OPEN

Utilizing Machine Learning Methods for Preoperative Prediction of Postsurgical Mortality and Intensive Care Unit Admission

Calvin J. Chiew, MBSS, MPH, Nan Liu, PhD,†‡ Ting Hway Wong, MB, BChir, MPH,*§ Yilin E. Sim, MBBS, MMed,¶ and Hairil R. Abdullah, MBBS, MMed¶*

Objective: To compare the performance of machine learning models against the traditionally derived Combined Assessment of Risk Encountered in Surgery (CARES) model and the American Society of Anaesthesiologists-Physical Status (ASA-PS) in the prediction of 30-day postsurgical mortality and need for intensive care unit (ICU) stay >24 hours.

Background: Prediction of surgical risk preoperatively is important for clinical shared decision-making and planning of health resources such as ICU beds. The current growth of electronic medical records coupled with machine learning presents an opportunity to improve the performance of established risk models.

Methods: All patients aged 18 years and above who underwent noncardiac and nonneurological surgery at Singapore General Hospital (SGH) between 1 January 2012 and 31 October 2016 were included. Patient demographics, comorbidities, preoperative laboratory results, and surgery details were obtained from their electronic medical records. Seventy percent of the observations were randomly selected for training, leaving 30% for testing. Baseline models were CARES and ASA-PS. Candidate models were trained using random forest, adaptive boosting, gradient boosting, and support vector machine. Models were evaluated on area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC).

Results: A total of 90,785 patients were included, of whom 539 (0.6%) died within 30 days and 1264 (1.4%) required ICU admission >24 hours postoperatively. Baseline models achieved high AUROCs despite poor sensitivities by predicting all negative in a predominantly negative dataset. Gradient boosting was the best performing model with AUPRCs of 0.23 and 0.38 for mortality and ICU admission outcomes respectively.

Conclusions: Machine learning can be used to improve surgical risk prediction compared to traditional risk calculators. AUPRC should be used to evaluate model predictive performance instead of AUROC when the dataset is imbalanced.

Keywords: forecasting, machine learning, postoperative complications, preoperative care, risk

(*Ann Surg* 2020;272:1133–1139)

From the *Health Services Research Unit, Division of Medicine, Singapore General Hospital, Singapore; †Health Services Research Centre, Singapore Health Services, Singapore; ‡Health Services and Systems Research, Duke-NUS Medical School, National University of Singapore, Singapore; §Department of General Surgery, Singapore General Hospital, Singapore; and ¶Department of Anaesthesiology, Singapore General Hospital, Singapore.

The authors disclose no conflicts of interest.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.annalsofsurgery.com).

Reprints: Hairil R. Abdullah, MBBS, MMed, Department of Anaesthesiology, Singapore General Hospital, Academia Level 5, 20 College Road, Singapore 169856. E-mail: hairil.rizal.abdullah@singhealth.com.sg.

ISSN: 0003-4932/20/27206-1133
DOI: 10.1097/SLA.0000000000003297

About 250 million surgeries are performed worldwide each year, and this number is increasing rapidly.[1] As access to surgery improves, the number of patients with postoperative complications will also increase.[2,3] Previous studies demonstrated that a large proportion of postoperative mortality occurs in a small, distinct group of patients with high-risk characteristics, yet less than 15% from this group were admitted to intensive care units (ICU) postoperatively.[4,5] In the preoperative assessment of a surgical patient, it is prudent to counsel the patient on the risks of postoperative mortality and need for critical care monitoring after surgery. Therefore, accurate preoperative prediction of surgical risks is important for clinical shared decision-making and for guiding the allocation of health resources such as ICU beds.

A number of risk stratification tools have been developed for this purpose, such as the American Society of Anaesthesiologists-Physical Status (ASA-PS), Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity (POSSUM), Surgical Outcome Risk Tool (SORT), and American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP). However, these tools have their own limitations, for example wide interuser variability (ASA-PS),[6] need for data which are not typically available during the preoperative period (POSSUM),[7,8] lack of validation outside the derived population's region (SORT, ACS-NSQIP) and complexity of the model itself (ACS-NSQIP).

We previously described a simple 9-variable surgical risk calculator, the Combined Assessment of Risk Encountered in Surgery (CARES), to predict both 30-day postoperative mortality and need for ICU stay >24 hours, using routinely available preoperative clinical and laboratory variables.[9] Unlike other existing tools, CARES was developed in an Asian majority population. It was also the first surgical risk stratification tool to incorporate the use of red-cell distribution width (RDW), a readily available hematological biomarker which has been shown to be associated with postoperative mortality independent of anaemia.[10–12] The model was developed by assigning rank scores to the odds ratios obtained from stepwise multivariate logistic regression. CARES achieved better predictive performance for both outcomes when compared against ASA-PS and ASA-PS with propensity scoring in terms of AUROC.[9]

In the development of CARES and other surgical risk calculators, constraints in analytic methods and concerns over usability have generally confined models to a small set of variables and to scoring systems that are easily calculated. Machine learning techniques capable of harnessing the large number of variables that are already captured in electronic health records (EHR) may offer better predictive performance and facilitate automation and deployment within clinical decision support systems.[13]

In this study, we aimed to demonstrate a local, EHR data-driven, machine learning approach for preoperative prediction of postsurgical mortality and ICU stay. We hypothesize that a machine learning model for surgical risk prediction would outperform traditional risk calculators, such as ASA-PS and CARES.

METHODS

Ethics approval for the study was obtained from SingHealth's Centralised Institutional Review Board (CIRB, Reference Number 2014/651/D), with waiver of patient consent. We conducted a retrospective analysis on all patients aged 18 years and above who underwent surgery under general or regional anesthesia at Singapore General Hospital (SGH) between January 1, 2012 and October 31, 2016. SGH is a 1700-bedded tertiary academic hospital in Singapore. We excluded patients who underwent cardiac surgery, neurosurgery, transplant and burns surgery, and only included the index surgery for patients who underwent multiple surgeries during the study period.

Data was sourced from the hospital's EHR system (Sunrise Clinical Manager, Allscripts, IL). Mortality data on the system was synchronized with Singapore's National Electronic Health Records (NEHR), ensuring a near complete follow-up. We collected data routinely captured during the preoperative anesthesia assessment visit, namely patient demographics, comorbidities, preoperative laboratory test results, and surgery details.

Patient demographics included age, sex, ethnicity, height, weight, and body mass index (BMI). Comorbidities were recorded as per the Revised Cardiac Risk Index (RCRI),[14] which consists of cerebrovascular accident, ischemic heart disease, congestive heart failure, diabetes mellitus (DM) on insulin and chronic kidney disease (CKD), as well as the ASA-PS class.[15] CKD, if present, was graded based on the estimated glomerular filtration rate (eGFR) by the Modification of Diet in Renal Disease (MDRD) equation according to the 2012 KDIGO guidelines.[16] Preoperative laboratory tests considered were the latest full blood count (FBC) and renal panel (RP) taken from 90 days before the surgery, and up to the day of surgery but before the start time of surgery. Laboratory test results included hemoglobin, platelet, mean corpuscular volume, red blood cell distribution width (RDW), hematocrit, activated partial thromboplastin time (aPTT), prothrombin time (PT), and serum creatinine. Surgery details included its description, surgical risk, priority (emergency or elective), type (inpatient or day surgery), department (surgical specialty), and anesthesia type (general or regional anesthesia). Surgical risk classification (low, moderate, or high) was based on the 2014 ESC/ESA guidelines.[17] Finally, we also recorded the number of transfusions the patient received preoperatively within 30 days if any.

Outcomes of interest were 30-day postsurgical mortality and ICU stay >24 hours. For each outcome, we compared the predictors between patients who met the outcome and patients who did not, using the Mann–Whitney *U* test for continuous variables and chi-square test for categorical variables.

Prior to modeling, all variables were scaled, and missing values were imputed using median. Free-text surgical descriptions were processed using bag-of-words analysis. We ignored stop words (such as ''the,'' ''and'') and only considered words that make up at least 0.5% of the corpus, which gave us 265 words. Each surgical description was transformed into 265 binary variables that indicated whether the description contained the particular word or not. Seventy percent of the observations were used to train the models, leaving 30% as a test set for subsequent model evaluation (the same training and testing cohorts used to develop CARES). Candidate models were trained using random forest, adaptive boosting, gradient boosting, and support vector machine algorithms. Baseline comparators were the 9-variable final combined CARES model and ASA-PS score.

To address the issue of class imbalance, training set patients without outcome were split into 10 roughly equal subsets. Each subset was then paired with all the training patients with outcome to create an ensemble of 10 datasets which were more balanced. A classifier was trained on each subset, producing 10 classifiers, whose predictions were then combined by majority vote. Additionally, class weighting was applied where relevant. Parameter tuning was performed via grid search 5-fold cross-validation with the aim of optimizing F1 score.

We used each model to predict on the test set and calculated its specificity, sensitivity (recall), positive predictive value (PPV) (precision), as well as F1 score, which is the weighted mean of precision and recall. (The formula for deriving F1 is 2 $*$ precision $*$ recall / [precision + recall].) For each model, we also generated a receiver operating characteristic curve (ROC) and precision-recall curve (PRC), and calculated the areas under both curves (AUROC and AUPRC respectively). We used F1 score and AUPRC as our main performance metrics for model comparison as they were more informative for evaluating binary classifiers on imbalanced datasets.[18]

To better understand how the GB model worked, we also visualized feature importance in terms of the total decrease in node impurity (indicated by Gini index) due to branching over a given predictor, averaged over all trees and aggregated across all classifiers in the ensemble.

Univariate statistical analysis was carried out in Stata version 13 (StataCorp 2013, College Station, TX). Machine learning models were developed in Python 3.6 (Python Software Foundation, Wilmington, DE) using the scikit-learn library.[19]

RESULTS

Supplemental Figure S1, http://links.lww.com/SLA/B609 shows the cohort selection process. A total of 90,785 patients were included in the study, of whom 539 (0.6%) died within 30 days and 1264 (1.4%) required ICU stay >24 hours postoperatively. 42,077 (46.3%) of them were male, with median age of 54 years (interquartile range [IQR] 39–65 yrs). Table 1 compares the patient demographics, comorbidities, laboratory test results, and surgery details of those who did and did not meet the outcome for each of the 2 outcomes. In addition, the last column indicates the percentage of missing data for each of the variables collected, which was between 0% and 43%.

All predictors were significantly different between patients who died and patients who did not, except for anesthesia type. Similarly, all predictors were significantly different between patients who were admitted to ICU and patients who were not, except for ethnicity, height, and mean corpuscular volume. In general, patients who met either of the adverse outcomes were older, more likely to be male, and had lower weight and BMI compared with patients who did not meet the corresponding outcome. They were also more likely to have the RCRI comorbidities, higher grade of CKD, and ASA-PS class. In terms of laboratory tests, they had lower hemoglobin, hematocrit, platelet count and eGFR, as well as higher RDW, aPTT, PT, and creatinine. Finally, they were more likely to undergo emergency surgeries and surgeries categorised as higher risk.

Figures 1 and 2 show the ROCs and PRCs respectively of the baseline and candidate models for the 30-day mortality outcome. Table 2 summarizes their specificities, sensitivities, PPVs, F1 scores, AUROCs, and AUPRCs. All models were able to achieve high AUROCs of between 0.89 and 0.96, including the baseline CARES model despite its poor sensitivity of 0.00. It achieved this by predicting all negative in a predominantly negative dataset, as evidenced by its specificity of 1.00 and PPV of 0.00. Gradient boosting (GB) was the best performing model with a F1 score of 0.28 and AUPRC of 0.23. Compared to the baseline CARES model (AUPRC 0.15), this translated to an improvement of 50% in sensitivity with only 2% loss in specificity.

TABLE 1. Summary of Predictor Variables by Presence of Outcome

Variable	30-d Mortality			ICU Stay >24 h			
	No (n = 90,246)	Yes (n = 539)	*P* Value	No (n = 89,521)	Yes (n = 1264)	*P* Value	% Missing
Patient demographics							
Age (yrs)	54 (38–65)	71 (60–79)	<0.001	54 (38–65)	65 (54–74)	<0.001	0%
Male sex	41,767 (46%)	310 (58%)	<0.001	41,292 (46%)	785 (62%)	<0.001	0%
Ethnicity			0.02			0.95	0%
Chinese	64,465 (71%)	396 (73%)		63,949 (71%)	912 (72%)		
Malay	8914 (10%)	65 (12%)		8858 (10%)	121 (10%)		
Indian	7969 (9%)	43 (8%)		7904 (9%)	108 (9%)		
Others	8892 (10%)	35 (6%)		8804 (10%)	123 (10%)		
Height (cm)	161 (155–168)	160 (152–166)	<0.001	161 (155–168)	161 (155–167)	0.47	18%
Weight (kg)	64 (55–74)	58 (48–69)	<0.001	64 (55–74)	62 (53–72)	<0.001	15%
BMI (kg/m^2)	25 (22–28)	23 (20–27)	<0.001	25 (22–28)	24 (21–27)	<0.001	18%
Comorbidities							
CVA	1501 (2%)	42 (13%)	<0.001	1488 (2%)	55 (7%)	<0.001	31%
IHD	4119 (7%)	126 (39%)	<0.001	4054 (7%)	191 (24%)	<0.001	31%
CHF	752 (1%)	35 (10%)	<0.001	727 (1%)	60 (7%)	<0.001	29%
DM on insulin	1964 (3%)	39 (12%)	<0.001	1945 (3%)	58 (7%)	<0.001	30%
CKD			<0.001			<0.001	12%
Grade 1	47,805 (60%)	143 (27%)		47,497 (60%)	451 (38%)		
Grade 2	23,532 (30%)	103 (20%)		23,342 (30%)	293 (25%)		
Grade 3	5007 (6%)	107 (20%)		4884 (6%)	230 (20%)		
Grade 4	1116 (1%)	83 (16%)		1083 (1%)	116 (10%)		
Grade 5	1968 (2%)	91 (17%)		1970 (3%)	89 (8%)		
ASA-PS class			<0.001			<0.001	5%
I	22,047 (26%)	0 (0%)		22,009 (26%)	38 (3%)		
II	49,362 (58%)	73 (16%)		49,105 (58%)	330 (29%)		
III	13,171 (15%)	234 (51%)		12,890 (15%)	515 (45%)		
IV–VI	926 (1%)	153 (33%)		818 (1%)	261 (23%)		
Laboratory tests							
Full blood count							
Hemoglobin (g/dL)	13 (12–15)	11 (9–12)	<0.001	13 (12–15)	12 (10–14)	<0.001	5%
MCV (fL)	89 (85–92)	90 (85–94)	0.002	89 (85–92)	89 (85–92)	0.72	8%
RDW (%)	13 (13–14)	15 (14–17)	<0.001	13 (13–14)	14 (13–16)	<0.001	8%
Hematocrit (%)	40 (37–43)	32 (28–37)	<0.001	40 (37–43)	36 (31–41)	<0.001	8%
Platelet (x10^9/L)	261 (219–313)	229 (158–321)	<0.001	261 (219–313)	244 (184–321)	<0.001	8%
aPTT (s)	28 (26–30)	31 (28–36)	<0.001	28 (26–30)	29 (27–32)	<0.001	43%
PT (s)	10 (10–11)	12 (11–13)	<0.001	10 (10–11)	11 (10–12)	<0.001	43%
Renal panel							
Creatinine (umol/L)	70 (57–86)	109 (67–233)	<0.001	70 (57–86)	86 (63–133)	<0.001	13%
eGFR (mL/min/1.73 m^2)	96 (79–114)	54 (21–95)	<0.001	96 (79–114)	77 (45–108)	<0.001	12%
Surgery details							
Surgical risk			<0.001				0%
Low	47,901 (53%)	148 (27%)		47,847 (53%)	202 (16%)		
Moderate	38,712 (43%)	302 (56%)		38,369 (43%)	645 (51%)		
High	3633 (4%)	89 (17%)		3305 (4%)	417 (33%)		
Priority of surgery			<0.001			<0.001	0%
Elective	72,148 (80%)	183 (34%)		71,655 (80%)	676 (53%)		
Emergency	18,098 (20%)	356 (66%)		17,866 (20%)	588 (47%)		
Anesthesia type			0.30			<0.001	0%
GA	75,997 (84%)	445 (83%)		75,234 (84%)	14,287 (96%)		
RA	14,249 (16%)	94 (17%)		1208 (1%)	56 (4%)		
No. of preoperative blood transfusions within 30 d	0 (0–0)	0 (0–0)	<0.001	0 (0–0)	0 (0–1)	<0.001	0%

For continuous variables, data is presented in medians and interquartile ranges. Mann–Whitney *U* test was used to test for differences.

For categorical variables, data is presented in frequencies and percentages. Chi square test was used to test for association.

aPTT indicates activated partial thromboplastin time; CHF, congestive heart failure; CVA, cerebrovascular accident; DM, diabetes mellitus; eGFR, estimated glomerular filtration rate; GA, general anesthesia; IHD, ischemic heart disease; MCV, mean corpuscular volume; PT, prothrombin time; RA, regional anesthesia.

Figures 3 and 4 show the ROCs and PRCs respectively of the baseline and candidate models for the ICU admission outcome. Table 3 summarizes their evaluation metrics. Similar to the results for mortality, the baseline CARES model obtained relatively high AUROC of 0.84 despite poor sensitivity of 0.00. Candidate models achieved higher F1 scores and AUPRCs across the board than for the mortality outcome. The best performing model was again GB with F1 score of 0.36 and AUPRC of 0.38. Compared to the baseline CARES model (AUPRC 0.18), this translated to an improvement of 58% in sensitivity with only 3% loss in specificity.

Supplemental Figures S2, http://links.lww.com/SLA/B609 and S3, http://links.lww.com/SLA/B609 show the most predictive features in the GB ensemble and their relative importance for the 30-day mortality and ICU admission outcomes respectively. Top

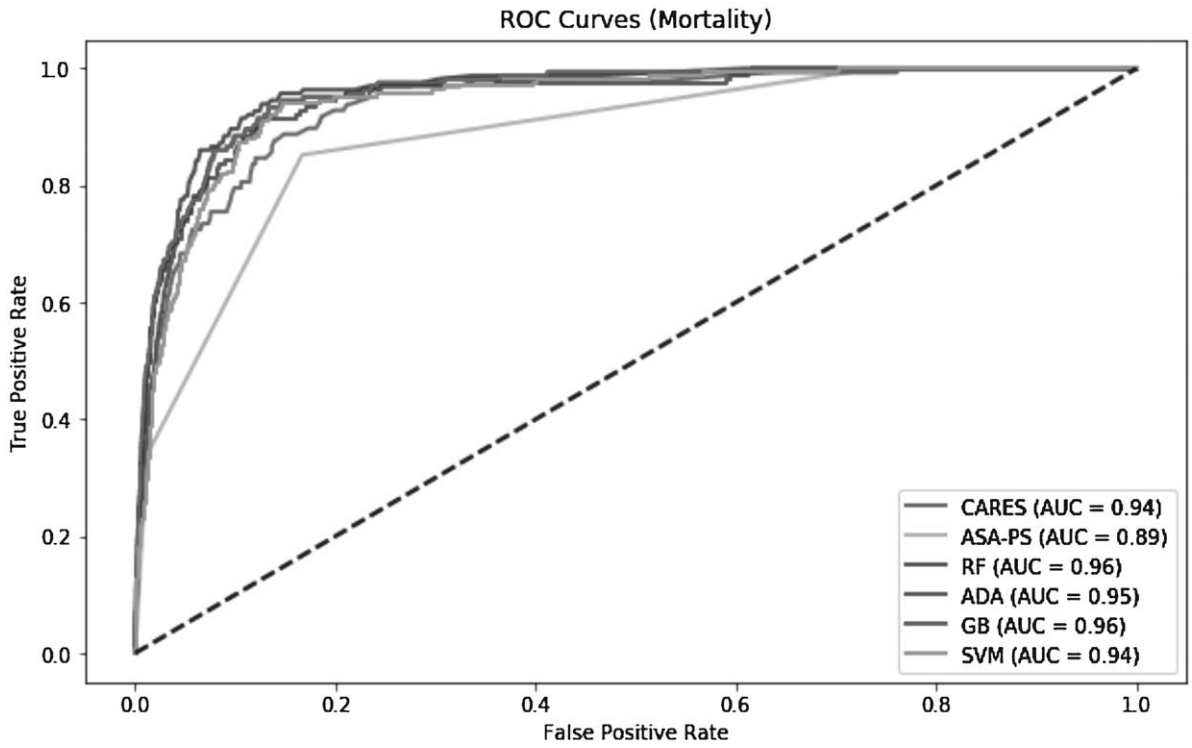

FIGURE 1. Receiver operating curves of baseline and candidate models for mortality.

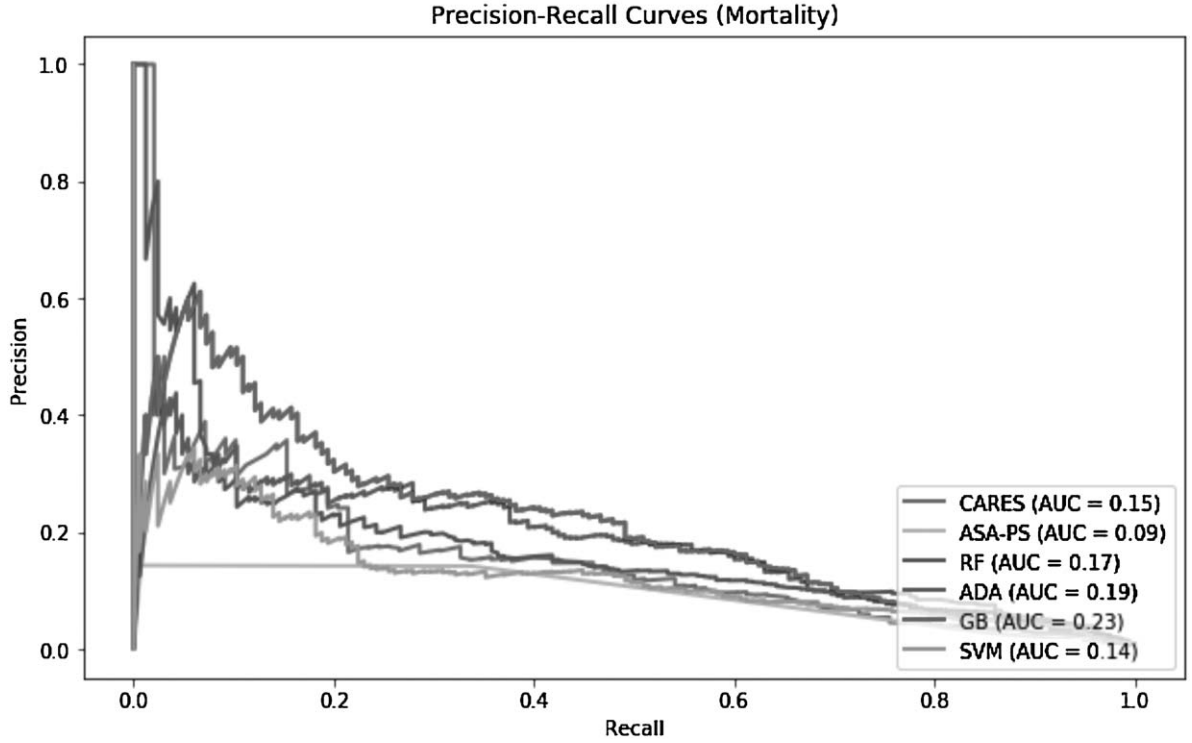

FIGURE 2. Precision-recall curves of baseline and candidate models for mortality.

TABLE 2. Results of Model Evaluation for Mortality

Model	Specificity	Sensitivity/Recall	PPV/Precision	F1 Score	AUROC	AUPRC
Baseline models						
CARES	1.00	0.00	0.00	0.00	0.94	0.15
ASA-PS	–	–	–	–	0.89	0.09
Candidate models						
Random forest (RF)	0.99	0.21	0.21	0.21	0.96	0.17
Adaptive boosting (ADA)	0.98	0.50	0.18	0.27	0.95	0.19
Gradient boosting (GB)	0.98	0.50	0.20	0.28	0.96	0.23
Support vector machine (SVM)	0.94	0.70	0.07	0.13	0.94	0.14

predictors for mortality were age, creatinine, platelet, eGFR and PT. Top predictors for postoperative ICU admission >24 hours were eGFR, aPTT, PT, weight and platelet count.

DISCUSSION

In this study, we applied machine learning methods to improve the predictive performance of CARES, a Singapore-derived surgical risk calculator predicting both 30-day mortality and need for ICU admission > 24 hours. An ensemble of gradient boosting models had significantly better sensitivity (recall) and PPV (precision) than the original CARES, as well as the ASA-PS, one of the traditionally used risk stratification methods.

The CARES surgical risk calculator uses only variables available from routine preoperative evaluation, and the machine learning version allows for the calculation of individualised predicted probabilities of outcomes, as opposed to categorizing patients into risk bands as in the original CARES.[9] Many hospitals and clinics already employ EHR systems, on which predictive analytics could be deployed, making them even more convenient to use than traditional manual scoring tools.

Postsurgical 30-day all-cause mortality is a widely accepted and relevant outcome measure of surgical care.[20] It is of interest to both surgeons and patients and its accurate prediction aids in shared decision-making.[21] For this study, postoperative mortality data was obtained from the National Registry of Death, ensuring integrity and completeness of the data. The prediction of ICU admission risk postoperatively is novel and not available in most current risk stratification tools. The ability to predict need for ICU stay after surgery could aid clinicians in determining a patient's postoperative disposition plan before surgery. This could improve patient outcomes by reducing failure to rescue events[22] and efficiency in allocation of valuable ICU resources. While ICU admission by itself would not be a useful measure of morbidity, length of stay in ICU may be seen as an indirect measure of morbidity-related outcomes.[23] We defined ICU admission for >24 hours as a significant clinical outcome upon observation that patients who were discharged from ICU within the

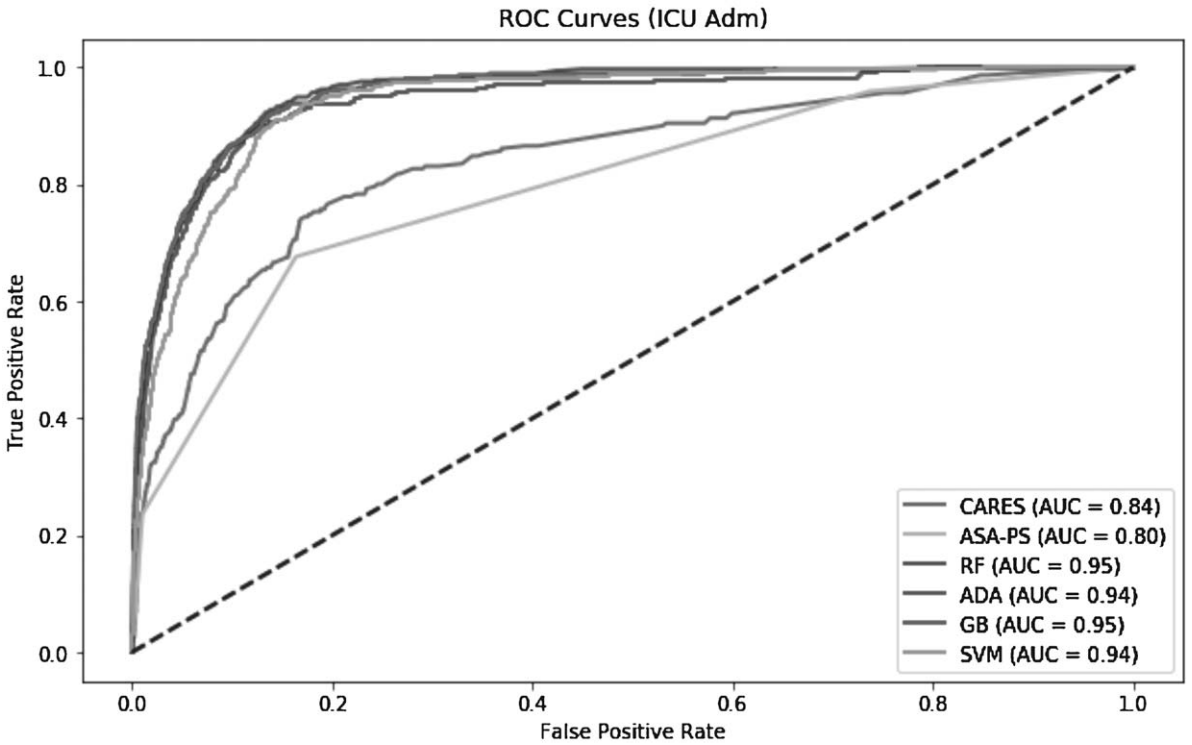

FIGURE 3. Receiver operating curves of baseline and candidate models for ICU admission.

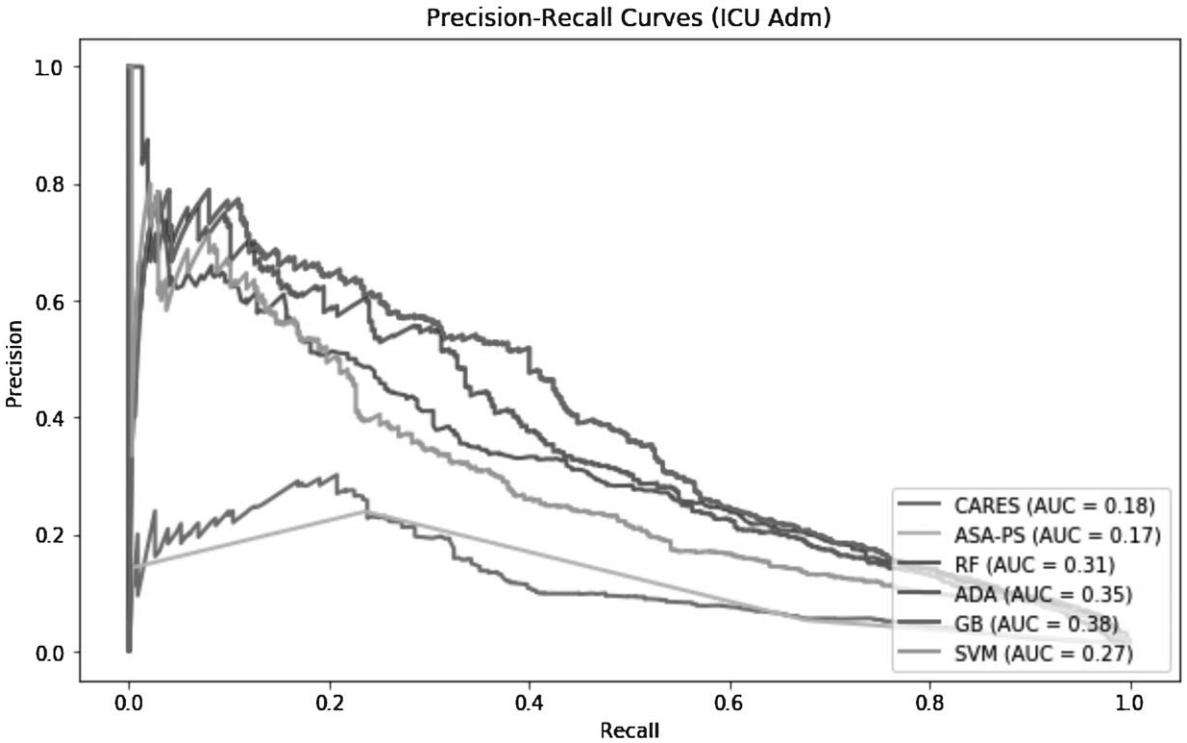

FIGURE 4. Precision-recall curves of baseline and candidate models for ICU admission.

first 24 hours may have been safely monitored postoperatively in a lower intensity unit.

Our study had several strengths. We demonstrated a big data-driven, machine learning approach to predictive analytics in perioperative care, which has several notable advantages over traditional risk calculators. Our approach uses local real-world data to make predictions about the local population, with improved accuracy over traditionally derived models that tend to have poorer performance when applied to populations and settings outside of the derivation study centres.[24] As big data analytic methods are introduced into clinical practice, future efforts should seek to move from generalizable rules to generalizable methods that utilize the richness of local data.[13] The machine learning algorithm also allows for evaluation of far more clinical variables than would be present in traditional modeling approaches, contributing to its superior performance. In addition, the model can be updated either in real-time or periodically as new data is acquired, reflecting a key component of the push toward a self-learning healthcare system.[25]

We chose a methodology that was appropriate for the severely imbalanced dataset, in which the number of negatives (nonevents) outweighs the number of positives (events) significantly. Such imbalanced datasets are common in perioperative studies with relatively rare outcomes such as postoperative mortality. In general, it was difficult to train the models due to the strongly imbalanced nature of the dataset. We addressed this by creating an ensemble of more balanced subsets, and training a classifier on each subset, producing an ensemble of classifiers, whose predictions are combined by majority vote. It was easier to predict ICU admission than mortality as the former outcome is slightly less rare (1.4% compared to 0.6%). We also used F1 score and AUPRC as our main evaluation metrics as they do not give credit for predicting true negatives and are thus robust to imbalanced datasets.[18] While ROC plots and AUROCs are popularly used in the literature to evaluate the performance of binary classifiers, they can be misleading and deceptive with respect to conclusions about classifier performance in the context of imbalanced datasets, as evidenced by our results. F1 scores, PRC plots, and

TABLE 3. Results of Model Evaluation for ICU Admission

Model	Specificity	Sensitivity/Recall	PPV/Precision	F1 Score	AUROC	AUPRC
Baseline models						
CARES	1.00	0.00	0.00	0.00	0.84	0.18
ASA-PS	–	–	–	–	0.80	0.17
Candidate models						
Random forest (RF)	0.98	0.45	0.32	0.37	0.95	0.31
Adaptive boosting (ADA)	0.97	0.57	0.23	0.33	0.94	0.35
Gradient boosting (GB)	0.97	0.58	0.27	0.36	0.95	0.38
Support vector machine (SVM)	0.91	0.78	0.10	0.18	0.94	0.27

AUPRCs, on the other hand, express the susceptibility of classifiers to imbalanced datasets and allow for more accurate interpretation of practical classifier performance. These metrics are based on precision (equivalent to PPV) which measures the fraction of correct predictions among the positive predictions, intuitively revealing differences in performance that go unnoticed when just using accuracy to evaluate classifiers on imbalanced datasets. Our findings have potential implications for similar studies involving predictive analytics on imbalanced datasets, which are very common in medicine.

Our study had several limitations. First, our model was developed and internally validated using data from a single institution in Singapore, and thus might not be generalizable to other settings. Nonetheless, we believe the importance of our work lies in demonstrating a generalizable method which could be replicated in other EHR systems, rather than in a predictive model to be applied globally. Second, the choice of 10 classifiers for each ensemble was somewhat arbitrary. We searched across a range of candidate values and observed that as the number of models in the ensemble increases, the recall improves but precision decreases (ie, less false negatives are obtained at the expense of more false positives). In our case, an ensemble of 10 classifiers optimized the F1 score, but any decision on this value is likely to be data-driven and dependent on which metric is deemed most important in the given context. Third, we chose to handle missing values by simply imputing them with the observed median for that variable, in order to reduce the computational complexity of our model, which is a consideration for eventual clinical translation on the ground. Other more sophisticated imputation approaches could have improved predictive performance. Lastly, we acknowledge that a machine learning approach carries issues of interpretability and logistical challenges for implementation.[26] We attempted to understand how our GB ensemble worked by visualizing variable importance plots and we believe hospitals are increasingly developing the infrastructure necessary to integrate predictive analytics into their EHR systems. Further studies are needed to compare our approach with provider judgment, to determine whether it influences physician behavior, and to assess how patient outcomes may be impacted.

In conclusion, machine learning can be used to improve surgical risk prediction compared to traditional risk calculators. Our study serves as an example that could be automated, applied to other clinical outcomes of interest, and integrated in EHRs to enable locally relevant clinical predictions. However, methods for model building and evaluation must be carefully considered. In particular, AUPRC should be used to evaluate model predictive performance instead of AUROC when the dataset is imbalanced.

REFERENCES

1. Weiser TG, Haynes AB, Molina G, et al. Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes. *Lancet*. 2015;385(suppl 2):S11.
2. Alkire BC, Raykar NP, Shrime MG, et al. Global access to surgical care: a modelling study. *Lancet Glob Health*. 2015;3:e316–e323.
3. Tevis SE, Kennedy GD. Postoperative complications and implications on patient-centered outcomes. *J Surg Res*. 2013;181:106–113.
4. 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.
5. Pearse RM, Moreno RP, Bauer P, et al. Mortality after surgery in Europe: a 7 day cohort study. *Lancet*. 2012;380:1059–1065.
6. Cohen ME, Bilimoria KY, Ko CY, et al. Effect of subjective preoperative variables on risk-adjusted assessment of hospital morbidity and mortality. *Ann Surg*. 2009;249:682–689.
7. Brooks MJ, Sutton R, Sarin S. Comparison of surgical risk score, POSSUM and P-POSSUM in higher-risk surgical patients. *Br J Surg*. 2005;92:1288–1292.
8. Copeland GP, Jones D, Walters M. POSSUM: a scoring system for surgical audit. *Br J Surg*. 1991;78:355–360.
9. Chan DXH, Sim YE, Chan YH, et al. Development of the Combined Assessment of Risk Encountered in Surgery (CARES) surgical risk calculator for prediction of postsurgical mortality and need for intensive care unit admission risk: A single-center retrospective study. *BMJ Open*. 2018;8:e019427.
10. Borne Y, Smith JG, Melander O, et al. Red cell distribution width in relation to incidence of coronary events and case fatality rates: a population-based cohort study. *Heart*. 2014;100:1119–1124.
11. Chen PC, Sung FC, Chien KL, et al. Red blood cell distribution width and risk of cardiovascular events and mortality in a community cohort in Taiwan. *Am J Epidemiol*. 2010;171:214–220.
12. Sim YE, Wee HE, Ang AL, et al. Prevalence of preoperative anemia, abnormal mean corpuscular volume and red cell distribution width among surgical patients in Singapore, and their influence on one year mortality. *PLoS One*. 2017;12:e0182543.
13. Taylor RA, Pare JR, Venkatesh AK, et al. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad Emerg Med*. 2016;23:269–278.
14. Lee TH, Marcantonio ER, Mangione CM, et al. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. *Circulation*. 1999;100:1043–1049.
15. Hackett NJ, De Oliveira GS, Jain UK, et al. ASA class is a reliable independent predictor of medical complications and mortality following surgery. *Int J Surg*. 2015;18:184–190.
16. Kidney Disease: Improving Global Outcomes (KDIGO) CKD-MBD Work Group. KDIGO clinical practice guideline for the diagnosis, evaluation, prevention, and treatment of Chronic Kidney Disease-Mineral and Bone Disorder (CKD-MBD). *Kidney Int Suppl*. 2009;S1–130.
17. Kristensen SD, Knuuti J, Saraste A, et al. The Joint Task Force on non-cardiac surgery: cardiovascular assessment and management of the European Society of Cardiology (ESC) and the European Society of Anaesthesiology (ESA). *Eur Heart J*. 2014;35:2383–2431.
18. 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.
19. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
20. Birkmeyer JD, Dimick JB, Birkmeyer NJ. Measuring the quality of surgical care: structure, process, or outcomes? *J Am Coll Surg*. 2004;198:626–632.
21. Yek JL, Lee AK, Tan JA, et al. Defining reasonable patient standard and preference for shared decision making among patients undergoing anaesthesia in Singapore. *BMC Med Ethics*. 2017;18:6.
22. Ghaferi AA, Birkmeyer JD, Dimick JB. Complications, failure to rescue, and mortality with major inpatient surgery in Medicare patients. *Ann Surg*. 2009;250:1029–1034.
23. Abelha FJ, Castro MA, Landeiro NM, et al. Mortality and length of stay in a surgical intensive care unit. *Rev Bras Anestesiol*. 2006;56:34–45.
24. Toll DB, Janssen KJ, Vergouwe Y, et al. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol*. 2008;61:1085–1094.
25. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Aff*. 2014;33:1163–1170.
26. Amarasingham R, Patzer RE, Huesch M, et al. Implementing electronic health care predictive analytics: considerations and challenges. *Health Aff*. 2014;33:1148–1154.