Beyond Thresholds: Learning Perioperative Risk from Intraoperative Physiological Trajectories

Anonymous Authors[1]

Abstract

Clinical risk scores for surgical patients are computed once from preoperative variables and consequently fail to capture the physiological dynamics that unfold during surgery. We ask whether intraoperative vital-sign time series, routinely recorded but rarely used for real-time risk stratification, provide predictive information that static models fail to capture. Using 130,960 operations from the INSPIRE perioperative dataset, we compare four modelling arms across progressively richer feature regimes (static baseline, physiological summaries, HMM-derived trajectory features, and their combination) and evaluate performance across four postoperative outcomes using grouped cross-validation. We find that learned trajectory representations capture complementary prognostic structure to threshold-based summaries for ICU-related outcomes. Predictive signal is detectable as early as 15 minutes after anaesthesia onset, and gains are largest in patients whose static risk profile underestimates true intraoperative vulnerability. These results demonstrate that temporal modelling of multivariate intraoperative time series provides clinically meaningful information that threshold-based approaches systematically miss.

1. Introduction

Surgery is performed at massive scale, with over 300 million procedures worldwide each year (Weiser et al., 2015). While most patients have an uneventful perioperative course, a substantial minority develop serious complications in the hours and days that follow, including unplanned intensive care unit (ICU) admission and death. Identifying which patients will deteriorate remains a central and unresolved challenge in perioperative medicine. The European Surgical Outcomes Study found that 73% of patients who died after non-cardiac surgery had never been admitted to ICU (Pearse et al., 2012), indicating that adverse outcomes often arise from failures to recognise risk in time.

Current approaches to risk identification are fundamentally static. Preoperative tools such as the Surgical Outcome Risk Tool (SORT) (Protopapa et al., 2014) assign a scalar risk estimate before the patient enters the operating room, based on American Society of Anesthesiologists (ASA) physical status grade, urgency, surgical severity, specialty, cancer status, and age. While these tools perform well on average, achieving AUROC of 0.90 for 30-day mortality (Wong et al., 2020), the resulting estimate is not designed to be updated once surgery begins. It does not reflect the patient's response to anaesthesia and subsequent surgical stress. Once the operation starts, risk assessment shifts to threshold-based alarm systems that flag individual vital signs when they breach pre-defined limits. These systems treat vital signs as independent scalar quantities, discarding the multivariate temporal context that encodes the patient's evolving physiological state. As a result, they provide no representation of the physiological trajectory over the course of the procedure.

During surgery, each patient generates a continuous stream of physiological data at minute-level resolution, yet this information is rarely incorporated into predictive models and is absent from formal risk estimates. Machine learning provides a natural framework for modelling multivariate physiological time series, where clinically relevant information is encoded in temporal patterns and cross-signal dependencies rather than in isolated measurements. Prior studies using gradient-boosted models, deep neural networks, and sequence models have shown promise for postoperative risk prediction from intraoperative data (Xue et al., 2021; Fritz et al., 2019; Shickel et al., 2023). However, it is unclear whether predictive gains arise from detecting threshold violations or from modelling the structure of physiological trajectories. More broadly, fundamental questions remain: what feature representations capture the most prognostic information, when does the signal become reliable, and for whom does dynamic monitoring add most value over static preoperative assessment?

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

We address these questions using the INSPIRE dataset (Lim et al., 2024), the largest publicly available perioperative dataset, which comprises 130,960 operations with linked intraoperative vital signs and postoperative outcomes. We make four contributions. (1) We provide a structured comparison of static baseline risk, intraoperative physiological summaries, and hidden Markov model (HMM)-derived trajectory features, demonstrating that each adds incremental predictive value over the previous. (2) We show that HMM trajectory features capture complementary predictive structure to per-channel summaries, with the combined representation performing best. This indicates that prognostic signal is multivariate and contextual, rather than reducible to threshold crossings of individual signals. (3) We demonstrate that predictive signal is detectable as early as 15 minutes after anaesthesia onset. (4) We find that gains concentrate in patients whose static risk profile underestimates true physiological vulnerability, a pattern not explained by any single preoperative variable.

2. Problem Setup and Methods

Dataset. INSPIRE (Lim et al., 2024) contains 130,960 surgical operations from a single academic centre (2011–2020), with continuously recorded intraoperative vital signs, including heart rate, peripheral oxygen saturation (SpO_2), non-invasive mean arterial pressure (NIBP MAP), invasive mean arterial pressure (arterial MAP), and inspired oxygen fraction (FiO_2), alongside postoperative outcomes. The dataset is publicly available on PhysioNet and provides sufficient scale and outcome prevalence for robust evaluation. Cohort characteristics and outcome incidence are summarised in Appendix A.

Outcomes. We evaluate four binary outcomes: prolonged ICU stay >48 h (prevalence 2.2%), ICU admission (9.6%), early ICU escalation within 6 h of OR exit (9.4%), and in-hospital death (1.2%). The primary outcome is prolonged ICU stay, capturing sustained postoperative deterioration and providing a more specific marker of clinically meaningful risk than ICU admission alone. ICU admission captures all postoperative ICU transfers within 24 hours of surgery, while ICU within 6 h captures only immediate escalations. The high overlap between these definitions explains their near-identical predictive performance. Class imbalance motivates reporting both AUROC and AUPRC.

Feature regimes. We compare three feature regimes (baseline, summaries, and trajectories) and their combination, evaluated across four modelling arms. *Baseline* uses static preoperative variables (age, sex, BMI, ASA grade, emergency status, surgical specialty, anaesthesia type, procedure class). Anaesthesia and operation duration are excluded to maintain a strictly preoperative baseline. *Summaries* augment the baseline with per-channel statistics (mean, SD, min, max, and time beyond clinical thresholds (e.g. MAP <65 mmHg, SpO_2 <92%)) computed over the full intraoperative period. *Trajectories* augment the baseline with features derived from a Gaussian HMM fit to 5-minute binned multivariate intraoperative time series, replacing per-channel summaries. The four modelling arms are: baseline only (A), baseline plus summaries (B), baseline plus trajectories (C), and the combined model (D).

HMM design. Six channels were used: NIBP MAP, arterial MAP, SpO_2, heart rate, FiO_2, and a binary arterial line indicator. Signals were binned into 5-minute intervals. Short gaps of up to two consecutive bins (10 minutes) were forward-filled, and bins with no observations across all channels were excluded. The number of states K was selected using held-out BIC and silhouette criteria on a grouped 80/20 split of the training data; $K = 6$ was chosen to balance data fit and state separability. Operation-level features comprised state occupancy fractions, transition entropy, and transitions into and out of the dominant high-complexity state, defined as the state with highest transition entropy. The inclusion of an arterial line indicator may partly reflect monitoring intensity rather than physiology alone, as higher-risk patients are more likely to receive invasive monitoring. To assess robustness, we repeated the analysis using only universally available signals (NIBP MAP, SpO_2, heart rate, FiO_2).

Fold-safe evaluation. All models and preprocessing steps were fit within cross-validation folds using training operations only and applied to held-out data. HMMs were trained and decoded in a fold-safe manner, with cross-fold state alignment performed via nearest-neighbour emission remapping. All experiments use 5-fold grouped cross-validation (`GroupKFold`) by patient identifier to prevent leakage between multiple operations from the same patient.

Models. Primary models use logistic regression to transparently evaluate feature regime contributions. Sensitivity analyses with histogram gradient boosting, LightGBM, and XGBoost confirm that results are not dependent on model class.

3. Results

3.1. Trajectory features complement summaries

Table 1 shows the predictive comparison across four arms: baseline (A), baseline plus summaries (B), baseline plus HMM trajectories (C), and the combined model (D). For ICU-related outcomes, HMM features provide modest improvements over summaries on AUPRC, with the combined model performing best. For prolonged ICU stay, AUPRC increases from 0.244 (baseline) to 0.305 (summaries), remains similar for HMM features alone (0.303), and reaches 0.339 when combined. For ICU admission, AUPRC rises from

Table 1. Fold-safe predictive comparison. Arms: A = baseline, B = baseline + summaries, C = baseline + HMM trajectories, D = all combined. Best per outcome in bold.

Outcome	Arm	AUROC	AUPRC
ICU stay >48 h	A: Baseline	0.927	0.244
	B: + Summaries	0.944	0.305
	C: + HMM	0.947	0.303
	D: Combined	**0.954**	**0.339**
ICU admission	A: Baseline	0.903	0.520
	B: + Summaries	0.927	0.582
	C: + HMM	0.936	0.599
	D: Combined	**0.944**	**0.623**
ICU within 6 h	A: Baseline	0.903	0.513
	B: + Summaries	0.927	0.575
	C: + HMM	0.936	0.594
	D: Combined	**0.943**	**0.616**
In-hospital death	A: Baseline	0.863	0.151
	B: + Summaries	0.882	0.198
	C: + HMM	0.877	0.172
	D: Combined	**0.889**	**0.204**

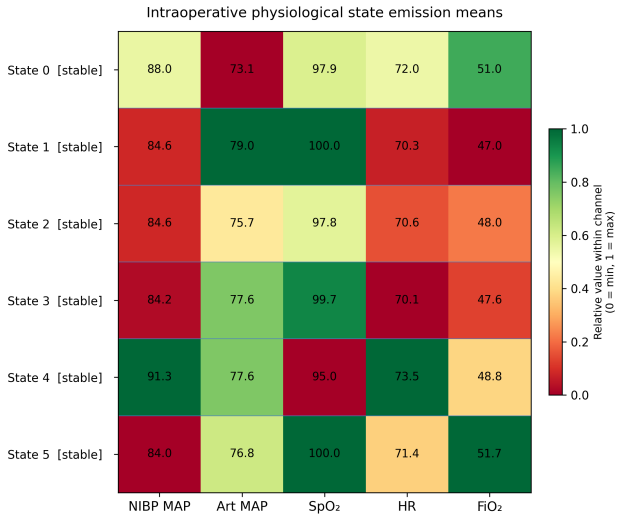

Figure 1. Emission means for the six HMM states across five continuous physiological channels. Colour encodes relative values within each channel (red = minimum, green = maximum). All states lie within standard clinical ranges, with no individual variable breaching typical instability thresholds. Differences between states arise from subtle multivariate patterns across channels rather than from isolated abnormalities. The binary arterial line indicator is omitted.

0.520 (baseline) to 0.582 (summaries) and 0.599 (HMM), reaching 0.623 when combined, with a near-identical pattern for early ICU escalation within 6 h. For in-hospital death, summaries outperform HMM features alone (0.198 vs 0.172), suggesting that mortality risk depends more on aggregate physiological burden, whereas ICU-related outcomes are more sensitive to temporal structure. Trajectory features therefore provide complementary information but are insufficient in isolation. These findings were consistent across ICU-related outcomes when restricting inputs to four universally available channels (NIBP MAP, SpO_2, heart rate, FiO_2), indicating that the observed gains are not driven by invasive monitoring alone.

3.2. Latent states are prognostic beyond threshold violations

The six inferred states differ in multivariate combinations of physiological signals while remaining within clinically normal ranges (Figure 1). Despite the absence of overt abnormalities in any individual variable, state occupancy is strongly associated with postoperative outcomes (Appendix B). For prolonged ICU stay, operations with adverse outcomes spend 57.1% of operative time in State 1 compared with 19.7% for operations without the event, with the opposite pattern observed for State 3 (1.3% vs 59.6%). This separation is consistent across outcomes. These states do not reflect overt physiological instability. Instead, prognostic signal arises from multivariate temporal structure, indicating that clinically relevant information is encoded in trajectories rather than isolated threshold violations.

Table 2. ΔAUROC at intraoperative checkpoints for the combined model over baseline. Signal is detectable at 15 minutes for all outcomes and accumulates progressively for ICU-related outcomes.

Outcome	15 min	30 min	60 min	Full
ICU stay >48 h	+0.014	+0.015	+0.017	+0.029
ICU admission	+0.017	+0.022	+0.027	+0.042
ICU within 6 h	+0.017	+0.022	+0.027	+0.042
In-hospital death	+0.023	+0.024	+0.024	+0.027

3.3. Predictive signal emerges within 15 minutes

Predictive signal is detectable within 15 minutes of anaesthesia onset for all outcomes (Appendix C). As shown in Table 2, for prolonged ICU stay, ΔAUROC increases from +0.014 at 15 minutes to +0.029 at full duration. For ICU admission and early ICU escalation within 6 h, gains increase from +0.017 at 15 minutes to +0.042 at full duration. For in-hospital death, gains plateau after 30 minutes, with early physiological response to anaesthesia induction carrying disproportionate prognostic weight. For ICU-related outcomes, gains accumulate progressively, supporting the value of continuous rather than single-point monitoring. These results indicate that a monitoring-augmented model can flag elevated risk early in the procedure.

3.4. Gains concentrate where static models fail

The incremental value of dynamic physiology varies substantially across patients. Stratifying operations by baseline predicted risk reveals marked heterogeneity. For prolonged ICU stay, ΔAUROC is +0.223 (95% CI +0.170 to +0.274) in the lowest-risk stratum compared with +0.106 in the medium-risk stratum (Appendix D). For in-hospital death, gains decrease monotonically from low to high baseline risk.

This heterogeneity is not explained by any single preoperative variable. When stratified by ASA physical status, ΔAUROC is nearly identical in ASA I–II and ASA III+ groups (difference <0.01 AUROC points). The differential benefit therefore reflects heterogeneity in the *joint* baseline risk profile, captured only by composite learned representations. Patient vulnerability is thus encoded in combinations of preoperative factors rather than in any individual measure.

4. Discussion

Three findings emerge. First, HMM-derived trajectory representations add complementary predictive value to per-channel summaries for ICU-related outcomes, indicating that prognostic structure is multivariate and contextual rather than threshold-based. The Gaussian HMM serves as a deliberately simple baseline. It makes strong parametric assumptions, uses fixed 5-minute bins, and cannot model long-range dependencies. That it nonetheless improves prediction suggests that temporal structure in intraoperative physiology is genuinely informative, rather than an artefact of model flexibility. This leaves substantial room for richer sequence models. Transformers could capture long-range dependencies across the operative course, state space models could better handle irregular sampling and missingness, and neural process models could represent uncertainty over physiological trajectories.

Second, predictive signal is available within 15 minutes of anaesthesia onset. A model need not wait until the procedure ends to provide useful risk stratification. This raises a representation learning question: if early physiology carries substantial predictive information, what features of anaesthesia induction are most informative, and can they be learned in a self-supervised manner from unlabelled data? The scale of intraoperative monitoring data in hospital systems makes such pretraining a promising direction.

Third, gains concentrate where static models fail, in patients whose preoperative profile appears reassuring but whose physiological trajectory reveals unexpected vulnerability. ASA grade is a coarse variable, whereas the model integrates multiple baseline factors jointly and is more sensitive to where physiology adds value. This motivates learning-based approaches to identifying which patients benefit most from monitoring. Rather than relying on clinical proxies, a learned representation of the joint preoperative risk profile may better identify patients for whom intraoperative data is most informative. More broadly, this suggests that the value of a data stream is not fixed, but depends on what the model already knows, a principle relevant to any setting where dynamic observations complement static risk.

Taken together, these findings suggest a shift from static risk estimation to modelling continuous physiological trajectories. Our results indicate that clinically relevant information lies in how physiological variables co-evolve over time, rather than in isolated threshold breaches, highlighting the limitations of feature engineering approaches based on per-channel summaries. From a machine learning perspective, this setting motivates representation learning methods that capture temporal structure in multivariate time series, where the central challenge is to extract predictive signal from noisy, partially observed trajectories. The INSPIRE dataset provides a reproducible benchmark for this agenda, with sufficient scale and clinically meaningful outcomes. We therefore view perioperative risk as a prototypical problem for understanding how temporal representations can complement static features in high-stakes prediction tasks.

Limitations. Data originate from a single centre in South Korea, and case-mix, monitoring practices, and outcome rates may differ elsewhere. External validation at an independent institution is required before clinical translation. Temporal validation was precluded by date anonymisation in INSPIRE, and direct comparison with established preoperative risk scores such as SORT was prevented by missing variables (e.g. surgical urgency). Bootstrap confidence intervals were computed at the observation level for the stratified analysis and should be interpreted as approximate. Grouped resampling by patient would be more principled.

Broader impact. This work explores how intraoperative monitoring data could support dynamic risk stratification during surgery, enabling earlier recognition of patient deterioration and more timely escalation of care. By leveraging existing data streams, such approaches could improve decision-making without requiring changes to workflows. However, deploying predictive models in this setting carries risks. Automation bias may lead clinicians to over-rely on model outputs, particularly in high-pressure environments such as the operating room. Model errors or miscalibration could result in inappropriate escalation or missed deterioration. In addition, performance may vary across patient subgroups, raising concerns about fairness and generalisability. Careful evaluation, calibration, and integration into workflows will be essential to ensure such systems augment rather than replace clinical judgement, and that benefits are realised equitably across patient populations.

References

Fritz, B. A., Cui, Z., Zhang, M., He, Y., Chen, Y., Kronzer, A., Ben Abdallah, A., King, C. R., and Avidan, M. S. Deep-learning model for predicting 30-day postoperative mortality. *British Journal of Anaesthesia*, 123(5):688–695, 2019. doi: 10.1016/j.bja.2019.07.025.

Lim, L., Lee, H., Jung, C.-W., Sim, D., Borrat, X., Pollard, T. J., Celi, L. A., Mark, R. G., Vistisen, S. T., and Lee, H.-C. INSPIRE, a publicly available research dataset for perioperative medicine. *Scientific Data*, 11(1):655, 2024. doi: 10.1038/s41597-024-03517-4.

Pearse, R. M., Moreno, R. P., Bauer, P., Pelosi, P., Metnitz, P., Spies, C., Vallet, B., Vincent, J.-L., Hoeft, A., Rhodes, A., and European Surgical Outcomes Study (EuSOS) group. Mortality after surgery in Europe: a 7 day cohort study. *Lancet*, 380(9847):1059–1065, 2012. doi: 10.1016/S0140-6736(12)61148-9.

Protopapa, K. L., Simpson, J. C., Smith, N. C. E., and Moonesinghe, S. R. Development and validation of the Surgical Outcome Risk Tool (SORT). *British Journal of Surgery*, 101(13):1774–1783, 2014. doi: 10.1002/bjs.9638.

Shickel, B., Loftus, T. J., Ruppert, M., Upchurch, G. R., Ozrazgat-Baslanti, T., Rashidi, P., and Bihorac, A. Dynamic predictions of postoperative complications from explainable, uncertainty-aware, and multi-task deep neural networks. *Scientific Reports*, 13(1):1224, 2023. doi: 10.1038/s41598-023-27418-5.

Weiser, T. G., Haynes, A. B., Molina, G., Lipsitz, S. R., Esquivel, M. M., Uribe-Leitz, T., Fu, R., Azad, T., Chao, T. E., Berry, W. R., and Gawande, A. A. Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes. *Lancet*, 385(Suppl 2):S11, 2015. doi: 10.1016/S0140-6736(15)60806-6.

Wong, D. J. N., Harris, S., Sahni, A., Bedford, J. R., Cortes, L., Shawyer, R., Wilson, A. M., Lindsay, H. A., Campbell, D., Popham, S., Barneto, L. M., Myles, P. S., SNAP-2: EPICCS collaborators, and Moonesinghe, S. R. Developing and validating subjective and objective risk-assessment measures for predicting mortality after major surgery: an international prospective cohort study. *PLOS Medicine*, 17(10):e1003253, 2020. doi: 10.1371/journal.pmed.1003253.

Xue, B., Li, D., Lu, C., King, C. R., Wildes, T., Avidan, M. S., Kannampallil, T., and Abraham, J. Use of machine learning to develop and evaluate models using preoperative and intraoperative data to identify risks of postoperative complications. *JAMA Network Open*, 4(3):e212240, 2021. doi: 10.1001/jamanetworkopen.2021.2240.

A. Cohort Characteristics

Table A1. Baseline characteristics of the INSPIRE study cohort. Unit of analysis is the surgical operation ($n = 130{,}960$; $99{,}886$ unique patients).

Characteristic	Value
Patient demographics	
Age, years (median [IQR])	60 (45–70)
Female sex	73,023 (55.8%)
BMI, kg/m^2 (median [IQR])	23.8 (21.5–26.0)
Surgical characteristics	
Emergency surgery	12,330 (9.4%)
General anaesthesia	102,790 (78.5%)
ASA I–II	115,139 (87.9%)
ASA III+	12,274 (9.4%)
ASA missing	3,547 (2.7%)
Intraoperative monitoring	
Operations with usable intraoperative physiology	130,888 (99.9%)
Measurements per operation (median [IQR])	69 (36–126)
Anaesthesia duration, min (median [IQR])	115 (70–195)
Outcomes	
ICU admission	12,540 (9.6%)
ICU within 6 h of OR exit	12,285 (9.4%)
ICU stay >48 h (primary)	2,871 (2.2%)
In-hospital death	1,555 (1.2%)

B. HMM State Occupancy by Outcome

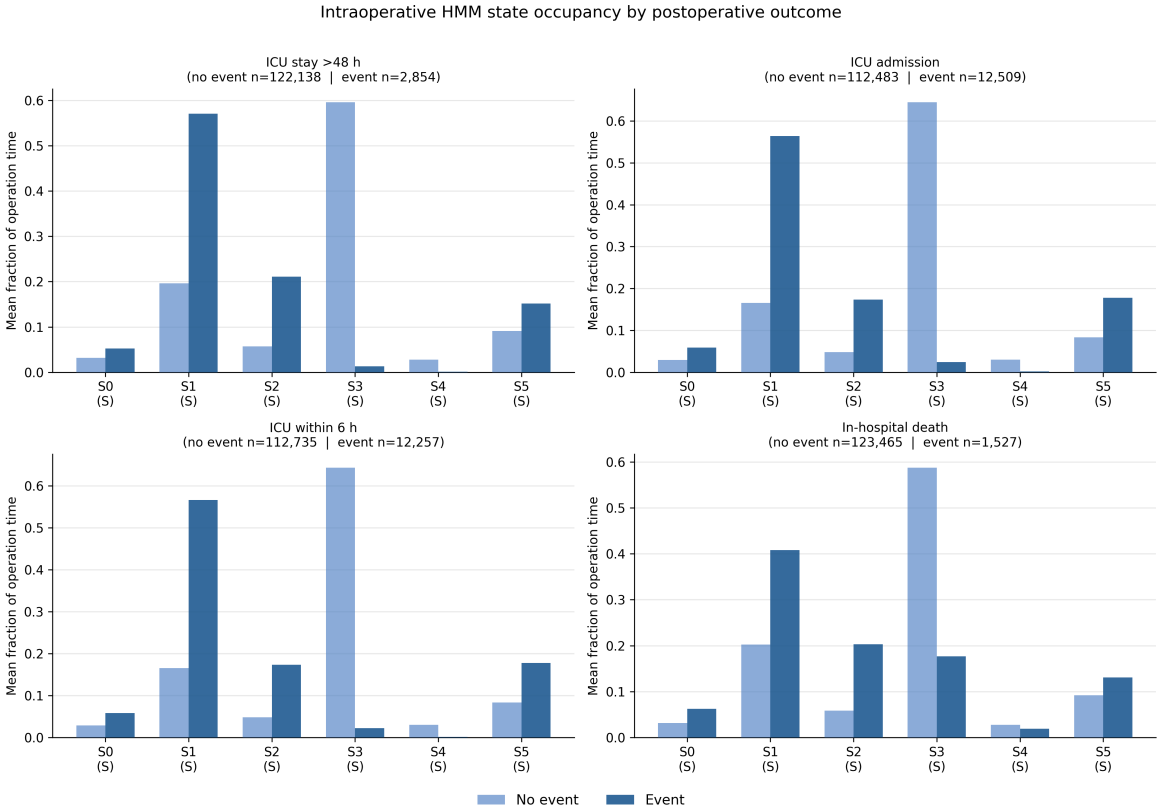

Figure B1. Mean fraction of operative time in each of the six HMM states, stratified by event (dark blue) and no-event (light blue) across four postoperative outcomes. Operations followed by adverse events spend substantially more time in State 1 and less in State 3 than operations without adverse outcomes. No state breaches any standard clinical instability threshold, yet the occupancy pattern is strongly and consistently prognostic across all four outcomes.

C. Temporal Emergence of Predictive Signal

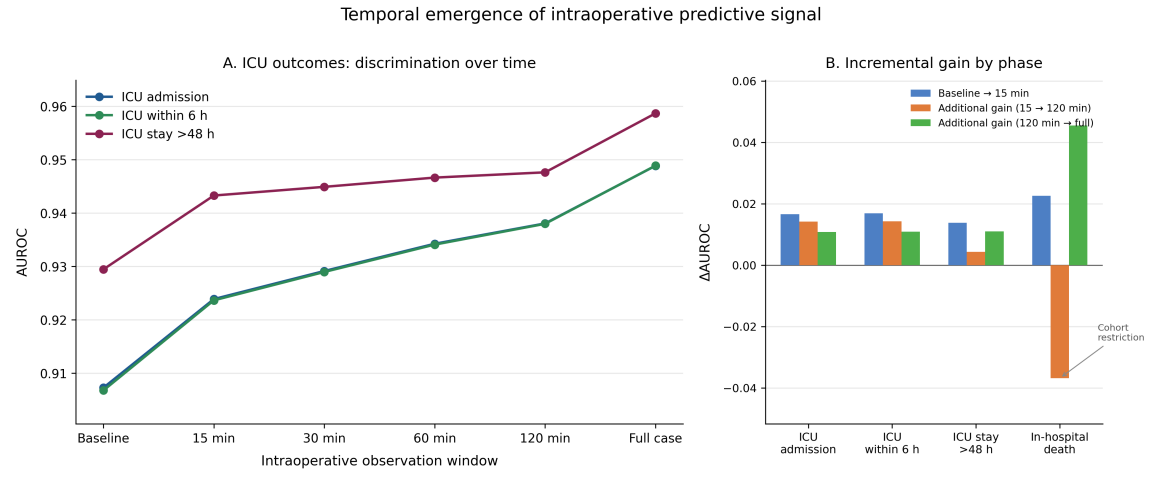

Figure C1. (A) AUROC for combined models across progressively longer observation windows for ICU-related outcomes. All three improve steadily with additional physiological data. (B) Incremental AUROC gains decomposed by phase across all four outcomes. For in-hospital death the early gain dominates; negative bars at 15–120 min are attributable to cohort restriction to longer operations rather than loss of signal.

D. Stratified Benefit Analysis

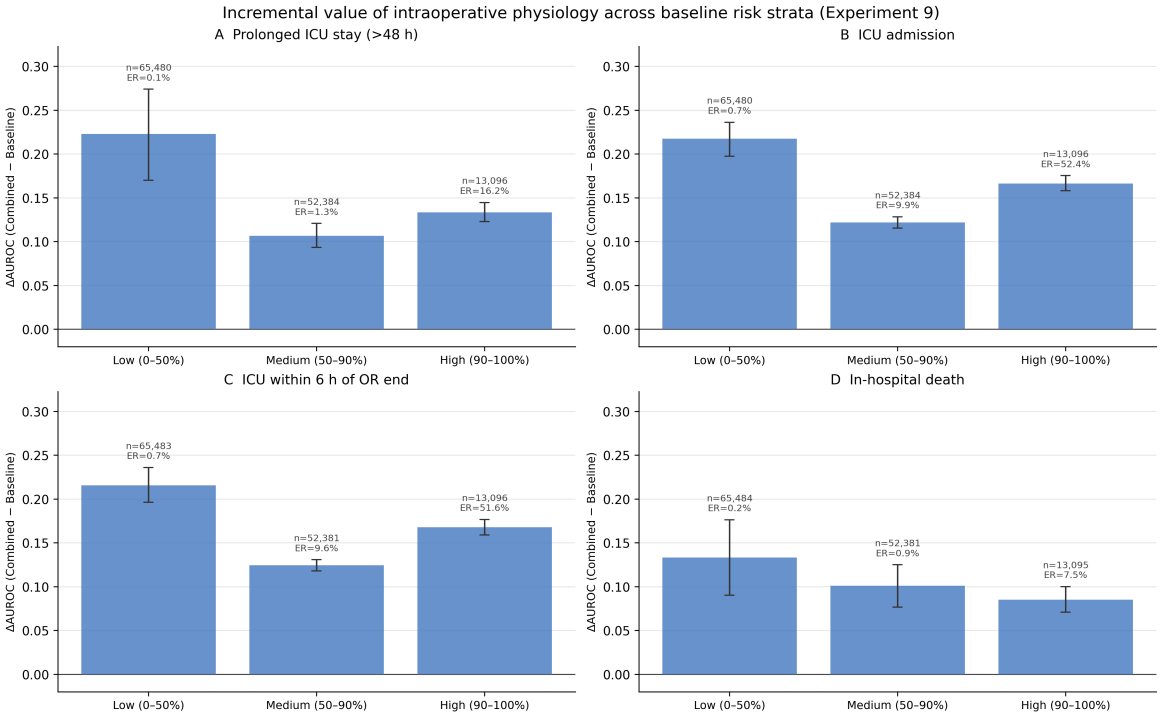

Figure D1. ΔAUROC from adding intraoperative physiology across baseline risk strata. Gains are largest in the lowest-risk stratum for all ICU outcomes, indicating that dynamic physiology adds most value where static models underestimate true risk. Error bars: 95% bootstrap CIs.