BJA

British Journal of Anaesthesia, 123 (6): 877–886 (2019)

doi: 10.1016/j.bja.2019.07.030
Advance Access Publication Date: 15 October 2019
Quality and Patient Safety

QUALITY AND PATIENT SAFETY

An automated machine learning-based model predicts postoperative mortality using readily-extractable preoperative electronic health record data

Brian L. Hill[1,†], Robert Brown[1,†], Eilon Gabel[2], Nadav Rakocz[1], Christine Lee[3,4], Maxime Cannesson[2], Pierre Baldi[4,15], Loes Olde Loohuis[5], Ruth Johnson[1], Brandon Jew[6], Uri Maoz[2,9,10,11,12,13], Aman Mahajan[14], Sriram Sankararaman[1,7,‡], Ira Hofer[2,*,‡] and Eran Halperin[1,2,7,8,‡]

[1]Department of Computer Science, University of California, Los Angeles, CA, USA, [2]Department of Anaesthesiology and Perioperative Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA, [3]Department of Anaesthesiology and Perioperative Care, University of California, Irvine, CA, USA, [4]Department of Biomedical Engineering, University of California, Irvine, CA, USA, [5]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behaviour, University of California, Los Angeles, CA, USA, [6]Bioinformatics Interdepartmental Program, University of California, Los Angeles, CA, USA, [7]Department of Human Genetics, University of California, Los Angeles, CA, USA, [8]Department of Biomathematics, University of California, Los Angeles, CA, USA, [9]Crean College of Health and Behavioural Sciences, Chapman University, Orange, CA, USA, [10]Schmid College of Science and Technology, Chapman University, Orange, CA, USA, [11]Institute for Interdisciplinary Brain and Behavioural Sciences, Chapman University, Orange, CA, USA, [12]Anderson School of Management at UCLA, Los Angeles, CA, USA, [13]Biology and Bioengineering, California Institute of Technology, Pasadena, CA, USA, [14]Department of Anaesthesiology and Perioperative Medicine, University of Pittsburgh, Pittsburgh, PA, USA and [15]Department of Computer Science, University of California, Irvine, CA, USA

*Corresponding author. E-mail: ihofer@mednet.ucla.edu

[†]These authors contributed equally to this work.
[‡]These authors contributed equally to this work.

Abstract

Background: Rapid, preoperative identification of patients with the highest risk for medical complications is necessary to ensure that limited infrastructure and human resources are directed towards those most likely to benefit. Existing risk scores either lack specificity at the patient level or utilise the American Society of Anesthesiologists (ASA) physical status classification, which requires a clinician to review the chart.

Methods: We report on the use of machine learning algorithms, specifically random forests, to create a fully automated score that predicts postoperative in-hospital mortality based solely on structured data available at the time of surgery. Electronic health record data from 53 097 surgical patients (2.01% mortality rate) who underwent general anaesthesia between April 1, 2013 and December 10, 2018 in a large US academic medical centre were used to extract 58 preoperative features.

Results: Using a random forest classifier we found that automatically obtained preoperative features (area under the curve [AUC] of 0.932, 95% confidence interval [CI] 0.910–0.951) outperforms Preoperative Score to Predict Postoperative Mortality (POSPOM) scores (AUC of 0.660, 95% CI 0.598–0.722), Charlson comorbidity scores (AUC of 0.742, 95% CI

Editorial decision date: 11 Dec 2018; **Accepted:** 29 July 2019

0.658–0.812), and ASA physical status (AUC of 0.866, 95% CI 0.829–0.897). Including the ASA physical status with the preoperative features achieves an AUC of 0.936 (95% CI 0.917–0.955).
Conclusions: This automated score outperforms the ASA physical status score, the Charlson comorbidity score, and the POSPOM score for predicting in-hospital mortality. Additionally, we integrate this score with a previously published postoperative score to demonstrate the extent to which patient risk changes during the perioperative period.

Keywords: electronic health record; hospital mortality; machine learning; perioperative outcome; risk assessment

Editor's key points

- Perioperative risk prediction models need to be accurate and locally calibrated, but also clinically accessible.
- This study evaluates machine learning using readily available healthcare data to improve risk prediction.
- Changes in patient condition throughout the perioperative period can be included to update risk assessment.

A small proportion of high-risk patients comprise the majority of patients with surgical complications.[1] Many studies have demonstrated that early interventions can help reduce or even prevent perioperative complications.[2,3] In the current value-based care environment, it is critical to have methods to rapidly identify patients who are at the highest risk for perioperative complications and most likely to benefit from labour or cost-intensive interventions. Unfortunately, many current methods of risk stratification either lack patient-level precision or require a trained clinician to review each patient's medical record and assess a score.

Existing preoperative patient risk scores generally fall into one of two groups. The first leverages International Statistical Classification of Diseases and Related Health Problems (ICD) codes in order to create models of risk.[4–6] Unfortunately, ICD codes are not available until after patient discharge. While these scores tend to perform well at the population level, they rely on data not available before surgery, and have been repeatedly shown to lack precision at the patient level.[7] The second group of models relies on subjective clinician judgment, as seen with the ASA physical status score (ASA score) alone or when incorporated into another model (such as the National Surgical Quality Improvement Program [NSQIP] risk calculator).[8] While these scores tend to have increased precision compared with ICD codes, they cannot be fully automated because of the need for a highly trained clinician to manually review the patient's chart before calculation.

Recently, attempts have been made to leverage machine learning techniques using healthcare data in order to improve the predictive ability of various models.[9,10] These methods have shown progress in leveraging increasingly complex data while still allowing for the full automation of the scoring system.

In this manuscript, we hypothesised that machine learning methods can be used to predict in-hospital post-surgical mortality using only features from the electronic medical record (EMR) readily available and automatically extracted before surgery. We compare the performance of our model with existing clinical risk scores (ASA score, POSPOM score,[4] and Charlson comorbidity score[6]). Lastly, we aim to integrate our model with a previously published model[11] that estimates in-hospital mortality at the end of surgery to quantify the change in risk during the perioperative period.

Methods

Data source and extraction

All data for this study were extracted from the perioperative data warehouse (PDW), a custom built, robust data warehouse containing all patients who have undergone surgery at UCLA Health since the implementation of UCLA's EMR (EPIC Systems, Madison, WI, USA) in March 2013. We have previously described the creation of the PDW, which has a two-stage design.[12] Briefly, in the first stage, data are extracted from EPIC's Clarity database into 29 tables organised around three distinct concepts: patients, surgical procedures, and health system encounters. These data are then used to populate a series of 4000 distinct measures and metrics such as procedure duration, readmissions, admission ICD codes, and others. All data used for this study were obtained from this data warehouse and institutional review board (IRB) approval (IRB#16-001768) was obtained with exemption status for this retrospective review.

Model endpoint definition

We trained classification models to predict in-hospital mortality as a binary outcome. This classification was extracted from the PDW and was set to true if a 'death date' was noted during the hospitalisation, or the final disposition was set to 'expired' and there were no future admissions for the patient and a clinician 'death note' existed. Because of the concern about the need to eliminate false positive results, the resulting labels using this definition were validated by trained clinicians in a subset of patients.

Inclusion and exclusion criteria

Patients were included in the study if they underwent a surgical procedure with general anaesthesia between April 1, 2013 and December 10, 2018. The type of anaesthesia was extracted from the post-anaesthesia hand-off note documented by the anaesthesia provider at the end of the case. Cases were excluded if they had an ASA physical status score of 6 (indicating organ donors), were not discharged at the time of data analysis, or were aged less than 18 yr, and patients older than 89 yr had their age redacted (because of institutional restrictions on data security). A Consolidated Standards of Reporting Trials (CONSORT)[13] diagram is shown in Supplementary Figure S1.

Some patients, particularly those of highest risk, underwent more than one surgery during the course of their hospital

admission. In these cases all surgeries that met the above criteria were included. We performed a subsequent analysis to ensure that their inclusion would not unduly affect the results of the entire population. This analysis is shown and described in Supplementary Appendix S1.

Model input features

The model was created using a set of features including basic patient information such as age, sex, BMI, BP, and HR; laboratory tests frequently obtained before surgery, such as sodium, potassium, creatinine, and blood cell counts; and surgery-specific information such as the surgical procedure codes. In total, 58 preoperative features (including ASA status) were selected by clinicians' consensus (IH, EG) as potentially useful for predicting the outcome, and a full list is available in Supplementary Table S1. For all variables, only the most recent value before surgery was included.

In order to help elucidate the relative predictive value of different types of features, five models were created. Model 1 included all the input features, including the ASA physical status score. The Model 2 included all input features except the ASA physical status score—as this score would not be able to be fully automated before review by a trained anaesthesia provider. In order to overcome this limitation of automation, Model 3 included all of the input features with an automated surrogate for the ASA score. The details of the generation of this surrogate score can be found below. Models 4 and 5 were variations of Models 1 and 3; however, they excluded the timestamps of the preoperative laboratory results (relative to the admission start time), though they included the actual results themselves. As the time between a laboratory result and a surgery is not a marker of the patient illness, we excluded this information so that the model would not incorrectly weight the significance of this feature.

Comparison of model performance

In order to assess the performance of our models against currently used risk stratification systems, we also tested the performance of three 'baseline' models: a model containing only the ASA physical status score, a model containing only the POSPOM score,[4] and a model containing only the Charlson comorbidity score.[6] Using a model with a single feature such as these has the effect of producing the same result format as our more complex models and allows a direct comparison.

Data preprocessing

Data points greater than four standard deviations from the mean were removed as they were assumed to be erroneous outliers. Categorical features were converted into indicator variables, and the first variable was dropped. Thus, if a categorical variable takes on k values, only k-1 values are converted into indicator variables, because the kth variable becomes the reference value. The cohort was divided into a training dataset and a testing dataset by selecting all surgeries that occurred between April 1, 2013 and February 28, 2018 for training, and surgeries from March 1, 2018 and December 10, 2018 as the test set. Any patients that appeared in the test set were removed from the training set to prevent information leakage. Temporally splitting the cohort allows us to estimate model performance on future surgical cases. The training data features were rescaled to have a mean of 0 and a standard deviation of 1, and the test data were rescaled using the training data means and standard deviations. Missing data were imputed in the training and testing sets separately using the SoftImpute algorithm,[14] which leverages the similarity of groups of patients to estimate missing values. The SoftImpute algorithm was implemented by the fancyimpute Python package (version 0.2.0; Python Software Foundation, Beaverton, OR, USA), with a maximum of 200 iterations.

The number of inpatient mortalities was much smaller than the number of survivors, resulting in extreme class imbalance (2.01% mortality rate). To overcome this issue, the training set was oversampled using the Synthetic Minority Over-sampling Technique (SMOTE) algorithm,[15] implemented in the imblearn Python package (Python Software Foundation),[16] using three nearest neighbours and the 'baseline1' method to create a balanced class distribution. The testing set was not oversampled and, therefore, maintained the natural outcome frequency.

Generating a surrogate for ASA physical status

While the ASA status is a strong predictor of patient health status,[17–20] this classification requires a clinician to look through the patient's chart and subjectively determine the score, consuming valuable time and requiring clinical expertise. In order to balance the value of this score with the desire for automation, we sought to generate a similar metric using readily available data from the EMR—a surrogate ASA score. Recent works have similarly attempted to develop machine learning approaches to predict ASA scores.[21,22] However, these methods have difficulty differentiating ASA scores of 4 and 5 because of the low frequency of occurrence of 5 scores, and resort to either grouping classes together or ignoring patients with an ASA status of 5. The goal in our work is not to predict the ASA score, but to estimate a measure of general patient health for use as a feature in our model to predict in-hospital mortality, without needing the time-consuming clinician chart review.

Using the existing ASA physical status classification extracted from the EMR data, we trained a gradient boosted tree regression model to predict the ASA status of new patients using preoperative features unrelated to the surgery. The model was implemented using the XGBoost package[23] with 2000 trees and a maximum tree depth of 7. We used five-fold cross-validation to generate predictions. This surrogate-ASA value is a continuous number, unlike the actual ASA status which is limited to integers. We call it the 'ASA surrogate' score to distinguish it from an actual ASA score. This score is a continuous score of patient risk that uses the ASA score to supervise parameter learning in the model.

Model creation, training, and testing

We evaluated four different classification models: logistic regression, Elastic Net[24] logistic regression, random forests, and gradient boosted trees. Logistic regression is a statistical model that assumes a binary outcome can be predicted as a weighted combination of independent variables. The Elastic Net[24] logistic regression adds additional constraints to a linear prediction model by forcing the weights to be both small and sparse. A random forest classifier uses an ensemble of independently-trained decision trees, which classify data based on a series of binary questions about the values of particular features, to determine the most likely outcome

based on a majority vote. Like random forests, gradient boosted tree classifiers predict using an ensemble of decision trees, but instead of building each decision tree independently, the trees are created sequentially such that each new tree is fit to the residual error remaining after the previous step.

Model hyperparameters were chosen using five-fold cross-validation on the training dataset, where surgeries from the same patient were grouped together such that they appeared only in a training or testing fold, but not both. In five-fold cross-validation, the dataset is divided into five partitions, where four-fifths of the data are used to train the models and the remaining one-fifth are used as the testing set. This process is repeated such that each partition is used as a testing set only once and a training set four times. Cross-validation provides a better assessment of model performance by averaging metrics over multiple trials. Logistic regression classifiers were trained with both an L2 penalty and an ElasticNet[24] penalty, where alpha (regularisation constant) and the L1/L2 mixing parameter were set using five-fold cross-validation. The random forest classifiers were trained with 2000 estimators, Gini impurity as the splitting criterion, and no maximum tree depth was specified. The gradient boosted tree classifiers were trained using 2000 estimators and a maximum tree depth of 5. The logistic regression and random forest classifiers were implemented using Scikit-learn,[25] and the gradient boosted tree classifiers were implemented using the XGBoost package.[23] All performance metrics were calculated on the held-out test set using methods implemented by Scikit-learn.[25]

We generate confidence intervals (CIs) for the test set performance metrics using block bootstrapping of the predictions. As patients in the test set can undergo multiple surgeries, their risk predictions for each surgery are correlated. However, the general bootstrapping procedure typically samples cases randomly, and assumes each case is independent, but under this assumption, the correlation structure would be lost. Therefore, instead of randomly sampling cases, we randomly sample patients, and include all predictions in the bootstrap sample. This block bootstrap procedure is repeated 1000 times. For each bootstrap sample we calculate performance metrics; these metrics are then sorted, and we select the 25th and 975th values of the sorted list of metrics to determine the 95% CI.

As described above, we compared our method with the Charlson comorbidity index scores,[6] a well-known and proved existing method for prediction of risk of postoperative mortality, for each patient in the cohort. We used the updated weights as described by Quan and colleagues.[26] Scores were calculated using the R package *icd* (R Foundation, for Statistical Computing, Vienna, Austria) on all ICD10 codes associated with each surgery admission.

Another respected preoperative risk score is the POSPOM score.[4] While the POSPOM risk score was shown to have excellent discriminative ability, the features used in the model present an issue when trying to implement such a model in a medical centre that does not use the French classification for medical procedures (Classification Commune des Actes Médicaux [CCAM]). The POSPOM score groups CCAM surgery codes to 25 categories, where each category has an associated risk score as determined by their model. In order to replicate their model on our dataset, HCUP (or CPT) surgery codes must be mapped to CCAM codes. However, a mapping between HCUP (or CPT) codes to CCAM codes currently does not exist. Therefore, we created an approximate mapping between the case service group and the POSPOM surgical category. For each patient, we generated the ICD-based POSPOM score and the approximate surgical POSPOM score to compare the predictive capability of the POSPOM risk score to our method.

To determine which features were most important to the classification models, we examined the model weights for linear models, the feature (Gini) importance for the random forest models, and the feature weight (number of times a feature appears in a tree) for the gradient boosted tree models.

Table 1 Patient characteristics. Patient characteristics for the cohort used for training and testing models. Number of patients and percent of the cohort are shown. The selected surgical services represent the top four most frequent surgical services.

Property	Training data	Testing data
Patients, *n*	46 400	6494
Admissions, *n*	54 813	6853
Surgeries, *n*	58 916	7378
Average number of surgeries per patient	1.27	1.14
Average number of admissions per patient	1.18	1.06
Average number of surgeries per admission	1.07	1.08
Patients with more than one admission, *n* (%)	6400 (13.79)	328 (5.05)
Admissions with more than one surgery, *n* (%)	2817 (5.14)	351 (5.12)
Mortalities, *n* (%)	1243 (2.11)	124 (1.68)
Mean age	55.99 (18–89)	56.07 (18–89)
Female patients, *n* (%)	29 770 (50.53)	3680 (49.88)
ASA physical status 1, *n* (%)	3592 (6.10)	383 (5.19)
ASA physical status 2, *n* (%)	21 093 (35.80)	2412 (32.69)
ASA physical status 3, *n* (%)	27 395 (46.50)	3751 (50.84)
ASA physical status 4, *n* (%)	6432 (10.92)	779 (10.56)
ASA physical status 5, *n* (%)	404 (0.69)	53 (0.72)
Ronald Reagan operating room, *n* (%)	39 599 (67.21)	4935 (66.89)
Santa Monica operating room, *n* (%)	19 317 (32.79)	2443 (33.11)
Types of surgery, *n* (%)		
- Orthopaedics	9113 (15.47)	1083 (14.68)
- General surgery	7456 (12.66)	958 (12.98)
- Urology	7255 (12.31)	931 (12.62)
- Neurosurgery	6404 (10.87)	843 (11.43)
- Other	28 688 (48.69)	3563 (48.29)

Model calibration

A well-calibrated binary classification model outputs probabilities that are close to the true label (in our case, either a 1 for patients who die in the hospital, or 0 for survivors). Model calibration is often measured using the Brier score, which is the average squared distance between the predicted probability of the outcome and the true label; thus, a lower Brier

Table 2 Mortality prediction performance using area under the receiver operating characteristic (AUROC) curve. AUROC curve values for each model and each of the eight input feature sets on the held-out test set. Models with the highest AUROC are shown in bold. The mean value of the AUROC is shown, along with the 95% confidence interval (CI) from bootstrapping the test predictions 1000 times shown in parenthesis. When using the ASA status or the Charlson comorbidity score as the only input feature, the linear models (logistic regression, ElasticNet) outperform the non-linear models (random forest, XGBoost). However, for the other feature sets, the non-linear models outperform the linear models. In particular, the random forest has the highest AUROC compared with the other models. POSPOM, preoperative score to predict postoperative mortality; Preop, preoperative.

Model/AUC (95% CI)	Logistic regression	ElasticNet classifier	Random forest	XGBoost classifier
POSPOM	0.653 (0.602–0.705)	0.653 (0.602–0.705)	**0.660 (0.598–0.722)**	**0.660 (0.598–0.722)**
Charlson comorbidity	**0.742 (0.658–0.812)**	**0.742 (0.658–0.812)**	0.740 (0.658–0.811)	0.740 (0.658–0.811)
ASA status	**0.866 (0.829–0.897)**	**0.866 (0.829–0.897)**	0.855 (0.819–0.888)	0.855 (0.819–0.888)
Preop features	0.900 (0.863–0.931)	0.919 (0.891–0.942)	**0.925 (0.900–0.947)**	0.920 (0.894–0.944)
Preop+ASA status	0.913 (0.880–0.940)	0.924 (0.895–0.947)	**0.936 (0.915–0.956)**	0.922 (0.894–0.948)
Preop+surrogate-ASA	0.908 (0.872–0.937)	0.923 (0.895–0.946)	**0.931 (0.909–0.952)**	0.929 (0.907–0.948)
Preop (no time)+ASA status	0.919 (0.887–0.944)	0.932 (0.908–0.951)	**0.936 (0.917–0.955)**	0.923 (0.895–0.950)
Preop (no time)+surrogate-ASA	0.911 (0.877–0.941)	0.924 (0.898–0.948)	**0.932 (0.910–0.951)**	0.915 (0.887–0.940)

score usually indicates a better performing model. We used this metric to assess the calibration of our models.

Precision and recall calculations

Receiver operating characteristic (ROC) curves are very informative of binary classification prediction performance in general, as they illustrate how the performance changes as the discriminative threshold varies. However, precision-recall curves can be more informative when the classes are highly imbalanced.[27] ROC curves show the true positive rate (recall, sensitivity) as a function of the false positive rate (1-specificity), but for imbalanced datasets, the false positive rate can be misleading. The false positive rate is inversely related to the total number of negative samples and, therefore, a model that predicts a large number of false positives (relative to the number of true positives) may still achieve a small false positive rate. Therefore, precision (or positive predictive value) which penalises a model for a large number of false positives relative to the number of true positives, is a useful metric. Precision-recall curves show the precision of a classifier as a function of recall. An optimal model would reach the point in the upper-right corner of the precision-recall plot (i.e. perfect recall and perfect precision).

Integration of preoperative risk with postoperative risk

Previous work[11] has shown that integrating a measure of preoperative risk, such as the ASA score, into a postoperative mortality risk prediction model increases the model performance. We aimed to conduct a similar approach, but instead of using the ASA status as a measure of preoperative risk, we replaced it with the preoperative predictions from our model. First, we used the deep neural network architecture and features described by Lee and colleagues.[11] However, we replaced the ASA status feature with the preoperative risk scores which were generated using the random forest model, which was trained using the preoperative features and surrogate ASA scores as described in the previous section. Next, we trained the postoperative model using the training cohort used for preoperative risk prediction using five-fold cross-validation, where the intraoperative data was pre-processed in the same manner as described by Lee and colleagues.[11] We then compared the area under the ROC of the postoperative model trained using the ASA status and intraoperative features to the model that was trained using our preoperative risk score and intraoperative features. Lastly, in order to attempt to assess the degree to which risk changes during the intraoperative period, we compared on a per-patient basis the risk scores generated by our preoperative model with those generated by the incorporation of our results with the model described by Lee and colleagues.[11]

Results

Patient characteristics

The patient dataset contained 66 294 surgical records encompassing 52 894 patients. Patients were between the ages of 18 and 89 yr, with a mean age of 56 yr, and were classified as either inpatients, same-day admits, emergencies, or overnight recoveries. The frequency of mortality in the dataset was approximately 2.01%. An ASA status of 3 was the most common, comprising 47% of the dataset. Detailed information on patient characteristics can be found in Table 1.

Model performance

Area under the ROC curve

The area under the ROC curve values for each model are shown in Table 2 and ROC curves are shown for the random forest model in Figure 1a and for all models in Supplementary Figure S2. For all models except the ASA status alone, the random forest model produced the best results, although these differences often did not reach statistical significance. Models using the preoperative features have higher area under the ROC values (0.925, 95% CI 0.900–0.947) than the models that use the Charlson comorbidity score (0.742, 95% CI 0.658–0.812), the POSPOM score (0.660, 95% CI 0.598–0.722), or the ASA status (0.866, 95% CI 0.829–0.897) alone. Adding the surrogate ASA status values to the preoperative features did not improve the area under the ROC (0.931, 95% CI 0.909–0.952) as compared with the preoperative features alone (0.925, 95%

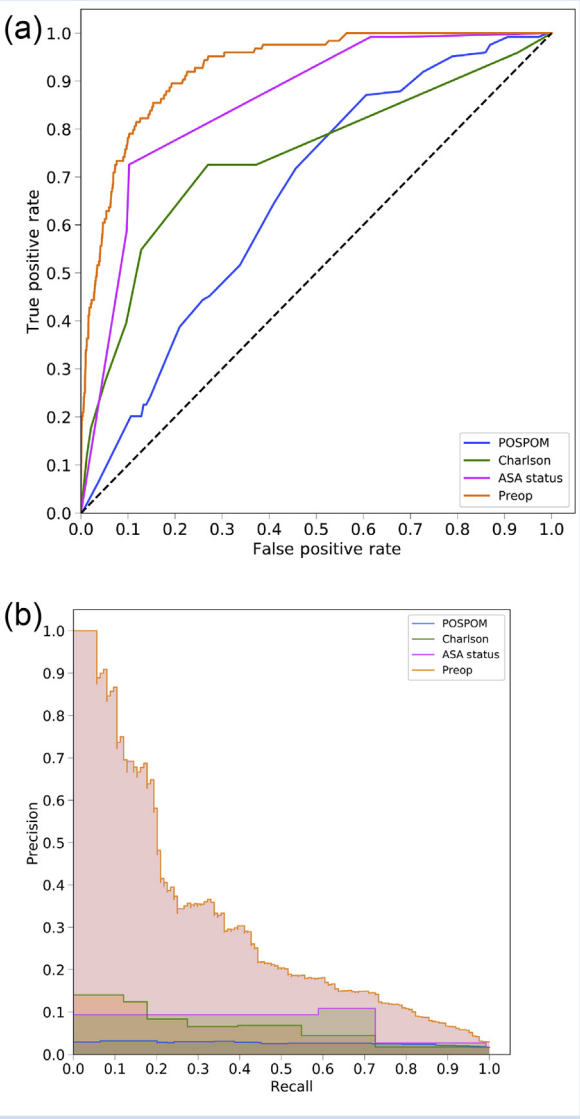

Fig 1. Receiver operating characteristic (ROC) and precision recall curves for the random forest model. Plots were generated using the predictions from the held-out test dataset. ROC curves (a) show the false positive rate on the X-axis and the true positive rate on the Y-axis. The optimal point is the upper-left corner. Precision-recall curves (b) show the recall on the X-axis and precision on the Y-axis. The optimal point is in the upper-right corner. POSPOM, preoperative score to predict postoperative mortality; Preop, preoperative.

CI 0.900–0.947). While adding the true ASA value assigned by anaesthesiologists to the preoperative features (0.936, 95%CI 0.915–0.956, $P>0.05$ as compared with preoperative features with surrogate ASA score) and reducing the preoperative feature set by removing variables indicating when the laboratory tests resulted ([0.932, 95% CI 0.910–0.951] and the preoperative features and true ASA status [0.936, 95% CI 0.917–0.955]) increased the AUC; these increases were not statistically significant ($P>0.05$). Table 3 contains the accuracy, F1 score, precision, recall, and specificity for all five random forest models.

Calibration

The non-linear models (random forest, XGBoost) had much lower (better) Brier scores compared with the linear models (logistic regression, ElasticNet). When using either the POSPOM score, the Charlson comorbidity score, or the ASA status as the only feature, the random forest and XGBoost classifiers had the lowest Brier score (0.098, 0.091, and 0.086, respectively). For the other five feature sets, the XGBoost models obtained the lowest Brier scores (0.015, 0.015, 0.016, 0.016, and 0.017, respectively). These data are shown in Supplementary Table S2.

Precision-recall

Using the random forest model, precision and recall curves for each of the sets of features are shown in Figure 1b and, for all models, in Supplementary Figure S3. Overall the various sets of preoperative features had better performance than the ASA score, the Charlson comorbidity score, and the POSPOM score.

Hospitals have limited resources and must decide how to allocate those resources. One option is to allocate prioritised care to individuals who are at the highest risk of adverse outcomes, particularly mortality. A hospital could choose to use the ASA score, the Charlson comorbidity score, the POSPOM score, or the score generated by the random forest model as an estimate of the risk. Our score is continuous and therefore has a definitive ordering of patients, while the ASA score, the Charlson comorbidity score, and the POSPOM score, being discrete, have random intra-score level ordering. To assess the effectiveness of the ordering based on the proposed score compared with the ASA score, the Charlson comorbidity score, and the POSPOM score, in Figure 2 we order the individuals by their risk of mortality and calculate the number of mortalities in our set of high-risk patients as we vary the size of the set. In other words, if we have a fixed set of resources such that we can allocate additional care to n patients, we would like to know how many of the n patients are true positives. While receiving prioritised care does not imply that a specific individual will not die, we argue that a population should have improved outcomes, as care levels are better matched to patients.

Feature importance

To determine the most important features for each of the models, we examined the feature weights of the linear models and feature importance of the non-linear models. In Supplementary Table S3, the feature importance for the random forest model is shown using four different sets of input features. For the feature sets that include laboratory result timestamps, many of the most important features are the laboratory result timestamp features for laboratories, such as brain natriuretic peptide (BNP) and bicarbonate. However, when these features are removed, the feature importance shifts to the laboratory results themselves, for example, albumin, international normalised ratio (INR), prothrombin time, haemoglobin, and total bilirubin. Surgery-specific features such as the patient class (inpatient, same-day admission) and the location of the patient in the hospital before surgery are also highly informative. Additionally, the ASA score is the most important feature in every model where it is contained.

Table 3 Performance metrics for random forest model. Random forest model performance metrics for predicting in-hospital mortality using different sets of features. Confidence intervals derived by bootstrapping the predictions using 1000 samples shown in parenthesis. Accuracy=(TP+TN)/(TP+TN+FP+FN). Precision=TP/(TP+FP). Recall=TP/(TP+FN). Specificity=TN/(TN+FP). F1 score=2/([1/Recall]+[1/Precision]). FN, false negatives; FP, false positives; Preop, preoperative; TN, true negatives; TP, true positives

Model	Accuracy	F1 score	Precision	Recall	Specificity
POSPOM	0.861 (0.851–0.869)	0.047 (0.021–0.078)	0.026 (0.012–0.045)	0.201 (0.097–0.318)	0.872 (0.864–0.881)
Charlson comorbidity	0.895 (0.885–0.904)	0.112 (0.064–0.165)	0.065 (0.037–0.098)	0.390 (0.240–0.538)	0.904 (0.895–0.913)
ASA status	0.897 (0.889–0.906)	0.160 (0.110–0.222)	0.093 (0.061–0.133)	0.587 (0.472–0.709)	0.903 (0.895–0.911)
Preop features	0.985 (0.981–0.988)	0.275 (0.115–0.446)	0.610 (0.333–0.814)	0.179 (0.069–0.315)	0.998 (0.997–0.999)
Preop features+ASA status	0.984 (0.980–0.988)	0.284 (0.119–0.464)	0.590 (0.333–0.810)	0.189 (0.074–0.329)	0.998 (0.997–0.999)
Preop+surrogate-ASA	0.984 (0.980–0.988)	0.280 (0.125–0.452)	0.541 (0.294–0.750)	0.191 (0.078–0.331)	0.997 (0.996–0.998)
Preop+ASA status, w/o lab times	0.982 (0.977–0.986)	0.302 (0.172–0.449)	0.420 (0.245–0.615)	0.239 (0.127–0.379)	0.994 (0.992–0.997)
Preop+surrogate-ASA status, w/o lab times	0.980 (0.976–0.985)	0.258 (0.127–0.412)	0.358 (0.180–0.551)	0.204 (0.094–0.342)	0.994 (0.992–0.996)

Integrating preoperative risk with postoperative risk

Replacing the ASA status with the preoperative risk predictions in the postoperative risk prediction model generated similar results to what were previously published by Lee and colleagues.[11] The postoperative risk model that was trained using the preoperative risk scores had an area under the ROC of 0.943 (95% CI 0.934–0.953), whereas the postoperative model trained using the ASA status had an area under the ROC of 0.935 (95% CI 0.926–0.947) ($P>0.05$). This is in line with the previously published results of this model.[11]

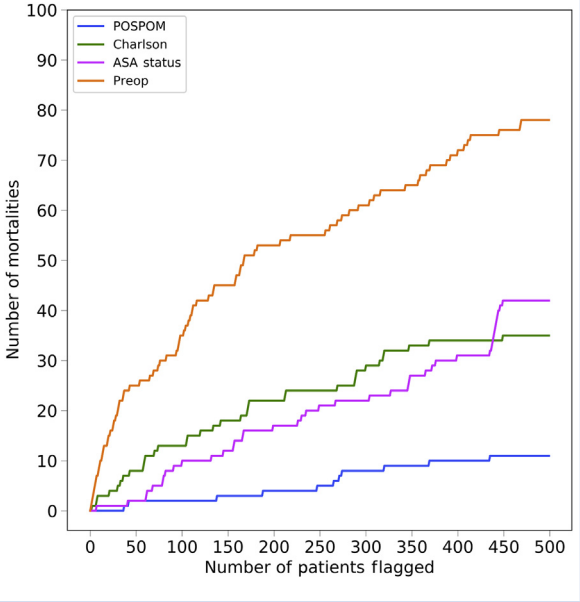

Fig 2. Number of in-hospital mortalities captured as a function of the number of patients flagged as high-risk. Using the random forest predicted probabilities for each set of features, surgeries were ranked from highest to lowest risk. For each feature set, we count the number of mortalities captured as we vary the number of high-risk patients flagged for additional resources. POSPOM, Preoperative Score to Predict Postoperative Mortality; Preop, preoperative.

In order to examine how mortality risk changes from immediately before surgery to after surgery, the pre- and postoperative risk scores for all patients were grouped by percentiles and the counts of each grouping are displayed in Figure 3a. For the majority of patients, we see a slight increase or decrease in their postoperative risk compared with the initial preoperative risk, as demonstrated by the colouring just above/below the diagonal line in Figure 3a. Figure 3b demonstrates the same plot but contains only those patients who eventually died during that admission. Most of these patients fall above the line indicating that their risk increased during the intraoperative period, and all patients in this cohort who had a preoperative risk below the 50th percentile had a postoperative risk that was substantially increased. Supplementary Tables S4a and b quantify this change in risk for the entire cohort and the in-hospital mortalities, respectively.

Discussion

We were able to successfully create a fully automated preoperative risk prediction score that can better predict in-hospital mortality than the ASA score, the POSPOM score, and the Charlson comorbidity score. In contrast to the ASA score, the POSPOM score, or the Charlson comorbidity scores, this score was built using purely objective clinical information that was readily available from the EMR before surgery and does not require a clinician's assistance for score calculation. Unlike previous models,[11] the results indicate that inclusion of the ASA score in the model did not improve the predictive ability. We were additionally able to integrate the results of our model into a previously developed postoperative risk prediction model and achieve a performance that was comparable to the use of the ASA physical status score in that model. Lastly, when using the preoperative and postoperative scores

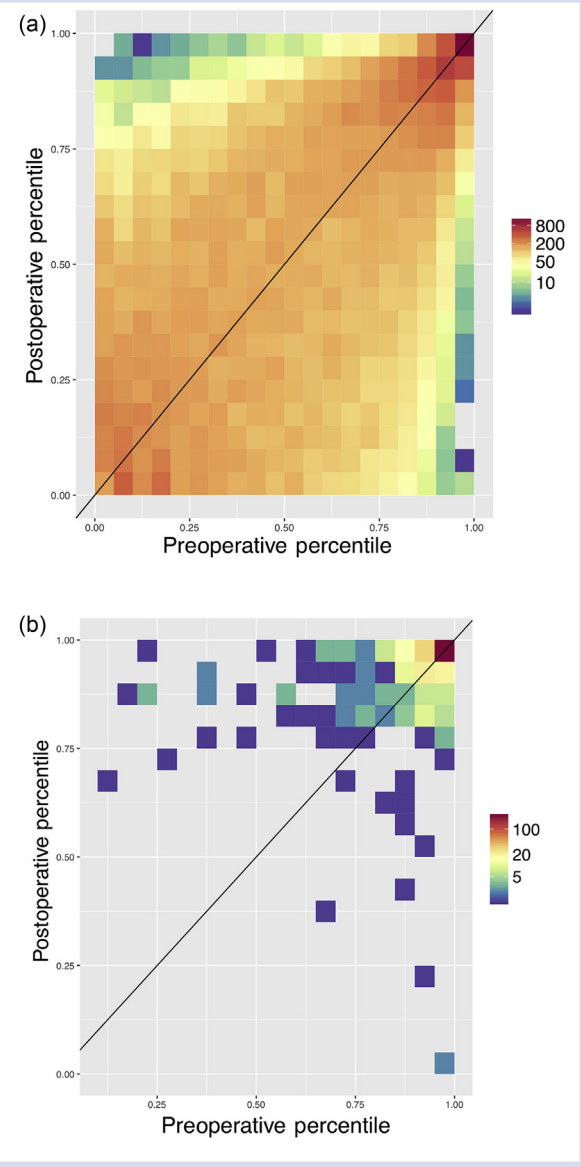

Fig 3. Heatmap of preoperative risk *vs* postoperative risk. Preoperative (X-axis) and postoperative (Y-axis) risk scores were binned by percentile, and the counts per bin visualised as a heatmap in log scale. Preoperative risk predictions were generated using the random forest model trained on the preoperative features, including laboratory times, and the surrogate-ASA status. In (a) all patients are displayed, and in (b) only the in-hospital mortalities are shown. Some 78% of patients who die and have a preoperative risk percentile below 95% have an increased postoperative risk percentile. This is substantially greater than the percent of matched patients from a null distribution who have an increased percentile.

together, we were able to demonstrate that, on a patient level, risk does change during the perioperative period—indicating that choices made in the operating room may have profound implications for our patients.

The challenge of perioperative risk stratification is certainly not new. In fact, the presence of so many varied risk scores (ASA score, Charlson comorbidity index,[6] POSPOM,[4] risk quantification index (RQI),[28] NSQIP risk calculator[8]) speaks to the importance with which clinicians view this problem. A major limitation of many of these models has been that they either rely on data not available at the time of surgery (i.e. ICD codes), or they require an anaesthesiologist to review the chart (those that contain the ASA score). Thus, the creation of a model that can be fully automated and perform better than these models implies that it may have broad applicability. Of note, in this study, the non-linear machine learning models outperformed logistic regression both regarding AUC and calibration (Brier score). This is different than what has been shown in other work[11] where the logistic regression performed similarly to non-linear machine learning approaches.

As demonstrated in many previous studies,[17–20] the ASA score itself remains a good predictor of postoperative outcomes. This is likely because the ASA score is essentially a predictor generated by the most advanced neural network known—the human brain. However, the ASA score alone did not perform as well as our baseline model. The discrepancy may have several possible explanations. One possibility is that the introduction of the EMR has led to an explosion of information, making it challenging for a clinician to consume everything. It would be essentially impossible for an anaesthesiologist to review every note, laboratory result, and pathology report before surgery. A second possible explanation is that the ASA score is not a predictor of mortality *per se*, but rather a marker of overall patient complexity. Thus, a score that is designed to predict a specific complication, such as mortality, will perform better than a measure of overall patient complexity. Regardless of the exact reason, we believe this highlights the advantage of an automated scoring system such as this—not as a replacement for physicians—but as a tool to help them better focus their efforts on those patients most likely to benefit.

Another advantage of an automated model such as this is that it allows for the continuous recalculation of risk longitudinally over time. As shown in Figure 3, most patients have either a minor increase or decrease in risk in the time from before to after surgery and, unsurprisingly, in patients who eventually die, that risk tends to increase. Given the challenges of continually monitoring the risk of all patients in the hospital, advanced analytical models, such as the one demonstrated in this manuscript, have great potential to act as early warning systems alerting clinicians to sudden changes in risk profiles and facilitating the use of rapid response teams.

More importantly, the frequency with which risk changed substantially during the operative period highlights the effect to which intraoperative interventions may have implications far beyond the operating room. Multiple specific interventions, including the avoidance of intraoperative hypotension and hypothermia, have been shown to have effects on longer-term outcomes, and currently enhanced recovery after surgery pathways have promoted the standardisation of intraoperative interventions. We believe that our findings should add to the evidence that a well-prescribed anaesthetic plan may be of significant long-term benefit to patient outcomes.

One potential promise of the use of machine learning in medicine is the ability to leverage these models in order to better understand what features are truly driving outcomes. In an effort to better understand this, we extracted the weights of the features in both the linear and non-linear models. Removing some features, specifically the relative

time of laboratory tests, actually improved the results of our model (though not to a level of statistical significance). This could potentially be caused by multiple correlated features tagging an underlying cause, and the correlation introduces noise in the model as the importance is distributed among multiple features rather than focused on a single feature. In theory, a machine learning model should be able to remove these features by setting a coefficient to zero. However, in practice, this may not always be the case—as illustrated here, where we force this behaviour by manually removing the features from the model. We believe that this finding highlights the importance of having collaborative relationships between experts in machine learning and clinicians who are able to help guide which features to include in a model. Simply entering large amounts of data from an EMR, without proper clinical context, is unlikely to create the most effective or efficient models.

There are several key limitations of this study. The most significant is the low frequency of the outcome in question—in-hospital mortality. The incidence of mortality in the testing set was less than 2%—implying that a model that blindly reports 'survives' every time will have an accuracy greater than 98%. Predicting such a rare outcome makes it highly challenging to produce results with very high precision. Nonetheless, the models presented in this paper do outperform other models currently in use, as measured by area under the ROC curve, and had precision-recall curves that were superior to the ASA score, POSPOM score, or Charlson comorbidity scores alone. Secondly, the large amount of missing data in the EMR makes imputation a complex task, in particular because the data are not necessarily missing at random. Many of the missing values are attributable to systematic reasons, such as forgoing a set of laboratory tests because the clinician believes the patient's laboratory values are relatively normal. In fact, creating optimal imputation algorithms is whole field of work on its own and suboptimal imputation algorithms will reduce the prediction performance. However, given the sparsity of the data, some form of imputation is necessary and our choice of imputation algorithm, while not optimal, is better than a trivial method such as mean imputation (see Supplementary Figs S4 and S5a–c). In fact, our algorithm performed better than the same algorithm using mean imputation (see Supplementary Fig. S6a and b). We believe that the overall strong performance of our models, despite these limitations, indicates the value of machine learning in predicting postoperative outcomes. Thirdly, the data used here are from a single large academic medical centre. Thus, it is possible, though unlikely, that this model will not perform similarly at another institution. More likely is that the model would require recalibration in order to be transferred from one institution to another. However, with such a recalibration, the exact weights of the various features might change. One last limitation lies not necessarily with the study itself but with the overall landscape of EMR data. While the promises of fully automated risk scores are great, the reality remains that most institutions still have trouble accessing the data stored in the EMRs. Thus, in order to truly automate processes such as these, robust data interoperability standards (such as Fast Healthcare Interoperability Resources) will be needed in order to allow access to data.

The promise of using machine learning techniques in healthcare is great. In this work we have presented a novel set of easily accessible (via EMR data) preoperative features that are combined in a machine learning model for predicting in-hospital post-surgical mortality, which outperforms current clinical risk scores; however, a model that incorporates both physician judgement (via the ASA score) with machine learning produces the best results. We have also shown that the risk of in-hospital mortality changes over time. It is our hope and expectation that the next few years will produce a plethora of research leveraging data obtained during routine patient care to improve care delivery models and outcomes for all of our patients.

Declarations of interest

MC is co-owner of US patent serial no. 61/432,081 for a closed-loop fluid administration system based on the dynamic predictors of fluid responsiveness which has been licensed to Edwards Lifesciences. MC is a consultant for Edwards Lifesciences (Irvine, CA, USA), Medtronic (Boulder, CO, USA), Masimo Corp. (Irvine, CA, USA). MC has received research support from Edwards Lifesciences through his Department and NIH R01 GM117622—Machine learning of physiological variables to predict diagnose and treat cardiorespiratory instability and NIH R01 NR013912—Predicting Patient Instability Noninvasively for Nursing Care—Two (PPINNC-2). IH is the founder and President of Clarity Healthcare Analytics Inc. a company that assists hospitals with extracting and using data from their electronic medical records. IH also receives research funding from Merck Pharmaceuticals. EG is founder and Secretary of Clarity Healthcare Analytics Inc. a company that assists hospitals with extracting and using data from their electronic medical records. No funding bodies had any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

Drafted the manuscript: BLH, RB.
Performed the experiments: BLH, RB, NR, CL, MC, RJ, BJ, PB.
Analysed results: BLH, RB.
Drafted the manuscript: RB.
Clinical evaluation of results: EG.
Set up infrastructure, defined clinical features: EG, IH.
Revised the manuscript: LOL, UM.
Proposed the problem: AM.
Supervised the statistical analysis: SS, EH.
Supervised the acquisition of clinical data: IH.

Funding

National Science Foundation (grant number 1705197) to BLH, NR, and EH. National Institute of Mental Health (award number K99MH116115) to LOL. National Institute of Health (grant numbers R00GM111744, R35GM125055), National Science Foundation (grant number III-1705121), an Alfred P. Sloan Research Fellowship, and Okawa Foundation to SS. National Institute of Neurological Disorders and Stroke of the National Institute of Health (award number T32NS048004) and a UCLA QCB Collaboratory Postdoctoral Fellowship directed by Matteo Pellegrini to RB. Bial Foundation to UM. National Science Foundation Graduate Research Fellowship Program (grant number DGE-1650604) to BJ.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bja.2019.07.030.

References

1. Pearse RM, Harrison DA, James P, et al. Identification and characterisation of the high-risk surgical population in the United Kingdom. *Crit Care* 2006; **10**: R81
2. Kang D-H, Kim Y-J, Kim S-H, et al. Early surgery versus conventional treatment for infective endocarditis. *N Engl J Med* 2012; **366**: 2466–73
3. Leeds IL, Truta B, Parian AM, et al. Early surgical intervention for acute ulcerative colitis is associated with improved postoperative outcomes. *J Gastrointest Surg* 2017; **21**: 1675–82
4. Le Manach Y, Collins G, Rodseth R, et al. Preoperative score to predict postoperative mortality (POSPOM). *Anesthesiology* 2016; **124**: 570–9
5. Sessler DI, Sigl JC, Manberg PJ, Kelley SD, Schubert A, Chamoun NG. Broadly applicable risk stratification system for predicting duration of hospitalization and mortality. *Anesthesiology* 2010; **113**: 1026–37
6. Charlson M, Szatrowski TP, Peterson J, Gold J. Validation of a combined comorbidity index. *J Clin Epidemiol* 1994; **47**: 1245–51
7. Sigakis MJG, Leffert LR, Mirzakhani H, et al. The validity of discharge billing codes reflecting severe maternal morbidity. *Anesth Analg* 2016; **123**: 731–8
8. Bilimoria KY, Liu Y, Paruch JL, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. *J Am Coll Surg* 2013; **217**: 833–42
9. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning for electronic health records. *NPJ Digit Med* 2018; **1**. http://arxiv.org/abs/1801.07860
10. Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. *NPJ Digit Med* 2018; **1**: 5
11. Lee CK, Hofer I, Gabel E, Baldi P, Cannesson M. Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality. *Anesthesiology* 2018; **129**: 649–62
12. Hofer IS, Gabel E, Pfeffer M, Mahbouba M, Mahajan A. A systematic approach to creation of a perioperative data warehouse. *Anesth Analg* 2016; **122**: 1880–4
13. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001; **357**: 1191–4
14. Mazumder R, Hastie T, Edu H, Tibshirani R, Edu T. Spectral regularization algorithms for learning large incomplete matrices. *J Mach Learn Res* 2010; **11**: 2287–322
15. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *J Artif Int Res* 2002; **16**: 321–57
16. Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res* 2017; **18**: 1–5
17. Wolters U, Wolf T, Stützer H, Schröder T. ASA classification and perioperative variables as predictors of postoperative outcome. *Br J Anaesth* 1996; **77**: 217–22
18. Daabiss M. American Society of Anaesthesiologists physical status classification. *Indian J Anaesth* 2011; **55**: 111–5
19. Vacanti CJ, Van Houten RJ, Hill RC. A statistical analysis of the relationship of physical status to postoperative mortality in 68,368 cases. *Anesth Analg* 1970; **49**: 564–6
20. Hackett NJ, De Oliveira GS, Jain UK, Kim JYS. ASA class is a reliable independent predictor of medical complications and mortality following surgery. *Int J Surg* 2015; **18**: 184–90
21. El Amine Lazouni M, El Habib Daho M, Settouti N, Chikh MA, Mahmoudi S. Machine learning tool for automatic ASA detection. In: Amine A, Otmane AM, Bellatreche L, editors. Cham: Springer International Publishing; 2013. p. 9–16
22. Zhang L, Fabbri D, Wanderer JP. Data-driven system for perioperative acuity prediction. *AMIA* 2016
23. Chen T, Guestrin C. *XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco, CA* 2016. p. 785–94
24. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol* 2005; **67**: 301–20
25. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; **12**: 2825–30
26. Quan H, Li B, Couris CM, et al. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am J Epidemiol* 2011; **173**: 676–82
27. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015; **10**, e0118432
28. Dalton JE, Kurz A, Turan A, Mascha EJ, Sessler DI SL. Development and validation of a risk quantification index for 30-day postoperative mortality and morbidity in noncardiac surgical patients. *Anesthesiology* 2011; **114**: 1336–44

Handling editor: P.S. Myles