Postoperative mortality risk prediction that incorporates intraoperative vital signs: development and internal validation in a historical cohort

Une prédiction du risque de mortalité postopératoire qui intègre les signes vitaux peropératoires : mise au point et validation interne dans une cohorte historique

Janny Xue Chen Ke, MD, MSc, FRCPC (ORCID) · **Daniel I. McIsaac, MD, MPH, FRCPC** · **Ronald B. George, MD, FRCPC** · **Paula Branco, PhD** · **E. Francis Cook, ScD** · **W. Scott Beattie, MD, PhD, FRCPC** · **Robin Urquhart, PhD** · **David B. MacDonald, MD, FRCPC**

Abstract

Purpose *Accurate risk reassessment after surgery is crucial for postoperative planning for monitoring and disposition. Existing postoperative mortality risk prediction models using preoperative features do not incorporate intraoperative hemodynamic derangements that may alter risk stratification. Intraoperative vital signs may provide an objective and readily available prognostic resource. Our primary objective was to derive and internally validate a logistic regression (LR) model by adding intraoperative features to established preoperative predictors to predict 30-day postoperative mortality.*

Methods *Following Research Ethics Board approval, we analyzed a historical cohort that included patients aged ≥ 45 undergoing noncardiac surgery with an overnight stay at two tertiary hospitals (2013 to 2017). Features included intraoperative vital signs (blood pressure, heart rate, end-tidal carbon dioxide partial pressure, oxygen saturation, and temperature) by threshold and duration of exposure, as well as patient, surgical, and anesthetic factors. The cohort was divided temporally 75:25 into derivation and validation sets. We constructed a multivariable LR model with 30-day all-cause mortality as the outcome and evaluated performance metrics.*

This article is accompanied by an editorial. Please see Can J Anesth 2022; this issue.

J. X. C. Ke, MD, MSc, FRCPC (✉)
Department of Anesthesia, Pain Management & Perioperative Medicine, Dalhousie University, Halifax, NS, Canada
e-mail: janny.ke@ubc.ca

T. H. Chan School of Public Health, Harvard University, Boston, MB, USA

Department of Anesthesia, St. Paul's Hospital, Providence Health Care, 1081 Burrard Street, Vancouver, BC V6Z 1Y6, Canada

Department of Anesthesiology, Pharmacology & Therapeutics, The University of British Columbia, Vancouver, BC, Canada

D. I. McIsaac, MD, MPH, FRCPC
Department of Anesthesiology and Pain Medicine, University of Ottawa, Ottawa, ON, Canada

R. B. George, MD, FRCPC
Department of Anesthesia, Pain Management & Perioperative Medicine, Dalhousie University, Halifax, NS, Canada

Department of Anesthesia and Perioperative Care, UCSF, San Francisco, CA, USA

P. Branco, PhD
School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON, Canada

E. F. Cook, ScD
T. H. Chan School of Public Health, Harvard University, Boston, MB, USA

W. S. Beattie, MD, PhD, FRCPC
Department of Anesthesia, University of Toronto, Toronto, ON, Canada

Results *There were 30,619 patients in the cohort (mean [standard deviation] age, 66 [11] yr; 50.2% female; 2.0% mortality). In the validation set, the primary LR model showed a c-statistic of 0.893 (99% confidence interval [CI], 0.853 to 0.927), a Nagelkerke R-squared of 0.269, a scaled Brier score of 0.082, and an area under precision-recall curve of 0.158 (baseline 0.017 for an uninformative model). The addition of intraoperative vital signs to preoperative factors minimally improved discrimination and calibration.*

Conclusion *We derived and internally validated a model that incorporated vital signs to improve risk stratification after surgery. Preoperative factors were strongly predictive of mortality risk, and intraoperative predictors only minimally improved discrimination. External and prospective validations are needed.*

Study registration www.ClinicalTrials.gov (NCT04014010); registered on 10 July 2019.

Résumé

Objectif *Une réévaluation précise des risques après la chirurgie est cruciale pour la planification postopératoire du monitorage et du congé. Les modèles existants de prédiction du risque de mortalité postopératoire utilisant des caractéristiques préopératoires n'intègrent pas les perturbations hémodynamiques peropératoires, lesquelles pourraient modifier la stratification du risque. Les signes vitaux peropératoires peuvent fournir une ressource pronostique objective et facilement disponible. Notre objectif principal était de dériver et de valider en interne un modèle de régression logistique (RL) en ajoutant des caractéristiques peropératoires aux prédicteurs préopératoires établis pour prédire la mortalité postopératoire à 30 jours.*

Méthode *À la suite de l'approbation du Comité d'éthique de la recherche, nous avons analysé une cohorte historique qui comprenait des patients âgés de ≥ 45 ans bénéficiant d'une chirurgie non cardiaque avec un séjour d'une nuit dans deux hôpitaux tertiaires (2013 à 2017). Les caractéristiques comprenaient les signes vitaux peropératoires (tension artérielle, fréquence cardiaque, pression télé-expiratoire en CO₂, saturation en oxygène et température) par seuil et durée d'exposition, ainsi que des facteurs propres au patient, chirurgicaux et anesthésiques. La cohorte a été divisée temporellement 75:25 en*

ensembles de dérivation et de validation. Nous avons élaboré un modèle de RL multivariée avec la mortalité toutes causes confondues à 30 jours comme critère, et évalué les mesures de performance.

Résultats *Il y avait 30 619 patients dans la cohorte (âge moyen [écart type], 66 [11] ans; 50,2 % de femmes; 2,0 % de mortalité). Dans l'ensemble de validation, le modèle de RL primaire a montré une statistique c de 0,893 (intervalle de confiance [IC] à 99 %, 0,853 à 0,927), un R carré de Nagelkerke de 0,269, un score de Brier mis à l'échelle de 0,082 et une aire sous la courbe de rappel et précision de 0,158 (ligne de base 0,017 pour un modèle non informatif). L'ajout de signes vitaux peropératoires aux facteurs préopératoires a amélioré de façon minimale la discrimination et l'étalonnage.*

Conclusion *Nous avons dérivé et validé en interne un modèle qui incorporait des signes vitaux pour améliorer la stratification des risques après la chirurgie. Les facteurs préopératoires étaient fortement prédictifs du risque de mortalité, et les prédicteurs peropératoires n'ont que très peu amélioré la discrimination. Une validation externe et prospective est nécessaire.*

Enregistrement de l'étude www.ClinicalTrials.gov (NCT04014010); enregistrée le 10 juillet 2019.

R. Urquhart, PhD
Department of Community Health and Epidemiology, Dalhousie University, Halifax, NS, Canada

D. B. MacDonald, MD, FRCPC
Department of Anesthesia, Pain Management & Perioperative Medicine, Dalhousie University, Halifax, NS, Canada

For every 1,000 major surgeries performed in Canada, 17 patients die postoperatively in hospital.[1] Postoperative deaths account for 7.7% of global deaths.[2] Preoperative risk stratification has been a standard of care for surgical and disposition planning. Several validated scores exist for use before surgery, including the Portsmouth-Physiology and Operative Severity Score for the Enumeration of Mortality (POSSUM), Risk Stratification Index (RSI), Surgical Outcome Risk Tool, Surgical Risk Scale, National Surgical Quality Improvement Program Universal Risk Calculator, and Revised Cardiac Risk Index.[3–5] Nevertheless, preoperative risk stratification does not incorporate intraoperative derangements, such as hypotension,[6,7] tachycardia,[8,9] hypocapnia,[10] desaturation,[11] and hypothermia,[12] which have been associated with postoperative morbidity and/or mortality. There is scant information on how risk prediction changes with evolving perioperative data.

Vital signs can be automatically recorded intraoperatively and, with the increasing adoption of electronic health records (EHR),[13] may provide an objective, readily extractable prognostic resource. The Surgical Apgar Score (SAS) is a simple ten-point score

involving three intraoperative variables (estimated blood loss [EBL], blood pressure [BP], and heart rate [HR]) that predict a composite outcome of mortality and morbidity.[14] In a multicentre noncardiac surgery cohort of 5,909 patients, the *c*-statistic for SAS for mortality was 0.77.[15] Nevertheless, this level of performance has not been reproduced across surgical specialties,[16] and the addition of SAS has not meaningfully improved upon preoperative risk prediction.[17]

Existing postoperative mortality risk prediction models do not adequately address the combined inputs of pre- and intraoperative predictors. Machine learning techniques have been used to capture complex relationships.[18,19] Nevertheless, they can be computationally intensive for real-time prediction within existing EHR, and have not consistently outperformed the more interpretable logistic regression (LR).[20,21] Therefore, our primary objective was to build upon established preoperative predictors present in existing models to derive and internally validate a multivariable LR model that additionally incorporated intraoperative variables to predict 30-day mortality. Secondary objectives were to: 1) compare the performance of our primary model with that of the SAS score, and 2) quantify the added predictive value of intraoperative variables compared with mortality prediction using preoperative variables alone.

Methods

This historical cohort study used linked, routinely collected health data. We obtained research ethics approval (Nova Scotia Health Authority Research Ethics Board, Halifax, NS, Canada; file # 1024251), registered the protocol prior to analysis (www.ClinicalTrials.gov: NCT04014010; registered on 10 July 2019), and followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement.[22]

Study population

To focus on the high-risk population, where risk reassessment may be particularly important, this study included all patients aged ≥ 45 undergoing their first noncardiac surgery with at least one postoperative overnight stay at the two adult tertiary academic hospitals (Victoria General Hospital and Halifax Infirmary) in Halifax, NS, Canada, between 1 January 2013, and 1 December 2017. We excluded organ donors and patients with unlinkable records (e.g., out-of-province patients).

Data extraction and linkage

Multiple data sets were linked by health card number and surgery date (Electronic Supplementary Material [ESM] eTable 1) and included the following: 1) hospitals' Anesthesia Information Management System (AIMS; Innovian IA5.1, Telford, PA, USA), containing automated recordings of standard monitors, time-stamped anesthesiology entries of anesthesia types, medications, fluids, interventions, and perioperative laboratory tests; 2) hospitals' perioperative EHR, containing administrative and clinical data; 3) the Nova Scotia Vital Statistics database, which records all deaths within Nova Scotia; and 4) the Canadian Institute for Health Information Discharge Abstract Database, which collects clinical and administrative data from each hospitalization using standardized format by trained abstractors, including preoperative diagnoses, surgery types, and postoperative complications. To include recent and relevant preoperative comorbidities, we included all diagnostic codes recorded within three years before surgery. The final deidentified data set was extracted by and accessed through Health Data Nova Scotia (HDNS).

Outcome

Our primary outcome was binary all-cause 30-day postoperative mortality (in or out of hospital). All deaths within Nova Scotia are captured in the provincial database by legislation.

Candidate features

Candidate features were intraoperative vital signs (BP, HR, oxygen saturation, end-tidal carbon dioxide partial pressure, and temperature), as well as preoperative and intraoperative patient and surgical variables based on literature (ESM eTable 1). While studies have shown an association among abnormal intraoperative vital signs and postoperative morbidity, the optimal thresholds and durations of exposure remain controversial.[6] Thus, we specified several definitions to summarize each vital sign (please see ESM eTable 1 for details).

Statistical analysis

Pre-processing

Since intraoperative vital signs are automatically recorded and mortality reporting is mandatory, complete case analysis was performed for these variables. For features that were not vital signs, the prespecified approach to missing data were complete case analysis, imputation by group mean, and exclusion of the variable, where the variable had < 5, 5 to 20, or $> 20\%$ missing values, respectively. Our artifact removal protocol for vital signs is summarized in ESM eAppendix 1.

Cohort characteristics

Continuous data are displayed as mean and standard deviation (SD) if normally distributed and median and interquartile range [IQR] if non-normally distributed. Categorical data were presented as frequency and percentage.

Model development

To mirror real-life application and account for temporal changes, the prediction model was built using data from the earliest data and validated with the latest data. The cohort was temporally ranked using surgery date. The derivation set consisted of patients with the earliest 75% of the surgery dates, and the validation set consisted of patients with the latest 25% of the surgery dates.

Variable reduction by domain knowledge and principal component analysis

In-hospital mortality after major surgery in Nova Scotia between 2016 and 2017 was 1.7%.[23] We estimated a derivation sample size of 24,000, which would provide 408 mortality outcomes. Using the Events Per Variable criterion (EPV) of > 10 outcomes per parameter, this would mean that a maximum of 40 parameters could be used. Nevertheless, EPV has many limitations, and there is limited guidance for models involving > 30 predictors.[24]

In the derivation set, first we removed or combined nonvital sign features based on correlation matrix (collinearity), domain knowledge, and data quality. As there is no single best approach to summarizing vital sign time series as predictors, for each type of vital sign deviation we used principal component analysis (PCA) to select the feature definition among possible options that most correlated with the first principal component (ESM eTable 2), without assessing the association with outcome.[25]

Primary logistic regression model

In the derivation set, we built a multivariable LR model using all predictor variables (LR_{all}), with 30-day mortality as the outcome variable. The model was evaluated using Area Under the Receiver Operating Curve (AUROC), scaled Brier score, and Nagelkerke R-squared. These metrics were also evaluated in the validation set, along with estimated calibration index (ECI), loess-smoothed calibration curve, and area under the Precision-Recall Curve (AUPRC).[26] The AUPRC graphs precision (positive predictive value) on the y-axis against recall (sensitivity) on the x-axis, with the baseline being the percentage of true positives in the data set.[27] When the incidence of the outcome is extremely low, as in our data set, the AUPRC more correctly reflects the ability of the model to identify true positives than the AUROC does.[27] As an exploratory analysis, we examined the impact of modeling nonlinear relationships of the continuous variables using restricted cubic splines with five knots in the multivariable regression (LR_{all_spline}) and compared the discrimination and calibration of this model with LR_{all}.

Comparison to the Surgical Apgar Score

The performances of the primary model (LR_{all}) were evaluated against a univariable LR model, where the SAS linear predictor was the only independent variable (LR_{SAS}).

Relative contribution of vital signs

To explore the added predictive value of vital signs in the context of preoperative and intraoperative variables, we created a series of LR models using the following combinations of feature groups (Table 2): preoperative features only (LR_{preop}), vital signs only (LR_{vitals}), preoperative features and vital signs (LR_{preop_vitals}), and all features (LR_{all}). In the derivation set, the relative performances were compared using the likelihood ratio test ($P < 0.05$) for nested models, and Akaike Information Criterion (AIC) for non-nested models. In the validation set, the models were compared using LR_{preop} as the baseline via AUROC, scaled Brier score, R-squared, ECI, and loess-smoothed calibration curves. Using decision curves, we evaluated the net benefit (which incorporates true positive rate, false positive rate, and probability cut off) from each nested model across a range of prediction probability thresholds according to published recommendations.[28–30]

Software

The AIMS/EHR data were processed in Microsoft Structured Query Language Server 2008 R2 (10.50.1600.1; Microsoft Corporation, Redmond, WA, USA). Data were analyzed on the HDNS Citadel server using SAS 9.4 (SAS Institute, Inc., Cary, NC, USA) and R 4.1.0 software (R Foundation for Statistical Computing, Vienna, Austria).

Results

Cohort characteristics

The cohort consisted of 30,619 patients, with a mean (SD) age of 66 (11) yr; 50.2% were female (Fig. 1). Of the 624 patients who died, 493 (79.0%) deaths occurred in hospital. For patients who died, the median [IQR] for postoperative survival was 11 [5–18] days. The 30-day mortality rate was
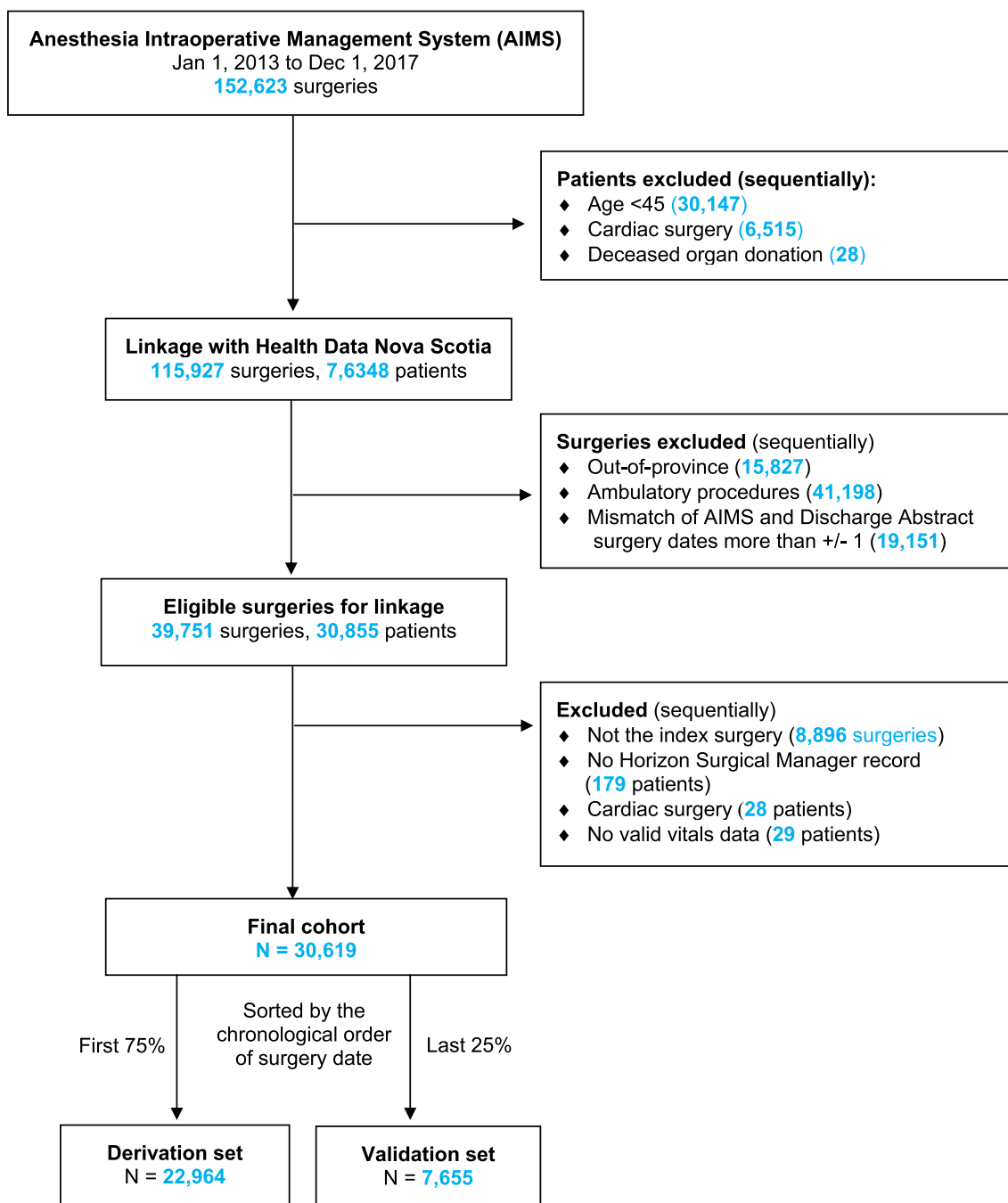

Fig. 1 Participant inclusion and exclusion flowchart

Table 1 Key cohort characteristics, stratified by 30-day mortality

Variable	No Mortality N = 29,995	Mortality N = 624	P value
Preoperative			
Age (yr)	65.6 (11.2)	74.6 (12.3)	< 0.001
Female sex	15,079 (50.3%)	284 (45.5%)	0.02
Body mass index > 30 kg·m^{-2}	9,073 (30.2%)	56 (9.0%)	< 0.001
Revised Cardiac Risk Index Score	0 [0–1]	0 [0–1]	< 0.001
	Range, 0–5	Range, 0–5	
Cerebrovascular disease	479 (1.60%)	60 (9.6%)	< 0.001
Ischemic heart disease	1,136 (3.79%)	64 (10.3%)	< 0.001
Congestive heart failure	267 (0.9%)	17 (2.7%)	< 0.001
Chronic kidney disease	752 (2.5%)	71 (11.4%)	< 0.001
Diabetes, and diabetes on insulin	1,863 (6.2%)	74 (11.9%)	< 0.001
High risk surgery	8,897 (29.7%)	141 (22.6%)	< 0.001
Elixhauser Comorbidity Index	0 [0–1]	1 [0–2]	< 0.001
Hospital Frailty Risk Score	0 [0–0]	1.6 [0–4.5]	< 0.001
Hypertension	2,403 (8.0%)	116 (18.6%)	< 0.001
Chronic obstructive lung disease	527 (1.8%)	41 (6.6%)	< 0.001
Main surgical service			< 0.001
General surgery	6,338 (21.1%)	152 (24.4%)	
Neurosurgery	1,861 (6.2%)	48 (7.7%)	
Obstetrics and gynecology	497 (1.7%)	< 6	
Orthopedic surgery	9,914 (33.1%)	148 (23.7%)	
Other	1,737 (5.8%)	12 (1.9%)	
Otolaryngology	1,882 (6.3%)	33 (5.3%)	
Plastic surgery	588 (2.0%)	< 6	
Thoracic surgery	1,959 (6.5%)	103 (16.5%)	
Urology	3,855 (12.9%)	41 (6.6%)	
Vascular surgery	1,340 (4.5%)	78 (12.5%)	
Procedural Index for Mortality Risk	0 [0–0]	0 [0–2]	< 0.001
Emergency surgery	14,581 (48.6%)	561 (89.9%)	< 0.001
Intraoperative			
Type of anesthesia			
General	25,373 (84.6%)	521 (83.5%)	0.45
Peripheral	1,504 (5.0%)	16 (2.6%)	0.005
Neuraxial	4,012 (13.4%)	52 (8.3%)	< 0.001
Laryngeal mask airway	2,441 (8.1%)	31 (5.0%)	0.004
Age-adjusted MAC, time-weighted average	1.0 [0.88–1.1]	0.88 [0.73–1.1]	< 0.001
	(N = 25,373)	(N = 521)	
Volume (mL) of crystalloid that exceeded 1 L	0 (0–1,700)	0 (0–1,500)	0.55
Any vasopressor or inotrope used	16,396 (54.7%)	434 (69.6%)	< 0.001
Any vasodilator used	2,346 (7.8%)	46 (7.4%)	0.68
Duration (min)	119.0 [85.0–173.0]	101.0 [64.0–154.5]	< 0.001
Laparoscopy without conversion to open	3,564 (11.9%)	103 (16.5%)	< 0.001
EBL (mL), missing value coded as zero	0 [0–0]	0 [0–0]	0.59
n/total N (%) with EBL > 750 mL	695/29995 (2.3%)	28/624 (4.5%)	< 0.001
Surgical Apgar Score	8 [7–9]	7 [6–8]	< 0.001
Vital signs			
First SBP in AIMS (mm Hg)	144 (27)	138 (31)	< 0.001

Table 1 continued

Variable	No Mortality $N = 29{,}995$	Mortality $N = 624$	*P* value
First MAP in AIMS (mm Hg)	103 (18)	96 (21)	< 0.001
First HR in AIMS	75 [65–86]	85 [71–99]	< 0.001
Any MAP < 80 mm Hg	27,534 (91.8%)	580 (93.0%)	0.30
Any MAP < 70 mm Hg	23,740 (79.2%)	534 (83.8%)	0.004
Any MAP < 65 mm Hg	20,215 (67.4%)	477 (76.4%)	< 0.001
Any MAP < 60 mm Hg	15,399 (51.3%)	407 (65.2%)	< 0.001
Max. % decrease in SBP relative to first AIMS SBP	40.8 (15.9)	41.2 (18.0)	0.53
Cumulative duration (min) MAP < 70 mm Hg	18.3 [3.3–45.8]	23.1 [5.8–61.5]	< 0.001
Max. change of HR in BPM above first AIMS HR	21 [10–34]	16 [7–28]	< 0.001
Max. change of HR in BPM below first AIMS HR	21 [13–30]	23 [14–36]	< 0.001
Cumulative duration (min) HR below 60 BPM	10.2 [0.3–47.4]	0.3 [0.0–9.9]	< 0.001
Cumulative duration (min) HR above 100 BPM	0.0 [0.0–4.2]	2.2 [0.0–22.1]	< 0.001
Cumulative duration (min) SpO_2 < 88%	0.0 [0.0–0.0]	0.0 [0.0–1.0]	< 0.001
	Range, 0–105.3	Range, 0–36.3	
Cumulative duration (min) temperature < 36°C	0 [0–12.5]	0 [0–0.75]	< 0.001
Cumulative duration (min) temperature > 38°C	0 [0–0]	0 [0–0]	< 0.001
Cumulative duration (min) $ETCO_2$ < 30 mm Hg	5.0 [2.5–9.0]	7.0 [2.8–18.5]	< 0.001
Cumulative duration (min) $ETCO_2$ > 45 mm Hg	2.8 [0.3–12.0]	0.8 [0.0–8.7]	< 0.001

Continuous data are displayed as mean (standard deviation) if normally distributed and median [interquartile range] if non-normally distributed. Categorical data are displayed as number (percentage). Student's *t* test, Wilcoxon test, and Chi square tests were performed for normal, non-normal, and categorical data, respectively.

AIMS = Anesthesia Information Management System; BPM = beats per minute; EBL = estimated blood loss; $ETCO_2$ = end-tidal carbon dioxide partial pressure; HR = heart rate; IQR = interquartile range; MAC = minimal alveolar concentration during general anesthesia; MAP = mean arterial pressure; Max. = maximum; SBP = systolic blood pressure; SpO_2 = oxygen saturation

2.2% (493/22,964) in the derivation set, and 1.7% (131/7,655) in the validation set, with similar cohort characteristics (standardized mean differences all < 0.2 [ESM eTable 2]). Cohort characteristics are listed in Table 1.

Candidate predictors

No imputation was required, since the candidate features with missing values had either < 0.1% missing and complete case analysis was performed, or > 20% missing and the features were excluded from modeling (eTable 1). Features with significant missing values were: American Society of Anesthesiologists Physical Status (81% missing, as it only became a mandatory field two months before the study end date); preoperative vital signs (38%, as these were only available for elective patients or inpatients who spent time in the preoperative area); and EBL (87%, as this was not a mandatory variable in the AIMS or EHR). For the calculation of SAS, patients with missing EBL values were given an EBL of zero (i.e., scored as "< 100 mL").[16] After feature reduction, we included 11 preoperative (with nine categories in the surgery type feature), nine

intraoperative, and 11 vital signs features (Table 2, ESM eTable 3) in the final LR$_{all}$ model.

LR$_{all}$ model

The formula and coefficients for calculating the LR$_{all}$ prediction formula are listed in Table 2. The apparent and optimism-corrected AUROC in the derivation set were 0.884 and 0.875, respectively (ESM eTable 4). The AUROC in the validation set was 0.893 (99% CI, 0.853 to 0.927) (Table 3). The validation set receiver operating curve and calibration are displayed in Figs. 2 and 3. While the model was well calibrated between predicted risks of 0 to 0.4, it overestimated risks at predicted probabilities above 0.4 (eFig. 1). The validation set AUPRC was 0.158, which showed improvement in mortality prediction compared with a model with no predictive value (which would have a baseline AUPRC of 0.017, the incidence of mortality in the validation set). Sensitivity analysis showed that the addition of splines to the continuous variables within LR$_{all}$ only minimally improved performance (ESM eTable 4, eFig. 1). Lastly, LR$_{all}$ outperformed LR$_{SAS}$ (Table 3, Figs. 2, and 3).

Table 2 The features and coefficients for the primary logistic regression (LR_{all}) formula are displayed

Feature	Coefficient	SE	P value (Wald Chi square)
Intercept	-8.2785	0.4637	< 0.001
Preoperative features			
Age (yr)	0.0564	0.0046	< 0.001
Female sex	-0.1852	0.1023	0.07
Emergency surgery	1.1724	0.1816	< 0.001
Procedural Index for Mortality Risk	0.1618	0.0377	< 0.001
Surgery type (compared with general surgery)			
Neurosurgery	0.3484	0.2497	0.16
Obstetrics and gynecology	-0.5112	1.0228	0.62
Orthopedic surgery	-0.2778	0.1896	0.14
Other	-1.578	0.4745	< 0.001
Otolaryngology	0.6906	0.2562	0.007
Plastic surgery	0.155	0.4973	0.76
Thoracic surgery	0.7485	0.1908	< 0.001
Urology	-0.5898	0.235	0.01
Vascular surgery	0.6579	0.2124	0.002
Hypertension	-0.5739	0.1701	< 0.001
Chronic obstructive pulmonary disease	-0.1053	0.2083	0.61
Revised Cardiac Risk Index	-0.0417	0.0718	0.56
Elixhauser Comorbidity Index	0.371	0.0461	< 0.001
Hospital Frailty Risk Score	0.0215	0.0144	0.14
Obesity	-0.3235	0.1648	0.05
Vital sign features			
Maximum decrease of SBP relative to first AIMS SBP (%)	-0.6087	0.3494	0.08
Cumulative duration of MAP < 70 mm Hg (min)	0.0057	0.0013	< 0.001
Maximum change of HR in 10 BPM above first AIMS HR	-0.1194	0.0332	< 0.001
Maximum change of HR in 10 BPM below first AIMS HR	0.034	0.0343	0.32
Cumulative duration of HR < 60 BPM (min)	-0.0031	0.0015	0.03
Cumulative duration of HR > 100 BPM (min)	0.0062	0.0014	< 0.001
Cumulative duration of SpO_2 < 88% (min)	0.025	0.0069	< 0.001
Cumulative duration of temperature < 36°C (hr)	-0.00237	0.0804	0.98
Cumulative duration of temperature > 38°C (hr)	-0.588	0.4858	0.23
Cumulative duration of $ETCO_2$ < 30 mm Hg, GA (min)	0.0071	0.0017	< 0.001
Cumulative duration of $ETCO_2$ > 45 mm Hg, GA (min)	0.0017	0.002	0.39
Other intraoperative features			
Duration of surgery (hr)	-0.2801	0.0656	< 0.001
General anesthesia	1.1923	0.2564	< 0.001
Neuraxial anesthesia	-0.2819	0.2216	0.20
Peripheral nerve block	-0.6181	0.3501	0.08
Laparoscopic surgery with no conversion to open	0.3753	0.1826	0.04
Age-adjusted MAC during GA, time averaged	-1.2849	0.194	< 0.001
Crystalloid in excess of 1 L (mL)	0.0463	0.0419	0.27
Use of vasopressors and inotropes	0.134	0.1238	0.28
Use of vasodilators	-0.1381	0.1936	0.48

Features are listed according to preoperative, intraoperative vital signs, and other intraoperative groups. For the surgery type, indicator variables were created, with the reference group being general surgery.

AIMS = Anesthesia Information Management System; BPM = beats per minute; $ETCO_2$ = end-tidal carbon dioxide partial pressure; GA = general anesthesia; HR = heart rate; IQR = interquartile range; MAC = minimal alveolar concentration; MAP = mean arterial pressure; SE = standard error; SBP = systolic blood pressure; SpO_2 = oxygen saturation

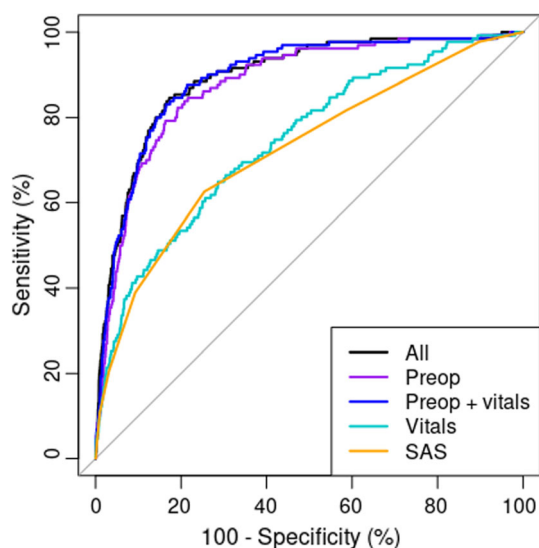**Table 3** Performances of models in the validation set (n = 7655): all features (LR_{all}), preoperative features only (LR_{preop}), vital signs only (LR_{vitals}), preoperative features and vital signs (LR_{preop_vitals}), and Surgical Apgar Score (LR_{SAS}).

Model	N	AUROC (99% CI)	Nagelkerke R-squared	Brier score	Scaled Brier score	ECI
LR_{all}	22,943	0.893 (0.853 to 0.927)	0.269	0.015	0.082	0.011
$LR_{splines}$	22,943	0.900 (0.859 to 0.934)	0.287	0.015	0.090	0.044
LR_{preop}	22,943	0.877 (0.835 to 0.914)	0.229	0.016	0.056	0.019
LR_{preop_vitals}	22,943	0.893 (0.851 to 0.927)	0.259	0.016	0.072	0.014
LR_{vitals}	22,964	0.744 (0.688 to 0.801)	0.086	0.017	0.018	0.014
LR_{SAS}	22,963	0.724 (0.663 to 0.787)	0.088	0.016	0.021	0.006

The 99% confidence intervals (CI) were calculated using 2,000 stratified bootstrap replicates.

AUROC = area under receiver operating characteristic curve; CI = confidence interval; ECI = estimated calibration index; N = total number of patients

Fig. 2 Receiver operating curves of models in the validation set (n = 7,655): all features (LR_{all}) in black, preoperative features only (LR_{preop}) in purple, preoperative features and vital signs (LR_{preop_vitals}) in dark blue, vital signs only (LR_{vitals}) in turquoise, and Surgical Apgar Score (LR_{SAS}) in orange

Nested model comparisons

The relative performances of the nested models are displayed in Table 3, Figs. 2, and 3, and the formulas in ESM eTable 6. In the derivation set, the incremental addition of each feature group improved prediction statistically (decreased AIC [ESM eTable 4], likelihood ratio test $P < 0.001$). In the derivation and validation sets, the addition of vital sign features (LR_{preop_vitals}) and both vital sign and other intraoperative features (LR_{all}) only minimally improved discrimination, but improved calibration compared with LR_{preop}. Examining the components of Fig. 3 in detail, the LR_{preop} has a very limited calibration range. Moreover, this model overestimated risk above an observed mortality risk of approximately 0.3. Note that these issues are even more apparent in the LR_{SAS} calibration

curve. In contrast, all models involving intraoperative vitals showed much improved calibration curves not only in terms of less deviation from a calibration slope of 1 but also in terms of an expanded range of calibration.

Decision curve analysis (eFig. 2) revealed that, below a mortality probability threshold of 10%, an intervention (e.g., disposition, monitoring strategies) that would be given based on risk prediction (i.e., the results of any of the LR models) provided higher net benefit, compared with providing intervention to no one or everyone without risk stratification. Compared with LR_{preop}, LR_{preop_vitals} led to increased net benefit up until a mortality probability threshold of 20%. Between risk thresholds of 0.15 and 0.25, the further addition of intraoperative variables (LR_{all}) increased benefit compared with LR_{preop_vitals}. At risk thresholds above 25%, none of the models added benefit.

Discussion

The adoption of AIMS and EHRs[13] has led to interest in personalized risk prediction that is timely and responsive to evolving perioperative information. To help guide risk reassessment at the end of surgery, we developed and internally validated a prediction model for 30-day postoperative mortality by incorporating intraoperative features to known preoperative predictors. This model (LR_{all}) provided strong discrimination, and good calibration under predicted mortality risks of 0.4 (a likely clinically relevant threshold for a change in monitoring or disposition). Our analyses suggest that the addition of intraoperative variables to preoperative variables only minimally improved discrimination. The clinical impact of improved calibration,[31] net benefit,[28] and risk restratification based on intraoperative vital signs requires prospective and external validation. If the model improves outcomes, the formula could be incorporated into EHR for automated, real-time risk prediction.
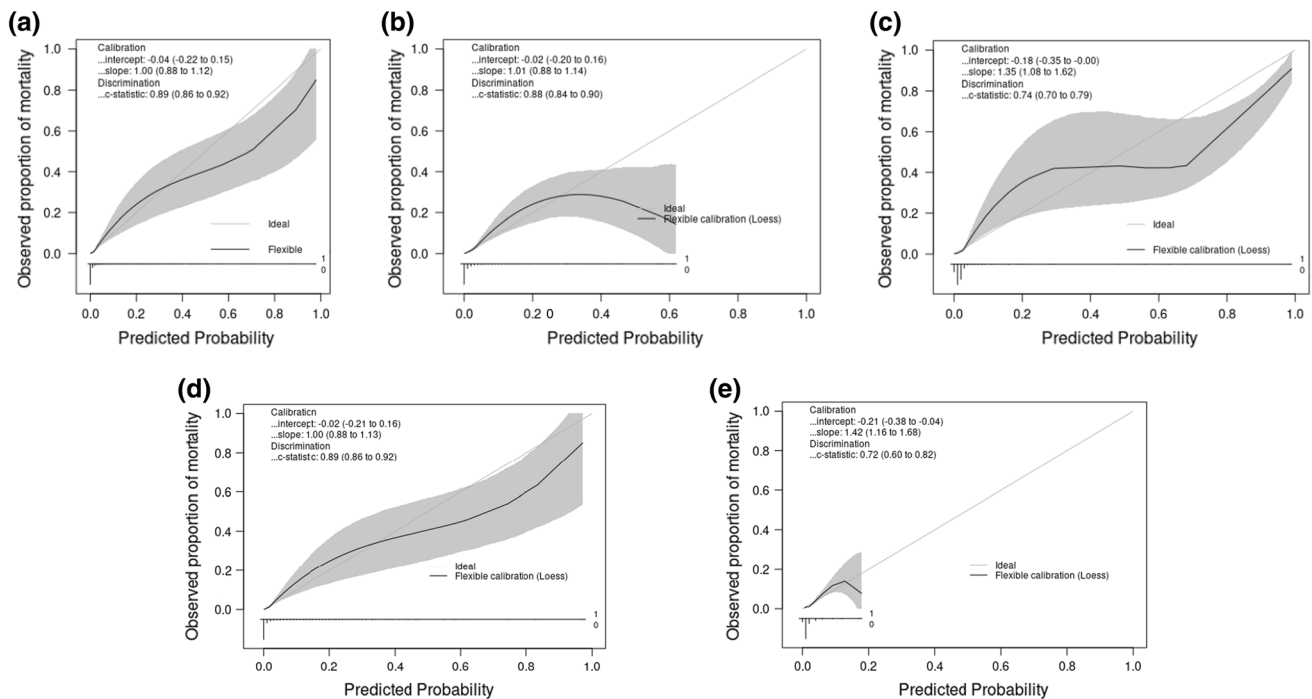
Fig. 3 Loess-smoothed calibration curves of observed *vs* predicted probability of models in the validation set (*n* = 7,655), with 95% confidence interval calculated based on 2,000 bootstrapped samples. The models were: a) all features (LR_all), b) preoperative features only (LR_preop), c) vital signs only (LR_vitals), d) preoperative features and vital signs (LR_preop_vitals), and e) Surgical Apgar Score (LR_SAS)

One of the most frequently studied risk scores that relies on intraoperative information is the SAS.[14] In the present study, LR_SAS performed similarly as a previous external validation (AUROC 0.72).[14] Our model that incorporated preoperative and intraoperative features (LR_all) outperformed LR_SAS in internal validation, a finding to be confirmed by external validation.

Building upon the literature on perioperative risk prediction, our study incorporated multiple EHR and AIMS features and examined the predictive contribution of intraoperative vital signs with more details and interpretability. Lee *et al.* studied a cohort of 59,985 inpatient surgeries under general anesthesia to predict in-hospital mortality,[18] and modeled intraoperative vital signs using various descriptive statistics over both the entire duration and only the last ten minutes of the case.[18] Their best neural network model obtained an AUROC of 0.91 (95% CI, 0.88 to 0.93). Similarly, Fritz *et al.* included 95,907 patients undergoing surgeries with tracheal intubation to predict mortality, and modeled vital signs as time series using a neural network.[19] Only one vital sign feature (mean temperature) was among the top ten most important features of the neural network model for three randomly selected patients.[19] Their best model obtained an AUROC of 0.867 (95% CI, 0.835 to 0.899) but with a low AUPRC of 0.095 (baseline 0.010 for an uninformative

model), and the calibration curve was only reported in the range of 0.2 for probabilities of mortality.

In contrast, our analysis may be more generalizable since we included all anesthesia techniques, evaluated a variety of vital sign definitions, and used duration-based definitions that would be less susceptible to artifacts.[32] To reduce bias during feature selection, we used PCA to select vital sign features among correlated definitions, ignoring any association with the outcome. These approaches are thought to decrease overfitting and maintain the performance of our model in external data. The LR approach is transparent for clinical use and may aid more in expedient implementation for real-time prediction than the more complex machine learning models do.

Our study adds to the literature by exploring the relative predictive contribution of features that would be available to clinicians before versus at the end of surgery. Preoperatively, the goals of risk prediction include informing the decision for whether to proceed to surgery and disposition planning. Current evidence suggests that models based only on preoperative data, such as the RSI, can be highly discriminative when applied to heterogeneous populations.[18] Similarly, we found that preoperative variables alone showed strong performance. Importantly, the inclusion of intraoperative data provides an opportunity to update risk assessment at the end of surgery. This new information could be incorporated into

EHR's as an automatic alert, and redirect patients who were deemed to have low risk preoperatively but exceeded the risk threshold for intraoperative hemodynamic derangements. Our results suggest that adding intraoperative vital sign features (LR_{preop_vitals}) increased net benefit, which likely reflects improvements in calibration and the expanded range of predicted probabilities (which supports addition of meaningful new information to the model).[33] The finding that our LR_{all} and LR_{preop_vitals} had similar performances suggest that the intraoperative variables that are not vital signs variables may not provide meaningful additional predictive value in this setting.

While preoperative variables have excellent predictive value, there are many situations where they may not be readily extractable for real-time EHR prediction. Nevertheless, our results showed that exclusive reliance on intraoperative vital signs was insufficient for accurate postoperative risk prediction. Similar to the SAS, our model of only intraoperative vital signs displayed moderate discrimination and poor calibration, emphasizing the importance of patient and procedural factors when assessing postoperative risk.

Strengths and limitations

Our study has several strengths. First, our data set had high data quality with complete capture of the mortality outcome variable. Second, we applied artifact algorithms,[34–36] and included duration-based hypotension definitions that were less affected by the choice of artifact filtering algorithms than definitions involving magnitude of change from thresholds were.[32] Third, we focused on objective features that are routinely collected, while models such as SAS and POSSUM include measures such as EBL, which has high variation and missingness depending on the technique of measurement.[37] Finally, using PCA, we selected vital sign definitions without considering outcome associations, to reduce bias.

This study has several limitations. The cohort involved two tertiary academic hospitals, necessitating external validation. The historical cohort is subject to misclassification and missing data. Models using retrospective data are biased perioperative interventions that can modify risk, for example when a clinician admits a comorbid patient to a high dependency unit, or maintains higher BPs for at-risk patients. We could not compare the performances of our models to the RSI because we could not cross walk procedure codes. Because of sample size limitations, we did not evaluate model performance in subgroups. Similar to previous studies, the data set had significant class imbalance due to low mortality rates.

Lastly, while mortality is an important outcome with high data quality, other outcomes may be more relevant to postoperative planning.

We have ongoing exploratory analysis on prediction of morbidity outcomes and application of machine learning techniques particularly in settings of rare event rates. Another future direction would be to evaluate the predictive value of intraoperative variables in certain subgroups (e.g., emergency surgery), particularly when other predictive information may not be readily available or extractible for modeling by the end of surgery.

Conclusion

We developed and internally validated a prediction model for 30-day postoperative mortality by adding intraoperative features to established preoperative predictors. While preoperative factors are strongly predictive of mortality risk, our analysis showed that the addition of intraoperative predictors slightly improved postoperative risk reassessment. Importantly, intraoperative vital signs alone were insufficient for prediction. Prospective and external validations are needed.

Although this research analysis is based on data obtained from the Nova Scotia Department of Health and Wellness, the observations and opinions expressed are those of the authors and do not represent those of either Health Data Nova Scotia or the Department of Health and Wellness.

References

1. *Canadian Institute for Health Information*. Your health system: hospital deaths following major surgery details for Nova Scotia 2022. Available from URL: https://yourhealthsystem.cihi.ca/hsp/indepth?lang=en#/indicator/020/2/C2000/ (accessed March 2022).

2. *Nepogodiev D, Martin J, Biccard B, Makupe A, Bhangu A, National Institute for Health Research Global Health Research Unit on Global Surgery*. Global burden of postoperative death. Lancet 2019; 393: 401. https://doi.org/10.1016/s0140-6736(18)33139-8.

3. *Moonesinghe SR, Mythen MG, Das P, Rowan KM, Grocott MP*. Risk stratification tools for predicting morbidity and mortality in adult patients undergoing major surgery: qualitative systematic review. Anesthesiology 2013; 119: 959–81. https://doi.org/10.1097/aln.0b013e3182a4e94d

4. *Wong DJ, Harris S, Sahni A, et al*. Developing and validating subjective and objective risk-assessment measures for predicting mortality after major surgery: an international prospective cohort study. PLOS Med 2020; 17: e1003253. https://doi.org/10.1371/journal.pmed.1003253.

5. *Sigakis MJ, Bittner EA, Wanderer JP*. Validation of a risk stratification index and risk quantification index for predicting patient outcomes in-hospital mortality, 30-day mortality, 1-year mortality, and length-of-stay. Anesthesiology 2013; 119: 525–40. https://doi.org/10.1097/aln.0b013e31829ce6e6

6. *Wesselink EM, Kappen TH, Torn HM, Slooter AJ, van Klei WA*. Intraoperative hypotension and the risk of postoperative adverse outcomes: a systematic review. Br J Anaesth 2018; 122: 706–21. https://doi.org/10.1016/j.bja.2018.04.036

7. *Roshanov PS, Sheth T, Duceppe E, et al*. Relationship between perioperative hypotension and perioperative cardiovascular events in patients with coronary artery disease undergoing major noncardiac surgery. Anesthesiology 2019; 130: 756–66. https://doi.org/10.1097/aln.0000000000002654

8. *House LM, Marolen KN, St Jacques PJ, McEvoy MD, Ehrenfeld JM*. Surgical apgar score is associated with myocardial injury after noncardiac surgery. J Clin Anesth 2016; 34: 395–402. https://doi.org/10.1016/j.jclinane.2016.05.009

9. *Abbott TE, Pearse RM, Archbold RA, et al*. A prospective international multicentre cohort study of intraoperative heart rate and systolic blood pressure and myocardial injury after noncardiac surgery: results of the VISION study. Anesth Analg 2018; 126: 1936–45. https://doi.org/10.1213/ane.0000000000002560

10. *Mutch WA, El-Gabalawy R, Girling L, Kilborn K, Jacobsohn E*. End-tidal hypocapnia under anesthesia predicts postoperative delirium. Front Neurol 2018; 9: 678. https://doi.org/10.3389/fneur.2018.00678.

11. *Abdelmalak BB, Cata JP, Bonilla A, et al*. Intraoperative tissue oxygenation and postoperative outcomes after major non-cardiac surgery: an observational study. Br J Anaesth 2013; 110: 241–9. https://doi.org/10.1093/bja/aes378

12. *Riley C, Andrzejowski J*. Inadvertent perioperative hypothermia. BJA Educ 2018; 18: 227–33. https://doi.org/10.1016/j.bjae.2018.05.003

13. *Kazemi P, Lau F, Simpao AF, Williams RJ, Matava C*. The state of adoption of anesthesia information management systems in Canadian academic anesthesia departments: a survey. Can J Anesth 2021; 68: 693–705. https://doi.org/10.1007/s12630-021-01924-4

14. *Gawande AA, Kwaan MR, Regenbogen SE, Lipsitz SA, Zinner MJ*. An apgar score for surgery. J Am Coll Surg 2007; 204: 201–8. https://doi.org/10.1016/j.jamcollsurg.2006.11.011

15. *Haynes AB, Regenbogen SE, Weiser TG, et al*. Surgical outcome measurement for a global patient population: validation of the surgical apgar score in 8 countries. Surgery 2011; 149: 519–24. https://doi.org/10.1016/j.surg.2010.10.019

16. *Reynolds PQ, Sanders NW, Schildcrout JS, Mercaldo ND, St Jacques PJ*. Expansion of the surgical apgar score across all surgical subspecialties as a means to predict postoperative mortality. Anesthesiology 2011; 114: 1305–12. https://doi.org/10.1097/aln.0b013e318219d734

17. *Terekhov MA, Ehrenfeld JM, Wanderer JP*. Preoperative surgical risk predictions are not meaningfully improved by including the surgical apgar score: an analysis of the risk quantification index and present-on-admission risk models. Anesthesiology 2015; 123: 1059–66. https://doi.org/10.1097/aln.0000000000000858

18. *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. https://doi.org/10.1097/aln.0000000000002186

19. *Fritz BA, Cui Z, Zhang M, et al*. Deep-learning model for predicting 30-day postoperative mortality. Br J Anaesth 2019; 123: 688–95. https://doi.org/10.1016/j.bja.2019.07.025

20. *Austin PC, Harrell FE Jr, Steyerberg EW*. Predictive performance of machine and statistical learning methods: impact of data-generating processes on external validity in the "large N, small p" setting. Stat Methods Med Res 2021; 30: 1465–83. https://doi.org/10.1177/09622802211002867

21. *Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B*. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019; 110: 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

22. *Collins GS, Reitsma JB, Altman DG, Moons KG*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Ann Intern Med 2015; 162: 55–63. https://doi.org/10.7326/m14-0698

23. *Canadian Institute for Health Information*. Your health system 2022. Available from URL: https://yourhealthsystem.cihi.ca/hsp/?lang=en (accessed March 2022).

24. *Riley RD, Ensor J, Snell KI, et al*. Calculating the sample size required for developing a clinical prediction model. BMJ 2020; 368: m441. https://doi.org/10.1136/bmj.m441.

25. *SAS Institute Inc*. The FACTOR procedure: Example 44.1 principal component analysis 2021. Available from URL: https://documentation.sas.com/doc/en/pgmsascdc/v_009/statug/statug_factor_examples01.htm (accessed March 2022).

26. *Branco P, Torgo L, Ribeiro RP*. A survey of predictive modeling on imbalanced domains. ACM Comput Surv 2016; 49: 1–50.

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. https://doi.org/10.1371/journal.pone.0118432.

28. *Vickers AJ, Van Calster B, Steyerberg EW*. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016; 352: i6. https://doi.org/10.1136/bmj.i6.

29. *Vickers AJ, Van Calster B, Steyerberg EW*. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res 2019; 3: 18. https://doi.org/10.1186/s41512-019-0064-7.

30. *Brown M*. rmda: risk model decision analysis 2018. Available from URL: https://CRAN.R-project.org/package=rmda (accessed March 2022).

31. *Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prdiction models' of the STRATOS Initiative*. Calibration: the Achilles heel of predictive analytics. BMC Med 2019; 17: 230. https://doi.org/10.1186/s12916-019-1466-7.

32. *Pasma W, Peelen LM, van Buuren S, van Klei WA, de Graaff JC*. Artifact processing methods influence on intraoperative hypotension quantification and outcome effect estimates. Anesthesiology 2020; 132: 723–37. https://doi.org/10.1097/aln.0000000000003131

33. *Harrell FE*. Statistically efficient ways to quantify added predictive value of new measurements 2020. Available from URL: https://www.fharrell.com/post/addvalue/ (accessed March 2022).

34. *Sun LY, Chung AM, Farkouh ME, et al.* Defining an intraoperative hypotension threshold in association with stroke in cardiac surgery. Anesthesiology 2018; 129: 440–7. https://doi.org/10.1097/aln.0000000000002298

35. *Salmasi V, Maheshwari K, Yang D, et al.* Relationship between intraoperative hypotension, defined by either reduction from baseline or absolute thresholds, and acute kidney and myocardial injury after noncardiac surgerya retrospective cohort analysis. Anesthesiology 2017; 126: 47–65. https://doi.org/10.1097/aln.0000000000001432

36. *Shanks A*. Automated identification and validation to detect physiologically implausible pulsatile BP artifacts 2018. Available from URL: http://mpog.org/wp-content/uploads/2018/10/9_Shanks_BP_MPOG.pdf (accessed March 2022).

37. *Tran A, Heuser J, Ramsay T, McIsaac DI, Martel G*. Techniques for blood loss estimation in major non-cardiac surgery: a systematic review and meta-analysis. Can J Anesth 2021; 68: 245–55. https://doi.org/10.1007/s12630-020-01857-4