscript{1} Despite efforts to provide a safe and successful perioperative experience, complications and deaths occur in as many as 32\% and 2\%, respectively, increasing costs by as much as $11,000 per major complication.\textsuperscript{2,3} Making accurate, personalized predictions of these complications can inform shared decisions between patients and surgeons regarding the appropriateness of surgery, prehabilitation strategies targeting modifiable risk factors (e.g., smoking cessation), and perioperative resource use (e.g., intense, high-frequency postoperative surveillance).

Historically, perioperative predictive analytic decision-support tools have been hindered by suboptimal performance, inability to incorporate intraoperative data, lack of clinical workflow integration, and time constraints imposed by manual data entry requirements.\textsuperscript{4–9} These challenges may be overcome by automated deep learning models that capture latent, nonlinear data structure and relationships among raw feature representations in large datasets,\textsuperscript{10} now widely available in electronic health records (EHRs).\textsuperscript{11} Despite these potential advantages and other successful examples in healthcare,\textsuperscript{12–19} deep learning using the full spectrum of patient-specific EHR data to predict postoperative complications has not been previously reported.

Using a large retrospective cohort of 43,943 patients who underwent 52,529 inpatient surgeries, we tested the hypotheses that deep learning models predicting postoperative complications would outperform logistic regression models, and that predictions using both preoperative and intraoperative physiological time series input data would outperform predictions using preoperative data alone. We also explored the utility of multi-task learning\textsuperscript{20} by training a single deep learning model on several postoperative complications simultaneously, as well as integrated gradients methods to promote model interpretability.

\section*{RESULTS}

\textit{Participant Baseline Characteristics and Outcomes}
Cohort characteristics are described in Table 1. The overall study population had mean age 57 years, and 49% were female. Among the validation cohort of 11,969 surgical procedures, the incidence of complications was as follows: 28.5% prolonged ICU stay (for 48 hours or more), 6.4% prolonged mechanical ventilation (for 48 hours or more), 19.8% neurological complications, 17.6% acute kidney injury, 18.7% cardiovascular complications, 7.9% venous thromboembolism, 25.4% wound complications, 12.2% sepsis, and 2.3% in-hospital mortality. The distribution of complications was similar between development and validation cohorts.

Results are presented in subsections corresponding to study aims. We begin by comparing deep learning-based postoperative models, which use the full spectrum of preoperative and intraoperative data, that were trained either individually or simultaneously on all nine outcomes. Next, we compare deep learning models with traditional baseline logistic regression models. Finally, we compare performance metrics for the deep preoperative models, which use only data prior to surgery, with the deep learning-based postoperative models that use all data available up to the end of surgery.

**Multi-Task Learning Outperformed Individual Complication Modeling**

In this section, we compare results between two types of model training scenarios: one in which a single model was simultaneously trained on all nine complication outcomes (multi-task learning), and one in which a separate model was individually trained for each outcome. Full results are described in Table 2.

Among the deep learning models trained on preoperative data alone, there were no significant differences between the multi-task model and outcome-specific models across all performance metrics. In terms of AUROC, the multi-task model showed non-significant improvements for all outcomes other than sepsis, for which AUROC was approximately equivalent, and prolonged ICU duration, for which the individual model was slightly better (0.89 (0.89-0.90) vs. 0.88 (0.88-0.89)).
For models trained only on intraoperative time series, the multi-task model yielded significantly higher AUROC for prolonged mechanical ventilation (0.92 [95% confidence interval 0.91-0.92] vs. 0.88 [0.87-0.89]), sepsis (0.80 [0.78-0.81] vs. 0.75 [0.74-0.76]), acute kidney injury (0.76 [0.74-0.77] vs. 0.72 [0.70-0.73]), neurological complications (0.77 [0.76-0.78] vs. 0.72 [0.71-0.73]), cardiovascular complications (0.84 [0.84-0.85] vs. 0.81 [0.80-0.82]), and wound complications (0.64 [0.63-0.65] vs. 0.59 [0.58-0.60]). Additionally, the intraoperative multi-task models had significantly higher AUPRC for prolonged mechanical ventilation (0.51 [0.47-0.55] vs. 0.39 [0.35-0.42]), sepsis (0.41 [0.39-0.44] vs. 0.34 [0.32-0.37]), and neurological complications (0.48 [0.46-0.50] vs. 0.39 [0.37-0.41]). The multi-task model also yielded significantly higher NPV for wound complications (0.83 [0.81-0.84] vs. 0.79 [0.78-0.80]). For the remainder of metrics and outcomes, the multi-task model generally demonstrated non-significant performance advantages over individual models.

Using all available preoperative and intraoperative data, the multi-task postoperative model yielded significantly higher AUPRC for wound complications (0.62 [0.60-0.64] vs. 0.58 [0.56-0.59]). Across all other metrics and outcomes, the multi-task model generally demonstrated non-significant performance advantages over individual models.

Given that multi-task models had globally stronger performance and have reduced model footprint and training times compared with nine individual models, the multi-task approach is used for the remainder of this study. Henceforth, when referring to our deep learning-based postoperative model without qualifications, we are referring to results obtained by using the multi-task version that was simultaneously trained on all nine complication outcomes.

A full AUROC comparison between individual models and multi-task learning is shown in Figure 1a-c.
Deep Learning Outperformed Logistic Regression

In this section, we compare our deep learning models with baseline logistic regression models for preoperative, intraoperative, and postoperative prediction points. Preoperative prediction uses only data available before surgery; intraoperative prediction uses only data generated during surgery; postoperative prediction uses all available data. Both deep learning and baseline models used exactly the same feature sets with one exception: due to the nature of sequential deep learning methods, our deep intraoperative models processed the entire physiological time series minute-by-minute, whereas the baseline intraoperative model required extraction of summary statistics. A full list of baseline time series features is described in Table 3.

For a more succinct comparison, in this section we exclusively report AUROC results. A full comparison among all models, metrics, and complication outcomes is described in Table 2.

The deep preoperative model trained only on static pre-surgical descriptors yielded significantly higher AUROC compared with baseline logistic regression models for all nine outcomes, including prolonged ICU duration (0.88 [95% confidence interval 0.88-0.89] vs. 0.86 [0.85-0.86]), prolonged mechanical ventilation (0.91 [0.90-0.92] vs. 0.84 [0.82-0.85]), sepsis (0.87 [0.86-0.88] vs. 0.79 [0.77-0.80]), acute kidney injury (0.81 [0.80-0.82] vs. 0.74 [0.73-0.75]), neurological complications (0.87 [0.86-0.88] vs. 0.81 [0.80-0.82]), venous thromboembolism (0.83 [0.81-0.84] vs. 0.71 [0.69-0.73]), cardiovascular complications (0.82 [0.81-0.83] vs. 0.75 [0.74-0.76]), wound complications (0.80 [0.79-0.81] vs. 0.76 [0.75-0.77]), and in-hospital mortality (0.90 [0.89-0.92] vs. 0.76 [0.73-0.79]).

When using intraoperative time series input data alone, deep learning yielded significantly higher AUROC for seven of the nine outcomes, including prolonged ICU duration (0.89 [0.88-0.89] vs 0.87 [0.86-0.87]), prolonged mechanical ventilation (0.92 [0.91-0.92] vs. 0.88 [0.86-0.89]), neurological complications (0.77 [0.76-0.78] vs. 0.72 [0.71-0.74]), venous thromboembolism (0.72 [0.70-0.74] vs. 0.64 [0.62-0.66]), cardiovascular complications (0.84 [0.84-0.85] vs. 0.81 [0.81-0.82]), wound complications (0.64 [0.63-0.65] vs. 0.61 [0.60-0.62]), and in-hospital mortality (0.89 [0.87-0.90] vs. 0.76 [0.72-0.79]).
The deep postoperative model trained on all available data had significantly higher AUROC compared with logistic regression baseline models for all nine outcomes, including prolonged ICU duration (0.92 [0.91-0.92] vs. 0.88 [0.87-0.89]), prolonged mechanical ventilation (0.93 [0.92-0.94] vs. 0.86 [0.85-0.88]), sepsis (0.89 [0.88-0.90] vs. 0.79 [0.78-0.81]), acute kidney injury (0.83 [0.82-0.84] vs. 0.74 [0.73-0.75]), neurological complications (0.88 [0.87-0.89] vs. 0.80 [0.79-0.81]), venous thromboembolism (0.83 [0.82-0.85] vs. 0.68 [0.66-0.70]), cardiovascular complications (0.87 [0.86-0.88] vs. 0.78 [0.76-0.79]), wound complications (0.80 [0.79-0.81] vs. 0.75 [0.74-0.77]), and in-hospital mortality (0.92 [0.91-0.93] vs. 0.77 [0.74-0.80]).

A full AUROC comparison between deep learning and logistic regression is shown in Figure 1a-c.

**Deep Postoperative Models Outperformed Deep Preoperative Models**

In this section we compare performance between preoperative deep learning models, which use only data available before surgery, with postoperative deep learning models, which use both preoperative and intraoperative data.

Compared with deep preoperative models, deep postoperative models had significantly higher AUROC for prolonged ICU duration (0.92 [95% confidence interval 0.91-0.92] vs. 0.88 [0.88-0.89]) and cardiovascular complications (0.87 [0.86-0.88] vs. 0.82 [0.81-0.83]), slightly higher AUROC for prolonged mechanical ventilation, sepsis, acute kidney injury, neurological complications, and in-hospital mortality, and similar AUROC for venous thromboembolism and wound complications. The postoperative model had significantly higher AUPRC for prolonged ICU duration (0.83 [0.82-0.84] vs. 0.77 [0.76-0.78]), and significantly better performance by other metrics for cardiovascular complications in terms of PPV (0.47 [0.44-0.52] vs. 0.38 [0.36-0.42]), accuracy (0.80 [0.77-0.82] vs. 0.72 [0.70-0.76]), and AUPRC (0.66 [0.64-0.68] vs. 0.56 [0.54-0.58]). A full comparison is shown in Figure 1d.

Using preoperative predictions as a benchmark, the postoperative models made significant overall reclassification improvements for four of nine outcomes, including
prolonged ICU stay (overall, correctly reclassified 1.75% of all surgical encounters, p<0.001), prolonged mechanical ventilation (overall, correctly reclassified 5.28%, p<0.001), sepsis (overall, correctly reclassified 1.71%, p<0.01), and cardiovascular complications (overall, correctly reclassified 7.69%, p<0.001).

The postoperative models made reclassification improvements that were not statistically significant for acute kidney injury (overall, correctly reclassified 5.41%, p=0.070), venous thromboembolism (overall, correctly reclassified 2.56%, p=0.501), and in-hospital mortality (8.60%, p=0.274). The postoperative models made reclassification declines that were not statistically significant for neurological complications (overall, incorrectly reclassified 1.63% of all surgical encounters, p=0.151) and wound complications (overall, incorrectly reclassified 1.11%, p=0.293). Full net reclassification results are shown in Table 4.

In some cases, deep models for individual complications yielded better net reclassification indices than multi-task models, including prolonged ICU stay (3.27% vs. 1.75%, p<0.001), neurological complications (1.45% vs. -1.63%, p<0.001), and venous thromboembolism (4.47% vs. 2.56%, p=0.102).

Model Interpretability

We applied integrated gradients to our multi-task deep learning postoperative prediction model. The top 20 features per complication outcome for every sample in the validation cohort are illustrated in Figure 2. The x-axis represents the magnitude of feature attribution, both away from a given prediction (left) or towards a given prediction (right). Each point represents a sample from the validation cohort, and are colored by their input value from low (blue) to high (red) values. Attributions were visualized using techniques derived by Lundberg, et al31. Full feature importance scores are shown in Table 5.

The presence of a body mass index measurement in the EHR within one year prior to surgery was one of the four most important predictive features for eight of the nine complications (second most important predictor for prolonged ICU stay and sepsis; third most important predictor for prolonged mechanical ventilation, acute kidney injury, venous
thromboembolism, and in-hospital mortality; fourth most important predictor for neurological complications and wound complications). For each complication, the presence of a body mass index measurement was associated with higher overall risk. Body mass index values themselves were the twelfth most important predictor of in-hospital mortality; higher values were associated with lower mortality risk. Body mass index values were not a top 20 feature for any other complication.

The frequency of intraoperative SpO\textsubscript{2} measurements was one of the three most important predictive features for seven of the nine complications (most important predictor for prolonged ICU stay, prolonged mechanical ventilation, and cardiovascular complications, second most important predictor for acute kidney injury and in-hospital mortality; third most important predictor for sepsis and neurological complications). For each complication, higher frequency of intraoperative SpO\textsubscript{2} measurements was associated with lower overall risk. Intraoperative SpO\textsubscript{2} values themselves were not a top 20 feature for any of the nine complications.

Most other top predictive features were value-based and consistent with known, recognized pathophysiologic patterns. Advanced age was the most important predictor for acute kidney injury and in-hospital mortality. Higher peak inspiratory pressures were the second most important predictor of prolonged mechanical ventilation. Intraoperative tachycardia was the most important predictor of sepsis. Cerebrovascular disease and undergoing neurosurgery were the first and second most important predictors of neurological complications, respectively. Emergency admission and frequent hemoglobin measurements in the preceding week were the first and second most important predictors of venous thromboembolism, respectively. Intraoperative blood pressure and heart rate measurement frequency and values were the second through sixth most important predictors of cardiovascular complications. Advanced age and emergency admission were the second and third most important predictors of wound complications, respectively.
DISCUSSION

In predicting postoperative complications among adult patients undergoing major inpatient surgery, deep neural networks outperformed logistic regression classifiers, exhibiting strongest performance when leveraging the full spectrum of preoperative and intraoperative EHR data. Intraoperative physiological time-series had meaningful associations with postoperative patient outcomes, suggesting that prediction models augmented with intraoperative data may have utility for routine clinical tasks such as sharing prognostic information with patients and caregivers and making clinical management decisions regarding triage destination and resource use immediately after surgery. Deep models performed best when using multi-task methods predicting nine complications simultaneously, rather than predicting individual complications with separate models that require extra training time. Finally, applying integrated gradients interpretability methods elucidated feature importance. Strikingly, the presence of a body mass index measurement within the prior year and the frequency of intraoperative SpO\textsubscript{2} measurements were consistently important, while their actual values were of negligible importance, appearing in in the top 20 feature lists only once. Most other top predictive features were biologically plausible and generally consistent with known risk factors, supporting the validity of applying integrated gradients in this study. These methods also demonstrated the importance of including data missingness itself as a model feature, rather than ignoring or imputing all missing values. Therefore, deep, multi-task models using both preoperative and intraoperative data and integrated gradients techniques made accurate, interpretable, patient-level predictions of postoperative complications and provided potentially useful insights regarding the importance of data missingness in modeling risk for postoperative complications.

Previous studies have established that for many clinical prediction tasks, deep neural networks outperform logistic regression classifiers.\textsuperscript{32,33} Parametric regression equations often fail to accurately represent complex, non-linear associations among input variables, limiting their predictive performance. More than thirty years ago, Schwartz et al.\textsuperscript{34} suggested that human disease is too broad and complex to be accurately represented by rule-based algorithms, and that machine learning models obviate this.
limitation by learning from data. As EHR data volumes expand, deep learning healthcare applications gain greater potential for clinical application. However, this will require integration with real-time clinical workflow. Therefore, it seems prudent to design models that make updated predictions as EHR data become available. We sought to achieve this objective by using causal convolutional neural networks (CNN) with exponentially increasing kernel dilations to ensure that at any point in the network, timestep representations depend exclusively on data available at that time. Our results suggest that these models would perform well in prospective clinical settings.

Multi-task methods yield performance advantages for some prediction tasks by simultaneously predicting multiple outcomes, providing opportunities for learning across outcomes. Specifically, multi-task learning can improve model generalizability by penalizing the exploration of certain regions of the available function space, thus reducing overfitting from the false assumption that data noise is sparse or absent. This has been demonstrated by Si and Roberts in applying CNN multi-task learning to word embeddings in MIMIC-III clinical notes data, demonstrating that multi-task learning models outperformed single-task models in predicting mortality within 1, 3, 5, and 20 different timeframes. In addition, multi-task learning can act as a regularizer for learning classifiers from a finite set of examples by penalizing complexity in a loss function, as demonstrated by Harutyunyan et al. in predicting mortality and physiological decompensation among ICU patients in the publicly available MIMIC-III database. However, multi-task learning was not advantageous for phenotyping acute care conditions; the authors postulated that this occurred because phenotyping is multi-task by nature, i.e., already benefits from regularization across phenotypes. This may not hold true for rare, complex phenotypes, for which multi-task learning can reduce neural network sensitivity to hyperparameter settings (i.e., parameters that are set before learning begins), as demonstrated by Ding et al. Properly applied, multi-task learning can improve model generalizability and classification in deep learning clinical prediction models, optimizing performance and usability across diverse settings and datasets, with the added advantage of reduced model training times relative to training multiple individual models.

One barrier to clinical adoption of deep learning clinical prediction models is difficulty interpreting outputs. Patients, caregivers, and clinicians may be more willing to
incorporate model predictions in shared decision-making processes if they understand how and why a prediction was made. Integrated gradients techniques attempt to explain predictions made by deep learning models, usually by feeding perturbed inputs to the model, evaluating effects on outputs, and using this information to quantify and convey feature importance. Sayres et al.\textsuperscript{39} used integrated gradients to identify retinal image regions contributing to deep learning-based diabetic retinopathy diagnoses, which was associated with improved ophthalmologist diagnostic accuracy and confidence. Beyond interpretability, end users may want to know whether a model is confident that outputs are accurate. One approach uses an activation function on the final layer with a softmax function that maps network activations to (0,1), with lower values suggesting lower confidence that predicted probabilities match true probabilities. Unfortunately, a model may be uncertain of its predictions even when the softmax output is high.\textsuperscript{40} These methods have the potential to facilitate clinical adoption of deep learning prediction models by allowing patients, caregivers, and clinicians to understand how and why and output was produced, and whether it should be trusted enough to incorporate in shared decision-making processes.

This study was limited by its single-institution, retrospective design. Although multi-task functions reduce over-fitting, this study used data from a single institution, limiting its generalizability. The CNN ensured that model representations were not influenced by data from future timesteps, but our models have not been tested using prospective, real-time data, which may present data pre-processing challenges. Additionally, while our approach provides semi-quantitative justification for predictions in the form of input feature importance, it is limited by the lack of prediction confidence assessment. Future research should seek prospective, multi-center validation of these findings. This will be difficult to perform until cloud sharing of standardized EHR data is achieved.\textsuperscript{41}

METHODS

All analyses were performed on a single-center retrospective longitudinal cohort of surgical patients that included data from both preoperative and intraoperative phases of
care. We used deep learning and logistic regression models to predict the onset of nine major postoperative complications following surgery.

This study had three primary objectives: (1) compare deep learning techniques with logistic regression models in predicting postoperative complications, (2) compare deep learning predictions made at two phases of perioperative care: immediately before surgery (using preoperative data alone), and immediately after surgery by two different methods: (a) using intraoperative data alone, and (b) using both preoperative and intraoperative data, and (3) explore the potential benefits of multi-task learning by training a single deep learning model on several postoperative complications compared with training separate models for each individual complication.

The University of Florida Institutional Review Board and Privacy Office approved this study with waiver of informed consent (IRB # 201600223). Recommendations were followed from both Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD\textsuperscript{21}) and from best practices for prediction modeling from Leisman, et al.\textsuperscript{22}

**Data Source**

The University of Florida Integrated Data Repository was used as an honest broker to build a single-center longitudinal dataset from a cohort of patients admitted to University of Florida Health between June 1\textsuperscript{st}, 2014 and March 1\textsuperscript{st}, 2019 who were at least 18 years of age and underwent at least one surgical procedure during their hospitalization. The dataset was constructed by integrating electronic health records with other clinical, administrative, and public databases.\textsuperscript{9} The resulting dataset included detailed information on patient demographics, laboratory values, vital signs, diagnoses, medications, blood product administration, procedures, and clinical outcomes, as well as detailed intraoperative physiologic and monitoring data.
Participants

We excluded patients with intraoperative mortality or incomplete electronic health records. If a single patient’s hospital encounter included more than one surgery, only the first surgery during that encounter was included in our study. Our final dataset included 43,943 patients who underwent 52,529 surgeries. Figure 3 illustrates derivation of the study population.

Outcomes

We used several different machine learning methods to model the risk of nine postoperative complications: prolonged (> 48 hours) intensive care unit stay (> 48 hours), prolonged mechanical ventilation requirement (> 48 hours), neurological complications, cardiovascular complications, acute kidney injury, sepsis, venous thromboembolism, wound complications, and in-hospital mortality.

Predictor Features

Our final cohort included electronic health record data from both before and during surgery. Preoperative models were trained on data available between one year prior to surgery and the day of surgery, prior to surgery start time (i.e., preoperative features alone). Intraoperative models were trained on data created during the surgical procedure (i.e., intraoperative features alone). Postoperative models were trained on data available between one year prior to surgery through the end of the surgical procedure (i.e., both preoperative and intraoperative features).

We identified 134 preoperative features, including demographic and socioeconomic indicators, planned procedure and provider information, Charlson comorbidities, and summary statistics of select medications, laboratory tests, and physiological measurements (e.g., vital signs such as heart rate and blood pressure)
taken prior to a surgical procedure. We calculated Charlson comorbidity indices using International Classification of Diseases (ICD) codes.\textsuperscript{23} We modeled procedure types on ICD-9-CM codes with a forest structure in which nodes represent groups of procedures, roots represent the most general groups of procedures, and leaf nodes represent specific procedures. Medications were derived from RxNorm codes grouped into drug classes as previously described.

Intraoperative data consisted of 14 physiological measurements taken during surgery: systolic blood pressure, diastolic blood pressure, end-tidal carbon dioxide (EtCO2), fraction of inspired oxygen (FiO2), heart rate, minimum alveolar concentration (MAC), oxygen flow rate, positive end-expiratory pressure (PEEP), peak inspiratory pressure (PIP), respiratory rate, blood oxygen saturation (SpO2), temperature, urine output, and operative blood loss. These variables were presented to deep learning models as variable-length multivariate time series. For logistic regression models, a set of 49 statistical features were extracted from each encounter’s intraoperative measurements.

Table 6 shows a summary of all input features and relevant statistical characteristics.

**Sample Size**

We chronologically divided our perioperative cohort into a development set of 40,560 surgeries occurring between June 1\textsuperscript{st}, 2014 through March 1\textsuperscript{st}, 2018, and a validation set of 11,969 surgeries occurring between March 1\textsuperscript{st}, 2018 through March 1\textsuperscript{st}, 2019. All models were trained on the development set; all results were reported for the validation set. For deep learning models, we used 10% of the development set for early stopping.

Using a validation cohort of 11,969 surgeries, the overall sample size allows for a maximum width of the 95% confidence interval for area under the receiver operating characteristic curve (AUROC) to be between 0.02 to 0.04 for postoperative complications.
with prevalence ranging between 5% and 30% for AUROC of 0.80 or higher. The sample size allows for a maximum width of 0.07 for hospital mortality given 2% prevalence.

**Predictive Analytic Workflow**

Our deep learning system consists of a dynamic model that updates preoperative risk predictions using data collected during surgery. This workflow emulates clinical scenarios in which patients' preoperative information is enriched by the influx of new data from the operating room. The model consists of two main preoperative and intraoperative layers, each containing a data transformer core and a data analytics core. The data transformer integrates data from multiple sources, including EHR data with zip code links to US Census data for patient neighborhood characteristics and distance from the hospital. The data transformer then performs preprocessing and feature transformation steps to optimize the data for analysis.

The 134 preoperative features contained 88 continuous features, 32 binary features, and 16 nominal features. Of the 16 nominal features, 12 contained fewer than 20 levels and were one-hot encoded as zero vectors of dimension equal to number of levels, with level indicators equal to one. The remaining four nominal features (ZIP code, attending surgeon, primary procedure, and scheduled operating room) were represented as unique integer identifiers ranging from zero to the number of levels minus one, and implicit variable representations were learned as part of the model training process.

Continuous preoperative feature observations that fell below the 1st or above the 99th percentiles were capped to the 1st and 99th percentile values, respectively.

Temporal preoperative features denoting the day and month of admission were transformed into two individual continuous features each through the use of sinusoidal functions based on the respective frequency of days or months, which encoded relative differences between time points (e.g., Sunday is close to Monday, and December is close to January).

Intraoperative measurements were identified as those falling between anesthesia start and stop times for a given procedure. Fixed-interval multivariate physiological time
series were constructed for each procedure by resampling measured values to a frequency of one minute, which represented the highest recorded frequency across all intraoperative features. For a given surgical procedure which had at least one measurement of a given feature, any gaps in that feature’s time series were imputed via linear interpolation in both directions. As surgeries vary in duration, each sample included a multivariate time series of length $T$ minutes.

Missing continuous features were imputed with the median of each feature value in the development cohort. For static preoperative descriptors, this represented a single number; for intraoperative time series, this was only performed when a single feature value did not exist, and the median value was imputed at every one-minute time step for the full duration of surgery. Missing preoperative nominal features were replaced with a distinct “missing” category.

To preserve patterns of missingness which may be informative$^{24}$, for each sample we derived a preoperative binary presence mask over all continuous and binary input variables that indicated whether a given value was observed or imputed. These missingness indicators were concatenated with their respective original measurements. For a given cohort set of size $N$ encounters, initial continuous and binary preoperative features were represented as a matrix of descriptors $p^{N \times 120}$. With a missingness mask of size $p^{N \times 120}_{mask}$, concatenation resulted in a final continuous and binary preoperative feature set of 240 numerical preoperative descriptors for each sample. Nominal preoperative features did not require a missingness mask, as missing values were transformed into a distinct categorical level. The 12 nominal variables that were one-hot encoded were concatenated with the above numerical preoperative representation to yield a final numerical preoperative feature set of 295 features. The 4 nominal features with greater than 20 levels were internally embedded by the model.

Multivariate time series missingness masks were computed and concatenated at each one-minute intraoperative timestep; for a single surgical time series $x^{T \times 14}$ of length $T$ including our 14 physiological measurements, the concatenation of these per-timestep masks resulted in a final input time series $x^{T \times 28}$ of 28 intraoperative predictors at each timestep.
All continuous input variables, both preoperative and intraoperative, were z-normalized to zero mean and unit variance based on values from the development set.

Following these processing steps, each surgical encounter was represented by three distinct sets of variables: a set of 295 numerical preoperative features, a set of 4 nominal preoperative features to be internally embedded by the model, and a multivariate time series of length $T$ composed of 28 intraoperative features. The length of intraoperative time series varied depending on surgery duration, and our deep learning models were designed to process the full scope of intraoperative physiological measurements.

In the data analytics core, deep learning and logistic regression models were trained to predict nine postoperative complications following a surgical procedure. Clinically, predictions made by preoperative models can inform patients, caregivers, and surgeons regarding risks of undergoing surgery, and estimate the utility of risk reduction strategies for specific complications (e.g., preoperative smoking cessation, perioperative renal protection bundles, and wound closure techniques). Intraoperative events can influence risk for complications (e.g., operative blood loss requiring allogenic blood transfusion increases risk for septic complications, intraoperative hypotension increases risk for acute kidney injury). Therefore, we generated intraoperative models to predict complications using data obtained during surgery. At the end of surgery, clinicians must reassess the patient’s prognosis, convey this information to the patient and their caregivers, and make clinical management decisions accordingly (e.g., a patient at high risk for cardiovascular complications may benefit from postoperative admission to an intensive care unit or continuous cardiac telemetry on a general hospital ward). At the end of surgery, it seems prudent to consider both baseline preoperative risk as well as the potential influence of intraoperative events to make updated predictions of postoperative complications. This is accomplished by our postoperative models.

As a technical explanation of deep learning fundamentals is beyond the scope of this study, we refer interested readers to the comprehensive work by Goodfellow et al.\textsuperscript{25} Our final postoperative deep learning model can be conceptualized as a composition of two sub-models: one for processing preoperative features, and one for processing intraoperative features. Reported preoperative results (i.e., predicting postoperative
complications using preoperative features alone) were obtained by only using the data representation from the preoperative sub-model; likewise, reported intraoperative results (i.e., predicting postoperative complications using intraoperative features alone) were obtained by only using the data representation from the intraoperative sub-model. The postoperative model (i.e., predicting postoperative complications using both preoperative and intraoperative features) used a transformed concatenation of both preoperative and intraoperative data representations (Figure 4).

The preoperative sub-model was composed of a dual pipeline for processing and representing numerical features and nominal features with greater than 20 levels. A representation of all four index-encoded nominal input features was obtained by concatenating individual nominal feature representations – each of which were the result of a learned, multidimensional per-feature embedding lookup table – and passing the concatenated result through a fully-connected layer. A representation of all 259 numerical preoperative variables was obtained by passing the input features through a fully-connected layer. A complete preoperative encounter representation was obtained by concatenating both continuous and nominal input feature representations and passing the result through a final fully-connected layer.

In the multi-task setting, this preoperative data representation was passed through nine branches corresponding to our nine postoperative complication outcomes. Each branch contained two outcome-specific fully-connected layers followed by a sigmoid activation function to produce a per-outcome prediction score, interpreted as the probability of a preoperative patient developing a given postoperative complication.

The primary driving force behind the intraoperative sub-model was a 7-layer causal convolutional neural network (CNN) with exponentially increasing kernel dilations (Figure 5), which was used to learn representations of the multivariate intraoperative time series. Causal convolutions refer to a careful implementation of layer-wise input padding to ensure that at any point in the network, the representation of a given timestep depends exclusively on data at or before the current time step. This method of operation is functionally similar to a recurrent neural network (RNN), but for longer sequence lengths such as ours, we preferred the convolutional network for its improved performance and processing speed. By ensuring that model representations are not influenced by data
from future timesteps, this approach allows inference of model performance in prospective, real-time clinical application, in which data from future timesteps are truly unknown.

Dilated convolutions refer to expanding the coverage of convolutional sliding window kernels by a given factor – without changing its size – by implicitly inserting zeros into the expanded kernel. In our multilayer network, each layer $L$ (beginning with $L = 0$) used a dilation factor of $2^L$. This process allows greater receptive field over the input sequence and reduces the number of layers required to include the entire sequence in the uppermost layer’s final timestep representation. Additionally, given exponential dilations, each layer of the convolutional network views the sequence on an increasing temporal scale.

At each level of the convolutional network, we apply an attention mechanism to derive a per-layer context based on variable temporal resolutions. Briefly, a fundamental attention mechanism for classification allows a model to assign importance scores to individual timesteps of a representation sequence such that the importance-weighted sequence is summed into a single context vector that is an optimal representation for a given predictive task. In essence, attention allows a model to learn to focus exclusively on timesteps that are important for classification decisions. In our model, an intermediate timestep corresponds to the representation of an input subsequence of variable length. Therefore, the importance-weighted sum of timesteps at each layer yields seven representations of the same input sequence taken at multiple resolutions. These seven representations were concatenated and fed into a final fully-connected layer prior to predicting complication outcomes. Similar to the preoperative sub-model, when reporting intraoperative model results, we used a separate branch for each complication outcome.

Our complete deep learning model, which we refer to as the postoperative model, includes both the preoperative and intraoperative sub-models described in this section. The postoperative model is composed of the sub-models, but it is trained separately and end-to-end. The postoperative model consists of concatenating both the static preoperative representation (the output of the preoperative sub-model) with the outcome-specific intraoperative representation (the output of the intraoperative sub-model for a
given outcome), and passing this combined feature representation through the same set of nine classification branches as the sub-models.

In our experiments and reported results, we use a nominal preoperative variable embedding size of 10, fully-connected layers size of 64 (except for final task output layers, which have size 1), kernel size of 3, 7 causal convolutional layers, Adam optimizer with learning rate of 0.001, L2 regularization of 0.01, batch size of 64, ELU activation, and patience of 4 used for early stopping based on the validation data set.

To determine whether the deep learning models offered a performance advantage over traditional predictive analytic methods, we assessed the performance of a baseline logistic regression classifier using the same preoperative and intraoperative input feature sets as the deep learning models, with predictions made at the same time points. Nominal preoperative features, which were index-encoded before passing through the deep model, were instead one-hot encoded before feeding into the baseline logistic regression model. Intraoperative time series were fed to the baseline model by way of 49 summary statistics, capturing static attributes and patterns of variability for each variable. These features are described in Table 3.

To account for class imbalance among the nine postoperative complication outcomes, both deep learning and baseline models were trained using outcome-specific class weights that were inversely proportional to their respective frequencies in the training set. Functionally, this ensures greater model focus on the minority class samples.

We apply the method of integrated gradients to our final postoperative model to illuminate specific input features that yielded the largest impact on predicting each of our nine complication outcomes. A complete discussion of this technique is beyond the scope of this study; we refer interested readers to the work of Sundarajan et al\textsuperscript{26}. Briefly, integrated gradients is a comparative technique for local interpretability, centered around the analysis of model outputs based on a given input and corresponding baseline values, and assigns attributions values to every input feature. In theory, features most influential to a given prediction will receive larger attribution values, and taken over an entire population, this can reveal the importance of certain features which drive the model predictions. In our work, we use a zero-vector reference value for such computations, and
as all variables are Z-normalized to zero mean and unit variance, such a reference can be viewed as the per-variable mean value across the entire cohort.

**Model Validation**

All models were trained on the development set of 40,560 surgeries occurring between June 1\textsuperscript{st}, 2014 through March 1\textsuperscript{st}, 2018. Models were evaluated on the validation set of 11,969 surgical procedures occurring between March 1\textsuperscript{st}, 2018 through March 1\textsuperscript{st}, 2019. For each model performance metric, ninety-five percent nonparametric confidence intervals were calculated using 1,000 bootstrapped samples with replacement.

**Model Performance**

Model performance was evaluated by sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, area under the precision-recall curve (AUPRC), and area under the receiver operating characteristic curve (AUROC). Reported metrics include class predictions based on Youden’s index threshold on predicted risk scores, which maximizes sensitivity and specificity, as the cutoff point for low versus high risk.\textsuperscript{27}

When predicting rare events, models can exhibit deceivingly high accuracy by predicting negative outcomes in predominantly negative datasets.\textsuperscript{28} False negative predictions of postoperative complications may be especially detrimental because patients, caregivers, and surgeons could unknowingly agree to perform prohibitively high-risk surgery, miss opportunities to mitigate preventable harm through prehabilitation and other risk-reduction strategies, and under-triage high-risk patients to general hospital wards with infrequent monitoring when close monitoring in an intensive care unit would be safer. Therefore, model performance was evaluated by calculating area under the precision-recall curve (AUPRC), which is adept at evaluating the performance of models predicting rare events.\textsuperscript{29}
For all performance metrics, we used bootstrap sampling and non-parametric methods to obtain 95% confidence intervals. Net Reclassification Improvement (NRI) indices were used to describe and quantify correct and incorrect reclassifications by deep learning models.30

REFERENCES

1. Elixhauser, A. & Andrews, R. M. Profile of Inpatient Operating Room Procedures in US Hospitals in 2007. *Arch. Surg.* **145**, 1201–1208 (2010).
2. Dimick, J. B. *et al.* Hospital Costs Associated with Surgical Complications: A Report from the Private-sector National Surgical Quality Improvement Program. *J. Am. Coll. Surg.* **199**, 531–537 (2004).
3. Healey, M. A., Shackford, S. R., Osler, T. M., Rogers, F. B. & Burns, E. Complications in Surgical Patients. *Arch. Surg.* **137**, 611–618 (2002).
4. Cohen, M. E., Liu, Y., Hall, B. L. & Ko, C. Y. An Examination of American College of Surgeons NSQIP Surgical Risk Calculator Accuracy. *J. Am. Coll. Surg.* **224**, 787–795 (2017).
5. Leeds, I. L. *et al.* Eye of the beholder: Risk calculators and barriers to adoption in surgical trainees. *Surgery* **164**, 1117–1123 (2018).
6. Raymond, B. L. *et al.* Use of the American College of Surgeons National Surgical Quality Improvement Program Surgical Risk Calculator During Preoperative Risk Discussion: The Patient Perspective. *Anesth. Analg.* **128**, 643–650 (2019).
7. Bilimoria, K. Y. *et al.* Development and Evaluation of the Universal ACS NSQIP Surgical Risk Calculator: A Decision Aid and Informed Consent Tool for Patients and Surgeons. *J. Am. Coll. Surg.* **217**, 833–842 (2013).
8. Bertsimas, D., Dunn, A., Velmahos, G. C. & Kaafarani, H. M. A. Surgical Risk Is Not Linear: Derivation and Validation of a Novel Trees in Emergency Surgery Risk (POTTER) Calculator. *Ann. Surg.* **268**, 574–583 (2018).
9. Bihorac, A. *et al.* MySurgeryRisk: Development and Validation of a Machine-
learning Risk Algorithm for Major Complications and Death After Surgery. *Ann. Surg.* **269**, 652–662 (2019).

10. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

11. Henry, J., Pylypchuk, Y., Searcy, T. & Patel, V. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008–2015. *ONC data Br.* **35**, 1–9 (2016).

12. Shickel, B., Tighe, P. J. & Bihorac, A. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J. Biomed. Heal. Informatics* **22**, 1589–1604 (2018).

13. Miotto, R., Wang, F., Wang, S., Jiang, X. & Dudley, J. T. Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinform.* 1–11 (2017). doi:10.1093/bib/bbx044

14. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).

15. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).

16. Lipton, Z. C., Kale, D. C., Elkan, C. & Wetzell, R. Learning to Diagnose with LSTM Recurrent Neural Networks. *arXiv* (2016).

17. Fauw, J. De *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, (2018).

18. Shickel, B. *et al.* DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. *Sci. Rep.* **9**, 1879 (2019).

19. Tomašev, N. *et al.* A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116–119 (2019).

20. Harutyunyan, H., Khachatrian, H., Kale, D. C. & Steeg, G. Ver. Multitask learning and benchmarking with clinical time series data. *Sci. Data* **6**, 1–18 (2019).

21. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. *Br. J. Surg.* **102**, 148–158 (2015).

22. Leisman, D. E. *et al.* Development and Reporting of Prediction Models: Guidance for Authors From Editors of Respiratory, Sleep, and Critical Care Journals. *Crit.
Care Med. (2020). doi:10.1097/CCM.0000000000004246

23. Charison, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. J. Chronic Dis. 40, 373–383 (1987).

24. Lipton, Z. C., Kale, D. C. & Wetzel, R. Modeling Missing Data in Clinical Time Series with RNNs. 56, (2016).

25. Goodfellow, I., Bengio, Y. & Courville, A. Deep Learning. (MIT Press, 2016).

26. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. in Proceedings of the 34th International Conference on Machine Learning 3319–3328 (2017).

27. Youden, W. J. Index for rating diagnostic tests. Cancer 3, 32–35 (1950).

28. Saito, T. & Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS One 10, (2015).

29. Chiew, C. J., Liu, N., Wong, T. H., Sim, Y. E. & Abdullah, H. R. Utilizing Machine Learning Methods for Preoperative Prediction of Postsurgical Mortality and Intensive Care Unit Admission. Ann. Surg. (2019). doi:10.1097/SLA.0000000000003297

30. Pencina, M. J., Sr, B. D. A. & Steyerberg, E. W. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat. Med. 30, 11–21 (2011).

31. Lundberg, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat. Biomed. Eng. 2, (2018).

32. Dybowski, R., Weller, P., Chang, R. & Gant, V. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. Lancet 347, 1146–1150 (1996).

33. Sujin Kim, Woojae Kim & Rae Woong Park. A Comparison of Intensive Care Unit Mortality Prediction Models through the Use of Data Mining Techniques. Healthc. Inform. Res. 17, 232–243 (2011).

34. Schwartz, W. B., Patil, R. S. & Szolovits, P. Artificial Intelligence in Medicine Where Do We Stand? N. Engl. J. Med. 316, 685–688 (1987).
35. Minor, L. B. *Harnessing the Power of Data in Health.* (2017).
36. Si, Y. & Roberts, K. Deep Patient Representation of Clinical Notes via Multi-Task Learning for Mortality Prediction. in *AMIA Summits on Translational Science Proceedings* 779–788 (2019).
37. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).
38. Pfohl, S., Kale, D. C., Jung, K. & Shah, N. H. The Effectiveness of Multitask Learning for Phenotyping with Electronic Health Records Data. in *Pacific Symposium on Biocomputing* 18–29 (2019).
39. Sayres, R. *et al.* Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. *Ophthalmology* **126**, 552–564 (2019).
40. Gal, Y. Uncertainty in Deep Learning. (2016).
41. Adibi, A., Evaluation, R. & Program, S. Validation and Utility Testing of Clinical Prediction Models Time to Change the Approach. *J. Am. Med. Assoc.* (2020). doi:10.7326/M14-0698
Figure 1. Postoperative complication prediction accuracy of logistic regression baseline and deep learning models expressed as area under the receiver operating characteristic curve (AUROC) using preoperative data only (a), intraoperative data only (b), and a combination of preoperative and intraoperative data (c). (d) Comparison between all three deep learning models shown in panels (a-c).
Figure 2. Postoperative complication-specific feature importances assigned by the method of integrated gradients on the trained postoperative deep learning model for every sample in the validation set. Input feature values are colored from low (blue) to high (red). Importance values given along the x-axis of each panel, where a low importance value drives the prediction towards the negative class (no complication), and a positive value pushes the prediction towards the positive class (predicted complication).
Figure 3. Cohort selection criteria for both training (development) and testing (validation) cohorts.

Figure 4. Data pipeline and model architecture for final postoperative deep learning model.
Figure 5. Multilayer dilated causal convolutional network used in our deep learning models. The attention-weighted sums of each layer’s sequential hidden representations were concatenated before passing to a final fully-connected output layer. $x$: input multivariate time series of variable length; $h$: fixed-length intraoperative representation vector.
TABLES

Table 1. Demographic overview of select patient characteristic differences between training (development) and testing (validation) cohorts.

|                          | Training | Testing |
|--------------------------|----------|---------|
| Date ranges              | June 2014-Feb 2018 (n=40560) | March 2018-Feb 2019 (n=11969) |
| Average age (years)      | 56.5     | 57.5    |
| Ethnicity, n (%)         |          |         |
| Not Hispanic             | 38116 (93.9) | 11210 (93.6) |
| Hispanic                 | 1772 (4.4)  | 599 (5)  |
| Missing                  | 717 (1.8)   | 171 (1.4) |
| Race, n (%)              |          |         |
| White                    | 31399 (77.3) | 9376 (78.3) |
| African American         | 6136 (15.1) | 1739 (14.5) |
| Other                    | 2483 (6.1)  | 702 (5.9) |
| Missing                  | 587 (1.5)   | 163 (1.4) |
| Gender, n (%)            |          |         |
| Male                     | 20614 (50.8) | 6072 (50.7) |
| Female                   | 19991 (49.2) | 5908 (49.3) |
| Primary Insurance, n (%) |          |         |
| Medicare                 | 18581 (45.8) | 5774 (48.2) |
| Private                  | 12463 (30.7) | 3308 (27.6) |
| Medicaid                 | 6577 (16.2)  | 1928 (16.1) |
| Uninsured                | 2984 (7.4)   | 970 (8.1)  |
| Outcomes, n (%)          |          |         |
| ICU Stay > 48 hours       | 10213 (25.2) | 3382 (28.3) |
| MV Duration > 48 hours   | 2372 (5.9)   | 767 (6.4)  |
| Neurological Complications and Delirium | 5860 (14.5) | 2364 (19.8) |
| Sepsis                   | 3859 (9.5)   | 1459 (12.2) |
| Acute Kidney Injury      | 6098 (15)    | 2111 (17.6) |
| Cardiovascular Complication | 5866 (14.5) | 2240 (18.7) |
| Venous Thromboembolism   | 2283 (5.6)   | 943 (7.9)  |
| Wound Complications      | 7548 (18.6)  | 3044 (25.4) |
| Hospital Mortality       | 192 (2.3)    | 93 (2.6)   |
Table 2. Full performance results for each of our nine models when predicting each of the nine postoperative complications.

|                            | Sensitivity (95% CI) | Specificity (95% CI) | PPV (95% CI) | NPV (95% CI) | Accuracy (95% CI) | AUPRC (95% CI) | AUROC (95% CI) |
|---------------------------|----------------------|----------------------|--------------|--------------|-------------------|----------------|----------------|
| **Prolonged ICU Stay**    |                      |                      |              |              |                   |                |                |
| Baseline Preop            | 0.77 (0.72-0.80)     | 0.80 (0.77-0.84)     | 0.60 (0.58-0.65) | 0.90 (0.88-0.91) | 0.79 (0.78-0.81) | 0.74 (0.73-0.76) | 0.86 (0.85-0.86) |
| Deep Preop               | 0.80 (0.75-0.84)     | 0.81 (0.78-0.87)     | 0.63 (0.60-0.69) | 0.91 (0.90-0.92) | 0.81 (0.80-0.83) | 0.79 (0.78-0.81) | 0.89 (0.89-0.90) |
| Deep Preop (Multi-Task)  | 0.78 (0.76-0.84)     | 0.83 (0.76-0.84)     | 0.64 (0.58-0.66) | 0.90 (0.90-0.92) | 0.81 (0.78-0.82) | 0.77 (0.76-0.78) | 0.88 (0.88-0.89) |
| Baseline Intraop         | 0.77 (0.73-0.79)     | 0.82 (0.79-0.85)     | 0.63 (0.60-0.66) | 0.90 (0.89-0.91) | 0.80 (0.79-0.82) | 0.76 (0.75-0.78) | 0.87 (0.86-0.87) |
| Deep Intraop             | 0.76 (0.75-0.83)     | 0.85 (0.78-0.87)     | 0.67 (0.60-0.69) | 0.90 (0.89-0.92) | 0.83 (0.79-0.83) | 0.80 (0.78-0.81) | 0.88 (0.88-0.89) |
| Deep Intraop (Multi-Task)| 0.80 (0.76-0.83)     | 0.81 (0.79-0.85)     | 0.63 (0.61-0.68) | 0.91 (0.90-0.92) | 0.81 (0.80-0.83) | 0.79 (0.78-0.80) | 0.89 (0.88-0.89) |
| Baseline Postop          | 0.79 (0.77-0.83)     | 0.84 (0.80-0.85)     | 0.66 (0.62-0.68) | 0.91 (0.90-0.92) | 0.82 (0.81-0.83) | 0.79 (0.77-0.80) | 0.88 (0.87-0.89) |
| Deep Postop              | 0.84 (0.83-0.87)     | 0.84 (0.81-0.86)     | 0.68 (0.65-0.70) | 0.93 (0.92-0.94) | 0.84 (0.83-0.85) | 0.84 (0.83-0.85) | 0.92 (0.92-0.93) |
| Deep Postop (Multi-Task) | 0.83 (0.81-0.87)     | 0.83 (0.79-0.85)     | 0.66 (0.62-0.69) | 0.93 (0.92-0.94) | 0.83 (0.81-0.84) | 0.83 (0.82-0.84) | 0.92 (0.91-0.92) |
| **Prolonged Mechanical Ventilation** |                      |                      |              |              |                   |                |                |
| Baseline Preop            | 0.72 (0.66-0.79)     | 0.80 (0.74-0.86)     | 0.20 (0.17-0.25) | 0.98 (0.97-0.98) | 0.80 (0.74-0.85) | 0.36 (0.33-0.40) | 0.84 (0.82-0.85) |
| Deep Preop               | 0.80 (0.74-0.89)     | 0.82 (0.73-0.87)     | 0.23 (0.18-0.30) | 0.98 (0.98-0.99) | 0.82 (0.74-0.87) | 0.45 (0.41-0.49) | 0.90 (0.89-0.91) |
| Deep Preop (Multi-Task)  | 0.84 (0.76-0.89)     | 0.80 (0.75-0.88)     | 0.22 (0.19-0.30) | 0.99 (0.98-0.99) | 0.80 (0.75-0.87) | 0.47 (0.43-0.51) | 0.91 (0.90-0.92) |
| Baseline Intraop         | 0.77 (0.73-0.85)     | 0.84 (0.77-0.89)     | 0.25 (0.19-0.31) | 0.98 (0.98-0.99) | 0.83 (0.77-0.88) | 0.41 (0.37-0.45) | 0.88 (0.86-0.89) |
| Deep Intraop             | 0.84 (0.79-0.89)     | 0.79 (0.73-0.83)     | 0.21 (0.18-0.25) | 0.99 (0.98-0.99) | 0.79 (0.74-0.83) | 0.39 (0.35-0.42) | 0.88 (0.87-0.89) |
| Deep Intraop (Multi-Task)| 0.86 (0.82-0.89)     | 0.84 (0.81-0.87)     | 0.27 (0.24-0.31) | 0.99 (0.99-0.99) | 0.84 (0.82-0.87) | 0.51 (0.47-0.55) | 0.92 (0.91-0.92) |
| Baseline Postop          | 0.75 (0.71-0.82)     | 0.82 (0.77-0.86)     | 0.23 (0.19-0.26) | 0.98 (0.98-0.98) | 0.82 (0.77-0.85) | 0.40 (0.36-0.44) | 0.86 (0.85-0.88) |
| Deep Postop              | 0.81 (0.78-0.88)     | 0.85 (0.77-0.87)     | 0.27 (0.21-0.30) | 0.99 (0.98-0.99) | 0.84 (0.78-0.87) | 0.48 (0.45-0.53) | 0.91 (0.90-0.92) |
| Deep Postop (Multi-Task) | 0.84 (0.80-0.91)     | 0.86 (0.78-0.90)     | 0.28 (0.22-0.35) | 0.99 (0.98-0.99) | 0.85 (0.79-0.89) | 0.55 (0.51-0.58) | 0.93 (0.92-0.94) |
| **Sepsis**               |                      |                      |              |              |                   |                |                |
| Baseline Preop            | 0.64 (0.62-0.72)     | 0.84 (0.76-0.85)     | 0.36 (0.29-0.38) | 0.94 (0.94-0.95) | 0.82 (0.76-0.83) | 0.41 (0.39-0.44) | 0.79 (0.77-0.80) |
| Deep Preop               | 0.76 (0.73-0.81)     | 0.82 (0.77-0.84)     | 0.37 (0.33-0.40) | 0.96 (0.96-0.97) | 0.81 (0.77-0.83) | 0.56 (0.53-0.58) | 0.87 (0.86-0.88) |
| Deep Preop (Multi-Task)  | 0.76 (0.73-0.83)     | 0.83 (0.76-0.85)     | 0.38 (0.32-0.41) | 0.96 (0.96-0.97) | 0.82 (0.77-0.84) | 0.57 (0.54-0.60) | 0.87 (0.86-0.88) |
### Baseline Intraop

| Period          | Parameter 1 | Parameter 2 | Parameter 3 | Parameter 4 | Parameter 5 | Parameter 6 | Parameter 7 |
|-----------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Baseline Preop  | 0.60 (0.59-0.72) | 0.79 (0.66-0.80) | 0.38 (0.31-0.40) | 0.90 (0.90-0.92) | 0.76 (0.67-0.77) | 0.41 (0.39-0.44) | 0.74 (0.73-0.75) |
| Deep Preop      | 0.78 (0.72-0.82) | 0.67 (0.63-0.73) | 0.34 (0.32-0.37) | 0.93 (0.92-0.94) | 0.69 (0.67-0.73) | 0.47 (0.45-0.49) | 0.80 (0.79-0.81) |
| Deep Intraop (Multi-Task) | 0.77 (0.72-0.80) | 0.70 (0.69-0.76) | 0.36 (0.34-0.40) | 0.94 (0.93-0.94) | 0.72 (0.71-0.75) | 0.49 (0.46-0.51) | 0.81 (0.80-0.82) |

### Acute Kidney Injury

| Period          | Parameter 1 | Parameter 2 | Parameter 3 | Parameter 4 | Parameter 5 | Parameter 6 | Parameter 7 |
|-----------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Baseline Preop  | 0.60 (0.59-0.72) | 0.79 (0.66-0.80) | 0.38 (0.31-0.40) | 0.90 (0.90-0.92) | 0.76 (0.67-0.77) | 0.41 (0.39-0.44) | 0.74 (0.73-0.75) |
| Deep Preop      | 0.78 (0.72-0.82) | 0.67 (0.63-0.73) | 0.34 (0.32-0.37) | 0.93 (0.92-0.94) | 0.69 (0.67-0.73) | 0.47 (0.45-0.49) | 0.80 (0.79-0.81) |
| Deep Intraop (Multi-Task) | 0.77 (0.72-0.80) | 0.70 (0.69-0.76) | 0.36 (0.34-0.40) | 0.94 (0.93-0.94) | 0.72 (0.71-0.75) | 0.49 (0.46-0.51) | 0.81 (0.80-0.82) |

### Neurological Complications

| Period          | Parameter 1 | Parameter 2 | Parameter 3 | Parameter 4 | Parameter 5 | Parameter 6 | Parameter 7 |
|-----------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Baseline Preop  | 0.72 (0.70-0.75) | 0.80 (0.77-0.82) | 0.47 (0.44-0.49) | 0.92 (0.92-0.93) | 0.78 (0.77-0.80) | 0.60 (0.58-0.62) | 0.81 (0.80-0.82) |
| Deep Preop      | 0.80 (0.74-0.85) | 0.76 (0.71-0.82) | 0.45 (0.42-0.50) | 0.94 (0.93-0.95) | 0.76 (0.74-0.80) | 0.65 (0.63-0.67) | 0.86 (0.85-0.87) |
| Deep Intraop (Multi-Task) | 0.79 (0.76-0.83) | 0.80 (0.76-0.80) | 0.49 (0.45-0.50) | 0.94 (0.93-0.95) | 0.79 (0.77-0.80) | 0.67 (0.65-0.68) | 0.87 (0.86-0.88) |

### Venous Thromboembolism

| Period          | Parameter 1 | Parameter 2 | Parameter 3 | Parameter 4 | Parameter 5 | Parameter 6 | Parameter 7 |
|-----------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Baseline Preop  | 0.57 (0.51-0.62) | 0.78 (0.74-0.83) | 0.18 (0.16-0.21) | 0.95 (0.95-0.96) | 0.77 (0.73-0.81) | 0.23 (0.21-0.26) | 0.71 (0.69-0.73) |
| Deep Preop      | 0.81 (0.75-0.87) | 0.67 (0.61-0.73) | 0.17 (0.16-0.20) | 0.98 (0.97-0.98) | 0.68 (0.63-0.73) | 0.27 (0.25-0.30) | 0.81 (0.80-0.83) |
| Cardiovascular Complications |
|------------------------------|
| **Baseline Preop** | 0.68 (0.58-0.71) | 0.72 (0.70-0.82) | 0.36 (0.34-0.43) | 0.91 (0.89-0.91) | 0.71 (0.70-0.78) | 0.46 (0.44-0.49) | 0.75 (0.74-0.76) |
| **Deep Preop** | 0.77 (0.71-0.80) | 0.70 (0.67-0.76) | 0.37 (0.35-0.41) | 0.93 (0.92-0.94) | 0.72 (0.69-0.75) | 0.54 (0.52-0.56) | 0.81 (0.80-0.82) |
| **Deep Preop (Multi-Task)** | 0.78 (0.72-0.82) | 0.71 (0.67-0.76) | 0.38 (0.36-0.42) | 0.93 (0.92-0.94) | 0.72 (0.70-0.76) | 0.56 (0.54-0.58) | 0.82 (0.81-0.83) |
| **Baseline Intraop** | 0.70 (0.66-0.77) | 0.78 (0.72-0.82) | 0.42 (0.38-0.46) | 0.92 (0.91-0.93) | 0.77 (0.73-0.79) | 0.55 (0.53-0.58) | 0.81 (0.81-0.82) |
| **Deep Intraop** | 0.73 (0.67-0.78) | 0.74 (0.71-0.81) | 0.40 (0.37-0.45) | 0.92 (0.91-0.93) | 0.74 (0.72-0.78) | 0.56 (0.54-0.58) | 0.81 (0.80-0.82) |
| **Deep Intraop (Multi-Task)** | 0.75 (0.71-0.78) | 0.79 (0.77-0.83) | 0.46 (0.43-0.50) | 0.93 (0.93-0.94) | 0.79 (0.77-0.81) | 0.60 (0.58-0.62) | 0.84 (0.84-0.85) |
| **Baseline Postop** | 0.66 (0.63-0.69) | 0.79 (0.76-0.82) | 0.42 (0.40-0.45) | 0.91 (0.90-0.92) | 0.76 (0.75-0.79) | 0.52 (0.49-0.54) | 0.78 (0.76-0.79) |
| **Deep Postop** | 0.81 (0.76-0.84) | 0.77 (0.74-0.81) | 0.45 (0.42-0.49) | 0.95 (0.94-0.95) | 0.78 (0.76-0.80) | 0.63 (0.61-0.65) | 0.86 (0.86-0.87) |
| **Deep Postop (Multi-Task)** | 0.79 (0.75-0.83) | 0.80 (0.76-0.84) | 0.47 (0.44-0.52) | 0.94 (0.93-0.95) | 0.80 (0.77-0.82) | 0.66 (0.64-0.68) | 0.87 (0.86-0.88) |

| Wound Complications |
|---------------------|
| **Baseline Preop** | 0.66 (0.55-0.68) | 0.75 (0.73-0.85) | 0.47 (0.46-0.56) | 0.87 (0.85-0.87) | 0.73 (0.72-0.78) | 0.59 (0.57-0.61) | 0.76 (0.75-0.77) |
| **Deep Preop** | 0.69 (0.66-0.73) | 0.76 (0.72-0.79) | 0.49 (0.47-0.52) | 0.88 (0.87-0.89) | 0.74 (0.72-0.76) | 0.60 (0.58-0.62) | 0.79 (0.78-0.80) |
| **Deep Preop (Multi-Task)** | 0.70 (0.65-0.75) | 0.73 (0.69-0.79) | 0.48 (0.45-0.52) | 0.88 (0.87-0.89) | 0.73 (0.70-0.76) | 0.62 (0.60-0.64) | 0.80 (0.79-0.81) |
| **Baseline Intraop** | 0.65 (0.44-0.69) | 0.51 (0.47-0.71) | 0.31 (0.30-0.35) | 0.81 (0.79-0.82) | 0.54 (0.53-0.65) | 0.35 (0.33-0.36) | 0.61 (0.60-0.62) |
| **Deep Intraop** | 0.47 (0.44-0.55) | 0.69 (0.61-0.71) | 0.34 (0.32-0.35) | 0.79 (0.78-0.80) | 0.63 (0.59-0.65) | 0.33 (0.32-0.35) | 0.59 (0.58-0.60) |
| **Deep Intraop (Multi-Task)** | 0.66 (0.54-0.71) | 0.54 (0.50-0.66) | 0.33 (0.32-0.36) | 0.83 (0.81-0.84) | 0.57 (0.55-0.63) | 0.36 (0.35-0.38) | 0.64 (0.63-0.65) |
| **Baseline Postop** | 0.69 (0.57-0.73) | 0.71 (0.68-0.83) | 0.45 (0.43-0.54) | 0.87 (0.85-0.88) | 0.71 (0.69-0.77) | 0.58 (0.56-0.60) | 0.75 (0.74-0.77) |
| **Deep Postop** | 0.75 (0.67-0.78) | 0.67 (0.64-0.75) | 0.44 (0.42-0.48) | 0.89 (0.87-0.90) | 0.69 (0.68-0.73) | 0.58 (0.56-0.59) | 0.78 (0.77-0.79) |
| **Deep Postop (Multi-Task)** | 0.74 (0.66-0.78) | 0.71 (0.67-0.79) | 0.46 (0.44-0.52) | 0.89 (0.87-0.90) | 0.72 (0.70-0.76) | 0.62 (0.60-0.64) | 0.80 (0.79-0.81) |

| In-Hospital Mortality |
|----------------------|
| **Baseline Preop** | 0.56 (0.50-0.69) | 0.85 (0.74-0.87) | 0.08 (0.05-0.10) | 0.99 (0.99-0.99) | 0.84 (0.74-0.86) | 0.12 (0.09-0.15) | 0.76 (0.73-0.79) |
|          | Deep Preop | Deep Preop (Multi-Task) | Baseline Intraop | Deep Intraop | Deep Intraop (Multi-Task) | Baseline Postop | Deep Postop | Deep Postop (Multi-Task) |
|----------|------------|-------------------------|------------------|--------------|---------------------------|----------------|-------------|------------------------|
| accuracy | 0.87 (0.78-0.94) | 0.75 (0.68-0.83) | 0.08 (0.06-0.11) | 1.00 (0.99-1.00) | 0.75 (0.68-0.83) | 0.22 (0.18-0.27) | 0.89 (0.88-0.91) |
| precision | 0.88 (0.78-0.93) | 0.77 (0.73-0.86) | 0.08 (0.07-0.13) | 1.00 (0.99-1.00) | 0.77 (0.74-0.86) | 0.23 (0.18-0.28) | 0.90 (0.89-0.92) |
| recall    | 0.65 (0.52-0.78) | 0.76 (0.67-0.88) | 0.06 (0.05-0.10) | 0.99 (0.99-0.99) | 0.76 (0.67-0.87) | 0.15 (0.11-0.19) | 0.76 (0.72-0.79) |
| f1        | 0.74 (0.70-0.87) | 0.82 (0.71-0.84) | 0.09 (0.06-0.11) | 0.99 (0.99-1.00) | 0.82 (0.72-0.84) | 0.21 (0.16-0.26) | 0.86 (0.83-0.88) |
|         | 0.82 (0.78-0.88) | 0.82 (0.77-0.84) | 0.10 (0.08-0.11) | 0.99 (0.99-1.00) | 0.82 (0.78-0.84) | 0.25 (0.20-0.31) | 0.89 (0.87-0.90) |
| accuracy | 0.65 (0.59-0.80) | 0.77 (0.61-0.81) | 0.06 (0.05-0.08) | 0.99 (0.99-0.99) | 0.77 (0.61-0.81) | 0.14 (0.10-0.18) | 0.77 (0.74-0.80) |
| precision | 0.85 (0.80-0.93) | 0.81 (0.74-0.86) | 0.10 (0.07-0.13) | 1.00 (0.99-1.00) | 0.81 (0.74-0.86) | 0.27 (0.22-0.32) | 0.91 (0.89-0.92) |
| recall    | 0.82 (0.79-0.94) | 0.86 (0.74-0.89) | 0.12 (0.08-0.16) | 1.00 (0.99-1.00) | 0.86 (0.75-0.89) | 0.31 (0.26-0.37) | 0.92 (0.91-0.93) |
Table 3. Input feature characteristics for every clinical descriptor included in our experiments.

| Variable                        | Type of Variable | Data Source | Number of Categories | Type of Preprocessing |
|---------------------------------|------------------|-------------|----------------------|------------------------|
| **Demographic Variables**       |                  |             |                      |                        |
| Age (years)                     | Continuous       | Derived     |                      | Outlier adjustment a, Feature scaling c |
| Gender                          | Binary           | Raw         | 2                    |                         |
| Race                            | Nominal          | Raw         | 3                    | Missing value imputation b, One-hot encoding d |
| Body mass index                 | Continuous       | Raw         |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Marital status                  | Nominal          | Raw         | 3                    | One-hot encoding d     |
| Ethnicity                       | Binary           | Raw         | 2                    | Missing value imputation b |
| **Socioeconomic Variables**     |                  |             |                      |                        |
| Primary insurance               | Nominal          | Raw         | 4                    | One-hot encoding d     |
| Residency area characteristics e|                  |             |                      |                        |
| Zip code                        | Nominal          | Raw         | 1,908                | Missing value imputation b, Embedding representation f |
| Rural area                      | Binary           | Derived     | 2                    | Missing value imputation b |
| Total Population                | Continuous       | Derived     |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Median Income                   | Continuous       | Derived     |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Total Proportion of African-Americans | Continuous     | Derived     |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Total Proportion of Hispanic    | Continuous       | Derived     |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Population Proportion Below Poverty | Continuous    | Derived     |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Distance from Residency to Hospital (km) | Continuous     | Derived     |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| **Operative Characteristics**   |                  |             |                      |                        |
| Day of admission                | Continuous       | Derived     | 7                    | Cyclical embedding g, Feature scaling c |
| Month of admission              | Continuous       | Derived     | 12                   | Cyclical embedding g, Feature scaling c |
| Attending Surgeon               | Nominal          | Raw         | 311                  | Embedding representation f |
| Admission Source                | Binary           | Raw         | 2                    |                         |
| Admission Type (Emergent/Elective) | Binary        | Derived     | 2                    |                         |
| Admitting Type (Medicine/Surgery) | Binary         | Derived     | 2                    |                         |
| Night Admission                 | Binary           | Derived     | 2                    |                         |
| Scheduled Surgery Type          | Nominal          | Derived     | 18                   | One-hot encoding d     |
| Variable                                      | Type of Variable | Data Source | Number of Categories | Type of Preprocessing          |
|-----------------------------------------------|------------------|-------------|----------------------|--------------------------------|
| Scheduled Surgery Room                        | Nominal          | Raw         | 64                   | Embedding representation | |
| Scheduled post operation location             | Binary           | Derived     | 2                    |                                | |
| Scheduled room is trauma room                 | Binary           | Derived     | 2                    |                                | |
| Time of surgery from admission (days)         | Continuous       | Derived     |                      | Outlier adjustment, Feature scaling | |
| Scheduled primary surgical procedure          | Nominal          | Derived     | 2126                 | Embedding representation | |
| **Comorbidities**                             |                  |             |                      |                                | |
| Charlson's Comorbidity Index                 | Continuous       | Derived     |                      | Outlier adjustment, Feature scaling | |
| Myocardial Infarction                         | Binary           | Derived     | 2                    | Outlier adjustment, Feature scaling | |
| Congestive Heart Failure                      | Binary           | Derived     | 2                    | Outlier adjustment, Feature scaling | |
| Peripheral Vascular Disease                  | Binary           | Derived     | 2                    | Outlier adjustment, Feature scaling | |
| Cerebrovascular Disease                       | Binary           | Derived     | 2                    | Outlier adjustment, Feature scaling | |
| Chronic Pulmonary Disease                     | Binary           | Derived     | 2                    | Outlier adjustment, Feature scaling | |
| Diabetes                                      | Binary           | Derived     | 2                    | Outlier adjustment, Feature scaling | |
| Cancer                                        | Binary           | Derived     | 2                    | Outlier adjustment, Feature scaling | |
| Liver Disease                                 | Binary           | Derived     | 2                    | Outlier adjustment, Feature scaling | |
| Valvular disease                              | Binary           | Derived     | 2                    | Outlier adjustment, Feature scaling | |
| Coagulopathy                                  | Binary           | Derived     | 2                    | Outlier adjustment, Feature scaling | |
| Weight loss                                   | Binary           | Derived     | 2                    | Outlier adjustment, Feature scaling | |
| Alcohol or Drug Abuse                         | Binary           | Derived     | 2                    | Outlier adjustment, Feature scaling | |
| Smoking Status                                | Nominal          | Raw         | 3                    | One-hot encoding | |
| **Medications History**                       |                  |             |                      |                                | |
| Betablockers                                  | Binary           | Derived     | 2                    |                                | |
| Diuretics                                     | Binary           | Derived     | 2                    |                                | |
| Statin                                        | Binary           | Derived     | 2                    |                                | |
| Aspirin                                       | Binary           | Derived     | 2                    |                                | |
| Angiotensin-Converting-Enzyme Inhibitors      | Binary           | Derived     | 2                    |                                | |
| Pressors or Inotropes                         | Binary           | Derived     | 2                    |                                | |
| Bicarbonate                                   | Binary           | Derived     | 2                    |                                | |
| Antiemetic                                    | Binary           | Derived     | 2                    |                                | |
| Aminoglycosides                               | Binary           | Derived     | 2                    |                                | |
| Vancomycin                                    | Binary           | Derived     | 2                    |                                | |
| Nonsteroidal Anti-inflammatory Drug           | Binary           | Derived     | 2                    |                                | |
| **Preoperative Laboratory Results**           |                  |             |                      |                                | |
| Urine Protein, mg/dL                         | Nominal          | Derived     | 4                    | Missing value imputation, One-hot encoding | |
| Urine Hemoglobin, mg/dL                       | Nominal          | Derived     | 4                    | Missing value imputation, One-hot encoding | |
| Variable                                      | Type of Variable | Data Source | Number of Categories | Type of Preprocessing                              |
|-----------------------------------------------|------------------|-------------|----------------------|---------------------------------------------------|
| Urine Glucose, mg/dL                         | Nominal          | Derived     | 3                    | Missing value imputation a, One-hot encoding b     |
| Urine Erythrocytes, mg/dL                    | Nominal          | Derived     | 4                    | Missing value imputation a, One-hot encoding b     |
| Serum Glucose, mg/dL                         | Continuous       | Raw         |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Blood Urea Nitrogen test, mg/dL              | Continuous       | Raw         |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Serum Creatinine, mg/dL                      | Continuous       | Raw         |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Serum Calcium, mmol/L                        | Continuous       | Raw         |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Serum Sodium, mmol/L                         | Continuous       | Raw         |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Serum Potassium, mmol/L                      | Continuous       | Raw         |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Serum Chloride, mmol/L                       | Continuous       | Raw         |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Serum CO2, mmol/L                            | Continuous       | Raw         |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Serum White Blood Cell, thou/uL              | Continuous       | Raw         |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Mean Corpuscular Hemoglobin in Blood, g/dL   | Continuous       | Raw         |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Mean Corpuscular Hemoglobin Concentration in Blood, pg | Continuous | Raw         |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Erythrocyte Distribution Width Count, %      | Continuous       | Raw         |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Serum creatinine, mg/dL                      | Continuous       | Raw         |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Serum Platelet, thou/uL                      | Continuous       | Raw         |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Serum Hemoglobin, g/dL                       | Continuous       | Raw         |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Reference Estimated Glomerular Filtration Rate, mL/min/1.73 m² | Continuous | Derived     |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| Urea nitrogen-Creatinine ratio in Serum      | Continuous       | Derived     |                      | Outlier adjustment a, Missing value imputation b, Feature scaling c |
| **Physiologic Intraoperative Time Series**   |                  |             |                      |                                                   |
| Systolic blood pressure, mmHg                | Continuous       | Raw         |                      | Data cleaning f, Imputation of outliers g, Statistical features extraction h |

**Notes:**
- a: Outlier adjustment
- b: Missing value imputation
- c: Feature scaling
- d: One-hot encoding
- e: Missing value imputation
- f: Data cleaning
- g: Imputation of outliers
- h: Statistical features extraction
| Variable                                      | Type of Variable | Data Source | Number of Categories | Type of Preprocessing                                      |
|-----------------------------------------------|------------------|-------------|----------------------|------------------------------------------------------------|
| Diastolic blood pressure, mmHg                | Continuous       | Raw         |                      | Temporal processing i, Time series creation j, Feature scaling c, Baseline time series extraction k |
| Minimum alveolar concentration                | Continuous       | Raw         |                      | Temporal processing i, Time series creation j, Feature scaling c, Baseline time series extraction k |
| Heart rate, bpm                               | Continuous       | Raw         |                      | Temporal processing i, Time series creation j, Feature scaling c, Baseline time series extraction k |
| Temperature (°C)                              | Continuous       | Raw         |                      | Temporal processing i, Time series creation j, Feature scaling c, Baseline time series extraction k |
| Peripheral capillary oxygen saturation (SPO2)| Continuous       | Raw         |                      | Temporal processing i, Time series creation j, Feature scaling c, Baseline time series extraction k |
| End-tidal CO2 (ETCO2)                         | Continuous       | Raw         |                      | Temporal processing i, Time series creation j, Feature scaling c, Baseline time series extraction k |
| Peak Inspiratory Pressure (PIP)              | Continuous       | Raw         |                      | Temporal processing i, Time series creation j, Feature scaling c, Baseline time series extraction k |
| Positive End-expiratory Pressure (PEEP)       | Continuous       | Raw         |                      | Temporal processing i, Time series creation j, Feature scaling c, Baseline time series extraction k |
| Respiratory O2                                | Continuous       | Raw         |                      | Temporal processing i, Time series creation j, Feature scaling c, Baseline time series extraction k |
| Respiratory Rate                              | Continuous       | Raw         |                      | Temporal processing i, Time series creation j, Feature scaling c, Baseline time series extraction k |
| **Laboratory Results from Surgery**           |                  |             |                      |                                                           |
| Fraction of inspired oxygen (FIO2)            | Continuous       | Raw         |                      | Temporal processing i, Time series creation j, Feature scaling c, Baseline time series extraction k |
| **Other Characteristics**                     |                  |             |                      |                                                           |
| Estimated blood loss, mL                      | Continuous       | Raw         |                      | Temporal processing i, Time series creation j, Feature scaling c, Baseline time series extraction k |
| Variable               | Type of Variable | Data Source | Number of Categories | Type of Preprocessing                                                                 |
|------------------------|------------------|-------------|----------------------|--------------------------------------------------------------------------------------|
| Urine output, mL       | Continuous       | Raw         |                      | Temporal processing †, Time series creation ‡, Feature scaling ‡, Baseline time series extraction § |

*a For continuous variables, values that fell in the top and bottom 1% of its distribution were considered outliers and capped to the respective values given at the 1st and 99th percentiles.

*b Missing numerical values were replaced with the median from the development cohort, and missing nominal variables were assigned to a distinct “missing” category.

*c Continuous variables were standardized to zero mean and unit variance.

*d Nominal variables with less than 20 levels were represented as zero vectors of length equal to the number of levels, with level indicators equal to one.

*e Using residency zip code, we linked to US Census data to calculate residing neighborhood characteristics and distance from hospital.

*f Nominal variables with 20 levels or greater were transformed to a numeric integer identifier ranging from 0 to the number of unique levels minus one, where implicit variable representations were learned as part of the model training process.

*g To preserve relative proximity, temporally recurring features such as month and day of admission were cyclically embedded as two separate features by sine and cosine-based transformation. For example, December (12) is near January (1), and Sunday (7) is near Monday (1).

*h Medications were taken within one year timeframe prior to surgery using RxNorms data grouped into drug classes according to the US, Department of Veterans Affairs National Drug File-Reference Terminology 24.

*i Measurement values lying outside of expert-defined clinically normal value ranges for each variable were discarded. If two measurements existed at the same timestamp for a given patient, a random measurement was kept.

*j For each surgical procedure, a time series was constructed by arranging intraoperative measurements chronologically, resampling to one-minute frequency intervals, performing linear interpolation in both directions (except for blood loss and urine output, which were imputed with zero), and imputing the development median at every timestep for procedures lacking a single measurement of a particular variable.

*k For baseline models, a set of 49 statistical features was extracted from each intraoperative time series. This set included the following features: minimum, maximum, mean, median, standard deviation, sum of values, variance, kurtosis, skewness, absolute energy, absolute sum of changes, counts above and below mean, first and last locations of both minimum and maximum, sequence length, longest strike above and below mean, mean absolute change, mean change, ratio of unique values to sequence length, variance larger than standard deviation, 9 quantiles, 9 index mass quantiles, 10-binned entropy, number of peaks, and range count.
Table 4. Net reclassification index (NRI) results for both individually trained and multi-task deep learning models across each of the nine postoperative complications.

| Complication                  | Model Type   | NRI (95% CI) | p    | Event | Non-Event | Overall |
|-------------------------------|--------------|--------------|------|-------|-----------|---------|
| Prolonged ICU Stay            | Individual   | 0.07 (0.07-0.08) | < 0.001 | 3.87  | 3.03      | 3.27    |
|                               | Multi-Task   | 0.06 (0.04-0.07) | < 0.001 | 5.81  | 0.13      | 1.75    |
| Prolonged Mechanical Ventilation | Individual  | 0.04 (0.03-0.05) | 0.026 | 1.83  | 2.52      | 2.47    |
|                               | Multi-Task   | 0.06 (0.06-0.07) | < 0.001 | 0.39  | 5.62      | 5.28    |
| Sepsis                        | Individual   | 0.02 (0.02-0.02) | 0.098 | -1.23 | 2.77      | 2.28    |
|                               | Multi-Task   | 0.04 (0.02-0.04) | 0.001 | 1.92  | 1.68      | 1.71    |
| Acute Kidney Injury           | Individual   | 0.04 (0.04-0.05) | < 0.001 | -1.94 | 5.88      | 4.50    |
|                               | Multi-Task   | 0.02 (0.01-0.02) | 0.070 | -5.73 | 7.79      | 5.41    |
| Neurological Complications    | Individual   | 0.04 (0.03-0.04) | < 0.001 | 2.41  | 1.22      | 1.45    |
|                               | Multi-Task   | 0.01 (0.00-0.03) | 0.151 | 4.57  | -3.15     | -1.63   |
| Venous Thromboembolism        | Individual   | 0.03 (0.01-0.03) | 0.102 | -2.55 | 5.07      | 4.47    |
|                               | Multi-Task   | 0.01 (0.00-0.02) | 0.501 | -1.91 | 2.95      | 2.56    |
| Cardiovascular Complications  | Individual   | 0.11 (0.10-0.12) | < 0.001 | 4.11  | 6.64      | 6.17    |
|                               | Multi-Task   | 0.10 (0.09-0.11) | < 0.001 | 0.80  | 9.27      | 7.69    |
| Wound Complications           | Individual   | -0.03 (-0.03--0.02) | 0.002 | 6.01  | -8.87     | -5.09   |
|                               | Multi-Task   | 0.01 (0.01-0.02) | 0.293 | 3.84  | -2.80     | -1.11   |
| In-Hospital Mortality         | Individual   | 0.05 (0.03-0.06) | 0.133 | -1.78 | 6.33      | 6.14    |
|                               | Multi-Task   | 0.03 (0.02-0.05) | 0.274 | -5.69 | 8.94      | 8.60    |
Table 5. The twenty most influential clinical predictors per outcome using the method of integrated gradients applied to every sample in the validation set using the final trained deep learning postoperative model.

| Prolonged ICU Stay | Prolonged Mechanical Ventilation | Sepsis | Acute Kidney Injury | Neurological Complications | Venous Thromb. | Cardiovascular Complications | Wound Complications | In-Hospital Mortality |
|-------------------|---------------------------------|--------|---------------------|---------------------------|----------------|-----------------------------|---------------------|----------------------|
| SPO2, intraop freq (0.684) | SPO2, intraop freq (0.839) | Heart rate, intraop mean (0.323) | Age (0.385) | Cerebrovascular disease (0.447) | Admission type (non-emergency/emergency) (0.282) | SPO2, intraop freq (0.592) | UNCR, variance past week not missing (0.233) | Age (0.720) |
| BMI not missing (0.465) | PIP, intraop mean (0.721) | BMI not missing (0.298) | SPO2, intraop freq (0.373) | Surgery type = neurosurgery (0.330) | Hemoglobin, # tests past week (0.259) | Diastolic BP, intraop freq (0.380) | Age (0.164) | SPO2, intraop freq (0.642) |
| Diastolic BP, intraop freq (0.447) | BMI not missing (0.585) | SPO2, intraop freq (0.239) | BMI not missing (0.337) | SPO2, intraop freq (0.319) | BMI not missing (0.258) | Systolic BP, intraop freq (0.353) | Admission type (non-emergency/emergency) (0.164) | BMI not missing (0.424) |
| Systolic BP, intraop freq (0.441) | Hemoglobin, # tests past week (0.470) | Admission type (non-emergency/emergency) (0.224) | Heart rate, intraop freq (0.240) | BMI not missing (0.277) | Heart rate, intraop mean (0.233) | Systolic BP, intraop mean (0.305) | BMI not missing (0.140) | Diastolic BP, intraop freq (0.370) |
| Heart rate, intraop freq (0.356) | Diastolic BP, intraop freq (0.462) | Age (0.165) | Diastolic BP, intraop freq (0.228) | Age (0.237) | UNCR, variance past week not missing (0.200) | Heart rate, intraop freq (0.296) | Heart rate, intraop mean (0.356) | SPO2, intraop freq (0.136) |
| Heart rate, intraop mean (0.336) | Heart rate, intraop freq (0.453) | Systolic BP, intraop mean (0.162) | Systolic BP, intraop freq (0.217) | Gender (male/female) (0.226) | SPO2, intraop freq (0.197) | Heart rate, intraop mean (0.293) | Heart rate, intraop freq (0.353) | EGFR not missing (0.119) |
| PIP, intraop mean (0.288) | Systolic BP, intraop freq (0.428) | MAC, intraop mean (0.160) | EGFR not missing (0.204) | PIP, intraop mean (0.224) | Hemoglobin, # tests past year (0.167) | MAC, intraop mean (0.238) | Hemoglobin, min past week (0.111) | MAC, intraop mean (0.343) |
| Prolonged ICU Stay | Prolonged Mechanical Ventilation | Sepsis | Acute Kidney Injury | Neurological Complications | Venous Thromb. | Cardiovascular Complications | Wound Complications | In-Hospital Mortality |
|-------------------|---------------------------------|--------|--------------------|---------------------------|---------------|-----------------------------|---------------------|----------------------|
| Hemoglobin, # tests past week (0.279) | Heart rate, intraop mean (0.351) | Hemoglobin, # tests past week (0.148) | PIP, intraop mean (0.203) | Hemoglobin, # tests past week (0.202) | Systolic BP, intraop freq (0.141) | MAC, intraop freq (0.231) | Received pressors or inotropes in past year (0.110) | Systolic BP, intraop freq (0.334) |
| EGFR not missing (0.277) | FIO2, intraop mean (0.339) | Weight loss (0.146) | EGFR (0.190) | Systolic BP, intraop freq (0.191) | Heart rate, intraop freq (0.139) | BMI not missing (0.214) | MCHC, maximum past week (0.100) | PIP, intraop mean (0.320) |
| FIO2, intraop mean (0.250) | MAC, intraop mean (0.283) | Diastolic BP, intraop freq (0.145) | BUN, min past week (0.173) | Diastolic BP, intraop freq (0.190) | Hemoglobin, min past week (0.133) | Respiratory rate, intraop freq (0.204) | Blood loss, intraop mean (0.091) | Systolic BP, intraop mean (0.310) |
| Respiratory rate, intraop freq (0.239) | Weight loss (0.269) | Hemoglobin, min past week (0.143) | Heart rate, intraop mean (0.155) | Heart rate, intraop freq (0.185) | Diastolic BP, intraop freq (0.132) | PIP, intraop mean (0.181) | Surgery type = orthopedic (0.090) | Hemoglobin, # tests past week (0.303) |
| Scheduled postop location (non-ICU/ICU) (0.238) | EGFR not missing (0.263) | WBC, maximum past week (0.142) | Respiratory rate, intraop freq (0.138) | Elapsed time before surgery (0.176) | Hemoglobin, maximum past year (0.131) | Blood loss, intraop mean (0.181) | Urine glucose, # tests past year (0.088) | BMI (0.303) |
| MAC, intraop mean (0.227) | Age (0.246) | Charlson Comorbidity Index (0.141) | Ethnicity not missing (0.175) | Serum calcium, min past week not missing (0.123) | Age (0.179) | Heart rate, intraop mean (0.079) | Ethnicity not missing (0.258) |
| MAC, intraop freq (0.226) | ETCO2, intraop mean (0.238) | Systolic BP, intraop freq (0.140) | Admitting type = surgery (0.170) | Hemoglobin, min past year (0.122) | Hemoglobin, # tests past week (0.174) | PIP, intraop mean (0.076) | Weight loss (0.246) |
| Surgery type = orthopedic (0.180) | PIP, intraop freq (0.236) | Elapsed time before surgery (0.134) | ETCO2, intraop mean (0.133) | Heart rate, intraop mean (0.157) | Hemoglobin, variance past year (0.116) | ETCO2, intraop mean (0.128) | CPT = 27130 (0.076) | Charlson Comorbidity Index (0.230) |

42
| Prolonged ICU Stay | Prolonged Mechanical Ventilation | Sepsis | Acute Kidney Injury | Neurological Complications | Venous Thromb. | Cardiovascular Complications | Wound Complications | In-Hospital Mortality |
|--------------------|---------------------------------|--------|---------------------|---------------------------|-----------------|-----------------------------|---------------------|---------------------|
| Blood loss, intraop mean (0.163) | Respiratory rate, intraop freq (0.214) | Urine glucose, # tests past week (0.134) | MAC, intraop freq (0.121) | FIO2, intraop mean (0.154) | Urine red blood cells = negative past year (0.114) | Temperature, intraop freq (0.119) | Admission month (transformed) (0.075) | ETCO2, intraop mean (0.226) |
| Weight loss (0.152) | Surgery type = orthopedic (0.207) | Serum chloride, min past week (0.133) | FIO2, intraop mean (0.119) | RDW, min past week (0.143) | Hemoglobin, mean past year (0.113) | Weight loss (0.118) | Weight loss (0.074) | Serum chloride, min past week (0.217) |
| Age (0.141) | MAC, intraop freq (0.199) | Serum sodium, mean past week (0.119) | Surgery type = orthopedic (0.116) | Admitting type = medicine (0.141) | Serum glucose, # tests past week (0.111) | FIO2, intraop mean (0.109) | Diastolic BP, intraop freq (0.074) | MAC, intraop freq (0.212) |
| Admitting type = medicine (0.137) | Cerebrovascular disease (0.198) | Urine protein = negative past year (0.116) | MAC, intraop mean (0.112) | Serum calcium, mean past week not missing (0.109) | Serum creatinine, # tests past week (0.101) | Residential poverty % (0.072) | Respiratory rate, intraop freq (0.204) | |
| ETCO2, intraop mean (0.128) | Scheduled postop location (non-ICU/ICU) (0.187) | Hemoglobin, mean past week (0.109) | BUN, maximum past week (0.110) | Urine red blood cells, # tests past year (0.132) | Hemoglobin, variance past week not missing (0.108) | Hemoglobin, min past week (0.096) | CPT = 27447 (0.070) | Admission type (non-emergency/emergency) (0.195) |
Table 6. Input variable statistics for all included clinical features. Summaries shown for both training (development) and testing (validation) cohorts.

|                                | Training Cohort (n=40,560) | Test Cohort (n=11,969) |
|--------------------------------|----------------------------|------------------------|
| **Socio-economic features**    |                            |                        |
| Neighborhood characteristics   |                            |                        |
| Rural area, n (%)              | 13986 (34.4%)              | 4111 (34.3%)           |
| Total population, median (25\(^{th}\),75\(^{th}\)) | 17599 (10884, 27063) | 17599 (10923, 27063) |
| Median income, median (25\(^{th}\),75\(^{th}\))     | 40528 (35194, 48430)      | 40320 (35244, 48245)  |
| Total proportion of African-Americans (%) (SD) | 0.16 (0.15)               | 0.16 (0.15)            |
| Total proportion of Hispanic (%) (SD) | 0.08 (0.08)               | 0.08 (0.08)            |
| Population proportion below poverty (%) (SD) | 20 (9.5)                  | 20 (2.5)               |
| Distance from residency to hospital (km), median (25\(^{th}\),75\(^{th}\)) | 43.2 (22.1, 81.1) | 43.6 (22.3, 80.7) |
| **Comorbidity features**       |                            |                        |
| Cancer, n (%)                  | 11381 (28%)                | 3136 (26.2%)           |
| Peripheral vascular disease, n (%) | 9207 (22.7%)             | 3022 (25.2%)           |
| Cerebrovascular disease, n (%)  | 7048 (17.4%)               | 2221 (18.5%)           |
| Myocardial infarction, n (%)    | 3187 (7.9%)                | 1107 (9.2%)            |
| Liver disease, n (%)            | 5979 (14.7%)               | 1955 (16.3%)           |
| Weight Loss, n (%)              | 5754 (14.2%)               | 2200 (18.4%)           |
| Diabetes, n (%)                 | 10079 (24.8%)              | 2917 (24.4%)           |
| Alcohol/ Drug abuse, n (%)      | 6338 (15.6%)               | 1783 (14.9%)           |
| Condition                        | Training Cohort (n=40,560) | Test Cohort (n=11,969) |
|---------------------------------|-----------------------------|------------------------|
| Congestive heart failure, n (%) | 7020 (17.3%)                | 2343 (19.6%)           |
| Valvular Disease, n (%)         | 5814 (14.3%)                | 2117 (17.7%)           |
| Coagulapty, n (%)               | 5940 (14.6%)                | 1776 (14.8%)           |
| Smoking, n (%)                  |                             |                        |
| Never                           | 17325 (42.7%)               | 5218 (43.6%)           |
| Former                          | 13827 (34.1%)               | 4168 (34.8%)           |
| Current                         | 7385 (18.2%)                | 2001 (16.7%)           |
| Missing                         | 2068 (5.1%)                 | 593 (5%)               |
| Number of diagnoses, median (25th-75th) | 41 (20, 91) | 48 (25, 110) |

**Operative features**

| Feature                        | Training Cohort | Test Cohort |
|--------------------------------|-----------------|-------------|
| Night admission, n (%)         | 18968 (46.7%)   | 5718 (47.7%) |
| Admission day (top 3 categories), n (%) |  |   |
| Monday                         | 7944 (19.6%)    | 2406 (20.1%) |
| Tuesday                        | 7665 (18.9%)    | 2177 (18.2%) |
| Wednesday                      | 6896 (17%)      | 2081 (17.4%) |
| Admission month (top 3 categories), n (%) |  |   |
| October                        | 3799 (9.4%)     | 1083 (9%)   |
| January                        | 3720 (9.2%)     | 1059 (8.8%) |
| August                         | 3704 (9.1%)     | 1049 (8.8%) |
| Number of operating surgeons, n | 283             | 195         |
| Number of procedures per operating surgeon, n (%) |  |   |
| First rank                     | 1319 (3.3%)     | 316 (2.6%)  |
| Category                                                                 | Training Cohort (n=40,560) | Test Cohort (n=11,969) |
|--------------------------------------------------------------------------|-----------------------------|------------------------|
| Second rank                                                              | 1129 (2.8%)                 | 302 (2.5%)             |
| Third rank                                                               | 1044 (2.6%)                 | 300 (2.5%)             |
| Admission source, n (%)                                                 |                             |                        |
| Transfer                                                                 | 6668 (16.4%)                | 2046 (17.1%)           |
| Emergent at Admission status, n (%)                                      | 15348 (37.8%)               | 4855 (40.5%)           |
| Admission to surgical service, n (%)                                     | 19264 (47.4%)               | 5851 (48.8%)           |
| Time of surgery from admission (days), median (25th, 75th)               | 3 (2, 25)                   | 3 (2, 30)              |
| Type of Surgery, n (%)                                                   |                             |                        |
| Orthopedic surgery                                                       | 9983 (24.6%)                | 2895 (24.1%)           |
| Neurosurgery                                                             | 5625 (13.8%)                | 1906 (15.9%)           |
| Vascular Surgery                                                         | 4043 (10%)                  | 1190 (9.9%)            |
| Thoracic/Cardiovascular surgery                                          | 3214 (7.9%)                 | 1241 (10.4%)           |
| Urologic surgery                                                         | 3131 (7.7%)                 | 704 (5.9%)             |
| Trauma- Acute Care surgery                                              | 2953 (7.3%)                 | 1056 (8.8%)            |
| Gastrointestinal surgery                                                 | 2521 (6.2%)                 | 784 (6.5%)             |
| Ear, nose, throat surgery                                                | 2446 (6%)                   | 592 (4.9%)             |
| Gynecology obstetrics surgery                                            | 1672 (4.1%)                 | 320 (2.7%)             |
| Pancreas & Biliary, BMSE surgery                                         | 1451 (3.6%)                 | 337 (2.8%)             |
| Transplant surgery                                                       | 961 (2.4%)                  | 227 (1.9%)             |
| Plastic surgery                                                          | 946 (2.3%)                  | 216 (1.8%)             |
| Burn Surgery                                                             | 837 (2.1%)                  | 185 (1.5%)             |
| Pediatric surgery                                                        | 447 (1.1%)                  | 123 (1%)               |
| Other specialty surgeries                                                | 299 (0.7%)                  | 164 (1.4%)             |
| Ophthalmology surgery                                                    | 80 (0.2%)                   | 51 (0.4%)              |
|                     | Training Cohort (n=40,560) | Test Cohort (n=11,969) |
|---------------------|----------------------------|------------------------|
| **Medicine Gastroenterology** | 27 (0.1%) | 4 (0%) |
| **Preoperative and admission day laboratory results** | | |
| **median (25th,75th)** | | |
| Hemoglobin within 7 days prior to surgery, g/dl | | |
| Minimum | 12.3 (11.4, 13.3) | 12.3 (11.4, 13.3) |
| Maximum | 13 (12.2, 13.9) | 13 (12.3, 14) |
| Average | 12.7 (11.8, 13.5) | 12.7 (11.9, 13.6) |
| Variance | 0.4 (0.4, 0.4) | 0.4 (0.4, 0.4) |
| Hemoglobin within 8-365 days prior to surgery, g/dl | | |
| Minimum | 11.8 (11.3, 12.2) | 11.8 (11.3, 12.5) |
| Maximum | 13.6 (13.3, 13.8) | 13.6 (13.3, 14) |
| Average | 12.5 (12.2, 12.8) | 12.5 (12.2, 13) |
| Variance | 0.9 (0.9, 0.9) | 0.9 (0.9, 0.9) |
| Glucose in blood within 7 days prior to surgery, mg/dL | | |
| Minimum | 99 (91, 110) | 99 (93, 112) |
| Maximum | 121 (105, 141) | 121 (107, 145) |
| Average | 111 (100, 125) | 111 (102, 126.5) |
| Variance | 269.8 (269.8, 269.8) | 269.8 (269.8, 269.8) |
| Count | 1 (1, 3) | 1 (1, 3) |
| Urea nitrogen in blood within 7 days prior to surgery, mg/dL | | |
|                         | Training Cohort (n=40,560) | Test Cohort (n=11,969) |
|-------------------------|----------------------------|------------------------|
| Minimum                 | 14 (12, 17)                | 14 (12, 17)            |
| Maximum                 | 16 (14, 19)                | 16 (14, 20)            |
| Average                 | 15 (13, 18)                | 15 (13, 18.5)          |
| Variance                | 4.5 (4.5, 4.5)             | 4.5 (4.5, 4.5)         |
| Count                   | 1 (0, 2)                   | 1 (0, 2)               |

Serum creatinine within 7 days prior to surgery, mg/dL
- Minimum: 0.9 (0.8, 1)   Test: 0.9 (0.7, 0.9)
- Maximum: 0.9 (0.8, 1.1) Test: 0.9 (0.8, 1)
- Average: 0.9 (0.8, 1)   Test: 0.9 (0.8, 1)
- Variance: 0 (0, 0)      Test: 0 (0, 0)
- Count: 1 (0, 2)         Test: 1 (0, 3)

Serum Calcium within 7 days prior to surgery, mmol/L
- Minimum: 9.1 (8.8, 9.3)  Test: 9.1 (8.8, 9.4)
- Maximum: 9.3 (9.2, 9.5)  Test: 9.3 (9.2, 9.6)
- Average: 9.2 (9.4)       Test: 9.2 (9.4)
- Variance: 0.1 (0.1, 0.1) Test: 0.1 (0.1, 0.1)
- Count: 1 (0, 2)          Test: 1 (0, 2)

Serum Sodium ion within 7 days prior to surgery, mmol/L
- Minimum: 138 (137, 139)  Test: 138 (136, 139)
- Maximum: 140 (139, 141)  Test: 140 (138, 140)
- Average: 139 (138, 140)  Test: 139 (137, 139)
- Variance: 2.6 (2.6, 2.6) Test: 2.6 (2.6, 2.6)
|                          | Training Cohort (n=40,560) | Test Cohort (n=11,969) |
|--------------------------|-----------------------------|-------------------------|
| **Count**                | 1 (0, 2)                    | 1 (0, 3)                |
| **Urea nitrogen-Creatinine ratio within 7 days prior to surgery** |                             |                         |
| Minimum                  | 15 (12.9, 17.4)             | 15 (13.7, 18.4)         |
| Maximum                  | 17.3 (15, 20)               | 17.3 (16, 21.3)         |
| Average                  | 16.2 (14, 18.6)             | 16.2 (15, 19.8)         |
| Variance                 | 1.6 (1.6, 1.6)              | 1.6 (0, 1.6)            |
| Count                    | 1 (0, 4)                    | 2 (0, 6)                |
| **Potassium in serum within 7 days prior to surgery, mmol/L** |                             |                         |
| Minimum                  | 3.9 (3.7, 4.1)              | 3.9 (3.7, 4)            |
| Maximum                  | 4.2 (4, 4.4)                | 4.2 (4, 4.3)            |
| Average                  | 4.1 (3.9, 4.2)              | 4.1 (3.9, 4.1)          |
| Variance                 | 0.1 (0.1, 0.1)              | 0.1 (0.1, 0.1)          |
| Count                    | 1 (0, 2)                    | 1 (0, 3)                |
| **Chloride in Serum within 7 days prior to surgery, mmol/L** |                             |                         |
| Minimum                  | 100 (99, 102)               | 100 (100, 104)          |
| Maximum                  | 102 (101, 104)              | 102 (102, 106)          |
| Average                  | 101 (100, 102.8)            | 101 (101, 104.7)        |
| Variance                 | 4 (4, 4)                    | 4 (4, 4)                |
| Count                    | 1 (0, 2)                    | 1 (0, 2)                |
| **Serum CO2 within 7 days prior to surgery, mmol/L** |                             |                         |
| Minimum                  | 24 (23, 25)                 | 24 (23, 26)             |
|                         | Training Cohort (n=40,560) | Test Cohort (n=11,969) |
|-------------------------|-----------------------------|------------------------|
| Maximum                 | 26 (25, 27)                 | 26 (25, 27)            |
| Average                 | 25 (24, 26)                 | 25 (24, 26)            |
| Variance                | 2.6 (2.6, 2.6)              | 2.6 (2.6, 2.6)         |
| Count                   | 1 (0, 2)                    | 1 (0, 3)               |

White Blood Cell in blood within 7 days prior to surgery, thou/uL

|                         | Training Cohort | Test Cohort |
|-------------------------|-----------------|-------------|
| Minimum                 | 7.6 (6.6, 8.6)  | 7.6 (6.6, 8.5) |
| Maximum                 | 8.8 (7.5, 10.3) | 8.8 (7.7, 10.2) |
| Average                 | 8.3 (7.1, 9.4)  | 8.3 (7.2, 9.3)  |
| Variance                | 1.7 (1.7, 1.7)  | 1.7 (1.7, 1.7)  |
| Count                   | 1 (0, 2)        | 1 (0, 2)      |

Mean Corpuscular Volume in blood within 7 days prior to surgery, fL

|                         | Training Cohort | Test Cohort |
|-------------------------|-----------------|-------------|
| Minimum                 | 90.6 (90.6, 90.6) | 90.6 (90.6, 90.6) |
| Maximum                 | 91.5 (91.5, 91.5) | 91.5 (91.5, 91.5) |
| Average                 | 91 (91, 91)     | 91 (91, 91)  |
| Variance                | 0.8 (0.8, 0.8)  | 0.8 (0.8, 0.8) |
| Count                   | 0 (0, 1)        | 0 (0, 0)     |

Mean Corpuscular Hemoglobin in blood within 7 days prior to surgery, g/dL

|                         | Training Cohort | Test Cohort |
|-------------------------|-----------------|-------------|
| Minimum                 | 29.8 (29, 30.6) | 29.8 (29.2, 30.8) |
| Maximum                 | 30.2 (29.4, 31) | 30.2 (29.6, 31.2) |
| Average                 | 30 (29.2, 30.8) | 30 (29.4, 31) |
| Variance                | 0.1 (0.1, 0.1)  | 0.1 (0.1, 0.1) |
| Count                   | 1 (0, 2)        | 1 (0, 2)     |
|                                | Training Cohort | Test Cohort |
|--------------------------------|----------------|-------------|
|                                | (n=40,560)     | (n=11,969)  |
| **Amount of hemoglobin relative to the size of the cell in blood, g/dL** |                |             |
| Minimum                        | 32.7 (32.1, 33.2) | 32.9 (32.7, 33.8) |
| Maximum                        | 33.5 (33, 33.9)  | 33.5 (33.5, 34.2) |
| Average                        | 33.1 (32.6, 33.5) | 33.2 (33.1, 34) |
| Variance                       | 0.2 (0.2, 0.3)   | 0.2 (0.2, 0.2) |
| Count                          | 2 (0, 4)        | 1 (0, 2)    |
| **Red cell distribution width in Blood within 7 days prior to surgery, %** |                |             |
| Minimum                        | 14.2 (13.7, 14.7) | 14.2 (13.7, 14.8) |
| Maximum                        | 14.5 (14, 15.1)  | 14.5 (13.9, 15) |
| Average                        | 14.3 (13.8, 14.9) | 14.3 (13.8, 14.9) |
| Variance                       | 0.1 (0.1, 0.1)   | 0.1 (0.1, 0.1) |
| Count                          | 1 (0, 2)        | 1 (0, 2)    |
| **Platelet in blood, within 7 days prior to surgery thou/uL** |                |             |
| Minimum                        | 219 (192, 248)   | 219 (194, 250) |
| Maximum                        | 239 (211, 269)   | 239 (215, 273) |
| Average                        | 228 (202, 258)   | 228 (205, 259.5) |
| Variance                       | 406.9 (406.9, 406.9) | 406.9 (406.9, 406.9) |
| Count                          | 1 (0, 2)        | 1 (0, 2)    |
| **Mean platelet volume in blood within 7 days prior to surgery, fL** |                |             |
| Minimum                        | 7.8 (7.8, 7.8)   | 7.8 (7.8, 7.8) |
| Maximum                        | 8.3 (8.3, 8.3)   | 8.3 (8.3, 8.3) |
|                           | Training Cohort  | Test Cohort   |
|---------------------------|------------------|--------------|
|                           | (n=40,560)       | (n=11,969)   |
| Average                   | 8 (8, 8)         | 8 (8, 8)     |
| Variance                  | 0.2 (0.2, 0.2)   | 0.2 (0.2, 0.2) |
| Count                     | 0 (0, 1)         | 0 (0, 0)     |
| Reference estimated glomerular filtration rate | 92.9 (83, 102.7) | 92.9 (82.5, 103.3) |

*Automated urinalysis, urine protein within 365 days prior to surgery (mg/dL), n (%)*

|                          | Training Cohort | Test Cohort |
|--------------------------|-----------------|-------------|
| Missing                  | 21410 (52.7%)   | 6631 (55.3%)|
| Negative                 | 12424 (30.6%)   | 3503 (29.2%)|
| Small (<30)              | 1189 (2.9%)     | 120 (1%)    |
| Moderate (300)           | 4423 (10.9%)    | 1352 (11.3%)|
| Large (>=300)            | 1194 (2.9%)     | 389 (3.2%)  |

*Automated urinalysis, urine glucose within 7 days prior to surgery (mg/dL), n (%)*

|                          | Training Cohort | Test Cohort |
|--------------------------|-----------------|-------------|
| Missing                  | 30673 (75.5%)   | 9299 (77.2%)|
| Negative                 | 8740 (21.5%)    | 2347 (19.6%)|
| Small (<499)             | 661 (1.6%)      | 200 (1.7%)  |
| Moderate (1000)          | 317 (0.8%)      | 138 (1.2%)  |
| Large (>1000)            | 249 (0.6%)      | 11 (0.1%)   |

*Automated urinalysis, urine glucose within 8 to 365 days prior to surgery (mg/dL), n (%)*

|                          | Training Cohort | Test Cohort |
|--------------------------|-----------------|-------------|
| Missing                  | 28151 (69.3%)   | 8492 (70.8%)|
| Negative                 | 11446 (28.2%)   | 3195 (26.6%)|
| Small (<500)             | 440 (1.1%)      | 137 (1.1%)  |
|                    | Training Cohort (n=40,560) | Test Cohort (n=11,969) |
|--------------------|----------------------------|------------------------|
| Moderate (<1000)   | 269 (0.7%)                 | 140 (1.2%)             |
| Large (>1000)      | 334 (0.8%)                 | 31 (0.3%)              |

bAutomated urinalysis, urine hemoglobin within 7 days prior to surgery (mg/dL), n (%)

|                    | Training Cohort (n=40,560) | Test Cohort (n=11,969) |
|--------------------|----------------------------|------------------------|
| Missing            | 34190 (84.2%)              | 11970 (99.9%)          |
| Negative           | 4030 (9.9%)                | 7 (0.1%)               |
| Small              | 1266 (3.1%)                | 0 (0%)                 |
| Moderate           | 620 (1.5%)                 | 1 (0%)                 |
| Large              | 499 (1.2%)                 | 2 (0%)                 |

bAutomated urinalysis, urine hemoglobin within 8 to 365 days prior to surgery (mg/dL), n (%)

|                    | Training Cohort (n=40,560) | Test Cohort (n=11,969) |
|--------------------|----------------------------|------------------------|
| Missing            | 32585 (80.3%)              | 11812 (98.6%)          |
| Negative           | 5860 (14.4%)               | 134 (1.1%)             |
| Small              | 956 (2.4%)                 | 17 (0.1%)              |
| Moderate           | 548 (1.4%)                 | 7 (0.1%)               |
| Large              | 656 (1.6%)                 | 10 (0.1%)              |

aAutomated urinalysis, urine erythrocytes within 365 days prior to surgery (mg/dL), n (%)

|                    | Training Cohort (n=40,560) | Test Cohort (n=11,969) |
|--------------------|----------------------------|------------------------|
| Missing            | 24724 (60.8%)              | 7163 (59.7%)           |
| Negative (<=4)     | 12657 (31.1%)              | 4099 (34.2%)           |
| Small (>4)         | 1423 (3.5%)                | 175 (1.5%)             |
| Moderate (>30)     | 411 (1%)                   | 192 (1.6%)             |
| Training Cohort (n=40,560) | Test Cohort (n=11,969) |
|----------------------------|------------------------|
| Large (>=50)               |                        |
| 1425 (3.5%)                | 366 (3.1%)             |
| Number of complete blood count tests, n (%) |                       |
| 29021 (71.5%)              | 8436 (70.4%)           |

**Medication history (1 year prior to Surgery)**

| Medication groups, n (%)                  | Training Cohort | Test Cohort |
|------------------------------------------|-----------------|-------------|
| Beta blockers                             | 6994 (17.2%)    | 2153 (18%)  |
| Diuretics                                | 4602 (11.3%)    | 1323 (11%)  |
| Statins                                  | 3676 (9.1%)     | 1259 (10.5%)|
| Aspirin                                  | 5708 (14.1%)    | 1807 (15.1%)|
| Angiotensin-converting-enzyme inhibitors  | 4139 (10.2%)    | 1204 (10.1%)|
| Vasopressors and inotropes               | 8427 (20.8%)    | 2799 (23.4%)|
| Bicarbonate                              | 4582 (11.3%)    | 1420 (11.9%)|
| Anti-emetics                             | 11788 (29%)     | 3694 (30.8%)|
| Aminoglycosides                          | 1371 (3.4%)     | 463 (3.9%)  |

**Intraoperative Variables**

| Diastolic Blood Pressure, mm Hg          | Training Cohort | Test Cohort |
|------------------------------------------|-----------------|-------------|
| Minimum, mean (SD)                       | 40.54 (12.99)   | 41.44 (14.25)|
| Maximum, mean (SD)                       | 98.91 (25.69)   | 101.43 (26.18)|
| Average, mean (SD)                       | 63.57 (9.69)    | 65 (10.15)  |
| Long Term Variability, mean (SD)         | 119.45 (119.92) | 120.78 (105.91)|
| Short Term Variability, mean (SD)        | 34.85 (53.71)   | 33.58 (54.05) |

| Systolic Blood Pressure, mm Hg           | Training Cohort | Test Cohort |
|------------------------------------------|-----------------|-------------|
| Minimum, mean (SD)                       | 76.43 (20.97)   | 76.52 (22.41)|
| Measure                              | Training Cohort (n=40,560) | Test Cohort (n=11,969) |
|-------------------------------------|-----------------------------|------------------------|
| Maximum, mean (SD)                  | 165.05 (32.27)              | 167.86 (33.11)         |
| Average, mean(SD)                   | 115.09 (14.66)              | 116.36 (14.83)         |
| Long Term Variability, mean (SD)    | 315.68 (248.99)             | 324.53 (247.35)        |
| Short Term Variability, mean (SD)   | 45.97 (67.87)               | 45.63 (65.34)          |

**Heart Rate, bpm**

| Measure                              | Training Cohort | Test Cohort |
|-------------------------------------|-----------------|-------------|
| Minimum, mean (SD)                  | 60 (13.41)      | 59.99 (13.87)|
| Maximum, mean (SD)                  | 109.32 (29.49)  | 107.99 (27.88)|
| Average, mean(SD)                   | 77.73 (13.35)   | 77.87 (13.5) |
| Long Term Variability, mean (SD)    | 108.93 (243.74) | 96.84 (160.45)|
| Short Term Variability, mean (SD)   | 8.46 (16.55)    | 8.09 (15)   |

**Respiratory Rate**

| Measure                              | Training Cohort | Test Cohort |
|-------------------------------------|-----------------|-------------|
| Minimum, mean (SD)                  | 2.6 (2.34)      | 3.11 (3.77) |
| Maximum, mean (SD)                  | 26.03 (8.59)    | 25.91 (8.72)|
| Average, mean(SD)                   | 11.49 (2.86)    | 12.1 (3.09) |
| Long Term Variability, mean (SD)    | 13.32 (14.56)   | 13.68 (16.34)|
| Short Term Variability, mean (SD)   | 1.37 (3.6)      | 1.43 (3.67) |

**Peripheral capillary oxygen saturation (SpO2)**

| Measure                              | Training Cohort | Test Cohort |
|-------------------------------------|-----------------|-------------|
| Minimum, mean (SD)                  | 88.18 (9.37)    | 87.78 (9.88)|
| Maximum, mean (SD)                  | 99.88 (0.65)    | 99.87 (0.7) |
| Average, mean(SD)                   | 98.12 (1.81)    | 98.05 (1.86)|
| Long Term Variability, mean (SD)    | 4.84 (13.78)    | 5.46 (24.26)|
| Short Term Variability, mean (SD)   | 0.15 (1.1)      | 0.16 (1.42)|

**End-tidal CO2 (ETCO2)**

| Measure                              | Training Cohort | Test Cohort |
|-------------------------------------|-----------------|-------------|
| Minimum, mean (SD)                  | 15.71 (5.85)    | 17.52 (7.62)|
|                                      | Training Cohort | Test Cohort |
|--------------------------------------|----------------|------------|
|                                      | \(n=40,560\)   | \(n=11,969\) |
| Maximum, mean (SD)                   | 46.72 (8.25)   | 45.28 (8.93) |
| Average, mean (SD)                   | 34.15 (4.67)   | 34.12 (5.24) |
| Long Term Variability, mean (SD)     | 26.06 (25.99)  | 26 (25.84)  |
| Short Term Variability, mean (SD)    | 1.26 (7.64)    | 1.18 (2.54) |
| **Respiratory O2**                   |                |            |
| Minimum, mean (SD)                   | 2.74 (1.79)    | 2.83 (1.67) |
| Maximum, mean (SD)                   | 2.79 (1.84)    | 2.89 (1.77) |
| Average, mean (SD)                   | 2.76 (1.8)     | 2.86 (1.7)  |
| Long Term Variability, mean (SD)     | 0.72 (4.91)    | 0.75 (5.74) |
| Short Term Variability, mean (SD)    | 0 (0)          | 0 (0)      |
| **Fraction of inspired oxygen (FiO2)** |          |            |
| Minimum, mean (SD)                   | 30.33 (9.09)   | 25.23 (7.58) |
| Maximum, mean (SD)                   | 39.8 (10.16)   | 35.83 (13.09) |
| Average, mean (SD)                   | 37.76 (5.77)   | 31.82 (9.13) |
| Long Term Variability, mean (SD)     | 5.85 (17.2)    | 8.06 (33.41) |
| Short Term Variability, mean (SD)    | 0.45 (9.12)    | 0.4 (4.64)  |
| **Positive end-expiratory pressure (PEEP)** |          |            |
| Minimum, mean (SD)                   | 6.14 (2.22)    | 6.2 (2.07)  |
| Maximum, mean (SD)                   | 6.19 (2.24)    | 6.24 (2.1)  |
| Average, mean (SD)                   | 6.16 (2.21)    | 6.22 (2.07) |
| Long Term Variability, mean (SD)     | 2.2 (4.11)     | 2.86 (5.69) |
| Short Term Variability, mean (SD)    | 0 (0)          | 0 (0)      |
| **Peak Inspiratory Pressure (PIP)**  |                |            |
| Minimum, mean (SD)                   | 0.08 (0.89)    | 0.07 (0.83) |
| Maximum, mean (SD)                   | 25.42 (10.04)  | 24.93 (10.08) |
|                                      | Training Cohort        | Test Cohort           |
|--------------------------------------|------------------------|-----------------------|
|                                      | (n=40,560)             | (n=11,969)            |
| Average, mean(SD)                    | 15.42 (7.08)           | 14.73 (6.84)          |
| Long Term Variability, mean (SD)     | 47.01 (36.57)          | 45.8 (36.73)          |
| Short Term Variability, mean (SD)    | 2.89 (4.24)            | 2.68 (4.16)           |
| Minimum alveolar concentration (MAC) |                        |                       |
| Minimum, mean (SD)                   | 0.07 (0.07)            | 0.04 (0.07)           |
| Maximum, mean (SD)                   | 1.06 (0.38)            | 1.02 (0.43)           |
| Average, mean(SD)                    | 0.61 (0.22)            | 0.55 (0.26)           |
| Long Term Variability, mean (SD)     | 0.05 (0.04)            | 0.06 (0.06)           |
| Short Term Variability, mean (SD)    | 0 (0.01)               | 0 (0.01)              |
| Temperature, °C                       |                        |                       |
| Minimum, mean (SD)                   | 35.49 (2.62)           | 35.62 (2.78)          |
| Maximum, mean (SD)                   | 37.59 (0.69)           | 37.57 (0.7)           |
| Average, mean(SD)                    | 36.96 (0.82)           | 36.98 (0.86)          |
| Variance, mean(SD)                   | 0.47 (2.08)            | 0.68 (2.52)           |

*a Result of both numeric data and text extraction

*b Result of text extraction; no numerical extraction was performed