BJA

British Journal of Anaesthesia, 128 (4): 623–635 (2022)

doi: 10.1016/j.bja.2021.10.052
Advance Access Publication Date: 17 December 2021
Cardiovascular

Intraoperative prediction of postanaesthesia care unit hypotension

Konstantina Palla[1,†], Stephanie L. Hyland[1,†], Karen Posner[2], Pratik Ghosh[1], Bala Nair[2], Melissa Bristow[1], Yoana Paleva[1], Ben Williams[1], Christine Fong[2], Wil Van Cleve[2], Dustin R. Long[2], Ronald Pauldine[2], Kenton O'Hara[1], Kenji Takeda[1] and Monica S. Vavilala[2,*]

[1]Microsoft Research, Cambridge, UK and [2]Department of Anaesthesiology and Pain Medicine, University of Washington, Seattle, WA, USA

*Corresponding author. E-mail: vavilala@uw.edu

[†]Joint first authors.

Abstract

Background: Postoperative hypotension is associated with adverse outcomes, but intraoperative prediction of postanaesthesia care unit (PACU) hypotension is not routine in anaesthesiology workflow. Although machine learning models may support clinician prediction of PACU hypotension, clinician acceptance of prediction models is poorly understood.

Methods: We developed a clinically informed gradient boosting machine learning model using preoperative and intraoperative data from 88 446 surgical patients from 2015 to 2019. Nine anaesthesiologists each made 192 predictions of PACU hypotension using a web-based visualisation tool with and without input from the machine learning model. Questionnaires and interviews were analysed using thematic content analysis for model acceptance by anaesthesiologists.

Results: The model predicted PACU hypotension in 17 029 patients (area under the receiver operating characteristic [AUROC] 0.82 [95% confidence interval {CI}: 0.81–0.83] and average precision 0.40 [95% CI: 0.38–0.42]). On a random representative subset of 192 cases, anaesthesiologist performance improved from AUROC 0.67 (95% CI: 0.60–0.73) to AUROC 0.74 (95% CI: 0.68–0.79) with model predictions and information on risk factors. Anaesthesiologists perceived more value and expressed trust in the prediction model for prospective planning, informing PACU handoffs, and drawing attention to unexpected cases of PACU hypotension, but they doubted the model when predictions and associated features were not aligned with clinical judgement. Anaesthesiologists expressed interest in patient-specific thresholds for defining and treating postoperative hypotension.

Conclusions: The ability of anaesthesiologists to predict PACU hypotension was improved by exposure to machine learning model predictions. Clinicians acknowledged value and trust in machine learning technology. Increasing familiarity with clinical use of model predictions is needed for effective integration into perioperative workflows.

Keywords: data science; hypotension; machine learning; postanaesthesia care unit; risk prediction

Received: 5 May 2021; **Accepted:** 18 October 2021

Editor's key points

- Postoperative hypotension is associated with adverse outcomes, but there is currently no reliable method for predicting postoperative hypotension.
- This study demonstrates that a machine learning model, combining preoperative and intraoperative data, can predict hypotension in the recovery area better than clinicians using readily available clinical information without access to the machine learning predictions.
- Clinicians were willing to modify their predictions regarding hypotension based on the machine learning predictions, suggesting that incorporation of interpretable machine learning algorithms into clinical practice could increase the accuracy with which postoperative hypotension is anticipated and potentially prevented.

Perioperative hypotension is common and associated with adverse postoperative outcomes, including myocardial infarction, acute kidney injury, and stroke.[1–10] Machine learning (ML) models may support clinician prediction of perioperative hypotension, enabling improved approaches to prevention, detection, and treatment of these events.[11–16] However, anaesthesiologist acceptance of and interaction with ML models of hypotension remain poorly understood, limiting efforts to translate these tools into meaningful clinical workflows.

To date, efforts to develop and validate predictive models for perioperative hypotension have focused predominantly on the intraoperative period.[11–14,16] Although intraoperative hypotension has demonstrated a reproducible association with adverse outcomes, such as acute kidney injury, myocardial injury, and death,[17] a large burden of the organ dysfunction and mortality attributable to perioperative hypotension has been shown to arise from hypotensive events occurring during the early postoperative period,[2,17–19] when the reduced intensity of monitoring routinely results in delayed or missed detection. For example, a sub-study of the Perioperative Ischemic Evaluation-2 trial demonstrated that hypotension occurring during the intraoperative period *vs* the remainder of the operative day had similar time-weighted associations with myocardial infarction and mortality,[17] and a subsequent study revealed that hypotensive events are particularly prolonged and severe in the postoperative period, but often go unrecognised and untreated in this setting.[19]

Intraoperative predictions of postoperative hypotension therefore have potential to improve patient- and population-level outcomes by prompting interventions that reduce the severity or duration of postoperative hypotension.[11,12] For example, this information could prompt more frequent vital sign measurements in selected patients during recovery and continuation of intraoperative infusions or monitoring, or facilitate identification of actionable factors in individual cases (e.g. hypovolaemia, neuraxial anaesthetic medication dosing, or adrenal insufficiency). However, no validated clinical decision support tools that help predict or prevent hypotension in the postoperative setting exist. Additionally, the interaction of anaesthesiologists with ML interfaces has been examined in other perioperative contexts, such as prediction of impending intraoperative hypoxaemia,[20] but human factors impacting the clinical implementation of ML decision-support systems in the PACU may be context specific and have not previously been examined.

In this study, we developed a clinically informed and interpretable ML model to perform end-of-surgery predictions of PACU hypotension and identify clinical factors contributing to individual risk. We then studied the interaction of experienced anaesthesiologists with this system by measuring the impact of the model on predictions of PACU hypotension and eliciting structured feedback regarding necessary considerations for future successful integration into the clinical environment. Overall, we aspired to extend our understanding of clinician acceptance of ML models in clinical anaesthesiology workflow.

Methods

This study was approved by the University of Washington (UW) Institutional Review Board (STUDY00005331 and STUDY00011636 [exempt]) and by the UW School of Medicine for legal, privacy, and compliance.

This mixed methods study was a partnership between the UW Department of Anaesthesiology and Pain Medicine and the Microsoft Corporation (MSR) with three phases: (i) hypotension ML model development, (ii) clinician validation of the model, and (iii) qualitative examination of clinician acceptance of the model. The partnership included MSR team visits to UW Medicine to learn about clinician challenges with PACU hypotension, observations of patient care, and in-person interviews with healthcare providers to understand the clinical and organisational context of hypotension management. The UW–MSR team also held weekly teleconference meetings for acculturation to the issue of PACU hypotension. All this informed co-development of the clinically informed ML model and feature identification.

Machine learning model development

Inclusion and exclusion criteria

Surgical procedures performed on adults (≥18 yr) from 2015 through 2019 resulting in PACU stays were included. Cardiac, obstetric, non-operating theatre, and procedures missing more than 40% of key features were excluded (Supplementary Table 1).

Data description and feature derivation

The UW used an integrated perioperative information management system (Merge AIMS, Inc., Hartland, WI, USA and Cerner, Inc., North Kansas City, MO, USA). Data were extracted from two UW hospitals and de-identified using the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor method.[21] Static and time-dependent model features were then derived from this limited data set composed of routinely captured clinical variables from the electronic health record (Supplementary section 3).

For time-varying measures, we hypothesised that physiological responses to anaesthetic and surgical interventions and the interactions of these changes with postoperative hypotension would vary over the course of surgery. We, therefore, split the surgical period into three clinically relevant phases: (i) *induction* of anaesthesia to procedure start, (ii) *maintenance* (procedure start to procedure end), and (iii) *emergence* (procedure end to last recorded BP [proxy for departure from operating

theatre]). The complete set is listed in Supplementary Table 2. Artifacts in the vitals data were detected and filtered from analysis, described in Supplementary section 2. Predictions of PACU hypotension were made at the time of operating theatre departure (end of emergence).

Hypotension outcome definition

Blood pressure data comprise nurse-validated recordings of BP (invasive or noninvasive) made approximately every 15 min within PACU. We reviewed the histogram of postoperative BP recording frequency and arrived upon a threshold of 10 min, an estimated upper time limit within which a nurse would typically make a new recording of BP measurement if the previous one is an artifact.

PACU hypotension was defined as at least one measurement of MAP <65 mm Hg[18] in the first 6 h of PACU admission. If MAP was not recorded, the patient was labelled hypotensive if systolic BP was measured <90 mm Hg, as these definitions are aligned with clinical escalation workflows in the hospitals, and even brief periods of mild hypotension are associated with poor outcomes.[18] Patients for whom hypotension was prevented counted as negative cases; hence, we are considering 'unanticipated' or 'unmanaged' hypotension. Approximately 5% of the study cohort had BP monitored through an arterial line in the PACU. For all patients, including patients who had noninvasive BP monitoring, BPs were measured every 5–15 min. Spurious values caused by transiently low MAP readings were removed. Although use of a binary labelling system does not capture the dynamic range of hypotension experienced by patients, these definitions are aligned with recent Perioperative Quality Initiative consensus statements on perioperative hypotension[18,22] and are consistent with triggers commonly included in postoperative 'rapid response' protocols in the

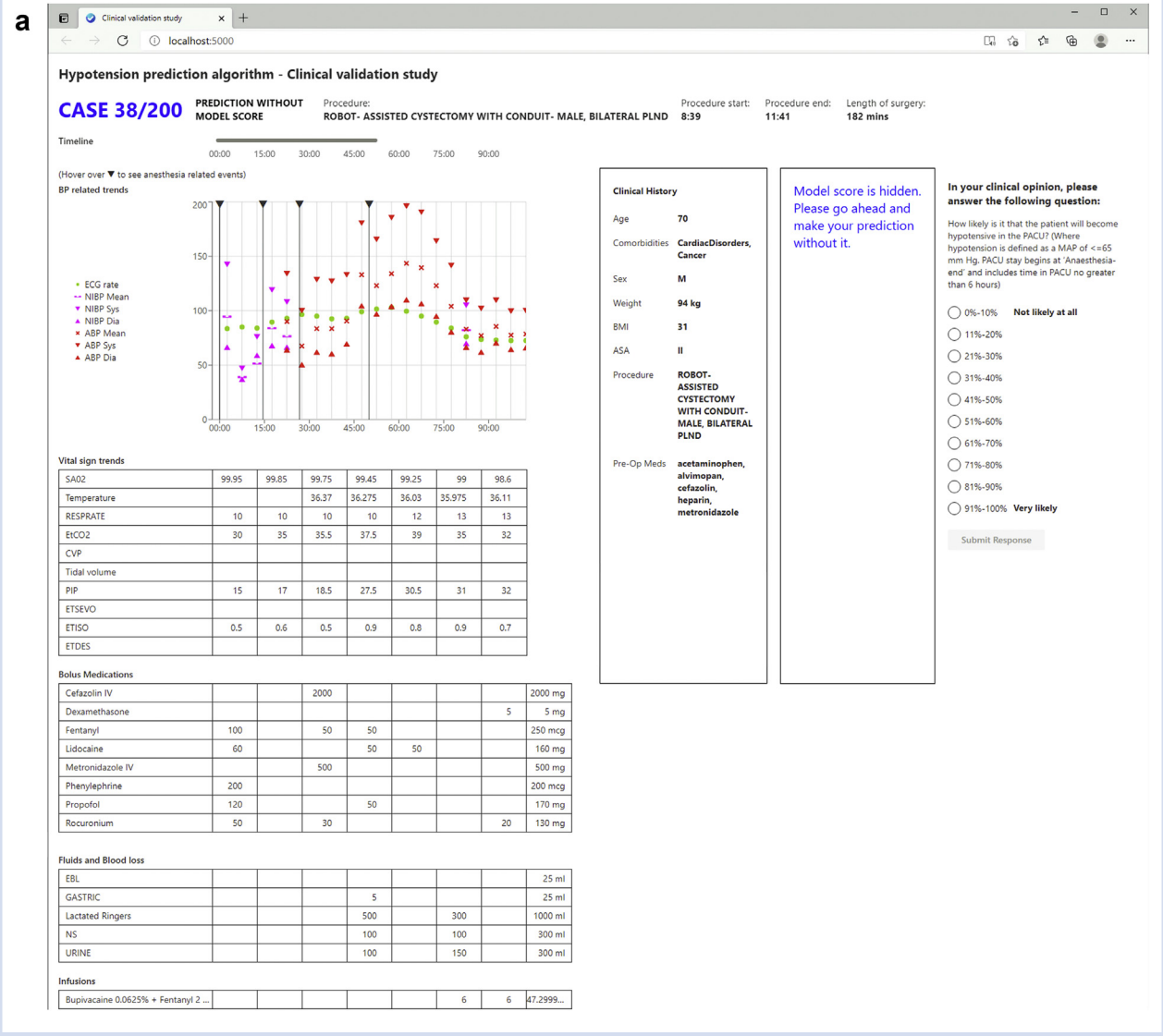

Fig 1. Clinician validation tool interface: (a) without and (b) with model score.

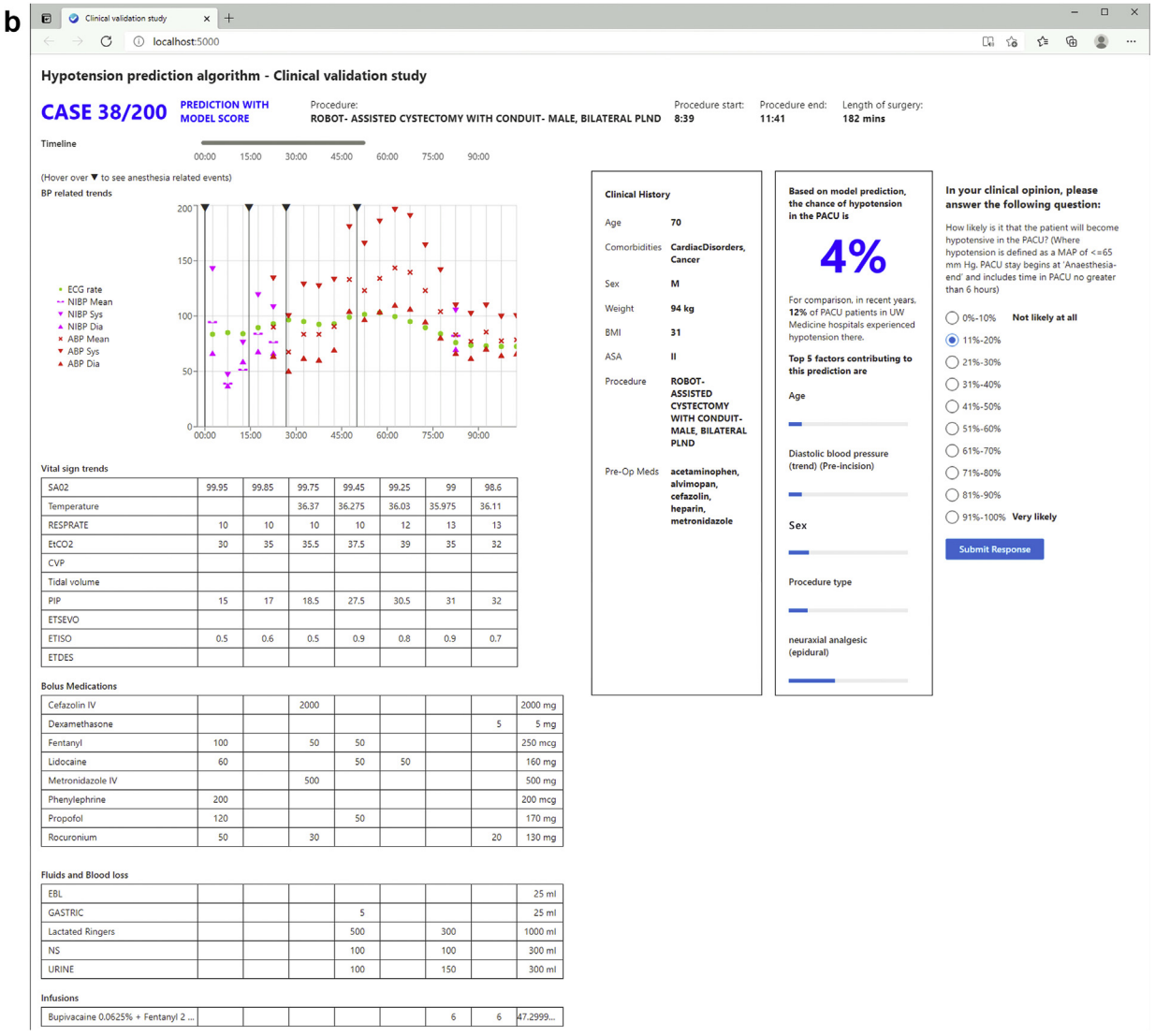

Fig 1. *(continued)*.

USA.[23] More details are in Supplementary section 'Hypotension labelling'.

Model creation, training, and testing

We divided the data from the set of included procedures into training, validation, and test sets, ensuring that patients with multiple procedures were assigned to only one set. The test set was defined as the most recent 16% of procedures, to emulate the expected use of the model and provide the closest approximation of prospective performance.[24] We trained a gradient-boosted tree model using the LightGBM library v3.1.1[25] (Supplementary Table 3), which previously demonstrated strong performance on clinical tasks, is amenable to interpretation,[26,27] and utilises computational resources potentially compatible with prospective implementation. The validation set was used for final model selection and *post hoc* calibration, and the test set was used to estimate performance. Missing-value imputation was not performed, as it is not necessary for this model class.

The model was evaluated for accuracy by calculating the area under the receiver operating characteristic (AUROC) curve and area under the precision–recall curve (AUPRC) and average precision. To evaluate calibration, we calculated the Brier score and reliability diagram. To quantify variability in the model's performance, we constructed 95% confidence intervals (CIs) using 1000 bootstrap samples from the predictions for the AUROC, AUPRC, and average precision. We processed the data set and implemented the model in Python 3.7.

Hypotension risk and feature importance

The model outcome was a probability given the vector representation of the static and time-varying features and used

Table 1 Patient characteristics used in the full and test data sets. Values are mean (SD) or number of procedures (%). 'Induction' refers to the period between the start of anaesthesia and the start of surgery. 'Maintenance' spans the surgery duration. 'Emergence' is the time between the end of surgery and the end of anaesthesia. The choice of vitals appearing here is based on the vitals that appear as most contributive to the model's predictions (see Fig. 3). Here, the average value of the depth of MAP is below 65 mm Hg, each weighted by the time until the next measurement. For medications that are represented as counts (number of times given during surgery or initiated for infusions), means refer to the mean of the count across the patient sample. ENT, ear, nose and throat; IQR, inter-quartile range; SD, standard deviation; TWA, time-weighted average.

	Full data set, n=104 875	Test data set, n=17 029
Demographics		
PACU hypotension, *n* (%)	12 601 (12.0)	1943 (11.4)
Age, *yr*, median (IQR)	54 (26)	54 (25)
Facility, *n* (%)	52 516 (50)	8613 (51)
Female sex, *n* (%)	50 279 (48)	8112 (48)
Race, *n* (%)		
White	83 721 (80)	13 832 (81)
Black/African American	7922 (8)	1263 (7)
Asian	6907 (7)	1040 (6)
American Indian	2827 (3)	448 (3)
Other	3190 (3)	427 (3)
Procedure details		
Emergency status, *n* (%)	6356 (6)	838 (5)
ASA physical status, *n* (%)		
1	11 372 (11)	1322 (8)
2	47 773 (46)	7041 (41)
3	41 334 (39)	7766 (46)
≥4	4396 (4)	900 (5)
Surgical specialty, *n* (%)		
Orthopaedic	30 515 (29)	4786 (28)
General	18 071 (17)	2748 (16)
Urology	10 738 (10)	1964 (12)
ENT	8667 (8)	1493 (9)
Gynaecology	7516 (7)	1082 (6)
Ophthalmology	6700 (6)	1016 (6)
Neurology	5795 (6)	1008 (6)
Plastic	4848 (5)	882 (5)
Vascular	2522 (2)	568 (3)
Oral and maxillofacial	2306 (2)	303 (2)
Thoracic	2304 (2)	532 (3)
Other	3887 (4)	481 (3)
Comorbidities		
BMI, kg m^{-2}, mean (SD)	28.8 (7.5)	28.9 (7.9)
Diabetes mellitus, *n* (%)	16 968 (16)	3214 (19)
Hypertension, *n* (%)	9138 (9)	1394 (8)
Vitals		
Diastolic BP (mm Hg)		
Induction (mean)	67 (12)	67 (12)
Maintenance (mean)	64 (12)	65 (12)
Emergence (mean)	71 (16)	71 (16)
MAP (mm Hg)		
Induction (mean)	82 (13)	83 (13)
Maintenance (mean)	79 (12)	80 (12)
Emergence (mean)	87 (17)	87 (17)
Systolic BP (mm Hg)		
Induction (mean)	114 (18)	115 (18)
Maintenance (mean)	111 (16)	112 (16)
Emergence (mean)	121 (22)	121 (22)
Drugs		
Vasopressor infusion i.v. count, median (IQR)	0 (1)	0 (2)
Opioid infusion i.v. count, median (IQR)	0 (0)	0 (0)
Anti-emetic bolus i.v. count, median (IQR)	1 (1)	2 (1)
Other		
Estimated blood loss (ml), median (IQR)	5 (50)	5 (50)
Intraoperative hypotension (TWA; mm Hg * h), median (IQR)	0.5 (1.6)	0.5 (1.5)

as the risk score of unanticipated or unmanaged PACU hypotension. The importance of each model feature was assessed using Shapley Additive Explanations.[27] Feature importance was determined by ranking each feature by its mean absolute Shapley value across all predictions in the test set.

Clinician validation of machine learning model

To assess the ML model's potential as an assistive tool, we compared the predictive performance of anaesthesiologists with and without exposure to model predictions. We developed a web-based visualisation tool to display preoperative and intraoperative information with content and format similar to actual anaesthesia records. This tool was used to 'play back' procedure data with and without assistance from ML model predictions (Fig. 1a and b). The tool had the capability to present the model's PACU hypotension risk score and associated clinical features for each anaesthesia record.

University of Washington anaesthesiologists with more than 2 yr of consultant-level experience were recruited (four from one hospital and five from the second hospital) to review a set of randomly selected electronic anaesthesia records. None of the study investigators participated in the validation component of this study. After enrolment and consent, we presented participants with the study purpose, how the model was developed, and with anaesthesia cases. Participants were informed that the incidence of hypotension in the cohort used to train the model was 12%.

Anaesthesia case records were randomly selected from the test data set sampled at a ratio of 1:2 hypotensive to non-hypotensive cases to ensure sufficient exposure to positive cases. These cases were selected at random, with distribution matching across age, gender, ASA, length of surgery, and facility, to ensure a representative cohort. We asked clinicians to estimate the probability of PACU hypotension at the end of surgery based on preoperative and intraoperative course in 10% increments between 1% and 100% (Fig. 1a and b). For each procedure, they first made a prediction without the model risk score. They were then presented with the model risk score, along with the top five contributory features, and made an updated prediction. No other clinical information was provided.

Statistical analysis

We estimated the model's impact on anaesthesiologist performance using the Obuchowski–Rockette[28] method to compare AUROC under a factorial design. This method is used to analyse multi-reader multi-case (MRMC) studies[29] (Supplementary section 6 details 'Statistical details of OR method for clinical validation'). For all other comparisons, we estimated the significance of differences in AUROC performance using the DeLong and colleagues[30] test of AUROC.

Sample size calculation

The validation study design was agreed upon before the study's initiation. The primary outcome was 'non-equivalence' of the two modalities: clinician AUROC when presented with the model and without it. A sample size for detecting an effect size of 0.05 (P-value=0.05) with 80% power was estimated as at least 125 cases for nine readers using the MRMC model and the software Multi-Reader Sample Size Program for Diagnostic Studies[31]; estimates of intra- and inter-rater variance were not known and estimated in Supplementary section 6.

Clinician acceptance of the model

After completing prediction tasks, each anaesthesiologist completed a four-question electronic questionnaire addressing validity, potential for clinical action, and potential for model adoption, which were consistent with the study aims: (i) how useful the PACU hypotension risk score would be for decision-making and communication concerning postoperative care, (ii) what actions prediction would enable, (iii) reasons why the model risk score might not be useful, and (iv) suggestions that would make it more likely for them to adopt it in their practise. We then interviewed four participants to elaborate on responses. These interviews were recorded and transcribed. The comments and transcripts were analysed by two investigators (PG and KOH) using an inductive grounded theoretical approach, in which excerpts are coded and clustered into an emerging set of themes.[32]

Results

Patient characteristics and primary outcome

The initial data set consisted of 144 468 procedures. After filtering and excluding 3183 procedures for data quality, 3478 for non-applicable procedure types, and 32 932 for recovery in non-PACU areas, the final data set included 104 875 cases. The most recent 17 029 (16%) procedures according to year were held out as a test set, with the remaining 87 846 (84%) used for model development (training and validation sets; Table 1). PACU hypotension was documented in 12% of cases, with a mean and median duration of 38 and 20 min, respectively. The median sampling frequency for BP measurements in the PACU was every 15 min, and 5% of patients had invasive BP monitoring postoperatively.

Model performance

The model achieved an AUROC of 0.82 (95% CI: 0.812–0.832) and AUPRC of 0.4 (95% CI: 0.38–0.42) with an average precision of 0.40 (95% CI: 0.377–0.420) in the held-out test set. Receiver operating characteristic and precision–recall curves are shown in Figure 2a. We illustrate screening implications of clinical triggers at different operating points along these curves in Table 2, based on a daily average of 44 cases going to PACU with six patients developing hypotension. At an operating point of 33% precision and 61.6% recall, 11 positive predictions would need be made to detect four true cases, whilst the remaining two would be missed. Predictions were well calibrated across the range of clinical risk, achieving a Brier score of 0.093, as seen in Figure 2b. (An improved Brier score of 0.083 was achieved with *post hoc* calibration using isotonic regression.) There was no significant difference in model performance based on age, ASA score, comorbidities, emergency status, length of surgery, race, or severity of postoperative hypotension (computed as the time-weighted average depth below 65 mm Hg).

Feature importance

The 25 most important features contributing to model prediction are reported in Figure 3. Amongst static features,

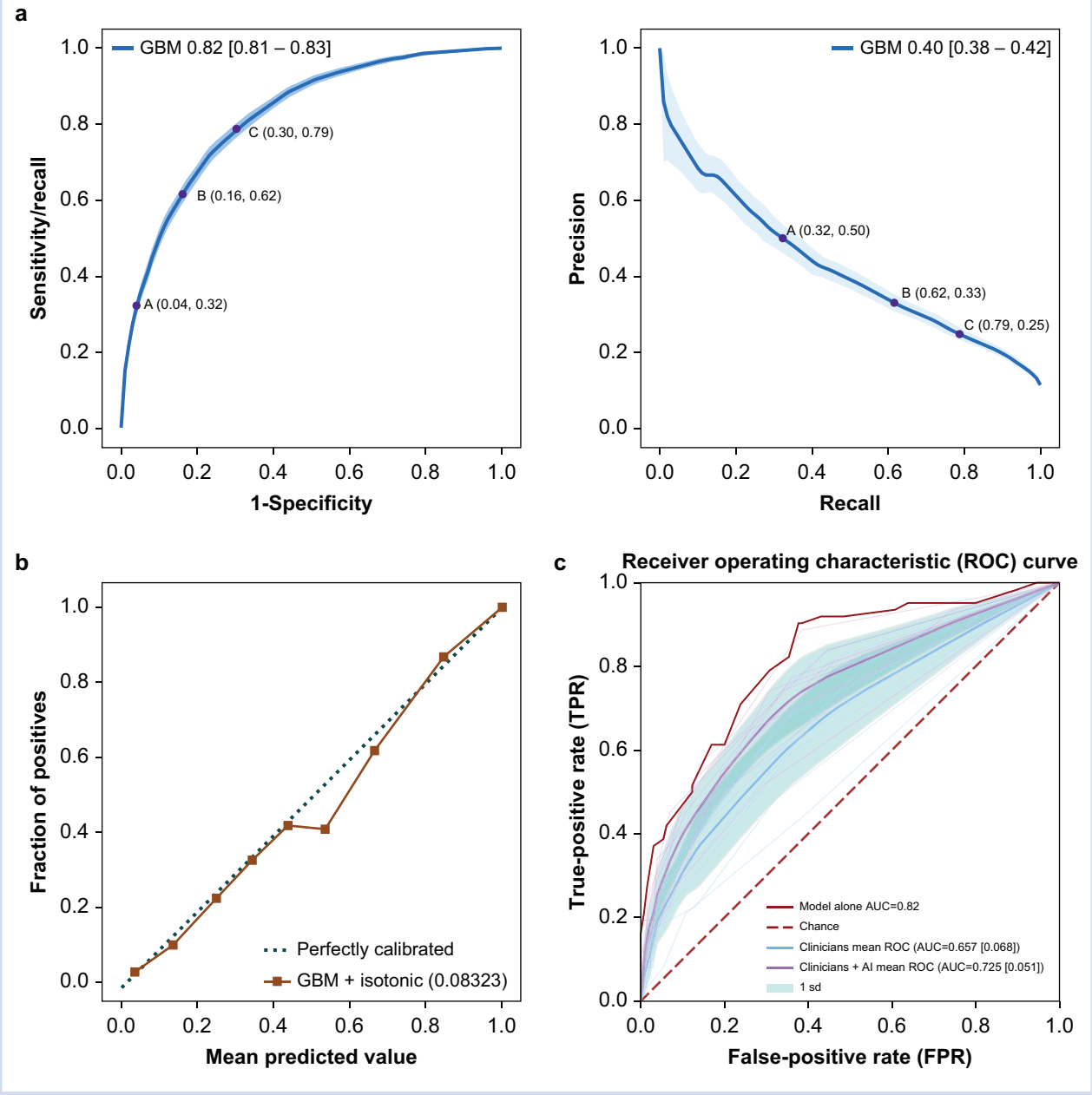

Fig 2. Gradient boosting model performance: (a) AUROC 0.82 (95% CI: 0.812–0.832), area under precision–recall 0.4 (95% CI: 0.38–0.42). (b) Calibration diagram after isotonic regression (diagonal line indicates perfect calibration). To produce this curve, the model's predictions are grouped into 10 buckets and compared with the fraction of positive labels in that bucket. The dark line plots are the mean of 1000 bootstrap samples. The 95% CI was determined from the 2.5th and 97.5th percentiles of the bootstrap sample statistics. (c) ROC curves of model alone, clinicians alone, and clinicians provided with model predictions, in the validation study cohort. The AUROC for 'clinicians' and 'clinicians+model' are computed by averaging the true positive rate of the nine clinicians over the false-positive rate points. The shaded area indicates[1] SD. AUC, area under the curve; AUROC, area under the receiver operating characteristic; CI, confidence interval; ROC, receiver operating characteristic; SD, standard deviation.

procedure type, sex, patterns of intraoperative medication administration, age, severity intraoperative hypotension, ASA physical status classification, and estimated blood loss significantly affected the prediction risk score. Amongst time-varying features, average volatile anaesthetic dose and phased intraoperative BP vectors were ranked highest. Across all three procedural phases, diastolic BP measurements were most predictive of PACU hypotension.

Table 2 Operating points for three different precision–recall and sensitivity–specificity pairs are shown on the curves and in the table.

Operating point	Precision	True positive: false positive	Sensitivity/recall	Specificity	Daily number of positive prediction	Daily false negatives	Daily true positives
A	50	1:1	32.3% [27.3–37.4]	96% [94.9–96.0]	3.9	4.1	1.9
B	33	1:2	61.6% [57.6–66.7]	83.8% [82.8–84.8]	11.2	2.3	3.7
C	25	1:3	78.8% [74.7–81.8]	68.7% [66.7–71.7]	18.9	1.3	4.7

Clinician validation of the ML model

Using the case playback tool, nine anaesthesiologists reviewed the same 192 cases each. ROCs for clinician prediction with and without support from the model are shown in Figure 2c. The positive impact of exposure to the model's predictions on clinician performance was significant ($P=0.0033$; Obuchowski–Rockette[28] test) with an average AUROC effect size of 0.067 (95% CI: 0.03–0.11), indicating better performance of clinician predictions with model support (AUROC 0.74; 95% CI: 0.68–0.79) *vs* without (AUROC 0.67; 95% CI: 0.60–0.73). Overall, performance of the model alone (AUROC 0.82; 95% CI: 0.76–0.88) was better than clinician performance both with ($P=0.014$; test of DeLong and colleagues[30]) and without ($P<0.0001$; test of DeLong and colleagues[30]) exposure to information from the model. The impact of the model on average estimated risk and the breakdown by individual clinician are further explored in Supplementary Figures 2 and 3. The additional difference after seeing the ML model was small (+2.5% to the risk estimate). Some participants made large adjustments, but most were small. Two participants made opposite sign adjustments on average. Further analyses are available in the Supplementary information.

Clinician acceptance of machine learning model

Qualitative feedback from questionnaires and interviews revealed five key themes detailed in Table 3. The key themes were (i) opportunity for prospective care planning and reflection on intraoperative care; (ii) hypotension risk prediction was of perceived value for PACU staff and handover discussions; (iii) clinician trust in model accuracy depends on alignment with clinician intuitions, and the model has greater value when the model risk score is unexpected; (iv) top-ranked feature list insufficient for clinicians to understand discrepancies with the model; and (v) use of a single threshold for defining hypotension may be overly simplistic.

Clinician feedback addressed the possible utility and situation of the model, highlighting the potential for informing decisions about PACU care alongside facilitating reflective awareness about intraoperative care decisions. Anaesthesiologists indicated that the hypotension risk score could influence handover communications with PACU nurses, potentially leading to increased monitoring frequency, creation of proactive orders for postoperative fluid administration, and establishment of early triggers for escalation of care.

Discussion

The novelty of this work lies in information learned about clinician perception and acceptance of a contextually relevant and clinically informed perioperative ML model. In this mixed methods study, an ML model for PACU hypotension utilising routinely collected electronic medical record data demonstrated predictive performance exceeding that of experienced anaesthesiologists when those anaesthesiologists were viewing a low-fidelity patient simulation on a web application. Exposure to information from model predictions (risk scores and ranking of contributing features) improved anaesthesiologist predictions of PACU hypotension. Clinicians expressed value and trust in ML model technology and themes important for successful clinical implementation. Compared with prior work on continuous prediction of impending intraoperative hypoxaemia,[20] we observed similar improvements in anaesthesiologist performance on this distinct clinical prediction task, but note that clinicians expressed the desire for more information about individual features underlying ML predictions of PACU hypotension, especially when these estimates diverged from their expectations. The presentation of such individual explanatory variables may be of particular importance in ML systems positioned at points of perioperative transitions of care, as this contextual information is essential to handoff communication surrounding anticipated future events, which must be managed by a receiving provider. Our study suggests that use of clinically informed, automated predictive systems may be well received by anaesthesiologists and used in clinical environments to facilitate improved detection, prevention, and management of PACU care. These results provide a use case for modifying current approaches to PACU hypotension and could be used to motivate and inform clinical trials.

Prior work has demonstrated the ability of ML to predict acute hypotension for patients receiving haemodialysis,[33] after the induction of anaesthesia,[13] during the intraoperative period,[14] and for critically ill patients.[24,26,34,35] ML-derived predictions of intraoperative hypotension based on arterial pressure waveform data provide strong near-term performance,[15] leading to reduction in hypotensive events.[14] Our study provides new information on how end-of-surgery predictions of postoperative hypotension may enhance the ability of anaesthesiologists to anticipate changes in patient status during recovery in the PACU. We purposely made the prediction at the end of surgery to account for the dynamic nature of the intraoperative period and to produce timely information at the point of operating theatre to PACU transitions of care. The focus on PACU hypotension is important because of the large volume of post-surgical patients who receive care in a PACU where they remain at risk of hypotension. Additionally, hypotension monitoring intensity is reduced in the PACU after surgery. Finally, organ dysfunction and mortality attributable to hypotension may occur in the postoperative period, which includes the PACU, amongst low-to moderate-risk surgical populations.[2,17,19,36]

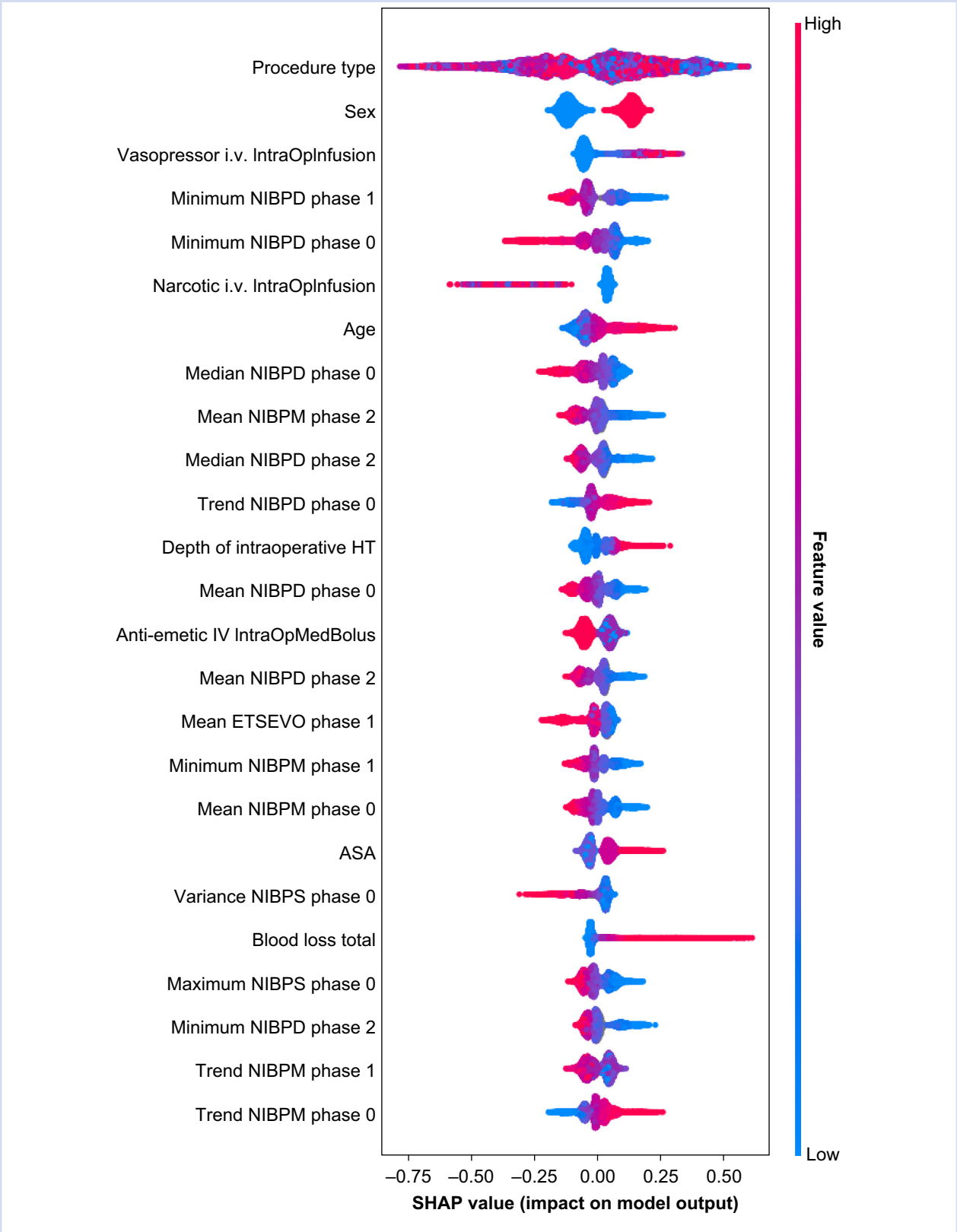

Fig 3. Top 25 important features for the test set of 17 029 cases: the y-axis indicates the features in order of importance from top to bottom. On the x-axis, the SHAP value indicates the change in log-odds. Gradient colour indicates the original value for that variable, and each point represents a sample from the test set. Procedure type is categorical, so colour gradation is not meaningful here. ASA, ASA physical status classification system; ETSEVO, end-expired sevoflurane concentration; HT, hypotension; NIBPD, noninvasive diastolic BP; NIBPM, noninvasive mean BP; NIBPS, noninvasive systolic BP; SHAP, Shapley Additive Explanations. For sex, blue/red indicates men/women, respectively. For medications (opioid i.v. and anti-emetic i.v.) represented as counts in the model blue/red indicates fewer/higher frequency (times) of administration throughout the anaesthesia.

Table 3 Thematic analysis of the feedback on the tool. Key themes revealed by analysing qualitative feedback from anaesthesiologist questionnaires and interviews.

Theme 1: hypotension prediction as a component of holistic postoperative risk assessment: opportunity for prospective care planning and reflection on intraoperative care
A 'Perhaps if the model predicts a high risk of PACU hypotension, that suggests I should have managed the anaesthetic differently for this patient, and the model could help me recognise that'.
B 'Potential uses of this risk model: (i) education of perioperative providers and improved awareness; (ii) potential strategies to prevent PACU hypotension, which may reduce PACU time and discharge to floor; and (iii) prevention of PACU hypotension with potential to prevent adverse events related to effects of hypotension on end-organ function'.
C 'It would typically take a lot more to be done before we could switch them to an ICU disposition, so the hypotension predictor alone would not move the dial to start talking about getting an ICU bed. It would have to be hypotension plus something else'.
Theme 2: hypotension risk prediction was of perceived value for PACU staff and handover discussion
A 'I would communicate that this patient is at increased risk and ask them to have a heightened awareness of that possibility and communicate with me if they are starting to see some degree of hypotension. It is reasonable to consider making sure that there is a fluid bolus option in the postoperative orders assuming that is a reasonable first step to treating hypotension. Asking the nurse to keep a closer eye and let me know if it trends downward'.
B 'Of course, the resource allocation in the PACU can be challenging. If the nurses knew [the risk], then they would need 1–1, dedicated, frequent monitoring of vital signs and assessment. Sometimes, we do that but not with any risk profile in mind. The score can help take it to the next step and define them better (e.g. check complete blood count, etc. Would help us be more proactive; minimise the AUC for blood pressure'.
C 'If everyone had similar levels of confidence in the model, that would have some value. Some PACU if they have been told that they have been assigned this patient coming out of operating theatre and if they have had time, they will have read over the chart and may have already brought a higher level of preparation to what they are doing. Like everywhere else, there are variations in ability and practice amongst PACU nurses so it might be most helpful for less experienced PACU nurses or nurses that did not have time to read the whole chart and think "oh yeah, when I see these patients, then I typically have to do this" or "I am expecting a longer PACU stay".
Theme 3: model has greater value when model risk score is unexpected
A 'It could help with identifying cases that might have an unrecognised high risk for PACU hypotension that I did not identify as being high risk'.
Theme 4: top-ranked feature list insufficient for clinicians to understand discrepancy with model
A 'It made me—when I was interpreting the indicators suggested a certain amount of weight—I would look at those and it still did not get me closer to confidence one way or the other, so I was not sure whether to throw out my estimate entirely or just trust the model. I suppose if I had enough info about how the model was generated and how deep the data were that were on it—so it kind of just left me just wondering when it was different from my prediction or experience was—who was wrong—I was not convinced I was entirely wrong and I was not convinced the model was entirely right'.
B 'Would need to see the data and peer review'
Theme 5: use of standard threshold labels may be overly simplistic
A 'Whilst the risk score is at times helpful, it seems to overpredict minor hypotension in teens and younger women who have normal BP for these age groups, so no treatment would be necessary'.

We uniquely developed a model that uses data from multiple phases of preoperative and intraoperative care to facilitate safer postoperative care transition. A key contribution of the study is in the examination of clinician acceptance and clinician validation of the model. This study indicates that anaesthesiologist performance in a clinically important prediction task could be improved by incorporating ML models during surgery. Despite our finding that the model alone outperformed clinician prediction both with and without the addition of ML prediction, we cannot conclude that ML models should replace clinician predictions. Clinical practice extends beyond simple predictions, and ML models have limitations. Clinician performance may have been worse than clinical practice because of lack of additional inputs (e.g. physical examination). Clinicians may outperform ML models if sufficiently trained to make predictions using ML models or if clinician variability in expertise and experience is decreased. Our finding that clinician performance can be augmented in the perioperative environment by exposure to model risk score is not only novel but offers a new tool for prevention of PACU hypotension by optimising end-of-surgery care and preparing PACU staff for patients at risk of deterioration. Specifically, implementing algorithmically informed decisions based on technology could guide anticipatory strategies, such as protocols for more frequent postoperative vital sign monitoring in select patients, proactive continuation of intraoperative fluids or vasopressors in PACU, optimising PACU nursing assignments for probable patient needs, and enhanced handover discussions between operating theatre and PACU teams, including establishment of treatment plans for hypotension and triggers for care escalation.

Integration of ML models into clinical decision support aids, either as part of an electronic medical record system or as an add-on technology, requires careful design considerations, expensive programming, and time-consuming development work. As an intermediate approach, we developed a web-based case playback tool to perform a cost-effective clinician validation of ML models. The study allowed us to assess clinician validation of ML models and provided initial feedback on usability that may motivate a clinical trial.

Qualitative interviews with experts highlight the complexity of human interactions with automated modelling approaches. Anaesthesiologists commented on the relatively simplistic prediction factors we used and expressed they and health systems would be unlikely to willingly cede ultimate patient care and decision-making authority to an algorithm.

Social determinants and unrecorded coexisting conditions, clinician biases, or technology flaws may all affect PACU hypotension predictions but would not be captured by the electronic health record, thereby affecting model performance. Importantly, thematic content analysis in this study revealed clinician acknowledgement of the value of using ML models to optimise postoperative outcomes, including PACU hypotension.

Our co-development and clinically informed approach to interpretable ML model development provides insight into clinical features associated with individual risk. The value of identifying the underlying causes of hypotension, in addition to prediction of overall risk, has recently been highlighted as an essential element of future ML models for perioperative hypotension that is needed to realise the clinical benefits of these technologies.[37] Although the factors we identified were consistent with clinical expectation, an unexpected finding was the strong influence of female sex associated with an increased likelihood of PACU hypotension. Sex differences in BP over the lifespan studied in the context of hypertension suggest that BP thresholds for hypertension-associated risk differ by sex and that lower treatment targets may be warranted for women.[38,39] Our study design did not allow us to determine the interaction of BP with sex on postoperative outcomes, but it demonstrates a need for future research on the interactions between sex, hypotension, and postoperative outcomes.

Our study has strengths and limitations. Strengths are the use of mixed methods to examine feasibility of end-of-surgery PACU hypotension prediction, examination of clinician acceptance of ML model implementation, use of large data set and electronic health record data, inclusion of diverse procedure types, use of state-of-the-art ML methods with and complementary approaches to understanding model performance, inclusion of a clinically diverse patient population, and use of the most recent subset of cases for internal validation. We used expert input and clinically and contextually relevant anaesthesia care phases to make the prediction. Although we use standard ML methods, the co-development approach used to create the model is novel. Limitations are use of single health system data, continued need for external validation and model comparison, temporal bias from our retrospective study design, the lack of prospective clinical evaluation, and use of a single absolute threshold for labelling hypotensive events. The use of an absolute threshold meant the model cannot predict severity of hypotension. The study tested clinicians using a web application in an unfamiliar environment without full access to data (such as text notes and visual and physical assessment of the patient), which likely resulted in an underestimate of their real-world performance on this task. Participants were informed of the 12% incidence of hypotension in the training data, whereas the prevalence in the study was chosen to be 33% to ensure sufficient numbers of positive cases. Presentation of these data might have affected their judgement, but learning effects were neither observed in the test cases or from the post-study feedback gathered from them. The experimental design for physicians making updated predictions may be biased compared with real-life behaviour, and that forcing an explicit estimate will anchor the participant to that original estimate. Despite limitations, this study provides new information to better understand clinician acceptance of a real-time ML model to advance clinical anaesthesia practice.

In conclusion, we developed and evaluated a clinically informed real-time end-of-surgery ML model that clinicians acknowledged may help them predict and prevent PACU hypotension. The risk score and list of contributing individual factors generated by the model enhanced clinician predictions, demonstrating the need for future model testing and potential value of this approach for handovers and proactive planning of PACU care. Clinician feedback suggests that ML models may be extended to encompass a holistic assessment of perioperative risk spanning different points in the perioperative journey. Optimising interactions between ML models and clinician decision makers will be key to successfully integrating model predictions into perioperative environments.

Authors' contributions

Problem proposal: MSV, WVC, DRL, RP, PG, KOH, KPo, KPa, SLH, BN
Setting up of infrastructure/defining of clinical features: MB, YP, BW, KPa, SLH, BN, CF, DRL, WVC, MSV, RP, MB, KT
Performance of experiments: MB, PG, KPa, BW, YP
Acquisition of clinical data: BN, CF
Analysis of results: KPa, SLH, WVC, MSV, KPo, RP, DRL, BN, KOH
Clinical evaluation of results: DRL, WVC, MSV, RP
Statistical analysis: KPa, SLH
Drafting of paper: KPa, SLH, WVC, MB, PG, MSV, KPo, DRL, KOH, RP, BN, KT
Revising of paper: KPa, SLH, WVC, MB, PG, MSV, KPo, DRL, KOH, RP, BN, KT

Acknowledgements

The authors thank Robert Fabiano, Roland Lai, and Tony Wieser for tool development; the University of Washington (UW) Anaesthesiology Faculty who participated in model validation and the qualitative study; and UW School of Medicine for data agreements. The authors also acknowledge Leanne Cornel for support of the conduct of the study and submission of paper.

Declarations of interest

KPo has been supported by a grant funding from Mathematica Policy Research for a validation study of an intraoperative hypotension quality measure. BN owns equity in Perimatics LLC and is its chief solution architect. The other authors declare that they have no conflicts of interest.

Funding

US National Institutes of Health/National Institute of General Medical Sciences 5T32GM086270-12 Anesthesiology and Perioperative Medicine Research Training to DRL; Microsoft Research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bja.2021.10.052.

References

1. 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 surgery: a retrospective cohort analysis. *Anesthesiology* 2017; **126**: 47–65
2. Sessler DI, Khanna AK. Perioperative myocardial injury and the contribution of hypotension. *Intensive Care Med* 2018; **44**: 811–22
3. Sun LY, Wijeysundera DN, Tait GA, Beattie WS. Association of intraoperative hypotension with acute kidney injury after elective noncardiac surgery. *Anesthesiology* 2015; **123**: 515–23
4. Futier E, Lefrant J-Y, Guinot P-G, et al. Effect of individualized vs standard blood pressure management strategies on postoperative organ dysfunction among high-risk patients undergoing major surgery: a randomized clinical trial. *JAMA* 2017; **318**: 1346–57
5. Monk TG, Bronsert MR, Henderson WG, et al. Association between intraoperative hypotension and hypertension and 30-day postoperative mortality in noncardiac surgery. *Anesthesiology* 2015; **123**: 307–19
6. 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
7. van Waes JAR, van Klei WA, Wijeysundera DN, van Wolfswinkel L, Lindsay TF, Beattie WS. Association between intraoperative hypotension and myocardial injury after vascular surgery. *Anesthesiology* 2016; **124**: 35–44
8. Vernooij LM, van Klei WA, Machina M, Pasma W, Beattie WS, Peelen LM. Different methods of modelling intraoperative hypotension and their association with postoperative complications in patients undergoing noncardiac surgery. *Br J Anaesth* 2018; **120**: 1080–9
9. Walsh M, Devereaux PJ, Garg AX, et al. Relationship between intraoperative mean arterial pressure and clinical outcomes after noncardiac surgery: toward an empirical definition of hypotension. *Anesthesiology* 2013; **119**: 507–15
10. Wesselink EM, Kappen TH, Torn HM, Slooter AJC, van Klei WA. Intraoperative hypotension and the risk of postoperative adverse outcomes: a systematic review. *Br J Anaesth* 2018; **121**: 706–21
11. Davies SJ, Vistisen ST, Jian Z, Hatib F, Scheeren TWL. Ability of an arterial waveform analysis-derived Hypotension Prediction Index to predict future hypotensive events in surgical patients. *Anesth Analg* 2020; **130**: 352–9
12. Schneck E, Schulte D, Habig L, et al. Hypotension Prediction Index based protocolized haemodynamic management reduces the incidence and duration of intraoperative hypotension in primary total hip arthroplasty: a single centre feasibility randomised blinded prospective interventional trial. *J Clin Monit Comput* 2020; **34**: 1149–58
13. Kendale S, Kulkarni P, Rosenberg AD, Wang J. Supervised machine-learning predictive analytics for prediction of postinduction hypotension. *Anesthesiology* 2018; **129**: 675–88
14. Wijnberge M, Geerts BF, Hol L, et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. *JAMA* 2020; **323**: 1052–60
15. Hatib F, Jian Z, Buddi S, et al. Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. *Anesthesiology* 2018; **129**: 663–74
16. Lee S, Lee H-C, Chu YS, et al. Deep learning models for the prediction of intraoperative hypotension. *Br J Anaesth* 2021; **126**: 808–17
17. Sessler DI, Meyhoff CS, Zimmerman NM, et al. Period-dependent associations between hypotension during and for four days after noncardiac surgery and a composite of myocardial infarction and death: a substudy of the POISE-2 trial. *Anesthesiology* 2018; **128**: 317–27
18. McEvoy MD, Gupta R, Koepke EJ, et al. Perioperative Quality Initiative consensus statement on postoperative blood pressure, risk and outcomes for elective surgery. *Br J Anaesth* 2019; **122**: 575–86
19. Turan A, Chang C, Cohen B, et al. Incidence, severity, and detection of blood pressure perturbations after abdominal surgery: a prospective blinded observational study. *Anesthesiology* 2019; **130**: 550–9
20. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2018; **2**: 749–60
21. Office for Civil Rights (OCR). *Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule | guidance portal* 2020 [Internet] Available from: https://www.hhs.gov/guidance/document/guidance-regarding-methods-de-identification-protected-health-information-accordance-0
22. Sessler DI, Bloomstone JA, Aronson S, et al. Perioperative Quality Initiative consensus statement on intraoperative blood pressure, risk and outcomes for elective surgery. *Br J Anaesth* 2019; **122**: 563–74
23. Agency for Healthcare Research and Quality. *Rapid response systems . Patient safety network* [Internet] Available from: https://psnet.ahrq.gov/primer/rapid-response-systems. [Accessed 12 March 2020]
24. Hyland SL, Faltys M, Hüser M, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat Med* 2020; **26**: 364–73
25. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. In: *Proceedings of the 31st international conference on neural information processing systems*. CA: Long Beach; 2017. p. 3149–57. [Accessed 24 January 2021]
26. Rocha T, Paredes S, de Carvalho P, Henriques J. Prediction of acute hypotensive episodes by means of neural network multi-models. *Comput Biol Med* 2011; **41**: 881–90
27. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. *ArXiv170507874 Cs Stat* 2017. Available from: http://arxiv.org/abs/1705.07874
28. Obuchowski N, Rockette H. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an anova approach with dependent observations. *Commun Stat Simul Comput* 1995; **24**: 285–308
29. Smith BJ, Hillis SL. Multi-reader multi-case analysis of variance software for diagnostic performance comparison of imaging modalities. *Proc SPIE Int Soc Opt Eng* 2020; **11316**: 113160K. [Accessed 12 March 2020]
30. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44**: 837–45

31. Hillis SL, Schartz KM. Multireader sample size program for diagnostic studies: demonstration and methodology. *J Med Imaging (Bellingham)* 2018; **5**, 045503
32. Glaser BG, Strauss AL. *The discovery of grounded theory: strategies for qualitative research*. New Brunswick: Aldine Transaction; 2010
33. Gómez-Pulido JA, Gómez-Pulido JM, Rodríguez-Puyol D, Polo-Luque M-L, Vargas-Lombardo M. Predicting the appearance of hypotension during hemodialysis sessions using machine learning classifiers. *Int J Environ Res Public Health* 2021; **18**: 2364
34. Lee J, Mark RG. An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care. *Biomed Eng Online* 2010; **9**: 62
35. Saugel B, Kouz K, Hoppe P, Maheshwari K, Scheeren TWL. Predicting hypotension in perioperative and intensive care medicine. *Best Pract Res Clin Anaesthesiol* 2019; **33**: 189–97
36. Roshanov PS, Rochwerg B, Patel A, et al. Withholding versus continuing angiotensin-converting enzyme inhibitors or angiotensin II receptor blockers before noncardiac surgery: an analysis of the vascular events in noncardiac surgery patients cohort evaluation prospective cohort. *Anesthesiology* 2017; **126**: 16–27
37. Kappen T, Beattie WS. Perioperative hypotension 2021: a contrarian view. *Br J Anaesth* 2021; **127**: 167–70
38. Ji H, Kim A, Ebinger JE, et al. Sex differences in blood pressure trajectories over the life course. *JAMA Cardiol* 2020; **5**: 19–26
39. Ji H, Niiranen TJ, Rader F, et al. Sex differences in blood pressure associations with cardiovascular outcomes. *Circulation* 2021; **143**: 761–3

Handling editor: Michael Avidan