Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prediction of ten intraoperative physiologic derangements using routine monitoring and drug administration data

Jake Stephen

js6781@princeton.edu

Princeton University

Mitchell Ng

Thomas Jefferson University Hospital

Kenneth Ng

Hackensack Meridian Health

Jonathan W. Pillow

Princeton University

Armita Kazemi

Princeton University

Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11446566/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Intraoperative hypotension, hypoxemia, and related derangements are associated with postoperative organ injury, yet existing predictors target a single event and require invasive or specialized monitoring. A temporal convolutional model using only flowsheet vital signs and the drug administration record was developed on 3,724 noncardiac surgical cases from the open VitalDB registry and evaluated, frozen, on 9,720 cases from an independent academic medical center. Ten hemodynamic, respiratory, and thermal events were predicted at 3-, 5-, and 10-min horizons. Mean area under the receiver operating characteristic curve at 5 min, restricted to windows in which the patient was not already breaching the threshold, was 0.887 internally and 0.843 externally. At two alarms per case-hour, sensitivity was 0.61 and positive predictive value 0.27 internally, with a median lead time of 3.8 min that was unchanged externally. Alarm precision was modest and unusable externally for hypoxemia, supporting feasibility rather than clinical readiness.

Introduction

Adverse physiologic events during anesthesia and surgery are common, are frequently subclinical at onset, and carry measurable postoperative consequences. Intraoperative hypotension has been repeatedly and independently associated with postoperative acute kidney injury, myocardial injury, and mortality after noncardiac surgery, with risk rising as a function of both the depth and the cumulative duration of exposure below mean arterial pressure (MAP) thresholds near 65 and 55 mmHg.[1–5] An international consensus statement synthesizes this evidence and recommends active avoidance of sustained hypotension.[6] Associations with adverse outcomes, less extensively quantified, have also been reported for intraoperative hypoxemia, dysrhythmia, hypocapnia and hypercapnia, and hypothermia.

Intraoperative monitoring nevertheless remains largely reactive. Clinicians are alerted once a monitored value has already crossed a threshold and therefore receive little advance warning of an impending derangement. The most mature predictive alternative uses high-fidelity arterial pressure waveform features to anticipate hypotension up to 15 min in advance.[7] External validation confirmed strong discrimination,[8] and a randomized trial found that index-guided management reduced the depth and duration of intraoperative hypotension relative to standard care.[9] A subsequent pilot trial in moderate- to high-risk surgery found no significant reduction,[10] and comparisons against arterial waveform monitoring alone have questioned the incremental value of the prediction layer.[11] A methodological critique further showed that much of the apparent performance arises from windows in which the patient is already hypotensive, so that the task shifts in part from anticipation to detection.[12]

Three limitations persist across this literature. First, most intraoperative predictors target a single event, usually hypotension,[7,8,13–15] with hypoxemia addressed separately,[16,17] whereas patients in the operating room are exposed to several concurrent physiologic hazards. Second, high-performing systems typically assume specialized instrumentation, most often an arterial catheter for waveform features, and existing

multi-event frameworks have likewise relied on inputs beyond routine flowsheet monitoring. Third, physiologic monitors already generate a substantial false-alarm burden, with 80% to 99% of alarms in intensive care and operating room settings estimated to be clinically insignificant,[18–20] so a predictor that improves discrimination while increasing alarm burden may have limited utility; and models validated only at their development site frequently degrade when applied elsewhere.[21–24]

The authors therefore developed a continuous early-warning model, here abbreviated PRISM (Perioperative Risk Index from Standard Monitoring), that predicts ten adverse physiologic events at 3-, 5-, and 10-min horizons using only flowsheet vital signs and the intraoperative drug administration record, requiring neither an arterial catheter, nor depth-of-anesthesia monitoring, nor infusion-pump telemetry. The hypothesis was that such a model would discriminate windows preceding an event from those not preceding one among patients whose physiology was stable at the time of prediction, and would retain that discrimination when applied without retraining at an institution contributing no training data. The work is presented as a feasibility study rather than as a validated clinical system.

Results

Cohorts

Patient and case characteristics for both cohorts are given in Table 1. Age was similar between cohorts, with a wider spread externally (mean 58.3 ± 15.3 year in the development cohort vs. 56.0 ± 19.6 year externally), and the external cohort contained a higher proportion of female patients (4,927/9,720 [50.7%] vs. 1,644/3,724 [44.1%]). Per-target episode prevalence differed substantially and is the main determinant of the positive predictive values reported below: the hypoxemia family was far less frequent externally (hypoxemia 1,588/9,720 [16.3%] vs. 2,837/3,724 [76.2%]; severe hypoxemia 692/9,720 [7.1%] vs. 2,219/3,724 [59.6%]), whereas hypercapnia was more than twice as frequent externally (2,770/9,720 [28.5%] vs. 456/3,724 [12.2%]).

Table 1
Patient and case characteristics, and number of qualifying episodes per target, in the development and external cohorts

Characteristic	Development cohort	External cohort
Cases, n	3,724	9,720
Age, yr, mean ± SD	58.3 ± 15.3	56.0 ± 19.6
Age, yr, median [25th, 75th]	60 [50, 69]	61 [44, 71]
Female sex, n (%)	1,644/3,724 (44.1)	4,927/9,720 (50.7)
Male sex, n (%)	2,080/3,724 (55.9)	4,793/9,720 (49.3)
ASA Physical Status I, n (%)	793/3,724 (21.3)	NA
ASA Physical Status II, n (%)	2,243/3,724 (60.2)	NA
ASA Physical Status III, n (%)	546/3,724 (14.7)	NA
ASA Physical Status IV, n (%)	43/3,724 (1.2)	NA
ASA Physical Status V or VI, n (%)	12/3,724 (0.3)	NA
Emergency surgery, n (%)	310/3,724 (8.3)	NA
Surgical service: general, n (%)	1,489/3,724 (40.0)	NA
Surgical service: orthopedic, n (%)	745/3,724 (20.0)	NA
Surgical service: other, n (%)	1,490/3,724 (40.0)	NA
Case duration, min, median [25th, 75th]	165 [115, 240]	NA
Evaluable windows, n	163,211	1,657,569
Cases with ≥ 1 hypotension episode, n (%)	3,125/3,724 (83.9)	5,875/9,720 (60.4)
Cases with ≥ 1 severe hypotension episode, n (%)	1,781/3,724 (47.8)	2,570/9,720 (26.4)
Cases with ≥ 1 relative hypotension episode, n (%)	2,864/3,724 (76.9)	8,001/9,720 (82.3)
Cases with ≥ 1 bradycardia episode, n (%)	2,661/3,724 (71.5)	5,743/9,720 (59.1)
Cases with ≥ 1 tachycardia episode, n (%)	2,380/3,724 (63.9)	4,124/9,720 (42.4)
Cases with ≥ 1 hypoxemia episode, n (%)	2,837/3,724 (76.2)	1,588/9,720 (16.3)
Cases with ≥ 1 severe hypoxemia episode, n (%)	2,219/3,724 (59.6)	692/9,720 (7.1)
Cases with ≥ 1 hypocapnia episode, n (%)	3,347/3,724 (89.9)	8,601/9,720 (88.5)

ASA = American Society of Anesthesiologists; NA = not available. Variables marked NA were not included in the data extract provided for the external cohort.

Characteristic	Development cohort	External cohort
Cases with ≥ 1 hypercapnia episode, n (%)	456/3,724 (12.2)	2,770/9,720 (28.5)
Cases with ≥ 1 hypothermia episode, n (%)	2,053/3,724 (55.1)	3,464/9,720 (35.6)

ASA = American Society of Anesthesiologists; NA = not available. Variables marked NA were not included in the data extract provided for the external cohort.

Discrimination

In the held-out development test partition, mean stable area under the receiver operating characteristic curve (AUROC; computed only over windows in which the patient was not already breaching the threshold, see Methods) across the ten targets was 0.905, 0.887, and 0.846 at the 3-, 5-, and 10-min horizons, with corresponding standard AUROC of 0.962, 0.943, and 0.903; discrimination declined monotonically as the horizon lengthened. The primary outcome, mean stable AUROC at 5 min, was 0.887.

Per-target discrimination, averaged across horizons, is given in Table 3. Performance was consistent across targets, with stable AUROC ranging from 0.825 for hypocapnia to 0.930 for severe hypoxemia and hypercapnia. The mean difference between standard and stable AUROC was 0.057, quantifying the component of apparent discrimination attributable to windows in which the derangement was already present. Stable and standard AUROC for all ten targets at all three horizons in both cohorts are given in Supplementary Table 1.

Table 3

Discrimination by target, averaged across the 3-, 5-, and 10-min horizons, in the development test partition and the external cohort

Target	Stable AUROC, development	Stable AUROC, external	Change	Standard AUROC, development	Standard AUROC, external
Hypotension	0.867 [0.858, 0.876]	0.846 [0.828, 0.864]	+ 0.021	0.916 [0.908, 0.925]	0.916 [0.900, 0.928]
Severe hypotension	0.853 [0.830, 0.875]	0.855 [0.824, 0.884]	−0.002	0.891 [0.868, 0.913]	0.906 [0.877, 0.925]
Relative hypotension	0.882 [0.869, 0.893]	0.830 [0.806, 0.852]	+ 0.052	0.959 [0.952, 0.965]	0.943 [0.935, 0.950]
Bradycardia	0.920 [0.909, 0.929]	0.879 [0.828, 0.931]	+ 0.041	0.972 [0.967, 0.977]	0.957 [0.914, 1.000]
Tachycardia	0.840 [0.818, 0.862]	0.856 [0.839, 0.872]	−0.016	0.927 [0.909, 0.942]	0.940 [0.932, 0.947]
Hypoxemia	0.915 [0.902, 0.929]	0.805 [0.785, 0.825]	+ 0.110	0.950 [0.942, 0.958]	0.865 [0.847, 0.886]
Severe hypoxemia	0.930 [0.918, 0.942]	0.810 [0.780, 0.840]	+ 0.120	0.954 [0.946, 0.963]	0.865 [0.832, 0.900]
Hypocapnia	0.825 [0.811, 0.842]	0.779 [0.762, 0.796]	+ 0.046	0.891 [0.880, 0.902]	0.882 [0.864, 0.896]
Hypercapnia	0.930 [0.913, 0.947]	0.875 [0.860, 0.889]	+ 0.055	0.944 [0.931, 0.957]	0.923 [0.911, 0.934]
Hypothermia	0.831 [0.808, 0.857]	0.840 [0.824, 0.853]	−0.009	0.957 [0.944, 0.968]	0.964 [0.951, 0.974]
Mean	**0.879**	**0.838**	**+ 0.042**	**0.936**	**0.916**

A positive change indicates worse external performance. AUROC = area under the receiver operating characteristic curve; values in brackets are 95% CIs obtained by percentile bootstrap with resampling at the level of the surgical case. Stable AUROC is computed only over windows in which the patient was not already breaching the threshold for that target at the time of prediction; standard AUROC is computed over all evaluable windows. CIs are not reported for the row means, which are arithmetic averages across targets.

Because several targets were highly imbalanced, the area under the precision-recall curve is also reported (Supplementary Table 2). Stable values were substantially lower than the corresponding AUROC throughout and diverged most for the least frequent targets: hypercapnia and hypothermia reached stable AUROC of 0.930 and 0.831 but stable precision-recall areas of 0.078 and 0.165. Ranking quality and attainable precision are therefore not interchangeable.

Alarm budget and horizon

Sensitivity and positive predictive value (PPV) across the three budgets and three horizons are given in Table 4. Two gradients were present and ran in opposite directions. Tightening the budget raised precision and lowered sensitivity. Lengthening the horizon also raised precision: mean PPV rose from 0.20 to 0.37 between the 3- and 10-min horizons at the standard budget, and from 0.27 to 0.46 at the conservative budget.

Table 4
Sensitivity and positive predictive value across three alarm budgets and three horizons in the development test partition, averaged across the ten targets

Horizon	Conservative (≤ 1 alarm/h)	Standard (≤ 2 alarms/h)	High-vigilance (≤ 4 alarms/h)
3 min	0.52 / 0.27	0.68 / 0.20	0.82 / 0.13
5 min	0.44 / 0.36	0.61 / 0.27	0.77 / 0.19
10 min	0.33 / 0.46	0.48 / 0.37	0.66 / 0.28

Values are sensitivity / positive predictive value. Decision thresholds were both selected and scored on pre-event windows only. Per-target values at the standard budget and 5-min horizon are given in Supplementary Table 3.

Precision remained modest in absolute terms. At the standard budget and 5-min horizon, mean PPV was 0.27, so approximately one alarm in four corresponded to a real event. Variation across targets was wide and was driven principally by episode prevalence rather than discrimination: relative hypotension and bradycardia reached PPV of 0.54 and 0.51, whereas hypercapnia, the least frequent target in the development cohort, reached 0.02 and did not exceed 0.07 at any of the nine horizon-by-budget combinations evaluated. Hypercapnia is therefore not deployable at any operating point examined. Per-target operating-point detail at the standard budget and 5-min horizon in both cohorts is given in Supplementary Table 3.

Warning time

Median lead time delivered by true-positive alarms at the standard budget was 2.5, 3.8, and 5.9 min at the 3-, 5-, and 10-min horizons. The proportion of warnings arriving at least 3 min before onset was 49.8%, 65.1%, and 75.4% across the same horizons. At 10 min the longest warnings were delivered for

the slowest-developing derangements: hypothermia (median 9 min [4, 14]), severe hypoxemia (8 min [4, 12]), and hypoxemia (7 min [4, 12]); the shortest were for hypotension and relative hypotension (4 min [2, 8] each). Median lead time with interquartile range by target and horizon in both cohorts is given in Supplementary Table 4.

External validation

Averaged across targets and horizons, stable AUROC fell from 0.879 in the development test partition to 0.838 externally, and standard AUROC from 0.936 to 0.916; at 5 min the external mean stable AUROC was 0.843. Per-target changes are given in Table 3. Eight of ten targets changed by 0.06 or less, and three (tachycardia, hypothermia, and severe hypotension) performed marginally better externally. The hemodynamic, thermal, and dyscapnia targets transported essentially intact. The exception was the hypoxemia family, which lost 0.110 and 0.120, two to three times the loss of any other target.

Calibration transported without degradation. Mean expected calibration error at 5 min was 0.0030 internally and 0.0022 externally, and mean Brier score 0.035 and 0.038 (Supplementary Table 5). Discrimination was similar across age, sex, and case-duration strata, the internal-to-external difference ranging only from 0.040 to 0.045 across all seven strata (Supplementary Table 6). It did not differ materially between female and male patients: mean stable AUROC at 5 min was 0.889 and 0.884 respectively internally, and 0.847 and 0.839 externally. Strata other than sex were not prespecified and are reported as consistency checks.

Deployment behavior at the standard budget in both cohorts is given in Table 5. Sensitivity and precision both declined moderately on transfer. At 5 min, sensitivity fell from 0.61 to 0.47 and PPV from 0.27 to 0.20; at 10 min, from 0.48 to 0.39 and from 0.37 to 0.27. Specificity was essentially unchanged (0.932 in both cohorts at the 5-min standard setting). The horizon-and-budget precision gradient reproduced closely, with external mean PPV rising from 0.15 at the 3-min standard setting to 0.32 at the 10-min conservative setting.

Table 5
Deployment behavior at the standard budget of two alarms per case-hour, averaged across the ten targets, in both cohorts

Horizon	Sensitivity, development	Sensitivity, external	PPV, development	PPV, external	Median lead, development, min	Median lead, external, min
3 min	0.68	0.53	0.20	0.15	2.5	2.3
5 min	0.61	0.47	0.27	0.20	3.8	3.8
10 min	0.48	0.39	0.37	0.27	5.9	5.8

PPV = positive predictive value. Decision thresholds were re-selected on the external cohort so that the same alarm budget was held fixed across cohorts; the external values therefore describe a locally calibrated system. Lead-time statistics are conditional on an alarm having been raised.

Lead time transported essentially intact. Median lead time was within 0.2 min of the internal value at every horizon (2.3 vs. 2.5, 3.8 vs. 3.8, and 5.8 vs. 5.9 min), and the proportion of warnings delivering at least 3 min of lead was within 1.5 percentage points throughout (48.4%, 63.8%, and 74.9% externally). The exception was again the hypoxemia family: externally, PPV did not exceed 0.07 for hypoxemia and 0.03 for severe hypoxemia at any budget or horizon, and both are unusable as alarms at that site. Hypercapnia moved in the opposite direction, improving from 0.06 internally to 0.20 externally at the 10-min conservative setting.

Contribution of pharmacologic inputs

The full model was compared with an otherwise identical model supplied only with the six physiologic channels. Internally the pharmacologic channels had little effect, mean stable AUROC moving from 0.892 to 0.879. Externally the effect was substantial and in the opposite direction: mean stable AUROC rose from 0.806 to 0.838 and standard AUROC from 0.897 to 0.916. The internal-to-external difference was consequently approximately halved, from 0.086 to 0.042 on the stable measure and from 0.043 to 0.020 on the standard measure.

An illustrative deployment configuration

A system running all ten targets simultaneously at the standard budget would generate up to 20 alarms per case-hour, so a deployable configuration must select a subset. Restricting it to the three most precise targets at the conservative budget and 10-min horizon, namely relative hypotension, bradycardia, and hypotension, consumed approximately three alarms per case-hour and yielded a mean PPV of 0.71 with a median lead time of 4 min and a sensitivity of 0.22 internally; re-thresholded externally to the same budget, these targets retained a sensitivity of 0.20 with a PPV of 0.54 (Supplementary Table 7). The configuration is illustrative rather than validated, because the choice of targets was informed by test-set performance.

Discussion

A model restricted to flowsheet monitoring and the intraoperative drug administration record anticipated ten distinct physiologic derangements several minutes before onset, with warning intervals long enough to permit intervention, and retained most of that capability at an institution it had not encountered; eight of ten targets changed by 0.06 or less in stable AUROC on transfer. These results indicate that the predictive signal is present in routinely collected data and is not an artifact of a single dataset, but not that such a system should be enabled in an operating room. Alarm precision was modest for most targets at every operating point examined and was unusable for hypercapnia internally and for the hypoxemia family externally, and the external deployment figures describe a locally re-thresholded system rather than one installed without site-specific tuning.

Reporting discrimination under two conventions clarifies what is being measured. The difference between standard and stable AUROC averaged 0.057 internally, which is the component of apparent performance attributable to windows in which the derangement was already present. This is the selection effect previously identified for the Hypotension Prediction Index,[12] and it is not specific to that system; any threshold-crossing predictor evaluated without restricting to pre-event windows will benefit from it. Both quantities are reported here because conventions differ across this literature.

The results also expose a trade-off that a discrimination analysis alone would conceal. Stable AUROC was highest at 3 min and fell monotonically as the horizon lengthened, whereas precision and lead time both moved in the opposite direction and were highest at 10 min. The two analyses answer different questions: discrimination asks how well the model separates windows that precede an event from those that do not, which is easier over a short window, whereas deployment asks what fraction of alarms are correct and how much warning they give, both of which improve when the window contains more true events. That the same gradient reproduced at an independent institution suggests it is a property of the task rather than of the development data. Much of this literature reports short horizons as though they were superior; these results suggest the choice is a trade rather than a ranking, and that for a system intended to prompt clinical action the longer horizon may be preferable.

A further dissociation appeared on transfer. Discrimination and alarm precision both declined moderately, but the interval by which a correct alarm preceded physiologic onset did not change at any horizon. A model that fires less often, and with more false alarms, at a new institution nevertheless gave the same warning when it was right. This matters because lead time determines whether an alarm can change management: an alarm that is frequently correct but arrives too late is of no use, whereas one that is correct less often but reliably delivers 4 min of warning can still be acted upon, provided precision is high enough for it to be trusted. Lead time is also, unlike sensitivity and precision, not a function of where the decision threshold is placed, which may partly account for its stability. It may therefore warrant reporting as an endpoint in its own right in external validation studies rather than being inferred from the prediction horizon.

Performance on transfer was not uniform across targets, and the pattern is informative. The hemodynamic and thermal targets transported essentially intact; these are among the derangements occurring most often in the development cohort, so the model encountered many thousands of examples of the physiology preceding them. Two mechanisms could plausibly account for the targets that transported least well. The first is training-data scarcity: VitalDB contains comparatively few desaturation episodes, and a model trained on few examples has less opportunity to learn the range of ways a derangement can develop. The second is between-institution difference in how densely the relevant channel is charted, which affects both the information available in the lookback window and the reliability with which events are identified. Neither accounts for the full pattern alone. Hypercapnia was the least frequent target in the development data by a wide margin yet transported better than relative hypotension; and the channels underlying the hypotension family were documented less densely than oxygen saturation measured by pulse oximetry (SpO_2) externally, yet those targets transported intact. A two-site design cannot adjudicate between these accounts.

Supplying the model with administered agents rather than observed physiology alone improved cross-institution transfer appreciably while leaving internal performance essentially unchanged. One interpretation is that pharmacologic practice is comparatively standardized, in that the hemodynamic effect of a propofol bolus follows from pharmacology rather than local convention, whereas vital-sign documentation varies between institutions, so a model supplied only with physiologic channels must rely in part on signal characteristics that reflect local practice. This is a single comparison on a single external cohort and should be regarded as suggestive. If confirmed, it would favor, in models intended to transport, inputs whose meaning is fixed by pharmacology or standardized practice over inputs shaped by local documentation convention. Discrimination was similar in female and male patients on both cohorts and no sex-specific modeling was undertaken; because the underlying derangements have known sex-related differences in incidence, this similarity is evidence that the model was not systematically worse in one sex rather than evidence that sex is irrelevant to intraoperative risk prediction.

Several limitations bound these conclusions. Targets were defined by monitor thresholds rather than chart adjudication, so the model predicts physiologic derangement and not diagnosed complication; the link between the two rests on the cited literature rather than being demonstrated here. All results derive from a single training run on a single split and should be read as point estimates for all secondary outcomes. The external evaluation, although genuinely independent, comprises a single additional institution; two sites establish that one transfer succeeded, not that the model generalizes. Thresholds were re-selected externally to hold the alarm budget fixed, so the external deployment metrics represent an upper bound on what an uncalibrated installation would achieve. Case mix, American Society of Anesthesiologists Physical Status, emergency status, and case duration were not available in the external data extract, which limits characterization of that cohort and therefore assessment of how far the transfer result extends to other case mixes. The alarm budgets were selected analytically rather than validated with clinicians. Lead-time statistics describe warning quality rather than coverage. The

illustrative configuration was selected by inspecting test-set results and requires prespecified confirmation.

The constraint bearing most directly on the remaining weaknesses is the size and completeness of available intraoperative monitoring data. Documentation is uneven across channels, pharmacologic records are partial, and the least frequent derangements appear few times, consistent with the specific failures observed. Progress therefore runs through data before architecture, requiring training across multiple institutions so that variation in population and recording practice is encountered during learning rather than at deployment, sufficient scale for rare events to be represented in the thousands, more complete pharmacologic records, and prospective silent deployment to measure alarm burden and lead time. Whether acting on such predictions improves patient-centered outcomes would require a randomized trial, following the precedent set for hypotension prediction.[9] In summary, non-invasive multi-event intraoperative prediction is feasible from routine monitoring, delivers warning intervals in which intervention is possible, and retains accuracy at an unseen institution, with the quantity and breadth of training data rather than the modeling approach appearing to be a principal limiting factor.

Methods

Ethics approval and reporting

This retrospective analysis used two previously collected cohorts and did not require institutional review board review, because neither dataset contained identifiable patient information and the research team had no access to identifiers or any means of re-identifying patients. The development cohort was drawn from VitalDB, a publicly available registry that the contributing institution de-identified before release; review board approval and consent for the registry were addressed at the time of collection (Seoul National University Hospital Institutional Review Board, protocol H-1408-012-599, approved 12 August 2014).[25] The external cohort was provided by a partner academic medical center, which de-identified the data before transfer under a data use agreement and whose review board approved the original data collection and waived written consent. Under that agreement the institution is not identified and the data cannot be released. Because both datasets were de-identified at the source and the analysis involved no identifiable private information, it did not constitute human subjects research, and review by the Princeton University Institutional Review Board was not sought. Reporting follows the TRIPOD + AI statement,[26] and risk of bias was self-assessed with PROBAST + AI.[27] This study is a retrospective analysis of previously collected data and is not a clinical trial; no trial registration applies.

Study design and clinical task

The model performs continuous early warning. At every timestep of a case, from a rolling 10-min window of monitoring and drug administration data, it predicts whether each of ten adverse events will occur within each of three future horizons (3, 5, and 10 min). An event was defined as a breach of the threshold

for its canonical channel sustained for at least 1 min within the horizon. A window was evaluable for a given target only if at least one sample of that channel was observed within the horizon; windows in which the channel was entirely unrecorded were excluded from that target's loss and metrics rather than scored as negatives, because scoring them as negatives would train the model to read sparse documentation as physiologic stability.

Prediction targets

The ten targets and their thresholds are given in Table 2, spanning hemodynamic, respiratory, and thermal derangements. Two nested severity pairs, hypotension with severe hypotension and hypoxemia with severe hypoxemia, encode the constraint that the probability of the severe event cannot exceed that of its base event; this was enforced during model fitting. Relative hypotension, a fall of more than 20% below a pre-induction baseline, is clinically important in chronically hypertensive patients, in whom an absolute MAP of 70 mmHg may already represent a dangerous relative fall.[1]

Table 2
The ten prediction targets

Target	Definition	Clinical basis
Hypotension	MAP < 65 mmHg	Standard intraoperative hypotension threshold
Severe hypotension	MAP < 55 mmHg	Severe tier; organ-injury risk
Relative hypotension	> 20% below pre-induction baseline	Relevant in chronic hypertension
Bradycardia	Heart rate < 60 beats/min	Standard threshold
Tachycardia	Heart rate > 100 beats/min	Standard threshold
Hypoxemia	SpO_2 < 90%	Clinically significant desaturation
Severe hypoxemia	SpO_2 < 85%	Severe tier
Hypocapnia	End-tidal CO_2 < 30 mmHg	Over-ventilation
Hypercapnia	End-tidal CO_2 > 50 mmHg	Requires management
Hypothermia	Core temperature < 35°C	Standard threshold

Each target is a breach of the stated threshold on its canonical channel sustained for at least 1 min within the horizon. All ten targets were predicted at 3, 5, and 10 min. CO_2 = carbon dioxide; MAP = mean arterial pressure; SpO_2 = oxygen saturation measured by pulse oximetry.

Data sources and cohorts

The development cohort comprised 3,724 noncardiac surgical cases from VitalDB, an open multi-parameter vital signs database.[25] Only numeric flowsheet-resolution physiologic channels and the drug administration record were used; the high-frequency waveform data VitalDB also provides were excluded, consistent with the objective of a model deployable in ordinary operating rooms. Cases were partitioned by patient into training (n = 2,979), validation (n = 372), and held-out test (n = 373) sets. All model fitting, hyperparameter selection, probability calibration, and threshold selection used the training and validation partitions only.

The external cohort comprised 9,720 consecutive intraoperative cases from a large academic medical center with a different patient population, case mix, anesthetic practice, and monitoring environment. No external case contributed to model fitting, feature normalization, or calibrator fitting, and model weights were not retrained or adjusted. Decision thresholds were re-selected externally to meet the same alarm budget, because the predicted-probability distribution differs between sites and a transferred threshold does not reproduce the intended alarm rate. Characteristics of both cohorts, and the number of qualifying episodes per target, are given in Table 1.

Model inputs

The input was a 20-channel time series sampled at 30-s resolution over a 10-min lookback window. Six channels carried physiologic measurements: MAP, heart rate, SpO_2, end-tidal carbon dioxide tension, respiratory rate, and core temperature. Each passed through a predefined artifact pipeline in which out-of-range values were dropped rather than clamped, isolated single-sample spikes were removed, and readings implausibly extreme on one channel while all corroborating channels read normal were treated as sensor artifact; this handling is necessary because a zeroed arterial transducer would otherwise generate a spurious hypotension label. Gaps were carried forward for up to 5 min, beyond which samples were masked as unobserved rather than imputed, with an indicator channel preserving the distinction. Ten further channels carried pharmacologic exposure, each representing one commonly administered agent as the dose or infusion rate given in that interval and zero otherwise, so that the model saw intervention as well as its physiologic consequence.

Input representation. Cases were resampled to a uniform 30-s grid. Gaps were forward-filled up to a 5-min limit, beyond which samples were masked as unobserved rather than imputed, with an explicit observed-indicator channel. The lookback window was 10 min (20 steps). Physiologic channels were z-score normalized using statistics computed from the training split only; identical statistics were applied unchanged to the internal test split and to the external cohort.

Pharmacologic channels. Ten agents were represented, each as a dedicated channel carrying the dose or infusion rate administered in each 30-s interval and zero otherwise. In channel order, the agents were propofol (bolus dose, mg; infusion rate, mg/h), phenylephrine (bolus dose, μg; infusion rate, μg/min), fentanyl (bolus dose, μg), ondansetron (bolus dose, mg), rocuronium (bolus dose, mg), ephedrine (bolus

dose, mg), sugammadex (bolus dose, mg), midazolam (bolus dose, mg), succinylcholine (bolus dose, mg), and remifentanil (bolus dose, μg; infusion rate, $\mu g \cdot kg^{-1} \cdot min^{-1}$). These ten channels were concatenated with the six normalized physiologic channels and the corresponding observation-indicator channels to give the 20-channel input.

Artifact pipeline. Applied in order during resampling: out-of-range values dropped rather than clamped; isolated-spike (median-of-three) filter; cross-channel disconnect filter flagging a reading implausibly extreme on one channel while all corroborating channels read normal; Hampel robust-outlier filter.

Model development

The model was a temporal convolutional network whose convolutions were causal and dilated, giving long effective memory over the lookback window while guaranteeing that no prediction could depend on information recorded after the time of prediction. A single classification head produced one probability for each of the 30 target-by-horizon pairs, so that a shared physiologic representation served all ten targets rather than a separate model being fitted for each. The training objective compensated for the severe class imbalance characteristic of these events, and an ordinal penalty enforced the monotonicity constraint described above. Probabilities were calibrated by isotonic regression fitted on the validation partition and applied unchanged to the internal test partition and the external cohort.

Architecture. Four residual temporal blocks with output channels of 128, 256, 512, and 512; kernel size 3; exponentially increasing dilations of 1, 2, 4, and 8. Left-padding with causal truncation was applied at each convolution to prevent look-ahead. The model used GELU activations, a dropout rate of 0.25, and layer normalization after each residual block. A two-layer classification head concatenated the final-timestep hidden representation with a one-dimensional convolved channel-wise observational mask representation and mapped them through a 256-unit fully connected layer with GELU and dropout to 30 logits, corresponding to ten targets at three horizons.

Training. The objective was a focal loss with focusing parameter $\gamma = 3.0$, addressing the extreme class imbalance across targets, combined with an ordinal hinge penalty (weight 0.5) over the nested severity pairs enforcing P(severe) ≤ P(milder). Optimization used AdamW with a constant learning rate of 1×10^{-5} and a batch size of 2,048, in single-precision arithmetic, with early stopping at a patience of 5 epochs monitored on mean validation AUROC.

Outcomes

Discrimination was summarized by two quantities computed from identical model outputs and differing only in the windows over which they were evaluated. Standard AUROC was computed over all evaluable windows; stable AUROC, equivalently pre-event AUROC, only over windows in which the patient was not already breaching that target's threshold. The distinction is consequential: a patient whose MAP is

already 60 mmHg is highly likely to remain below 65 mmHg 3 min later, and a model recognizing this is detecting an established derangement rather than anticipating a new one. Stable AUROC was therefore designated the primary measure, with standard AUROC reported alongside so that the difference quantifies how much apparent discrimination derives from concurrent state reading.[12]

Because tolerance for interruption varies by setting and acuity, three alarm budgets were evaluated: conservative (at most one alarm per case-hour), standard (at most two, the default), and high-vigilance (at most four). Thresholds were both selected and scored on pre-event windows only, so that every alarm counted represented anticipation rather than recognition of an established derangement. Sensitivity, specificity, PPV, and negative predictive value are reported at each.

Lead time was defined, for each true-positive episode, as the interval between the first qualifying alarm and physiologic onset, summarized as a median with interquartile range together with the proportion of warnings delivering at least 3 min of lead. These statistics are conditional on an alarm having been raised and therefore describe warning quality rather than coverage. A lead time of at least 3 min was specified as the threshold of clinical meaningfulness, being approximately the interval required to recognize an alert, form a plan, and intervene.

Operating-point convention. Decision thresholds were selected and all deployment metrics scored on pre-event windows only, matching the index set used for stable AUROC. Alarm budgets are expressed per case-hour of elapsed operative time. Lead time was measured from the first qualifying alarm to physiologic onset and is conditional on an alarm having been raised. Development thresholds were fitted on the development validation partition; external thresholds were re-selected on the external cohort so that the same alarm budget was held fixed across cohorts.

Statistics and reproducibility

The primary outcome was stable AUROC at the 5-min horizon, averaged across the ten targets, in the held-out development test partition; 5 min was designated primary to remain comparable with the existing literature. Secondary outcomes were stable AUROC at 3 and 10 min, standard AUROC at all horizons, area under the precision-recall curve, sensitivity and PPV at each budget, median lead time, calibration, and all corresponding external values. Because the number of reported comparisons is large, no adjustment for multiplicity was applied and all secondary findings are exploratory.

Estimates are reported with 95% confidence intervals (CIs) obtained by percentile bootstrap with 1,000 resamples drawn at the level of the surgical case rather than the individual window, because windows within a case are strongly correlated and window-level resampling would understate uncertainty. The internal-versus-external comparison is descriptive; the paired difference in stable AUROC is reported with its bootstrap CI rather than tested formally.

Descriptive statistics are reported as mean ± SD, median [25th, 75th percentile] for skewed variables, and n (%) with numerator and denominator given for all percentages. No a priori power calculation was performed, because cohort sizes were fixed by the available data; the rationale for adequacy is that the development test partition contained 163,211 evaluable windows and at least 666 qualifying episodes for its least frequent target, hypercapnia, with 1,657,569 evaluable windows and at least 948 qualifying episodes for its least frequent target, severe hypoxemia, externally (Table 1). Missing data were not imputed beyond the 5-min forward-fill described above. Consideration of sex as a biological variable was prespecified at the level of reporting: all discrimination analyses were repeated separately in male and female patients. Analyses were performed in Python version 3.13.13 (PyTorch 2.12.1, scikit-learn 1.9.0, NumPy 2.4.6, SciPy 1.18.0, Polars 1.42.1, joblib 1.5.3).

Use of large language models

The authors used AI-based tools (Claude, Anthropic) during preparation of this work to assist with writing training and evaluation scripts for the prediction models and to assist with formatting of the manuscript document. AI tools were not used to draft, generate, or edit the scientific content, analysis, interpretation, or text of the manuscript. The authors reviewed and take full responsibility for all content in this work.

Declarations

Competing interests

All authors declare no financial or non-financial competing interests.

Author Contribution

J.S. conceived and designed the study, developed and tuned the model, wrote the code, performed the analyses, interpreted the results and wrote the manuscript. M.N. and K.N. contributed to the conception and design of the study and obtained the data. J.W.P. contributed to the conception and design of the study and provided machine-learning supervision. A.K. contributed to building the model and writing the code. All authors read and approved the final manuscript.

Acknowledgements

This study received no external funding; support was provided solely from institutional and/or departmental sources.

Data availability

VitalDB is openly available at https://vitaldb.net. The external validation data were provided by a partner academic medical center under a data use agreement; under the terms of that agreement the institution is not identified and the data cannot be publicly released. All summary results supporting the findings of this study are included in this article and its Supplementary Information.

Code availability

The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

References

1. Salmasi, V. 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* 126, 47–65 (2017).
2. Sessler, D. I. 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* 128, 317–327 (2018).
3. Wesselink, E. M., Kappen, T. H., Torn, H. M., Slooter, A. J. C. & van Klei, W. A. Intraoperative hypotension and the risk of postoperative adverse outcomes: a systematic review. *Br. J. Anaesth.* 121, 706–721 (2018).
4. Gregory, A. et al. Intraoperative hypotension is associated with adverse clinical outcomes after noncardiac surgery. *Anesth. Analg.* 132, 1654–1665 (2021).
5. Wijnberge, M. et al. Association of intraoperative hypotension with postoperative morbidity and mortality: systematic review and meta-analysis. *BJS Open* 5, zraa018 (2021).
6. Saugel, B. et al. PeriOperative Quality Initiative (POQI) international consensus statement on perioperative arterial pressure management. *Br. J. Anaesth.* 133, 264–276 (2024).
7. Hatib, F. et al. Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. *Anesthesiology* 129, 663–674 (2018).
8. Davies, S. J., Vistisen, S. T., Jian, Z., Hatib, F. & Scheeren, T. W. L. Ability of an arterial waveform analysis-derived hypotension prediction index to predict future hypotensive events in surgical patients. *Anesth. Analg.* 130, 352–359 (2020).
9. Wijnberge, M. 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* 323, 1052–1060 (2020).
10. Maheshwari, K. et al. Hypotension Prediction Index for prevention of hypotension during moderate- to high-risk noncardiac surgery. *Anesthesiology* 133, 1214–1222 (2020).

11. Grundmann, C. D., Wischermann, J. M., Fassbender, P., Bischoff, P. & Frey, U. H. Hemodynamic monitoring with Hypotension Prediction Index versus arterial waveform analysis alone and incidence of perioperative hypotension. *Acta Anaesthesiol. Scand.* 65, 1404–1412 (2021).
12. Enevoldsen, J. & Vistisen, S. T. Performance of the Hypotension Prediction Index may be overestimated due to selection bias. *Anesthesiology* 137, 283–289 (2022).
13. Jo, Y.-Y. et al. Predicting intraoperative hypotension using deep learning with waveforms of arterial blood pressure, electroencephalogram, and electrocardiogram: retrospective study. *PLoS ONE* 17, e0272055 (2022).
14. Lee, S., Lee, M., Kim, S.-H. & Woo, J. Intraoperative hypotension prediction model based on systematic feature engineering and machine learning. *Sensors* 22, 3108 (2022).
15. Kapral, L. et al. Development and external validation of temporal fusion transformer models for continuous intraoperative blood pressure forecasting. *EClinicalMedicine* 75, 102797 (2024).
16. Lundberg, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* 2, 749–760 (2018).
17. Park, J.-B. et al. Machine learning-based prediction of intraoperative hypoxemia for pediatric patients. *PLoS ONE* 18, e0282303 (2023).
18. Schmid, F., Goepfert, M. S. & Reuter, D. A. Patient monitoring alarms in the ICU and in the operating room. *Crit. Care* 17, 216 (2013).
19. Drew, B. J. et al. Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. *PLoS ONE* 9, e110274 (2014).
20. Sendelbach, S. & Funk, M. Alarm fatigue: a patient safety concern. *AACN Adv. Crit. Care* 24, 378–386 (2013).
21. Steyerberg, E. W. & Harrell, F. E. Jr. Prediction models need appropriate internal, internal-external, and external validation. *J. Clin. Epidemiol.* 69, 245–247 (2016).
22. Riley, R. D. et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 353, i3140 (2016).
23. Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C. & van Diepen, M. External validation of prognostic models: what, why, how, when and where? *Clin. Kidney J.* 14, 49–58 (2021).
24. Wong, A. et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern. Med.* 181, 1065–1070 (2021).
25. Lee, H.-C. et al. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Sci. Data* 9, 279 (2022).
26. Collins, G. S. et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385, e078378 (2024).
27. Moons, K. G. M. et al. PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* 388, e082505

(2025).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PRISMnpjDigitalMedicineSupplementaryInformation.pdf