HEC Paris • École Polytechnique • Stanford University

MSc Data Science for Business

Research Paper

Academic Year 2025 – 2026

Machine Learning in Cardiology

Predicting Severe Pericardial Tamponade After Cardiac Surgery

Iliass Sijelmassi

In collaboration with Stanford University School of Medicine

Internship supervisor

Prof. Louise Sun, Stanford University

HEC Paris academic tutor

Vincent Fraitot

Jury

Vincent Fraitot

July 28, 2026 – Oral Presentation

☒ Public report ☐ Confidential report

Executive Summary

This paper asks what information predicts severe pericardial tamponade after adult cardiac surgery, and whether any data acquired after the pre-operative assessment adds value beyond the routine pre-operative clinical record. On a single-source Stanford cohort with one fixed outcome, severe incident post-operative tamponade, we build and evaluate four model families: a pre-operative covariate model, an early intensive-care structured model, a missingness-aware model, and a bedside-waveform model. The pre-operative covariate model, trained on 95 incident severe events and 6,761 audited controls, reaches an AUROC of 0.74 (95% CI 0.68 to 0.79) and is well calibrated, and its transportable core validates externally in MIMIC-IV, reaching an AUROC of 0.695 on 11,969 independent cardiac-surgery admissions, with the ranking preserved and the absolute risk level requiring recalibration to the local base rate. It is the result closest to clinical use. The three later-data models each carry some signal in isolation, yet none adds established value over the pre-operative model. Early intensive-care structured data give no significant increment; informative missingness, meaning the pattern of which tests were ordered rather than what they showed, since a test is ordered when a clinician is already concerned, contributes only a small care-intensity effect; and bedside-waveform features, although predictive on their own (numeric trends about 0.74 and raw 125 Hz features about 0.63 on the 27 patients with a recording), do not improve on the covariates once a monitoring-intensity confound and feature-selection optimism are removed. Two contributions follow. The clinical finding is that the predictive signal for severe post-operative tamponade is largely pre-operative: the later, richer sources re-express that same risk and do not add to it. The methodological contribution is a framework for drawing trustworthy conclusions from small, rare-event waveform cohorts: a bed-anchored method that matches each recording to the right patient, an explicit correction for the monitoring-intensity confound, and an evaluation protocol built to avoid the statistical traps that make small-sample waveform results look stronger than they are. We quantify the cohort size a confirmatory incremental test would require and set out an expansion path through a broadened diagnosis cohort and multi-center development, the external validation itself now being in hand.

Keywords. clinical prediction, cardiac surgery, pericardial tamponade, physiological waveforms, machine learning, informative missingness, leakage, permutation testing.

Preface and Acknowledgements

This research paper grew out of a research internship with the cardiac anesthesiology and outcomes-research group of Professor Louise Sun at Stanford University, where the question of whether the signals already recorded at every bedside could warn of post-operative deterioration first took shape.

I am deeply grateful to my internship supervisor, Dr. Louise Sun, Professor of Anesthesiology, Perioperative and Pain Medicine and Director of Cardiovascular Research at Stanford, for her guidance, her trust, and her generosity throughout this project.

I owe particular thanks to Osamu Yasui for his help throughout the project, and I thank my fellow collaborators at Stanford, Seung Hyun Kim and Sara El Baghdadi, for their work on the cohort and the data and for the many discussions that shaped this analysis.

I am also grateful, at HEC Paris and École Polytechnique, to Aymeric Dieuleveut and Aymeric Le Pennec for their teaching, to Vincent Fraitot, who directs the Data Science for Business program, and to Catherine Ly for her logistical support throughout the year.

Table of Contents

Executive Summary **2**

Preface and Acknowledgements **3**

1 Introduction **5**

2 Literature Review **11**

3 Data and Cohort Construction **16**

4 Exploratory Data Analysis **22**

5 Methodology **28**

6 Results **32**

7 Discussion and Recommendations **44**

8 Conclusion **50**

Bibliography **50**

Table of Annexes **52**

A Glossary of Clinical and Technical Terms **54**

B Extended Results and Figures **56**

C Feature Dictionary **61**

D Reproducibility and Code **63**

List of Figures **63**

List of Tables **64**

List of Annexes

Annex A Glossary of clinical and technical terms
Annex B Extended results and figures
Annex C Feature dictionary
Annex D Reproducibility and code

Chapter 1

Introduction

1.1 Context and Motivation

I carried out this research with the cardiac anesthesiology and outcomes-research group of Professor Louise Sun at Stanford University. The project joins Stanford's clinical-data platform and the hospital's continuous physiological-waveform archive and asks whether the signals already recorded at every bedside can warn of post-operative deterioration earlier than current care does.

Each monitor in an operating room or intensive-care unit produces a continuous stream of physiological signals. It shows most of them for a few seconds and archives the rest, where they sit unused. Turning those recordings into predictions costs almost nothing, since the sensors are installed and the data already exist.

Cardiac surgery carries some of the highest risk among routine procedures, and its complications lengthen hospital stays and raise mortality. Pericardial tamponade ranks among the rarer of these complications and the most lethal. It develops fast and hides behind the ordinary signs of recovery, so drainage often comes late. A model that flagged the patients most likely to progress would let the surgical team concentrate monitoring on them and drain early.

1.2 Clinical Background and Key Terms

Most readers of this paper work in data science, so this section defines the clinical terms it uses later.

1.2.1 Pericardial Tamponade

The heart sits inside a thin, stiff sac, the *pericardium*. *Pericardial tamponade* occurs when blood or fluid collects between the heart and this sac faster than the sac can stretch. The rising pressure squeezes the heart and stops it from filling between beats, so cardiac output and blood

pressure fall; an untreated patient can die within hours. After cardiac surgery, tamponade arises in two ways: acute bleeding into the pericardium soon after the operation, or an effusion that builds over days to weeks. Clinicians treat it by *drainage*, through a needle (pericardiocentesis) or surgical re-exploration. This study labels a patient positive when a drainage procedure, a cardiac-tamponade diagnosis, or a hemopericardium diagnosis first appears at least one day after the index surgery (Chapter 3).

1.2.2 The Perioperative Care Pathway

A cardiac-surgery patient moves through a fixed sequence of locations, shown in Figure 1.1. Surgery takes place in the *operating room* (OR), often with the heart stopped and a *cardiopulmonary bypass* machine maintaining the circulation. After major cardiac surgery the patient goes directly from the operating room to the *intensive-care unit* (ICU), where staff watch the vital signs. From the ICU most patients move directly to a hospital ward; sicker patients pass through a *step-down* unit first and reach the ward from there. The patient then goes home. The hospital's *admission, discharge and transfer* (ADT) record timestamps every move, and this study uses it to place each patient at each moment, and so to identify which monitor produced a given recording.

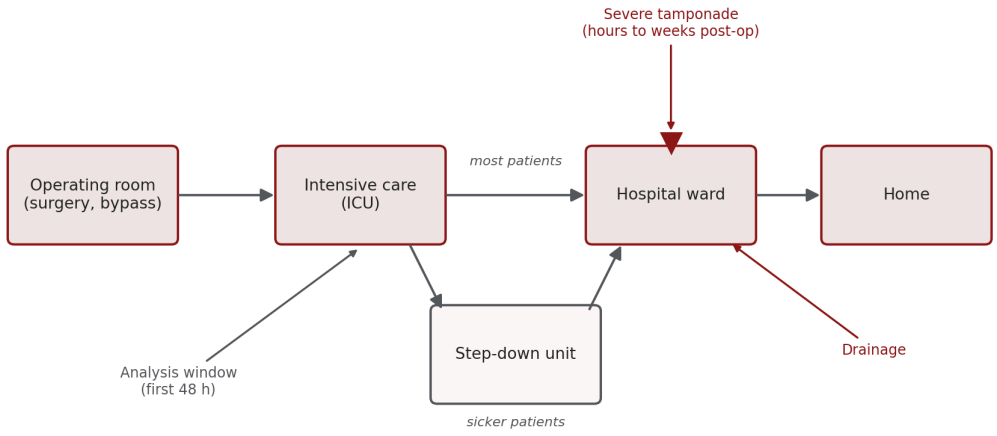

Figure 1.1: The perioperative care pathway and the timing of a tamponade event.

1.2.3 Bedside Waveforms and Invasive Monitoring

Throughout this pathway the bedside monitor records several continuous *waveforms*. Some come from *invasive lines*, thin catheters placed during surgery: an *arterial line* gives beat-to-beat arterial blood pressure (ABP); a *central venous catheter* gives central venous pressure (CVP); and a *pulmonary-artery catheter* gives pulmonary-artery pressure (PAP). Together these three form the *invasive hemodynamic signals*. Other signals are *non-invasive*: the elec-

trocardiogram (ECG), respiration, capnography (exhaled CO_2), and the *photoplethysmogram* (PPG, also called the pulse-oximetry or "Pleth" trace).

The invasive lines come out soon after the patient reaches the ICU, because they carry their own risks. The invasive hemodynamic signals (ABP, CVP, PAP) therefore exist only during and just after surgery, while the non-invasive signals continue throughout the rest of the hospitalisation, until hospital discharge. This clinical fact, not a modeling choice, sets which time window can carry an invasive-waveform signal, and Chapters 4 and 5 return to it.

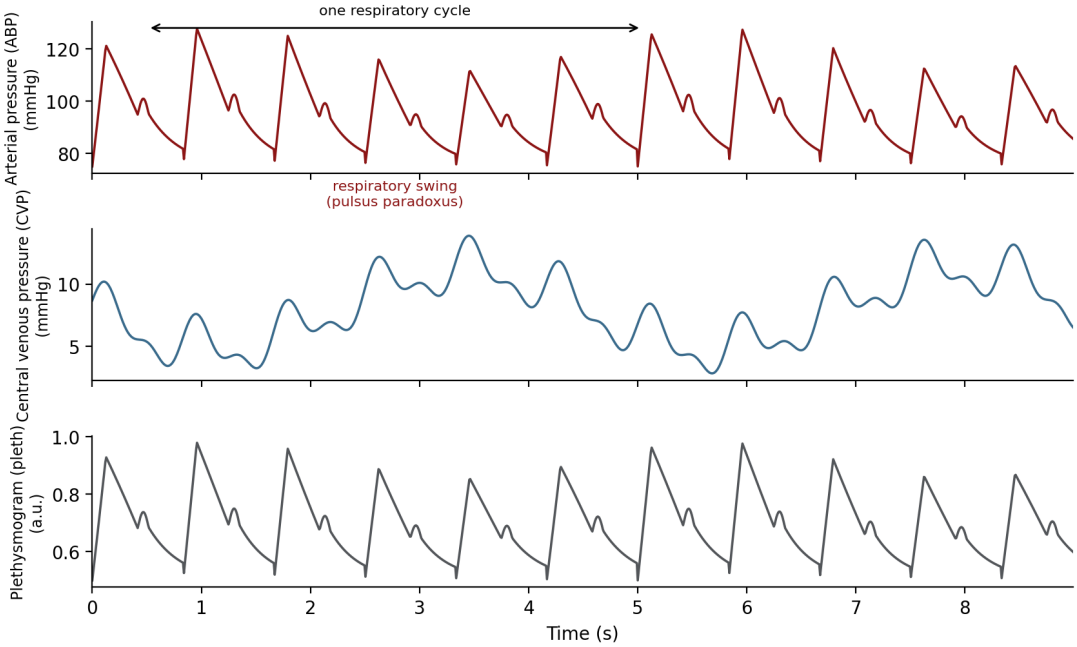

Figure 1.2: Illustrative bedside waveforms (schematic, not patient data): arterial pressure, central-venous pressure, and the plethysmogram, showing the respiratory swing that tamponade exaggerates as pulsus paradoxus.

1.3 Title Question and Objective

Title question. What information predicts severe post-cardiac-surgery tamponade? Specifically: how much of the risk is already visible in the routine pre-operative clinical record, how much is added by what happens in the operating room, and does any data acquired after the operation, whether early intensive-care measurements, the pattern of which measurements are taken, or the continuous bedside waveforms, add predictive value beyond both?

The aim is to measure whether each successive data source contributes information the earlier ones lack, while controlling the confounders and statistical artifacts that inflate small-cohort and waveform studies.

1.4 The Four Models

We answer the question with four model families, each evaluated on the same fixed outcome and compared against the same baseline.

- **Model 1, pre-operative and operative covariates.** Two nested specifications. *Model 1a, pre-operative*: age, body-mass index, ventricular function, renal function, anaemia, glycated haemoglobin, and frailty, every one of them available before the patient enters the operating room. *Model 1b, peri-operative*: Model 1a plus the operative-complexity variables, which are known only once the operation is under way or complete, namely cardiopulmonary-bypass duration, surgery length, and the circulatory-arrest and selective-antegrade-cerebral-perfusion flags. Reporting the two separately measures the contribution of the operation itself rather than assuming it, and keeps the label on each model honest: only Model 1a can be computed in clinic. Model 1b is the reference against which the later sources are compared, and the candidate for clinical use once externally validated.
- **Model 2, early intensive-care structured data.** The first forty-eight hours of post-operative laboratory values, chest-tube output, and vital-sign trends, asking whether early deterioration visible in structured data improves on the pre-operative picture.
- **Model 3, missingness-aware modeling.** Whether the pattern of which laboratory values were ordered, independent of the values themselves, carries information, since the decision to order a test reflects clinical concern.
- **Model 4, bedside waveforms.** Features derived from the continuous monitor signals, both the low-frequency numeric trends and the raw 125 Hz waveforms, asking whether the hemodynamic signature of developing tamponade is detectable before the event.

The unifying question across all four is incremental value: does the later or richer source improve discrimination beyond Model 1b, or merely re-encode the same underlying risk?

1.5 Hypotheses

The work is organized around four hypotheses, the last of which is methodological.

- **H1, the pre-operative record carries signal, and the operation adds to it.** Pre-operative clinical variables (age, ventricular function, anaemia, frailty, and so on) discriminate patients who will develop severe tamponade, because they capture baseline risk and physiological reserve. The operative-complexity variables, which describe how difficult the operation actually proved, should add to that. Separating Model 1a from Model 1b tests how much of the signal is available before the patient reaches the operating room and how much only the operation reveals.

- **H2, each post-operative source carries signal on its own.** Early intensive-care data, informative missingness, and waveform features each separate cases from controls in isolation. For the waveforms specifically, the clinical literature describes characteristic hemodynamic changes in tamponade, including an exaggerated respiratory swing in arterial pressure (*pulsus paradoxus*) and altered central-venous morphology.

- **H3, none adds incremental value (the central hypothesis).** Once compared against Model 1b, no data source acquired after the operation improves discrimination beyond it. The predictive signal is already carried by the pre-operative and operative record, and the post-operative sources re-express the same risk.

- **H4, apparent gains are confounded unless controlled.** Any naive advantage of the later or waveform models is inflated by a monitoring-intensity confound (sicker patients are monitored and measured more), by informative missingness acting as a care-intensity proxy, and by feature-selection optimism (hundreds of features screened on few events). Only a matched-cohort design, explicit presence modeling, and permutation testing reveal the true effect.

1.6 Issues and Challenges

Several obstacles shape every methodological choice that follows.

Record attribution. The waveform archive identifies recordings by a study identifier that is reused across many patients and dates, so a naive link can attach a recording to the wrong patient and silently corrupt the labels. We resolve this with a bed-anchored matching rule that accepts a recording for a patient only when its monitoring unit and bed appear in that patient's transfer log at the recording time (Chapter 3).

Jittered recording times. For privacy, the archive shifts every recording's date by a per-patient random offset, so the time stored in a file is not the real time. Aligning a recording to the surgery and to the event means recovering that offset and subtracting it, and a sign error there misplaces every window with no visible symptom. We recover the offset for each patient and check it against a hard fact: an operating-room recording must fall on the surgery date. An early version of the pipeline subtracted the offset twice, and this check caught it.

A rare event and a small waveform sample. Severe incident tamponade is uncommon. The structured cohort holds 95 events against 6,761 controls, an event rate near one and a half percent, while the waveform cohort, which also requires a usable recording, holds only 27. The rare event keeps the structured estimates modest, and the 27-patient waveform cohort caps how strong any waveform claim can be: complex models overfit at that size and the confidence intervals stay wide.

Monitoring-intensity confound. Because invasive lines are placed preferentially in sicker, more complex patients, the mere *presence* of a channel correlates with the outcome. A model can then appear to read the waveform while in fact detecting who was monitored. We neutralize this by matching controls to cases on procedure type and by modeling channel presence explicitly.

Feature-selection optimism. Screening hundreds of features for the best performers, on a small sample, yields apparently strong results even from noise. We quantify and remove this with in-fold feature selection and label-permutation testing.

Informative missingness. Whether a laboratory value exists is not random: a test is ordered because a clinician was concerned. A model given missingness indicators can therefore appear to predict the outcome while in fact detecting care intensity rather than physiology. We treat this as a confound to be measured, not a free signal to be exploited.

Temporal and label validity. For a genuine prediction rather than a description, the analysis window must close before the tamponade event; we verify that every case's window precedes its drainage.

Extraction-schema leakage. If cases and controls pass through even slightly different processing pipelines, spurious near-perfect performance can appear. We guard against this with symmetric processing and routine leakage audits.

1.7 Structure of the Paper

The remainder of the paper proceeds as follows. Chapter 2 reviews tamponade after cardiac surgery, predictive analytics in perioperative care, and the waveform-modeling literature, and positions the work against its closest comparator. Chapter 3 describes the data sources, the fixed outcome definition, the cohort funnel, and the bed-anchored attribution method, and explains why the four models are evaluated on different sample sizes. Chapter 4 reports the exploratory analysis, including the channel-coverage and confounder findings that shape the methodology. Chapter 5 sets out the four models, the feature extraction, the time anchor, and the evaluation protocol. Chapter 6 presents the results model by model and the convergent incremental-value finding. Chapter 7 discusses interpretation, clinical implications, limitations, and the expansion path, and Chapter 8 concludes.

Chapter 2

Literature Review

2.1 Pericardial Tamponade After Cardiac Surgery

Pericardial tamponade is a serious complication of cardiac surgery. In one single-center series of 556 heart-valve-surgery patients it occurred in 4.3 percent, every case requiring drainage, at a median of 17 days after the operation[1]. A larger consecutive series of 1,308 cardiac-surgery patients found that 6.2 percent required invasive treatment for a late pericardial effusion, split between pretamponade at 2.1 percent and frank tamponade at 4.1 percent, at a median of 11 days after surgery[2]. Both series place the delayed form in the first two post-operative weeks rather than the first hours. This paper's outcome spans both mechanisms, since it counts any severe event from one day after surgery onward. Two mechanisms dominate: early tamponade from post-operative bleeding into the pericardial space, and delayed tamponade from a progressively accumulating effusion, sometimes linked to post-operative anticoagulation or a post-pericardiotomy inflammatory response. The distinction matters for prediction, because the early, bleeding-driven form is the one a pre-operative risk profile is most likely to anticipate.

The underlying physiology is mechanical. As fluid collects faster than the stiff pericardium can stretch, intrapericardial pressure rises and equalises with the diastolic pressures of the cardiac chambers, which limits filling, so central venous pressure climbs while stroke volume and arterial pressure fall[3]. Detection rests on a combination of clinical signs and imaging. The classic findings, a falling arterial pressure, a rising central venous pressure, muffled heart sounds, and the exaggerated inspiratory fall in systolic pressure known as *pulsus paradoxus*, are specific but appear late, and echocardiography is the diagnostic standard. The practical difficulty, and the motivation for this work, is that the early signs are non-specific and easily attributed to

[1] You, S. C., Shim, C. Y., Hong, G. R., Kim, D., Cho, I. J., Lee, S., Chang, H. J., et al. (2016). *Incidence, Predictors, and Clinical Outcomes of Postoperative Cardiac Tamponade in Patients Undergoing Heart Valve Surgery.* PLoS ONE, 11(11), e0165754.

[2] Khan, N. K., Järvelä, K. M., Loisa, E. L., Sutinen, J. A., Laurikka, J. O., & Khan, J. A. (2017). *Incidence, Presentation and Risk Factors of Late Postoperative Pericardial Effusions Requiring Invasive Treatment after Cardiac Surgery.* Interactive CardioVascular and Thoracic Surgery, 24(6), 835–840.

[3] Spodick, D. H. (2003). *Acute Cardiac Tamponade.* New England Journal of Medicine, 349(7), 684–690.

the expected course of recovery, whereas deterioration once tamponade is established can be abrupt.

The closest prior attempts to predict tamponade come from electrophysiology rather than from surgery. In atrial-fibrillation catheter ablation, where tamponade follows catheter perforation rather than post-operative bleeding, Bansal et al.[4] reported the first machine-learning model for pericardial tamponade after ablation, and Bao et al.[5] trained gradient-boosted models on 1,481 ablation patients, reaching an internal-validation AUC of 0.908 with operator experience, D-dimer, total heparin dose, atrial-fibrillation type, and left-atrial diameter as the leading predictors. Two points matter here. Both models predict tamponade from clinical and procedural covariates rather than from the waveform, which is, by a different route and in a different setting, the conclusion this paper also reaches. But both describe a mechanically different event: an acute intra-procedural perforation, predicted largely by how the procedure was performed, whereas the post-surgical form develops over hours to weeks from bleeding into a pericardium that surgery has already opened. Their predictors do not transfer to this setting, and neither model has been externally validated. We are therefore not aware of a widely validated model that flags which post-*cardiac-surgery* patients will progress to severe tamponade, and that gap is what a prediction model would address.

2.2 Predictive Analytics in Perioperative Care

The use of routinely collected clinical data to predict perioperative outcomes is well established. The reference risk tools in cardiac surgery, EuroSCORE II[6] and the Society of Thoracic Surgeons models[7], are built and validated for operative mortality on hundreds of thousands of patients and are the yardstick for any new mortality model. No comparably validated score exists for a specific complication such as post-operative tamponade, which is the gap this work addresses; a general mortality or acuity score can at best serve as an indirect, off-label baseline. Beyond these scores, machine-learning models predict operative mortality after cardiac surgery from pre-operative and intra-operative variables in the electronic health record[8], and related work targets acute kidney injury, prolonged ventilation, and other complications. These models are usually judged by discrimination (the area under the receiver operating characteristic curve,

[4]Bansal, A., et al. (2022). *Machine Learning Prediction of Pericardial Tamponade After Atrial Fibrillation Ablation.* The American Journal of Cardiology, 175, 179–180.

[5]Bao, Z., et al. (2026). *Explainable Machine Learning for Risk Prediction of Acute Cardiac Tamponade during Atrial Fibrillation Ablation.* Scientific Reports, 16(1), 9476.

[6]Nashef, S. A., Roques, F., Sharples, L. D., et al. (2012). *EuroSCORE II.* European Journal of Cardio-Thoracic Surgery, 41(4), 734–744.

[7]Shahian, D. M., Jacobs, J. P., Badhwar, V., et al. (2018). *The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 1, Background, Design Considerations, and Model Development.* The Annals of Thoracic Surgery, 105(5), 1411–1418.

[8]Kilic, A., et al. (2020). *Predictive Utility of a Machine Learning Algorithm in Estimating Mortality Risk in Cardiac Surgery.* The Annals of Thoracic Surgery, 109(6), 1811–1819.

AUROC), by calibration[9], and increasingly by decision-curve analysis[10], the same conventions this paper adopts, and reported under the TRIPOD+AI guidance for prediction models[11]. The methodological literature is equally clear on how such models should be built and checked, from predictor coding through model estimation, internal validation, and presentation[12]. Rare outcomes raise the stakes at every step: with few events a model can post a respectable area under the curve while remaining poorly calibrated or over-fitted, so discrimination alone is an unreliable verdict, and the calibration and permutation checks this paper reports become necessary rather than optional.

A distinct strand uses the continuous physiological signals recorded in the operating room and intensive-care unit. The best-known example in routine practice is the Hypotension Prediction Index, a commercial model that anticipates intra-operative hypotension from the arterial-pressure waveform[13]. Such work shows that bedside waveforms can carry clinically actionable predictive information, but it also illustrates the hazard this paper guards against: the index's apparent performance was later shown to be inflated by a selection-bias artifact in how prediction windows are sampled[14]. Two features of the existing literature matter here. First, most successful waveform models predict a common, near-term physiological state on large datasets, whereas tamponade is a rare and delayed clinical event. Second, the methodological hazards of small-sample, high-dimensional waveform analysis, namely leakage, confounding by monitoring intensity, and feature-selection optimism, are seldom addressed explicitly, even though they can manufacture apparently strong results. This paper treats those hazards as first-class concerns.

2.3 Physiological Waveform Modeling

This paper belongs to a growing line of work that extracts a large bank of features from perioperative waveforms and feeds them to a gradient-boosted or penalized classifier. Pal et al.[15] sit closest to it. They applied a Light Gradient-Boosting Machine to 843 features from the photoplethysmography (PPG) waveform to identify patients with elevated central venous

[9]Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). *Calibration: the Achilles Heel of Predictive Analytics.* BMC Medicine, 17, 230.

[10]Vickers, A. J., & Elkin, E. B. (2006). *Decision Curve Analysis: A Novel Method for Evaluating Prediction Models.* Medical Decision Making, 26(6), 565–574.

[11]Collins, G. S., Moons, K. G. M., Dhiman, P., et al. (2024). *TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods.* BMJ, 385, e078378.

[12]Steyerberg, E. W., & Vergouwe, Y. (2014). *Towards Better Clinical Prediction Models: Seven Steps for Development and an ABCD for Validation.* European Heart Journal, 35(29), 1925–1931.

[13]Hatib, F., Jian, Z., Buddi, S., Lee, C., Settels, J., Sibert, K., Rinehart, J., & Cannesson, M. (2018). *Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis.* Anesthesiology, 129(4), 663–674.

[14]Enevoldsen, J., & Vistisen, S. T. (2022). *Performance of the Hypotension Prediction Index May Be Overestimated Due to Selection Bias.* Anesthesiology, 137(3), 283–289.

[15]Pal, R., Rudas, A., Chiang, J. N., Barney, A., & Cannesson, M. (2025). *Machine Learning-Based Identification of Patients with Elevated Central Venous Pressure Using Features Extracted from Photoplethysmography Waveforms.* medRxiv 2025.12.30.25343231.

pressure, reaching an AUC of 0.79 (95% CI 0.71 to 0.84) on 1,665 surgical patients, 483 of them with elevated CVP, from the UCLA perioperative dataset. They share this paper's feature-bank approach, its tree-based models, and its perioperative setting.

Two differences set this work apart, and the first is a distinction worth stating precisely because the word "predict" is used loosely in this literature. Pal et al. are *identifying*, not predicting: their model reads a photoplethysmogram and infers the central venous pressure measured at that same moment, so the target is a concurrent physiological state that a catheter could report directly. This paper predicts a future clinical event, tamponade anywhere from hours to weeks after the analysis window closes. Identification and prediction are different tasks with different ceilings, and an identification AUC is not a benchmark a prediction model should be expected to match. Their cohort also holds 483 positive cases against 27 here, so the problem that drives this paper, drawing reliable conclusions from a rare event, does not arise in theirs. Their result sets a useful upper bound on what the approach achieves when data are plentiful, and it motivates the leakage and optimism controls that form the methodological contribution here.

The wider field of physiological-signal modeling underlines the same point about scale. Open signal archives such as PhysioNet have made large volumes of waveform data available for research and have become the backbone of the field[16], and deep-learning methods trained on those volumes can reach specialist performance: a deep neural network trained on more than ninety thousand ambulatory recordings classifies arrhythmias at cardiologist level[17]. These results share a precondition the present setting lacks, namely tens of thousands of labeled examples. At a single center with tens of events, the same high-capacity methods overfit, which is why this paper favors interpretable models and spends its effort on confound control rather than model capacity. A positive waveform result here would be a feasibility signal to scale up, not a finished detector.

2.4 Gap Addressed by This Work

Three gaps define the contribution of this work. First, where the closest comparator predicts a concurrent physiological state on a large cohort, this paper predicts a future clinical event, severe tamponade occurring days after the analysis window, on a verified but small single-center cohort: a harder and more realistic prospective task. Second, the paper makes the methodological hazards explicit and builds the evaluation around them, through a bed-anchored attribution method that resolves the reuse of study identifiers across patients, an explicit correction for the monitoring-intensity confound, the treatment of informative missingness as a confound rather

[16]Goldberger, A. L., Amaral, L. A. N., Glass, L., et al. (2000). *PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals.* Circulation, 101(23), e215–e220.

[17]Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., et al. (2019). *Cardiologist-level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network.* Nature Medicine, 25(1), 65–69.

than a free signal, and a permutation-based protocol that removes feature-selection optimism. Third, rather than asking only whether waveforms help, the paper sets the question within a four-model design (Section 1.4) that compares the pre-operative record against three successively richer data sources on one fixed outcome, so that the central finding, that incremental value over the pre-operative record is not established at the current sample size, rests on several independent comparisons rather than one.

Chapter 3

Data and Cohort Construction

3.1 Data Sources

Three sources feed the study. Stanford's clinical-data platform supplies the structured clinical record: demographics, pre-operative laboratory values, surgical variables, diagnosis and procedure codes, intensive-care flowsheets, and the admission, discharge and transfer (ADT) log.[1] The hospital's bedside-monitor archive supplies the physiological signals: low-frequency Philips numeric trends (heart rate, beat-to-beat heart rate, respiratory rate, SpO_2, perfusion index, non-invasive blood pressure, and PVI, near one to two hertz) and the raw high-frequency waveforms (arterial, central-venous and pulmonary-artery pressure and the plethysmogram at 125 Hz, the electrocardiogram at 500 Hz, and respiration at a lower rate). Each recording carries a small companion file that names the monitoring unit and bed.[2] The ADT log links the two: it places each patient in a specific unit and bed over time, which lets the study attribute a recording to the patient who produced it.

3.2 Ethics and Data Governance

The study analyzes de-identified clinical and physiological data collected in routine care, under an approved Stanford University institutional review board protocol with a waiver of informed consent for this retrospective analysis. Two privacy measures shape the data. Recording dates carry a per-patient random offset, which the analysis removes only to align a recording to its own surgery and event, so no real calendar date leaves the secure environment. Race is recorded but excluded from every model by design, so the models cannot learn it as a predictor.

[1] The structured data come from a restricted-access hospital clinical-data warehouse, queried through Google BigQuery.

[2] The waveform recordings are stored as per-recording files in a secure Google Cloud Storage bucket.

3.3 Outcome Definition

One outcome holds for every model: severe, incident, post-operative tamponade. A patient counts as positive on a pericardial drainage procedure (ICD-10-PCS `0W9D3` or `0W9D4`; CPT `33016`, `33017`, `33020`) or a diagnosis of cardiac tamponade (`I31.4`) or hemopericardium (`I31.2`), recorded at least one day after the index surgery and within thirty days. The one-day rule is intended to keep the event incident, by separating a new post-operative event from bleeding handled during the operation itself. Its limitation is that it cuts on the calendar day rather than on elapsed time, so a patient operated in the morning and drained late the same evening is excluded while one operated in the evening and drained after midnight is kept, although both are post-operative events. The rule is therefore under review: we are re-anchoring it on the surgery end time, and the timing distribution this produces will decide whether the same-day group is reclassified.

Pericarditis (`I30.9`) is excluded from the study entirely. It is a different disease from tamponade rather than a milder form of it, and patients carrying it are neither cases nor controls. Pericardial effusion (`I31.3`) is likewise excluded from the positive class: it lies on the same pathophysiological axis as tamponade but without compression, so the paper holds it aside and revisits it only as a sensitivity arm (Chapter 6), where pooling it into the outcome is shown to lower discrimination. A negative is a cardiac-surgery patient with no pericardial code at all. The study audits the control pool and removes 158 patients who carry peri-operative pericardial codes.

3.4 The Cohort Funnel

The cohort is built as a short sequence of filters, each one a decision with a reason. Table 3.1 gives the count after each step. A single source population runs through one pipeline, one row per patient, which removes the cross-source missingness and duplication that can leak across cross-validation folds.

Source. The study starts from the 7,653 adult cardiac-surgery operations in the institutional surgical cohort, performed between 2016 and 2025.

Flag tamponade. A query joins that cohort to the procedure and diagnosis tables and keeps any patient with a pericardial drainage procedure or a pericardial diagnosis within thirty days of surgery. The window was widened from the original fourteen days to thirty so that delayed events, which typically appear one to three weeks after surgery, are not missed; and a procedure *or* a diagnosis qualifies, rather than requiring both, so a drained patient without a coded diagnosis is still caught. This flags 894 candidates (794 with a diagnosis only, 94 with both, 6 with a procedure only).

Keep severe. Each candidate is tagged from its codes as severe tamponade (cardiac tam-

ponade I31.4, hemopericardium I31.2, or a drainage procedure), effusion (I31.3), or pericarditis (I30.9). Effusion and pericarditis are different diseases and are set aside, which leaves 279 severe cases.

Keep incident. A case counts only if its first qualifying event falls at least one day after surgery. Dropping the 184 same-day cases leaves 95 incident severe cases. The paper previously read those 184 as hemopericardium present at or handled during the operation. That reading is under review: a same-day code marks the calendar date, not the hour, so it cannot by itself distinguish bleeding controlled in theatre from a tamponade that develops overnight and is drained before midnight. The audit described in Section 3.3 settles this on elapsed time from the surgery end, and the severe cohort is reported with and without the same-day group as a sensitivity analysis. This reading is reinforced externally: in MIMIC-IV, 40 of 109 pericardial drainages (37 percent) fall on the day of surgery, so the mixed, largely intra-operative character of same-day pericardial codes is a general feature of how the outcome is coded and not a Stanford artefact (Section 6.5).

Build the controls. The controls are the 6,761 cardiac-surgery patients carrying no pericardial code, after a control-pool audit removed 158 patients whose peri-operative pericardial codes the cohort query had missed. The audit also resolves the few patients who would otherwise sit in both the candidate and the control set, so the case and control groups do not overlap.

Require a recording. The structured models use all 95 cases. The waveform models additionally require a verified, bed-anchored recording in the analysis window; 27 procedure-confirmed severe cases have one, and that is the waveform cohort. These 27 are defined by the recording requirement, not carved out of the 95, which is why the two arms report different counts.

Table 3.1: Cohort attrition, with the count after each filter.

Step	**Filter**	**Count**
Source	cardiac-surgery operations, 2016–2025	7,653
Flag tamponade	drainage or pericardial diagnosis within 30 days	894
Keep severe	I31.4 / I31.2 / drainage (effusion and pericarditis set aside)	279
Keep incident	first event at least one day post-op (184 same-day removed)	95
Controls	cardiac-surgery patients with no pericardial code, audited	6,761
Waveform arm	severe cases with a verified bed-anchored recording	27

The same disease and the same labels yield different sample sizes, set by what each model needs (Table 3.2). The structured models run on 95 events and 6,761 controls across the full 2016–2025 era, because almost every patient has covariates. The strictest definition, the procedure-anchored drainage-confirmed cohort, depends on the window: the original fourteen-day query holds 44 cases, and widening the window to thirty days, consistent with the rest of the cohort, raises this to 88. The thirty-day count is the more coherent one, since every other filter in this study uses a thirty-day horizon. The waveform models, however, are narrower still: of the severe cases only 27 have a recording that both falls in a usable pre-event window and survives bed-anchored attribution, so 27 is the count of true positives the waveform arm can actually learn from, well below either drainage figure.

The timing of the 794 diagnosis cases sharpens this. Counting days from the index surgery to the tamponade diagnosis, 418 (53 percent) fall on the day of surgery and 60 on the next day, so 478 of 794 (60 percent) are coded within a day of the operation. Whether those are intra-operative or early post-operative events is exactly what the calendar-day rule cannot tell us, and what the elapsed-time audit resolves. The remaining 316 sit two or more days out, and 217 of those carry a recording date after the 2020 archive floor. The 217 set the ceiling for an expanded waveform analysis, although bed-matching recovers only a few of them.

3.5 Record Attribution: the Bed-Anchored Method

The waveform archive identifies each recording by a study identifier that recurs across many patients and dates. A join on that identifier attaches a recording to the wrong patient and corrupts the labels without warning. The bed-anchored method removes the risk: it accepts a recording for a patient only when the monitoring unit and bed named in the recording's companion file appear in that patient's ADT log during the recording window. The unit and bed, not the reused identifier, decide ownership.

Three independent checks confirm the attribution. The internal-consistency check requires the recording window to fall inside the patient's hospital stay. The encounter check maps each recording to a single encounter. The ADT check confirms against the admission log that the patient occupied the named bed at the recording time. A recording passes only when all three agree, which rules out a data-attribution artifact behind any later result.

A second hazard is temporal. For privacy, every recording's date carries a per-patient random offset, so the timestamp stored in a file is not the real calendar time, and aligning a recording to its surgery and to the event means recovering that offset and subtracting it. The offset is identified per patient from a fixed anchor: an operating-room recording must fall on the surgery date. The same anchor doubles as a validity check, because once the offset is removed the operating-room recording should land exactly on the day of surgery, and any window that does not is rejected. An early version of the pipeline applied the correction twice,

which this check caught before it could shift every window by a constant and silently invalidate the analysis.

3.6 The Modeling Cohorts

The four models use the cohorts in Table 3.2. Model 1 and the structured comparisons draw on the full 95 events and 6,761 controls, the latter being every cardiac-surgery patient with no pericardial code. The waveform models draw on the 27 verified positives, set against a non-tamponade group from the same pool, sampled to match the cases on procedure type and case classification and required to have a recording in the same post-operative window: 248 patients for the numeric-trend analysis and 120 for the raw-waveform analysis. The matching removes the monitoring-intensity asymmetry that an unmatched sample would carry. Pre-operative laboratory values are missing in roughly 15 to 45 percent of patients, more often in the urgent cases; the models impute them within the cross-validation folds (Chapter 5), and the missingness audit in Chapter 6 confirms the imputation does not drive the result. Severe-only stays the primary outcome; the pooled and procedure-anchored drainage-confirmed definitions serve as sensitivity arms.

Table 3.2: The modeling cohorts. Each sample size follows the data the analysis requires.

Analysis	**Severe positives**	**Comparators**
Pre-operative (Model 1)	95	6,761 audited controls
Early-ICU structured (Model 2)	78	sampled controls
Missingness-aware (Model 3)	95	6,761
Waveform, numeric trends (Model 4)	27	248
Waveform, raw 125 Hz (Model 4)	27	120
Drainage-confirmed, procedure-anchored (sensitivity)	44 (14-day) / 88 (30-day)	6,761
Pooled with effusion (sensitivity)	508	6,761

The cohort funnel appears in Figure 3.1.

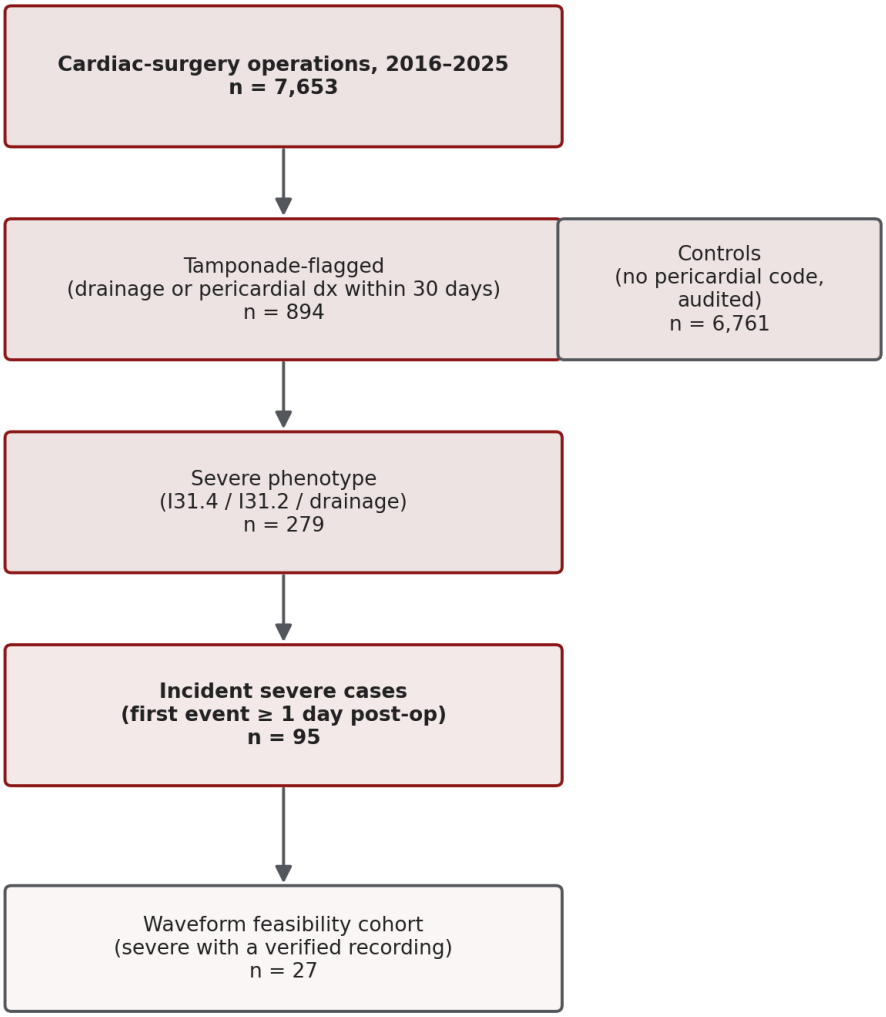

Figure 3.1: Cohort construction funnel.

Chapter 4

Exploratory Data Analysis

4.1 Clinical Covariates

Table 4.1 sets the severe cases against the controls. The groups separate on the variables that track surgical bleeding and physiological reserve. Severe cases have lower left-ventricular function, lower pre-operative hemoglobin, higher frailty, longer cardiopulmonary bypass, and longer surgery, each at $p < 0.001$, with lower eGFR as well ($p = 0.013$). Age and HbA1c do not differ. The patients who develop severe tamponade are frailer, bleed more, and undergo longer, more complex operations. This is a bleeding-and-frailty signature, not an age effect, which already hints that the predictive signal sits in the pre-operative record.

Table 4.1: Cohort characteristics by class, median [IQR]; p from the Mann-Whitney test. The final column states where each variable enters the model, so the table can be read as a specification and not only as a description: *1a* = pre-operative model, *1b* = added in the peri-operative model.

Variable	**Severe**	**Control**	p	**In model**
Age (years)	64 [56, 73]	65 [56, 73]	0.60	1a
BMI	25.6 [23.1, 29.1]	26.9 [23.8, 30.8]	0.05	1a
LVEF (%)	55 [42, 61]	59 [50, 64]	<0.001	1a
eGFR	75 [48, 96]	84 [66, 97]	0.013	1a
Pre-op hemoglobin	12.1 [9.5, 13.6]	13.6 [12.1, 14.7]	<0.001	1a
HbA1c	5.6 [5.3, 6.0]	5.6 [5.3, 6.1]	0.96	1a
Frailty score	11.8 [3.6, 20.5]	3.7 [1.4, 8.7]	<0.001	1a
Bypass time (min)	145 [110, 214]	118 [83, 177]	<0.001	1b
Surgery length (h)	8.3 [6.6, 10.1]	6.9 [5.7, 8.6]	<0.001	1b

Figure 4.1 shows the same covariates as distributions by class, which makes the separation on frailty, ejection fraction, hemoglobin, and surgical duration visible at a glance.

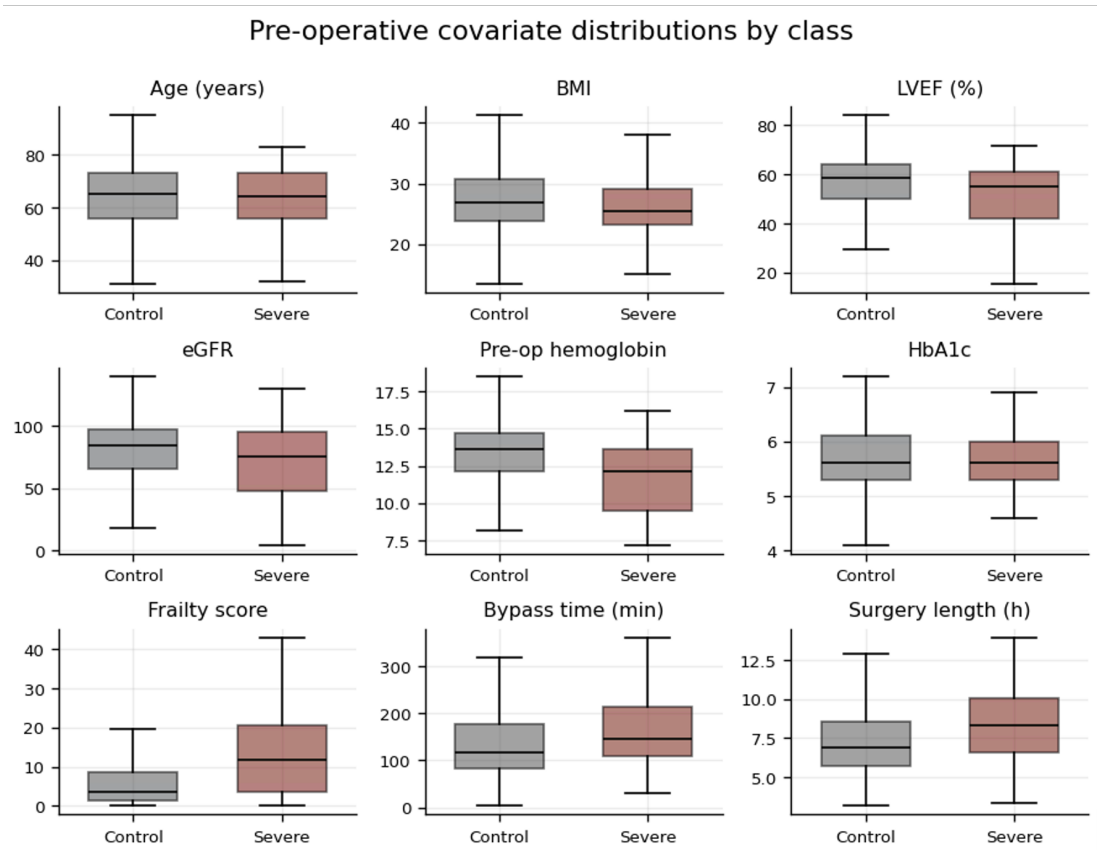

Figure 4.1: Pre-operative covariate distributions by class.

The continuous covariates are compared with the Mann-Whitney U test, a rank-based test of whether values drawn from one group tend to exceed those drawn from the other. It suits the data: the clinical variables are skewed and carry outliers, the two groups are of very unequal size (95 against 6,761), and a rank test assumes no particular distribution and resists both problems. In practice the test pools all observations, ranks them, sums the ranks within each group, and compares that sum to its expected value under the null of no difference; the reported p is the probability of a rank separation at least this large arising by chance. Binary surgical descriptors, such as the circulatory-arrest and selective-perfusion flags, which are covariates in Model 1b and not merely exploratory variables, call for the chi-squared test instead, which compares observed against expected counts in a contingency table. Because the control group is large, p-values fall easily, so the table reports medians and interquartile ranges next to each p and the reading leans on the size of the gap, not significance alone. These comparisons are exploratory and hypothesis-generating; the formal evaluation lives in the modeling chapters.

Missingness is modest and mixed, not uniformly higher in either class. Glycated hemoglobin is missing more often in the severe group (about 16 versus 7 percent), consistent with urgent admissions skipping parts of the pre-operative workup, whereas left-ventricular ejection fraction is missing more often in the controls (about 11 versus 3 percent); the remaining variables differ by a point or two. Chapter 5 tests whether this missingness carries signal of its own, and

Chapter 6 shows it does not change the headline.

4.2 Categorical Risk Factors

The categorical picture reinforces the bleeding-and-complexity reading (Table 4.2). Severe cases are markedly enriched for a pre-operative bleeding-and-instability profile: pre-operative shock (33 versus 8 percent), coagulopathy (43 versus 16 percent), and thrombocytopenia (33 versus 12 percent) each separate the groups at $p < 0.001$, as does selective antegrade cerebral perfusion (10 versus 3 percent), a marker of complex aortic surgery. Circulatory arrest is higher in severe cases but short of significance (19 versus 13 percent, $p = 0.10$), and sex and diabetes do not differ. The outcome itself is rare, near 1.4 percent, roughly one severe case for every seventy patients, which sets the rare-event challenge the evaluation in Chapter 5 is built around.

Table 4.2: Categorical characteristics by class, percent positive; p from the chi-squared test. Tables 4.1 and 4.2 are split by the test each variable requires, rank-based for continuous and chi-squared for binary, not by whether the variable is in the model; the *In model* column carries that information in both. *Tested* means the variable was added to the model and did not improve discrimination (Section 4.2).

Variable	**Severe (%)**	**Control (%)**	p	**In model**
Pre-operative shock	33	8	<0.001	Tested, not included
Coagulopathy	43	16	<0.001	Tested, not included
Thrombocytopenia	33	12	<0.001	Tested, not included
Selective cerebral perfusion	10	3	<0.001	1b
Circulatory arrest	19	13	0.10	1b
Diabetes	29	26	0.53	No
Male sex	72	69	0.65	No

Three of these flags, pre-operative shock, coagulopathy, and thrombocytopenia, are strong on their own yet sit outside the pre-operative model, and this is deliberate and tested. Added one at a time to the eleven-covariate model, none improves discrimination: the out-of-fold area under the curve moves from 0.74 to 0.74, 0.73, and 0.73 respectively, every confidence interval spanning zero. Their signal is already represented by frailty and ejection fraction, with which they correlate, so the parsimonious model loses nothing by excluding them. The exercise answers the obvious question, why not add these familiar risk factors, with evidence rather than assertion.

4.3 Collinearity

The pre-operative covariates are close to mutually independent, so the logistic coefficients stand for distinct effects rather than splitting across proxies. Every variance-inflation factor is below 2.1, well under the conventional threshold of 5, and the only sizable pairwise correlation is between cardiopulmonary bypass duration and surgery length (Spearman 0.68), expected because longer operations spend longer on bypass. Figure 4.2 gives the full correlation matrix.

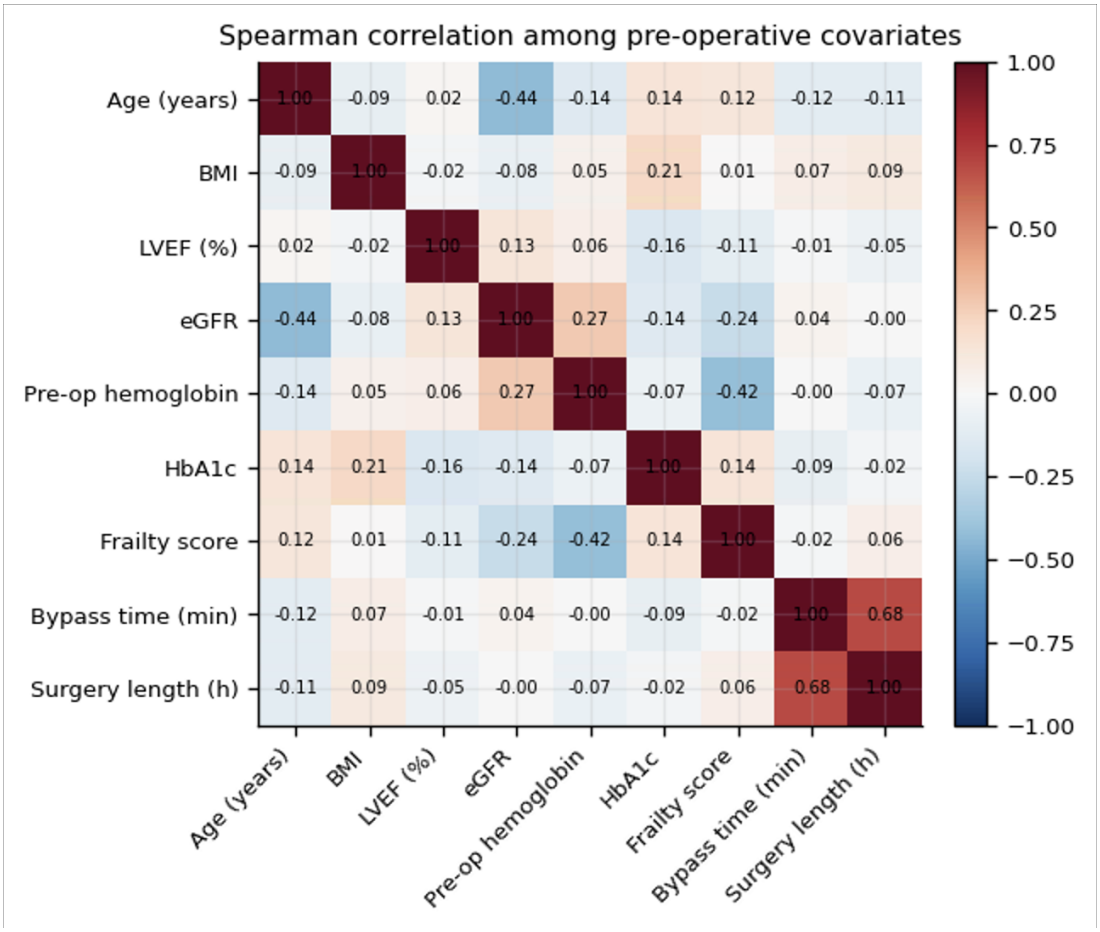

Figure 4.2: Correlation among the pre-operative covariates.

4.4 Waveform Data and Coverage

The bedside-monitor archive supplies two kinds of signal: low-frequency numeric trends near one to two hertz (heart rate, oxygen saturation, perfusion index, respiratory rate) and the raw high-frequency waveforms (arterial, central-venous and pulmonary-artery pressure and the plethysmogram at 125 Hz, the electrocardiogram at 500 Hz). The feasibility cohort is small by necessity, because it needs a verified, bed-anchored recording in the analysis window and the archive only begins in 2020. Coverage, not modeling, is the binding constraint: a usable invasive recording, the arterial and central-venous channels that carry tamponade physiology,

exists for only a minority of severe cases, and those recordings cluster in the first day or two after surgery while the severe events arrive a week or two later. The waveform data thin exactly where the diagnostic signal would appear. Among the cases that do carry a recording, no feature separates the classes reliably. The mechanistically expected signs of tamponade, the respiratory variation in central-venous and arterial pressure that underlies pulsus paradoxus, do not differ between classes at this sample size (Figure 4.3). A few non-specific features, reduced heart-rate variability and higher oxygen saturation and higher *non-invasive arterial* diastolic pressure, reach nominal significance, but with a few tens of patients and dozens of features tested these are exploratory and uncorrected. The diastolic pressure here is the non-invasive cuff measurement from the monitor's numeric channels, not pulmonary-artery diastolic pressure; arterial diastolic pressure is not a mechanism of tamponade, and we read its appearance as noise rather than signal. The mechanistically relevant quantity, the narrowing gap between pulmonary-artery diastolic pressure and central venous pressure, is a flowsheet measurement rather than a waveform feature and so does not appear in this figure. The waveform analysis is therefore reported as a feasibility study rather than a tested model, and the honest reading is a null at the current sample size.

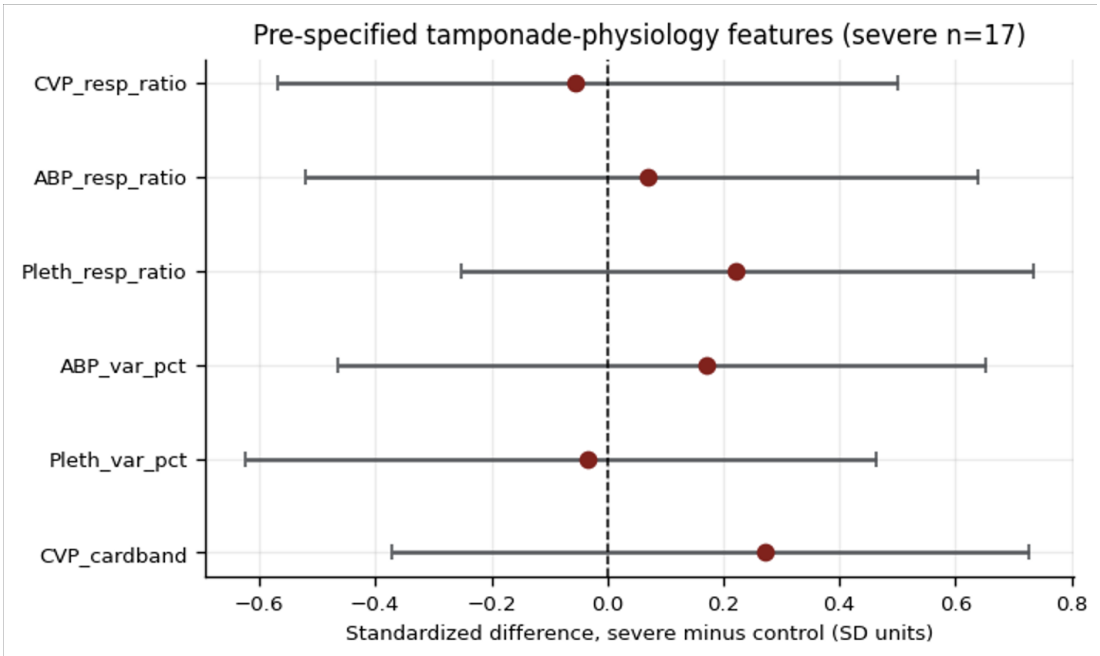

Figure 4.3: Standardized class difference for the pre-specified tamponade-physiology waveform features; all intervals span zero at the feasibility sample size.

4.5 Confounders

Two confounds shape every modeling choice that follows. Surgical acuity comes first: severe cases are frailer and undergo longer, more complex surgery, so a model earns credit only for what

it adds beyond LVEF, hemoglobin, frailty, and surgery duration, or it rewards acuity dressed up as physiology. Monitoring intensity comes second: invasive lines are placed preferentially in the sicker, more complex patients, so the severe cases carry *more* invasive channels in the recorded window than the controls do (about 70 percent against 50 percent), and the mere presence of a channel therefore tracks the outcome. The waveform models answer both by matching controls on the post-operative window and by modeling channel presence directly. Age, by Table 4.1, is not a confounder here. Figure 4.4 contrasts the two: age is flat across classes while surgery length separates them.

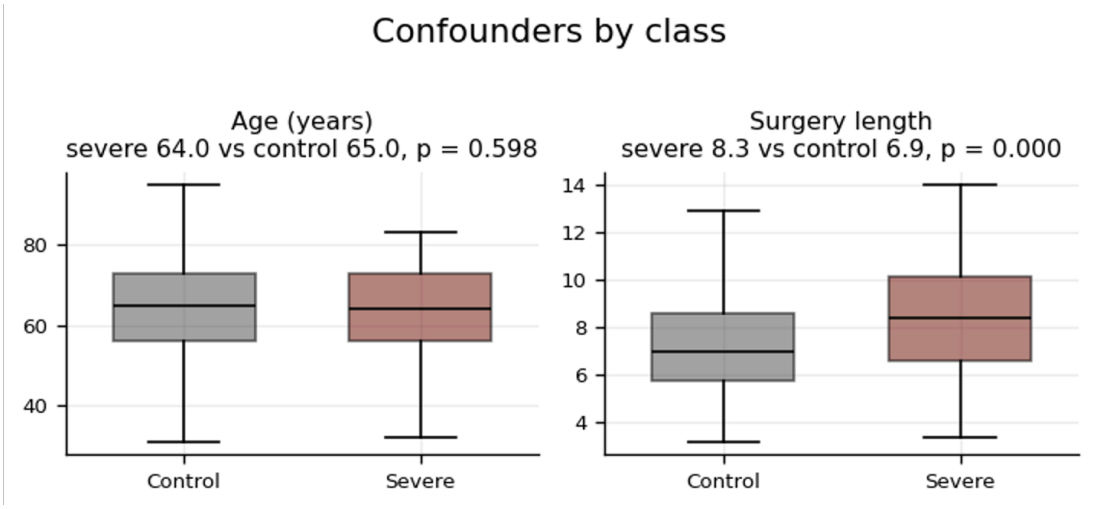

Figure 4.4: A non-confounder and a confounder: age versus surgery length by class.

Chapter 5

Methodology

5.1 Feature Extraction

Model 1 uses eleven covariates, in two nested specifications. *Model 1a*, the strictly pre-operative model, uses the seven variables available before the patient enters the operating room: age, BMI, LVEF, eGFR, pre-operative hemoglobin, HbA1c, and frailty score. *Model 1b*, the peri-operative model, adds the four that are known only once the operation is under way: cardiopulmonary bypass duration, surgery length, and the circulatory-arrest and selective-antegrade-cerebral-perfusion flags. Race is excluded by design.

These eleven were originally chosen by clinical judgment and held fixed, which is a legitimate basis but was not a stated rule. The selection is therefore pre-specified here. Three variables are forced in on clinical grounds, because the mechanism is bleeding into a pericardium and tolerance of that bleeding: LVEF, pre-operative hemoglobin, and frailty score, with bypass duration and surgery length forced in additionally for Model 1b. Every remaining candidate in the clinical record is then screened against the outcome *inside each cross-validation fold*, never on the full data, so that the screening cannot borrow information from the patients it is later evaluated on. The number of screened variables is chosen from the events-per-variable trade-off reported in Chapter 6 rather than by taking whichever count scores highest.

The early intensive-care model adds the first forty-eight hours of structured data: laboratory values (hemoglobin, hematocrit, lactate, creatinine, platelets), chest-tube output and its trend, and vital-sign trends, with the hour as the unit of analysis. The waveform model extracts features from the bedside signals: time-domain summaries, fast-Fourier-transform band powers, smoothed variability, pulse-pressure and systolic-pressure variation (the quantitative form of pulsus paradoxus), perfusion-index variability as a non-invasive surrogate, and central-venous-pressure morphology, computed on both the low-frequency numeric trends and the raw 125 Hz signals.

5.2 Time Anchor

Each waveform window is anchored at surgery start and extended forward over six, twelve, and eighteen hours. The first-48-hour window is the one the data support, because the recordings live there; the pre-event window where classic pulsus would appear holds only about nine patients and cannot be modeled (Chapter 4). The waveform analysis is therefore an early-warning model that predicts a later event from the first day or two of monitoring, not a last-hours detector.

5.3 Models

The reference model is an L2-penalized logistic regression, chosen for interpretability and calibration. It is compared against lasso, elastic net, gradient boosting, and random forest under *nested* cross-validation, and the nesting is what makes the comparison honest. An inner three-fold loop, run inside each outer training fold, selects each model's hyperparameters by grid search; the outer five-fold loop, repeated over random splits, does nothing but evaluate. The inner loop never sees the outer test fold, so the reported figures do not carry the selected configuration's luck, which a single loop that both tunes and evaluates would report as performance. The grids are deliberately narrow, because at 93 events a wide search selects noise rather than structure: regularization strength over four values for the penalized regressions, with the elastic-net mixing parameter over three; the number of trees, depth, and learning rate over two values each for gradient boosting; depth and minimum leaf size for the random forest.

The tuning describes the sample as much as it describes the models, and we report that rather than only the winning numbers. The L2 model selects the same regularization strength in all five outer folds. Gradient boosting selects four different configurations across five folds, and the random forest and elastic net three each, which is the signature of a search fitting noise rather than finding structure. That instability is a result in its own right: at this event count the data supports tuning a one-parameter penalized regression and does not support tuning a tree ensemble, so the tree models' selected hyperparameters are not interpreted. One caveat we state rather than bury: the L2 model selects the smallest regularization strength the grid offers, in every fold, so the grid's lower bound is binding and the true optimum may lie below it. Its reported figure is therefore a floor rather than a ceiling. Each later data source, early intensive care, informative missingness, and the waveform, is tested twice: on its own, and stacked onto the out-of-fold predictions of the pre-operative model, so the question is always incremental value rather than raw performance. The waveform models use controls matched on the post-operative window and an explicit channel-presence term, which strips out the monitoring-intensity confound.

The waveform cohort is defined differently from the structured one. It comprises the 27

severe cases with a verified bedside recording, set against sampled control recordings in a ratio between one to five and one to nine, rather than the full 95 cases and 6,761 controls of the pre-operative model. Because discrimination is measured by the area under the ROC curve, which is threshold-free and insensitive to the case-to-control ratio, no class-balancing or resampling scheme is applied and the conclusions do not depend on the exact ratio chosen. Separately, a convolutional neural network on the raw waveform was attempted but overfit at this sample size, consistent with the data-scale argument of Chapter 2, and is not reported as a formal comparison.

5.4 Evaluation Protocol

Discrimination is the area under the ROC curve, estimated with patient-grouped repeated stratified cross-validation: the data are split into five folds stratified on the outcome, the split is repeated several times with different seeds, and the fold-level estimates are pooled, so the figure does not hinge on one lucky partition and no patient appears in both training and test within a fold. Confidence intervals come from bootstrap resampling. Calibration, how closely predicted risks match observed rates, is summarized by the calibration slope and inspected on a decile calibration curve, which plots observed against expected event rates within tenths of predicted risk[1], and clinical usefulness is read from a decision-curve analysis, which weighs true against false positives across a range of risk thresholds[2].

Three metrics are reported rather than one, and the reason is the event rate. At 1.4 percent, accuracy is uninformative, and discrimination alone is not enough to justify clinical use: a model can rank patients well and still put the absolute risks in the wrong place, which is what calibration catches, and it can be both discriminating and calibrated while still doing more harm than good at any threshold a clinician would actually use, which is what decision-curve analysis catches. We do not report sensitivity, specificity, or predictive values as headline figures, because each depends on a chosen cut-point and no cut-point has been agreed for this outcome; the decision curve subsumes them by sweeping the whole threshold range instead of privileging one.

Incremental value, the central question, is judged two ways. A later source is stacked onto the out-of-fold predictions of the pre-operative model, and the difference in AUROC is tested with the DeLong method[3] and a paired bootstrap of the stacked-minus-base difference on the same patients. To guard against feature-selection optimism, the danger that screening many features on few events yields a strong-looking result from noise, the labels are permuted many times and the whole pipeline re-run to build a null distribution, against which the observed

[1] Van Calster et al. (2019); see Bibliography.

[2] Vickers & Elkin (2006); see Bibliography.

[3] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). *Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach.* Biometrics, 44(3), 837–845.

value is read. Feature selection and imputation run inside each training fold, never on the full data, so no information leaks from test to train. Every post-operative feature is timestamped strictly before the event, and any marker that is part of the event itself, the drainage and the vasopressors started in response, stays out of the primary model. Reporting follows the TRIPOD+AI guidance for prediction-model studies[4].

5.4.1 Rare Events and Class Imbalance

The outcome is rare, near one in seventy, which shapes the evaluation. Discrimination and calibration, not accuracy, are the reported metrics, because accuracy is uninformative when the negative class dominates[5]. The reference model is left unweighted so its predicted probabilities stay calibrated to the true event rate; class weighting, which rebalances the classes toward parity, inflates predicted risk far above the observed rate and is therefore not used. For the waveform arm the controls are sampled to the cases at a ratio between one to five and one to nine, and because the area under the curve is threshold-free this ratio does not bias the estimate, so no oversampling or synthetic-minority scheme is applied. The small event count is treated as a limit on power, reflected in wide confidence intervals, rather than concealed by resampling.

[4] Collins et al. (2024); see Bibliography.
[5] Steyerberg & Vergouwe (2014); see Bibliography.

Chapter 6

Results

6.1 The Pre-operative Model

The pre-operative covariate model discriminates incident severe tamponade with an AUROC of 0.74 (95% CI 0.68 to 0.79) on 95 events and 6,761 audited controls. It survives two checks that could have moved it and did not. Re-running it on the audited label, 93 events against 6,755 controls, gives 0.735; tuning it under nested cross-validation on that cohort gives 0.746 (95% CI 0.691 to 0.797). The three intervals overlap almost exactly, so neither the label correction nor the hyperparameter search moves the headline, and the figure quoted throughout is the original 0.74. Figure 6.1 shows the ROC curve with its bootstrap band. Its mean predicted risk matches the observed rate closely (1.40 against 1.39 percent), and it shows positive net benefit on decision-curve analysis across the one-to-five-percent risk range (Annex B).

Calibration is best judged by decile of predicted risk rather than by a single slope, and Table 6.1 gives observed against expected event rates with their confidence intervals. The expected rate lies inside the observed interval in every one of the ten deciles, so no part of the risk range is significantly miscalibrated at this sample size. There is nonetheless a systematic pattern: the observed rate sits below the expected in six of the seven lowest deciles and above it in all three of the highest, which is the compression a calibration slope above one describes, and the direction that matters clinically is the upper end. In the top decile the model expects 3.98 percent and observes 5.11, so it understates risk precisely where a clinician would act on it, by an amount this cohort cannot resolve but which external validation should be sized to detect.

Read the other way, the same table is the clearest statement of what the model is for. The top decile of predicted risk contains 35 of the 95 events, so 10 percent of patients carry 37 percent of the tamponades; the top fifth carries 55 percent; and the lower half of the distribution carries 18 percent. That concentration, rather than the AUROC, is what makes the model usable as a monitoring-triage tool.

Table 6.1: Observed against expected event rates by decile of predicted risk, out of fold. Intervals are Wilson 95 percent intervals on the observed rate, which stay valid when a decile contains only one or two events. The expected rate falls inside the observed interval in all ten.

Decile	**Patients**	**Events**	**Expected**	**Observed (95% CI)**
1	685	1	0.62%	0.15% (0.03–0.82)
2	685	5	0.73%	0.73% (0.31–1.70)
3	685	4	0.81%	0.58% (0.23–1.49)
4	684	3	0.89%	0.44% (0.15–1.28)
5	685	4	0.99%	0.58% (0.23–1.49)
6	685	5	1.10%	0.73% (0.31–1.70)
7	684	7	1.25%	1.02% (0.50–2.10)
8	685	14	1.53%	2.04% (1.22–3.40)
9	685	17	2.05%	2.48% (1.56–3.94)
10	685	35	3.98%	5.11% (3.70–7.02)

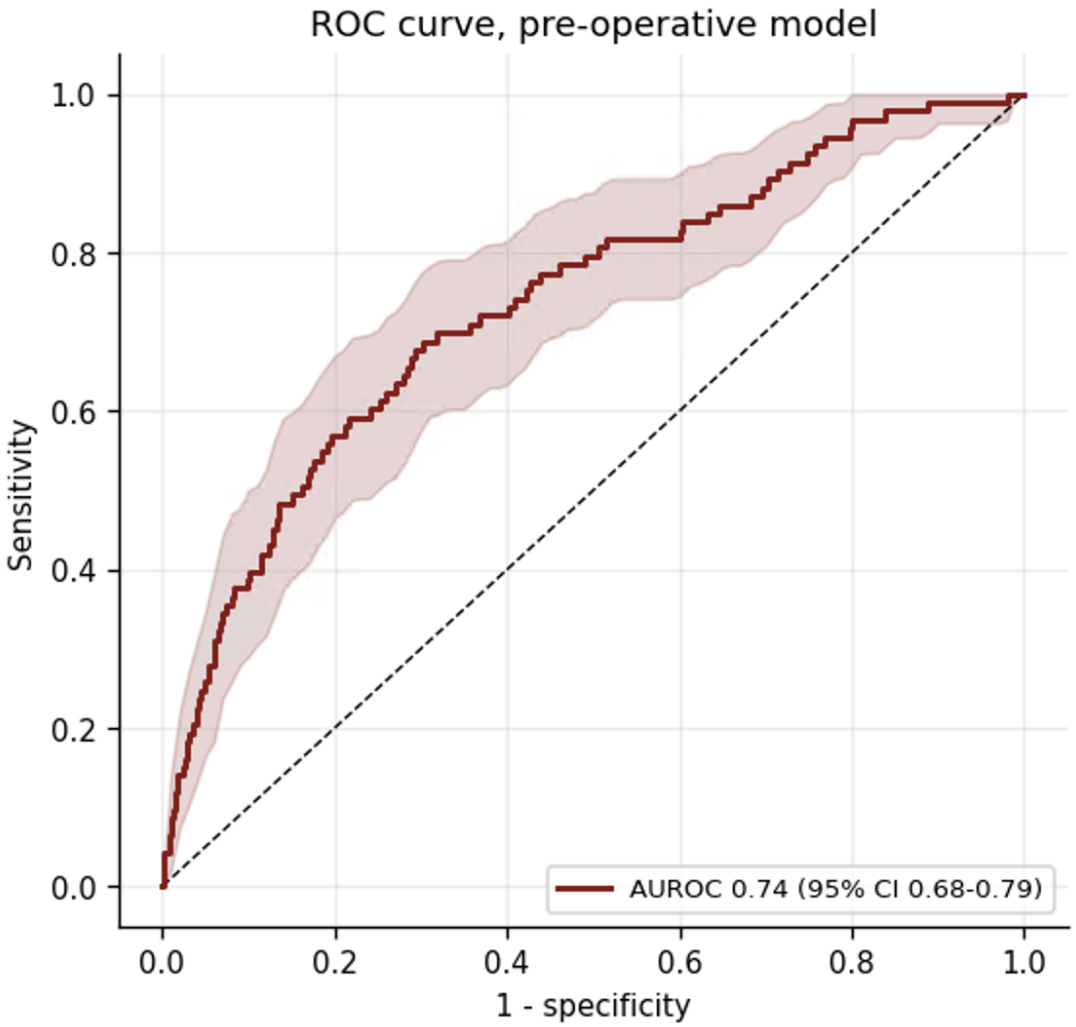

Figure 6.1: ROC curve of the pre-operative model.

A comparison against lasso, elastic net, gradient boosting, and random forest, each tuned by grid search inside the cross-validation, leaves the reference model on top at 0.746. On a paired bootstrap none of the four beats it: lasso is worse by 0.012 (95% CI −0.024 to −0.003) and gradient boosting by 0.035 (−0.074 to −0.000), an interval that only just excludes zero, while elastic net (−0.008) and random forest (−0.018) are indistinguishable from it. Tuning is worth little to the penalized regressions and most to the random forest, which gains 0.043 and rises from 0.685 to 0.728; an untuned comparison would have understated it by that margin. The logistic model is therefore the locked reference not because the alternatives were handicapped, but because they were tuned and still did not win.

Discrimination is not the only criterion, and for a rare outcome it is not the binding one, so the candidates were also compared on calibration (Table 6.2). The headline is that every model is left unweighted and every one recovers the event rate correctly: mean predicted risk lands within a hundredth of a percentage point of the observed 1.39 percent in all five, so none of them is systematically over- or under-stating risk across the cohort. The calibration slopes

differ, but by margins that a different random partition could plausibly move, so we do not rank the models on them. One difference is larger than that noise and worth stating: gradient boosting returns a scaled Brier score below zero, meaning its probabilities are less useful than quoting the base rate to every patient, while predicting up to 83 percent risk for individuals. The penalized regressions do not do that.

Table 6.2: Discrimination and calibration together, out of fold. Slope 1.0 is perfect; above 1 the predictions are too conservative, below 1 too extreme. Scaled Brier above 0 beats always predicting the base rate. Slopes are single-partition estimates and are not used to rank the models.

Model	**AUROC**	**Calib. slope**	**Mean pred.**	**Max pred.**	**Scaled Brier**
Logistic (L2), reference	0.745	1.34	1.40%	18.7%	0.016
Lasso (L1)	0.738	1.10	1.39%	24.0%	0.015
Elastic net	0.738	1.03	1.39%	29.0%	0.017
Gradient boosting	0.724	0.80	1.39%	83.0%	−0.008
Random forest	0.741	1.19	1.39%	13.0%	0.013

The reference model is therefore retained on discrimination and interpretability, and the comparison adds that no alternative is better behaved on calibration in a way this sample can demonstrate.

The same question can be asked of the variables rather than the algorithm, and it was: whether a larger, data-driven feature set would beat the eleven. Applying the pre-specified selection procedure of Section 5.3, forcing in ejection fraction, pre-operative haemoglobin and frailty on mechanistic grounds and screening every remaining candidate in the clinical record inside each fold, reaches 0.738, against 0.735 for the eleven. The long-standing set is therefore at or near the ceiling this sample supports, and the screen does not find structure the original judgment missed.

What the ladder shows more usefully is where the ceiling lies (Table 6.3). Discrimination peaks at about seven variables and then falls monotonically as more are admitted, so the binding constraint is the event count rather than the feature list. A label-permutation null on the selected configuration returns 0.505 against an observed 0.731, so the signal that is there is real; there is simply not enough of it to support a richer model.

Table 6.3: Discrimination against model size, forcing in the mechanistic variables and screening the rest in-fold. EPV is events per variable; below about ten a model of this kind begins to over-fit.

Screened variables added	**Total variables**	**EPV**	**AUROC**
0 (forced-in only)	5	18.6	0.734
2	7	13.3	**0.738**
4	9	10.3	0.731
6	11	8.5	0.722
8	13	7.2	0.717

One caution belongs with this result. The screen selects coagulopathy and thrombocytopenia ahead of everything else, in essentially every fold. Both are comorbidity-panel flags that may be coded from the whole admission rather than strictly pre-operatively, in which case a patient who bleeds into tamponade could acquire the code after the event. Neither is admitted to a reported model until that timing is confirmed, and the finding is noted here because a data-driven screen will reliably reach for exactly such variables. The strongest predictors form the bleeding-and-frailty picture of Chapter 4: higher frailty and longer, more complex surgery raise risk, while higher LVEF and higher pre-operative hemoglobin lower it. Figure 6.2 shows the standardized odds ratios for the nine continuous covariates, the interpretable analog of a feature-importance plot, where frailty, bypass duration, and surgery length push risk up while ejection fraction and hemoglobin pull it down. The age and glycated-hemoglobin estimates fall below one but are not interpreted: both are null in the univariate comparison (Table 4.1) and unstable in the adjusted model, glycated hemoglobin because of its higher missingness, so the robust signal is the frailty-and-complexity axis set against ventricular function and anaemia. The two binary surgical flags appear in the categorical analysis (Table 4.2) rather than here.

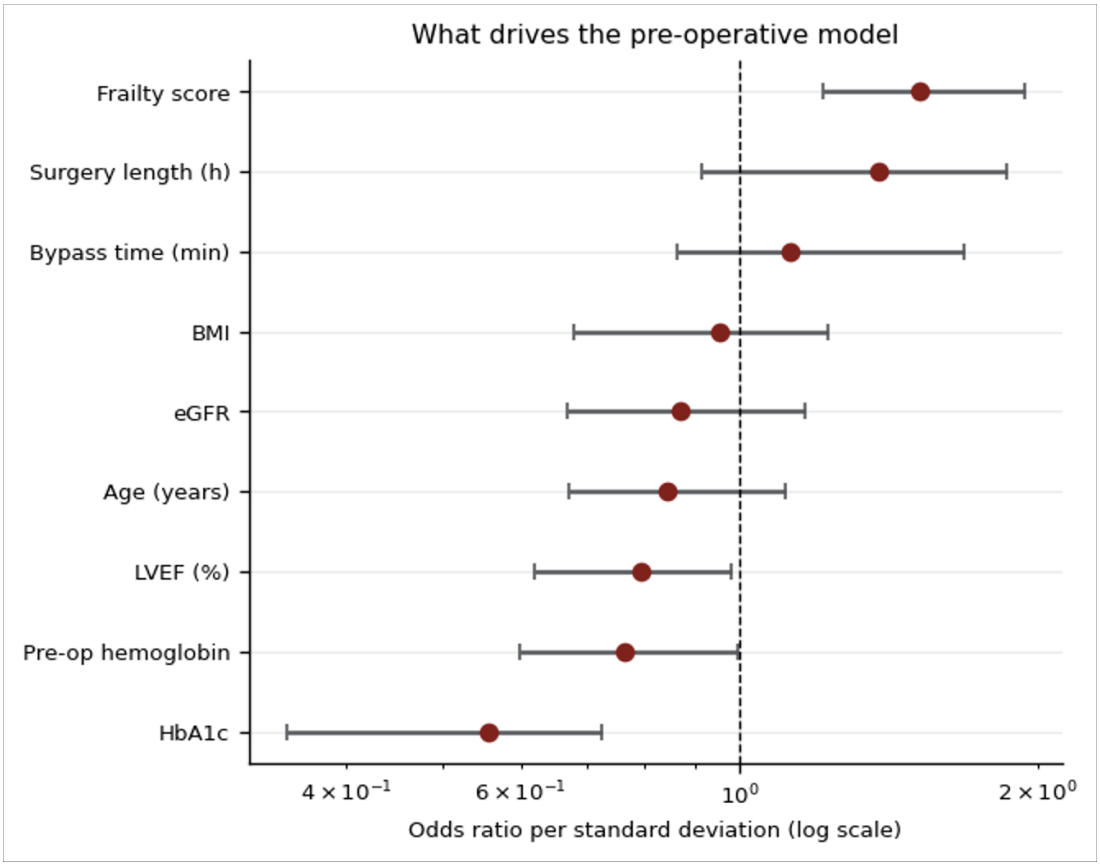

Figure 6.2: What drives the pre-operative model: standardized odds ratios per continuous covariate.

6.2 Does Later Data Add Value?

Table 6.4 sets each later data source against the pre-operative model on the same patients.

Table 6.4: Each later data source against the pre-operative model.

Model / analysis	**Severe n**	**AUROC**	**Adds over pre-op?**
Pre-operative covariates (Model 1)	95	0.74	baseline
Early-ICU structured (Model 2)	78	0.72	No (Δ+0.02, CI incl. 0)
Missingness-aware (Model 3)	95	Δ+0.02	Marginal, care-intensity
Bedside monitor, numeric trends (Model 4)	27	0.74	No
Bedside monitor, raw 125 Hz (Model 4)	27	0.63	No
Waveform plus pre-operative (stacked)	27	Δ+0.03	No (CI incl. 0)
Drainage-confirmed (sensitivity)	44 (14-day)	0.68	more specific
Pooled with effusion (sensitivity)	508	0.67	worse

6.2.1 Model 2: Early Intensive-Care Structured Data

The early intensive-care model adds the first forty-eight hours of post-operative laboratory values, chest-tube output, and vital-sign trends to the pre-operative covariates, on the 78 delayed-severe cases whose event falls at least two days after surgery so the analysis window precedes it. On its own it discriminates at an AUROC of 0.72, close to the pre-operative model. Stacked onto the pre-operative out-of-fold predictions it adds Δ+0.02, with a confidence interval that crosses zero. The early structured trajectory therefore restates the pre-operative risk rather than extending it: the patients flagged by a falling first-day hemoglobin or a rising lactate are, in the main, the patients the pre-operative profile already marks.

6.2.2 Model 3: Missingness-Aware Modeling

The missingness-aware model treats the pattern of which laboratory values were ordered as features in their own right, on the full 95-case cohort. It yields a small gain for the random forest (Δ+0.02), but the informative part is whether a test such as a lactate was ordered at all, not its value. That is a care-intensity signal: clinicians order more tests on the patients who worry them, so the indicator tracks clinical concern rather than new physiology. Treated as a confound rather than a free feature, it does not change the headline, and a model built from the missingness indicators alone scores below chance (AUROC 0.46), which confirms the indicators carry no independent predictive structure.

6.2.3 Model 4: Bedside Waveform

The waveform model runs on the 27 severe cases with a verified recording. The numeric monitor trends reach an AUROC of 0.74 and the raw 125 Hz signals 0.63, both comparable to the pre-operative model on the same patients, yet stacking the waveform features onto the pre-operative model adds only Δ+0.03 with a confidence interval through zero. Two cautions bound these numbers. First, the sample is small and the recordings cluster in the first day or two after surgery rather than the pre-event window, so the analysis is a feasibility study rather than a tested model. Second, a convolutional neural network on the raw waveform overfit at this size and is not reported, consistent with the scale argument of Chapter 2. The waveform layer establishes a working pipeline, but no waveform feature separates the classes reliably, and it adds no measurable value over the covariates at the current sample size.

6.3 Convergent Finding

Four independent later sources, intra-operative variables, early intensive-care data, informative missingness, and the bedside waveform, each add little or nothing over the pre-operative record.

The predictive signal is pre-operative. The later sources re-express the same risk; none extends it.

The strength of this conclusion is that it does not rest on a single comparison. The four sources differ in kind, in timing, and in cohort: intra-operative variables are measured in the operating room, early intensive-care data in the first forty-eight hours, informative missingness in the ordering pattern itself, and the waveform at the bedside on a separate recorded cohort. They are analyzed with different feature sets and, in the waveform case, a different control design, yet they converge on the same answer. Each increment over the pre-operative model is small and its confidence interval crosses zero: early intensive-care data Δ+0.02, the stacked waveform Δ+0.03, and informative missingness a marginal random-forest gain that, on inspection, reflects whether a test was ordered at all, not what it showed. A spurious null in one source could be chance; the same null across four sources that share only the outcome they predict is unlikely to be an artifact, and it points to a single substantive interpretation. Severe post-operative tamponade is largely written into the patient before the operation, in frailty, ventricular function, anaemia, and the length and complexity of the surgery. The later data restate that risk; they do not reveal new physiology. What the waveform analysis delivers, concretely, is therefore not an incremental model but a verified pipeline, an honest measurement of how far the current data can carry a waveform claim, and a quantified path to the cohort size a confirmatory test would need.

6.4 Sensitivity and Robustness

Discrimination is stable across every reasonable definition of the outcome, which matters more than any single figure: the choice of definition moves the sample size a great deal and the AUROC very little (Table 6.5).

Table 6.5: Discrimination under alternative outcome definitions, all on the same covariates and the same audited control pool. Only n moves. These counts are re-derived from the procedure and diagnosis tables and differ from the funnel in Chapter 3, which was built from a cached cohort file. The large difference is the any-timing severe count, 323 here against 279 there: the cached file assigned severity from diagnosis codes alone, so it missed patients whose only severe evidence is a drainage procedure. The primary count differs by two, from the drainage code list used. What the table establishes is unaffected either way, since discrimination barely moves across a fourfold change in n.

Positive definition	**Events**	**AUROC**
Drainage-confirmed incident (drainage $\geq$ 1 day)	82	0.729
Severe, new-onset $\geq$ 1 day (primary)	**93**	**0.746**
Severe, first severe event $\geq$ 1 day	119	0.732
Drainage $\geq$ 1 day or new-onset severe diagnosis	137	0.744
Severe, any timing 0–30 days (includes same-day)	323	0.751

The span is 0.729 to 0.751 while the event count varies almost fourfold, so the conclusion does not rest on the definitional choice. Two rows answer questions that arise naturally. Widening the drainage-confirmed arm from the original fourteen-day window to thirty days raises it from 44 events to 82 and gives 0.729; the fourteen-day figure of 0.68 was the smaller, noisier estimate. Including the same-day events, rather than excluding them as intra-operative, gives 323 events and 0.751, the highest of the five. We do not read that as evidence that the same-day cases belong in the primary outcome: 70 percent of them are aortic procedures, against 41 percent of the incident cases, and whether their pericardial codes record a complication of surgery or the pathology the patient presented with is a clinical question we have not resolved. Discrimination degrades only when pericardial effusion is pooled into the outcome (508 cases, 0.67), which confirms that effusion without compression is a different entity. Discrimination is stable across the all-era and post-2020 cohorts, across event-timing thresholds, and under the control-pool audit. The higher missingness in the severe group does not drive the result: a model built from the missingness indicators alone scores below chance (AUROC 0.46), and a complete-case model without imputation reaches discrimination comparable to the imputed model.

6.5 External Validation in MIMIC-IV

The single-center result invites the obvious question of whether it transports, and the recommendation this study made most insistently was to test it against an independent cohort before reading anything clinical into it. That test is now available. MIMIC-IV, the critical-care database from the Beth Israel Deaconess Medical Center, supplies a cardiac-surgery population with the diagnosis and procedure coding needed to reconstruct the same outcome, and the

pre-operative model was scored on it without refitting.

One constraint sets the terms of the comparison and must be stated first. The eleven-variable primary model cannot be scored in MIMIC-IV, because three of its strongest predictors, ejection fraction, frailty, and cardiopulmonary bypass duration, are absent from the hospital records that database exposes. What can be scored is the transportable core: a reduced model restricted to the features the two institutions share, age and twenty comorbidity indicators derived from the diagnosis codes. Refit on the Stanford cohort, that twenty-one-feature reduced model reaches an internal cross-validated AUROC of 0.634, below the 0.74 of the full model precisely because it is denied the covariates that carry most of the signal. The reduced model, not the headline model, is the object of the external test, and 0.634 is the internal benchmark it is measured against. The claim this section supports is therefore specific: the transportable core of the model validates externally. It is not a claim that the 0.74 itself has been reproduced elsewhere, which the available data cannot establish.

Scored on 11,969 MIMIC-IV cardiac-surgery admissions, with a coded tamponade rate of 1.5 percent against the Stanford 1.4, the reduced model behaves as Table 6.6 reports. The outcome is built three ways, from the loosest coding to the definition that matches the Stanford primary, and discrimination is highest under the matched definition: on 11,832 admissions with 63 incident, drainage-confirmed events, the model reaches an AUROC of 0.695 (95% CI 0.62 to 0.76), against the internal reference of 0.634 whose interval it contains. The point estimate sits above the internal benchmark, so the reduced model does not merely survive the move to a second institution; it discriminates there at least as well as it does at home.

Table 6.6: External validation in MIMIC-IV: discrimination of the twenty-one-feature reduced model under three outcome definitions, against its Stanford internal benchmark of 0.634. Only the case definition changes; the control pool is the audited clean-control set throughout. The matched definition reproduces the Stanford primary outcome. Confidence intervals are by bootstrap.

Outcome definition	**Admissions**	**Events**	**AUROC (95% CI)**
All coded tamponade (diagnosis or drainage)	11,969	183	0.674 (0.63–0.72)
Drainage-confirmed, any timing	11,878	109	0.654 (0.59–0.72)
Drainage-confirmed incident, 1–30 days (matched)	**11,832**	**63**	**0.695 (0.62–0.76)**

A second reading of the same table is independent evidence for a choice made in Chapter 3. Discrimination rises as the outcome is cleaned: from 0.654 on drainage-confirmed cases of any timing to 0.695 once the incident rule restricts them to events one to thirty days after surgery, on the same control pool. The incident definition, questioned as arbitrary when it was introduced, does measurable work at a second institution rather than merely shrinking the sample. The same-day group behaves as it does at Stanford: of 109 MIMIC-IV drainages, 40 (37 percent) fall on the day of surgery, against 56 at Stanford, so the mixed, largely intra-operative character

of same-day pericardial codes is a general feature of how this outcome is recorded and not a local artefact.

Calibration tells the other half of the story, and it is the half the study predicted in advance. The model transports on rank but not on level: it was fit at a 1.4 percent event rate and applied to a population whose incident rate is 0.5 percent, so it over-predicts absolute risk by a factor of 2.3 (observed 0.53 percent against a mean predicted 1.24). The calibration slope is 1.1, but at 63 events its confidence interval is too wide to interpret, so the level, not the slope, carries the calibration argument, and the level is corrected in the standard way by recalibration-in-the-large: a single intercept adjustment brings the mean predicted risk onto the observed rate and, being a monotone shift, leaves discrimination unchanged. On decision-curve analysis the uncorrected model shows slightly negative net benefit at the one-to-two-percent thresholds, because over-prediction flags too many patients, while the recalibrated model is positive across the acting range. This is the textbook external-validation signature, a model that ranks correctly and needs its level reset to the local base rate.

Three checks bound the result. First, the comorbidity mapping: discrimination is not an artefact of the approximate prefix map that builds the twenty indicators. Rebuilding them under a coarser three-character ICD grouping leaves the matched AUROC at 0.670, and a drastically simpler model using only age and the total comorbidity count transports at 0.691, both inside the confidence interval of the full reduced model. The external signal is coarse, carried by age and overall comorbidity burden, which is reassuring for transportability even as it tempers any claim that the specific coding matters. Second, subgroups: the matched-arm AUROC is 0.686 in women and 0.699 in men, and 0.712 below age 65 against 0.674 above it, with heavily overlapping intervals and a uniform over-prediction across all four groups, so a single recalibration serves every subgroup and no gross disparity is detected; with 20 to 43 events per group these are checks for large discrepancies, not powered subgroup claims. Third, mechanism: pre-operative haemoglobin, the one laboratory value worth adding to the reduced model at Stanford, was recoverable in MIMIC-IV for 99.8 percent of admissions. The anaemia association replicates, the incident cases being more anaemic than the controls (median 11.7 against 12.5 grams per deciliter, $p = 0.011$), but the discrimination increment does not transport at this sample size (a change of -0.009 whose interval contains both zero and the Stanford increment of 0.017), so the finding corroborates the bleeding-and-reserve mechanism across institutions without adding measurable external discrimination.

The external test carries its own limitations. MIMIC-IV diagnoses are not dated within an admission, so the matched arm is anchored on the drainage procedure rather than on a diagnosis date, and the comorbidity indicators rest on an approximate ICD prefix map. The event count, 63, is small, so the confidence intervals are wide and the calibration slope uninterpretable. And the validated object is the reduced model, not the eleven-variable primary, which the database cannot supply. Within those bounds the conclusion is clear and answers the study's chief

outstanding question: the pre-operative signal is not a Stanford artefact. Its transportable core discriminates severe post-operative tamponade at an independent institution, and what fails to transport is the absolute risk level, which recalibration restores.

6.6 Comparison with the Closest Study

Pal et al.[1] reach an AUC of 0.79 for a concurrent state, elevated central venous pressure, on 1,665 patients with 483 positive cases. The comparison needs its terms stated, because the two studies differ in almost every dimension that matters. Their cohort is 1,665 surgical patients from the UCLA perioperative dataset; their features are 843 derived from the photoplethysmogram alone, against our eleven clinical covariates; their model is a Light Gradient-Boosting Machine, against penalized logistic regression; their outcome is a physiological state measured at the same moment as the input, against a clinical event days later; and their event rate is 29 percent (483 of 1,665) against our 1.4, which alone makes the two discrimination figures hard to compare. That figure therefore sets an upper bound on what the feature-bank approach achieves when data are plentiful and the label is measured at the same moment as the signal. The task here is harder on both counts, a future event rather than a present state, and 95 events rather than 483, which is why the result here is a pre-operative model that works and a waveform increment that is not established at the current sample size.

[1] Pal et al. (2025); see Bibliography.

Chapter 7

Discussion and Recommendations

7.1 Principal Findings

A covariate model predicts severe post-operative tamponade at an AUROC of 0.74, stable across cohort definitions, and no data source acquired after the operation improves on it. Early intensive-care structured data, informative missingness, and the bedside waveform each restate the same risk; none extends it. This is the central result, and it arrives from independent comparisons rather than one, which is what makes it credible.

The precise version of the claim matters, and an earlier draft overstated it by asserting that the signal is present before the patient reaches the operating room. Four of the eleven covariates are operative rather than pre-operative, so the two must be separated before that sentence can be made. Splitting them (Section 5.3) gives 0.711 for the strictly pre-operative Model 1a against 0.737 for the peri-operative Model 1b. The difference of 0.026 is not significant on the same patients and the same folds (DeLong $p = 0.125$; paired bootstrap -0.006 to $+0.060$). So the defensible statement is narrower and, if anything, stronger for the paper's thesis: roughly nine tenths of the model's discrimination above chance is available before the patient enters the operating room, and the contribution of what happens in theatre cannot be demonstrated at this event count. It is not shown to be zero; it is not shown to exist.

7.2 Why the Pre-operative Record Dominates

The covariates that carry the model, frailty, ejection fraction, pre-operative hemoglobin, and the length and complexity of surgery, describe a patient who bleeds more and tolerates bleeding less. The later data sources add little because they measure downstream consequences of that same state. The categorical analysis makes the point sharply: pre-operative shock, coagulopathy, and thrombocytopenia are each strongly associated with the outcome on their own, yet none improves the model when added, because their signal is already represented by frailty and ejection fraction. Predicting this complication is therefore less about capturing a late physio-

logical cascade and more about recognizing, before surgery, which patients are most likely to bleed into the pericardium.

7.3 The Waveform Result in Context

The bedside waveform does not add measurable value here, but the limit is data, not method. The closest study, Pal et al.[1], reaches an AUC of 0.79 for a concurrent state, elevated central venous pressure, drawn from a perioperative dataset of 17,327 surgical patients of whom 1,665 had both photoplethysmography and central-venous-pressure waveforms available.

That comparison is worth stating carefully, because the gap in scale is narrower than the headline counts imply. Their 1,665 patients are the pool, not the evaluation: the data were split ninety to ten, and the reported AUC is measured on a held-out test set of 167 patients containing 48 elevated-CVP cases, from a single split rather than repeated cross-validation. Our waveform arm evaluates 31 positives under repeated cross-validation. On evaluation events the two studies are therefore of comparable size, 48 against 27, not the 483 against 27 a reading of the abstract would suggest. What genuinely differs is the training pool, the density of recordings, and above all the nature of the label: theirs is a physiological state measured at the same instant as the input, ours a clinical event days later. The hypotension-prediction work of Hatib et al.[2] shows the same pattern: waveforms perform when recordings are plentiful and the label is measured at the same moment as the signal. The deep-learning route is closed here for the same reason: a convolutional neural network attempted on the raw waveform overfit at this sample size, exactly as the scale of those successes would predict. Neither condition holds for post-operative tamponade at a single center, where a usable invasive recording exists for only a minority of severe cases and the recordings cluster in the first day or two after surgery, away from an event that arrives a week or two later. The waveform analysis reflects this: it is a feasibility study that establishes the pipeline and the physiological direction rather than a tested model.

7.4 Methodological Contribution

The contribution is as much in how the question was asked as in the answer. A bed-anchored attribution method resolves the reuse of study identifiers across patients, which would otherwise corrupt the waveform labels silently. Feature selection and imputation run inside the cross-validation folds, and incremental value is judged by paired bootstrap and label permutation, which removes the optimism of screening many features on few events. Markers that are part of the outcome, the drainage procedure and the vasopressors started in response, are held out

[1] Pal et al. (2025); see Bibliography.
[2] Hatib et al. (2018); see Bibliography.

of the predictors. The parsimony of the model is tested rather than assumed. These controls are why the headline is modest rather than inflated, and why an apparent waveform signal was correctly read as not yet established.

7.5 Limitations

The study is single-center and retrospective, and the outcome is rare, with 95 incident severe events, so confidence intervals are wide and an apparent increment of a few points cannot be distinguished from noise. The waveform cohort is small and its exact membership is not fully settled, which is the main reason the waveform arm is reported as feasibility only. The control pool was audited but residual misclassification cannot be excluded. Because EuroSCORE II and the Society of Thoracic Surgeons models[3][4] predict mortality rather than tamponade, no direct head-to-head benchmark exists; whether a general acuity score would capture the same pre-operative risk this model relies on is an open question that external validation should test. The outcome is defined from diagnosis and procedure codes rather than chart review, so some events may be misattributed or missed, which a clinician-adjudicated label would address. The count is also sensitive to the incident definition, moving between roughly 44 and 137 with the window and the code list, though discrimination stays near 0.74 across these choices, so the conclusion is robust even where the exact number is not. Because the data come from one center over one era, transportability was the study's chief outstanding question. It has now been tested: scored on 11,969 cardiac-surgery admissions in MIMIC-IV (Section 6.5), the transportable core of the model discriminates at an AUROC of 0.695, above its internal benchmark, although the absolute risk level requires recalibration to the local event rate. What remains untested is prospective and multi-center performance, and the binding constraint on further internal work is still the positive event count rather than the modeling. The model predicts risk, not timing, and was built on de-identified data with jittered dates, which the analysis realigns only within the secure environment. The study was conducted under an approved Stanford University institutional review board protocol.

7.6 Recommendations

Five recommendations follow from these findings, ordered by how soon they can act.

1. **Use the pre-operative model now as a risk-stratifier.** It is calibrated and needs only data available before surgery, so it can flag high-risk patients for closer post-operative monitoring without any new data capture. It should inform attention, not replace clinical judgment.

[3]Nashef et al. (2012), EuroSCORE II; see Bibliography.
[4]Shahian et al. (2018); see Bibliography.

2. **Complete external validation before any clinical deployment.** A first external test is now in hand: the transportable core validates in MIMIC-IV (Section 6.5), discriminating at 0.695 with the absolute risk level restored by recalibration. What remains before the model changes care is prospective confirmation, ideally multi-center, and validation of the full eleven-variable model at a site that records ejection fraction, frailty, and bypass duration, which MIMIC-IV does not. The reverse direction, taking a published model and validating it on this cohort, is attractive but is not currently available: Pal et al. do not publish model weights or a code repository, offering instead access to their dataset and their photoplethysmography featurization tool on request to the authors[5], and the two machine-learning models for tamponade after atrial-fibrillation ablation report variable importances rather than a coefficient set that could be applied elsewhere[6][7]. Even if weights were available, none of the three predicts this outcome: two describe intra-procedural perforation and the third a concurrent physiological state. Requesting the featurization tool is nonetheless worthwhile, because it would let our waveform features be computed on their definitions and make the two feature sets directly comparable.

3. **Expand the verified positive cohort.** This is the binding constraint. More incident-severe cases with clean, single-definition labels are the prerequisite for testing any waveform increment with adequate power; method refinement on the current sample will not settle the question.

4. **Capture recordings closer to the event.** The pre-event window where classic pulsus would appear is currently almost empty. Targeted capture in the day or two before drainage would let the physiological hypothesis be tested directly rather than inferred.

5. **Invest in the data pipeline.** A single, documented cohort definition and the bed-anchored attribution method should be standard for future work, so that the next waveform analysis starts from verified data rather than rebuilding it.

6. **Extend the operative and post-operative predictors, but not the feature count.** Two things are worth adding because they act on the bleeding mechanism directly and are not yet in the model: intra-operative transfusion, and the degree of hypothermia. Red-cell units, crystalloid volume and the minimum intra-operative haematocrit are already recorded for most patients in the surgical cohort and need only to be carried into the modelling table; estimated blood loss is recorded for too few patients to use, and no core-temperature field is currently available, so hypothermia requires a new extraction. Post-operative vasopressor escalation is a plausible early-warning signal but is recorded as a yes-or-no flag rather than a dose, and it must in any case be anchored to a window

[5] Pal et al. (2025); see Bibliography.
[6] Bansal et al. (2022); see Bibliography.
[7] Bao et al. (2026); see Bibliography.

that closes before the event, or it measures the response to tamponade rather than its antecedent.

The qualifier in the last recommendation is deliberate, because more predictors and better predictors are not the same thing. We tested whether a larger, data-driven feature set would help: the eleven covariates were replaced by a pre-specified procedure that forces in the variables with a mechanistic claim on the outcome, ejection fraction, pre-operative haemoglobin and frailty, then screens every remaining candidate in the clinical record against the outcome inside each cross-validation fold. That procedure reaches 0.738, against 0.735 for the eleven, so the long-standing set is at or near the ceiling this sample supports. Discrimination peaks at roughly seven variables and falls monotonically as more are added, to 0.731 at nine, 0.722 at eleven and 0.717 at thirteen. With 95 events the constraint is the event count, not the feature list, and investment in additional predictors should therefore be judged on whether it introduces a genuinely new mechanism rather than on whether it increases the number of variables.

7.7 Implications for Practice

A calibrated pre-operative flag fits naturally at the end of the surgical planning step and at handover to intensive care, where it can raise the threshold for echocardiography and for keeping invasive monitoring in place. It points attention toward patients whose pre-operative profile already marks them. Because it estimates who is at risk and not when the event will occur, it supports bedside judgment rather than replacing it.

7.8 Future Work

The priorities are an extended, cleanly labeled positive cohort, multi-center development building on the external validation now in hand, and, once the positive count supports it, a confirmatory test of whether dense pre-event waveform data add to the covariate baseline.

Growing the cohort is the precondition for the rest, and several routes are open. The first is chart-review confirmation: replacing the billing-code label with a clinician-adjudicated one, as a co-investigator is undertaking, would both clean the positives and recover cases the codes miss, enlarging the verified set beyond the current count. The second is multi-center pooling: the same outcome definition applied at a second institution, or against a public peri-operative waveform archive such as those distributed through PhysioNet[8], would multiply the number of recordings and supply the external test the single-center result needs. The third is prospective capture aimed at the pre-event window, so the day or two before drainage, where classic pulsus would appear, is recorded rather than absent. The fourth, relevant only once thousands of

[8]Goldberger et al. (2000); see Bibliography.

labeled recordings exist, is to pretrain a waveform model on a large public signal corpus and fine-tune it here, the transfer-learning path that makes high-capacity methods viable at small local scale.

Multi-center pooling deserves emphasis beyond its role in enlarging the cohort, because it is the only route that tests transportability rather than assuming it. This study made a specific prediction on that point, and the external validation bore it out. From a source-classifier trained on 269 independent arterial recordings, which separates the Stanford recordings from a public intensive-care archive at an AUROC of 0.72 on genuine physiological features, we argued that the pipeline transports but the populations differ enough that a model trained at one center should not be expected to arrive at another already calibrated. The MIMIC-IV validation (Section 6.5) confirmed exactly this: discrimination transported, at 0.695, while the absolute risk level did not, over-predicting 2.3-fold until recalibration reset it. The populations differ in prevalence, not in the direction of risk, which is an argument for pooled multi-center development that arrives already recalibrated, rather than for external validation alone.

The methodological extension worth naming separately is dynamic forecasting: a model that updates risk through the post-operative course as new data arrive, rather than scoring once from a fixed window. Clinically it is the right target, since it matches how deterioration is actually watched for. Whether it is feasible here is a question about coverage rather than about method, and the current answer is no. Of the incident severe cases, only about a third have any usable bedside recording at all, and the recordings that exist cluster in the first day or two after surgery, a median of ten days before the event, so there is no continuous signal through the window in which risk would need to update. The structured record is more promising, because flowsheet vitals and laboratory values are recorded for nearly every patient throughout admission and could support an hourly-updating model without any new data capture; that is the version of dynamic forecasting this dataset could sustain today, and the waveform version becomes feasible only after the targeted pre-event capture described above.

Chapter 8

Conclusion

This work asked whether machine learning can predict severe pericardial tamponade after cardiac surgery, and whether progressively richer data, intra-operative variables, early intensive-care measurements, informative missingness, and the bedside waveform, improve that prediction. The answer is that prediction is feasible and that its power is pre-operative. A penalized logistic model built from eleven variables known before surgery discriminates incident severe tamponade at an AUROC of 0.74, with good calibration and positive net benefit, and it survives a battery of leakage and optimism controls that were the methodological heart of the study. Against this baseline, none of the three richer data sources adds established value: each re-expresses the same bleeding-and-frailty risk rather than extending it.

The waveform component, the original motivation for the work, is reported as a feasibility result. The pipeline is built, but no waveform feature separates the classes reliably, and the recordings are too few and too far from the event to support a quantitative claim at one center. This limit comes from the data, not the method, and it sets a clear agenda: enlarge the verified positive cohort and capture monitoring closer to the event before the waveform question can be answered with power. The pre-operative model, by contrast, has already cleared its chief outstanding test: its transportable core validates on an independent institution's cardiac-surgery population (Section 6.5), discriminating at an AUROC of 0.695 with the absolute risk level restored by recalibration. The deliverable today is a calibrated, externally validated pre-operative risk-stratifier and a rigorous, reproducible framework for the harder prospective question that remains open.

Bibliography

Peer-reviewed articles

Bansal, A., et al. (2022). *Machine Learning Prediction of Pericardial Tamponade After Atrial Fibrillation Ablation.* The American Journal of Cardiology, 175, 179–180.

Bao, Z., et al. (2026). *Explainable Machine Learning for Risk Prediction of Acute Cardiac Tamponade during Atrial Fibrillation Ablation.* Scientific Reports, 16(1), 9476.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). *Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach.* Biometrics, 44(3), 837–845.

Enevoldsen, J., & Vistisen, S. T. (2022). *Performance of the Hypotension Prediction Index May Be Overestimated Due to Selection Bias.* Anesthesiology, 137(3), 283–289.

Goldberger, A. L., Amaral, L. A. N., Glass, L., et al. (2000). *PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals.* Circulation, 101(23), e215–e220.

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., et al. (2019). *Cardiologist-level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network.* Nature Medicine, 25(1), 65–69.

Hatib, F., Jian, Z., Buddi, S., Lee, C., Settels, J., Sibert, K., Rinehart, J., & Cannesson, M. (2018). *Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis.* Anesthesiology, 129(4), 663–674.

Khan, N. K., Järvelä, K. M., Loisa, E. L., Sutinen, J. A., Laurikka, J. O., & Khan, J. A. (2017). *Incidence, Presentation and Risk Factors of Late Postoperative Pericardial Effusions Requiring Invasive Treatment after Cardiac Surgery.* Interactive CardioVascular and Thoracic Surgery, 24(6), 835–840.

Kilic, A., et al. (2020). *Predictive Utility of a Machine Learning Algorithm in Estimating Mortality Risk in Cardiac Surgery.* The Annals of Thoracic Surgery, 109(6), 1811–1819.

Nashef, S. A., Roques, F., Sharples, L. D., et al. (2012). *EuroSCORE II.* European Journal of

Cardio-Thoracic Surgery, 41(4), 734–744.

Shahian, D. M., Jacobs, J. P., Badhwar, V., et al. (2018). *The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 1, Background, Design Considerations, and Model Development.* The Annals of Thoracic Surgery, 105(5), 1411–1418.

Spodick, D. H. (2003). *Acute Cardiac Tamponade.* New England Journal of Medicine, 349(7), 684–690.

Steyerberg, E. W., & Vergouwe, Y. (2014). *Towards Better Clinical Prediction Models: Seven Steps for Development and an ABCD for Validation.* European Heart Journal, 35(29), 1925–1931.

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). *Calibration: the Achilles Heel of Predictive Analytics.* BMC Medicine, 17, 230.

Vickers, A. J., & Elkin, E. B. (2006). *Decision Curve Analysis: A Novel Method for Evaluating Prediction Models.* Medical Decision Making, 26(6), 565–574.

You, S. C., Shim, C. Y., Hong, G. R., Kim, D., Cho, I. J., Lee, S., Chang, H. J., et al. (2016). *Incidence, Predictors, and Clinical Outcomes of Postoperative Cardiac Tamponade in Patients Undergoing Heart Valve Surgery.* PLoS ONE, 11(11), e0165754.

Preprints

Pal, R., Rudas, A., Chiang, J. N., Barney, A., & Cannesson, M. (2025). *Machine Learning-Based Identification of Patients with Elevated Central Venous Pressure Using Features Extracted from Photoplethysmography Waveforms.* medRxiv 2025.12.30.25343231.

Reporting guidelines

Collins, G. S., Moons, K. G. M., Dhiman, P., et al. (2024). *TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods.* BMJ, 385, e078378.

Table of Annexes

- Annex A. Glossary of clinical and technical terms
- Annex B. Extended results and figures
- Annex C. Feature dictionary
- Annex D. Reproducibility and code

Appendix A

Glossary of Clinical and Technical Terms

Pericardial tamponade
Compression of the heart by fluid or blood trapped between the heart and the pericardial sac, which stops the heart from filling and lowers cardiac output.

Pericardium
The thin, stiff sac around the heart.

Pericardiocentesis
Drainage of the pericardial space with a needle.

Pulsus paradoxus
An exaggerated fall in arterial pressure during inspiration, a classic sign of tamponade.

ABP, CVP, PAP
Arterial, central-venous, and pulmonary-artery blood pressure, each measured through an invasive line.

PPG (pleth)
Photoplethysmogram, the pulse-oximeter waveform, a non-invasive optical pulse signal.

Cardiopulmonary bypass
A machine that takes over the circulation while the heart is stopped during surgery.

LVEF
Left-ventricular ejection fraction, the share of blood the left ventricle pumps each beat.

eGFR
Estimated glomerular filtration rate, a measure of kidney function.

Frailty score
A composite measure of a patient's physiological reserve.

SACP
Selective antegrade cerebral perfusion, a brain-protection technique used in complex aortic surgery.

OR, ICU
Operating room and intensive-care unit.

ADT log
The admission, discharge and transfer record, which places each patient in a unit and bed over time.

AUROC
Area under the receiver operating characteristic curve, a measure of how well a model ranks cases above non-cases. 0.5 is chance and 1.0 is perfect.

Calibration
How closely predicted probabilities match observed event rates.

Permutation test
A test that shuffles the labels many times to measure how often a result of the observed size appears by chance.

Informative missingness (MIA)
Using whether a value was measured, not only its value, as a feature. Here it is treated as a care-intensity confound.

Pulse-pressure variation, PVI
Beat-to-beat respiratory swings in arterial pressure, or in the perfusion index, the quantitative form of pulsus paradoxus.

FFT
Fast Fourier transform, which decomposes a signal into its frequency components.

Monitoring-intensity confound
The tendency for sicker patients to receive more monitoring, so that the presence of a signal correlates with the outcome.

Appendix B

Extended Results and Figures

This annex collects the full result table, the calibration and permutation diagnostics, and the model comparison that the main text summarizes.

B.1 Full Result Table

Table B.1 restates each analysis with its discrimination and, where applicable, the paired increment over the pre-operative model with its confidence interval.

Table B.1: Extended results. AUROC with 95% confidence interval; increment is the paired stacked-minus-base difference.

Analysis	**Severe n**	**AUROC (95% CI)**	**Increment (95% CI)**	**Verdict**
Pre-operative (Model 1)	95	0.74 (0.68–0.79)	baseline	reference
audited label, untuned	93	0.735	n/a	confirms
nested CV, tuned	93	0.746 (0.69–0.80)	n/a	confirms
Early-ICU structured (Model 2)	78	0.72	+0.02 (incl. 0)	no add
Missingness-aware (Model 3)	95	n/a	+0.02 (RF)	care-intensity
Waveform numeric (Model 4)	27	0.74	n/a	feasibility
Waveform raw 125 Hz (Model 4)	27	0.63	n/a	feasibility
Waveform + pre-op (stacked)	27	n/a	+0.03 (incl. 0)	no add
Drainage-confirmed (sens.)	44	0.68	n/a	more specific
Pooled with effusion (sens.)	508	0.67	n/a	worse
Missingness indicators only	95	0.46	n/a	below chance
External validation, MIMIC-IV	63	0.695 (0.62–0.76)	n/a	transports

B.2 Calibration

The pre-operative model is well calibrated, with a calibration slope of 0.90 and a mean predicted risk of 1.4 percent against an observed 1.4 percent. Figure B.1 shows the calibration curve, and Figure B.2 the decision-curve analysis, which confirms positive net benefit over treat-all and treat-none across the one-to-five-percent risk range.

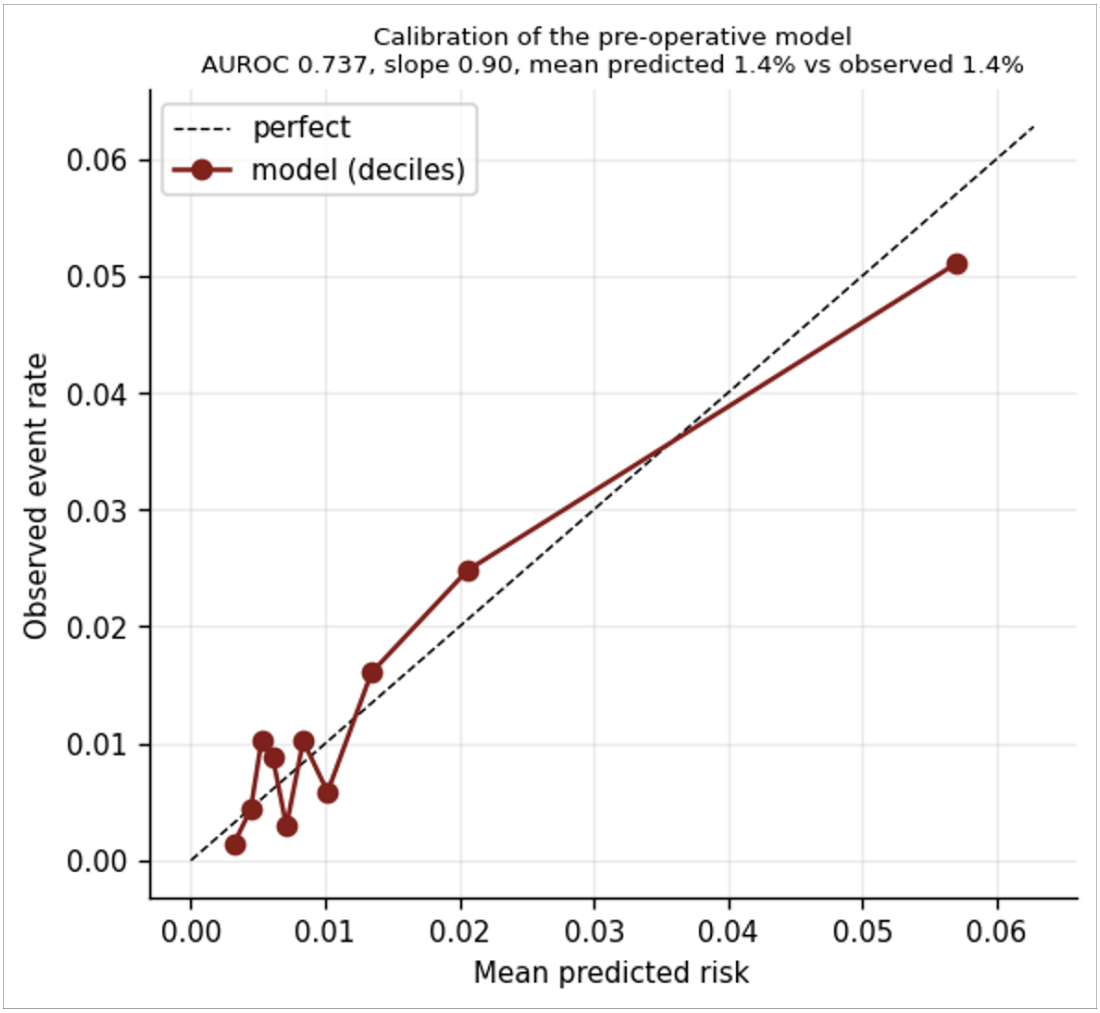

Figure B.1: Calibration of the pre-operative model.

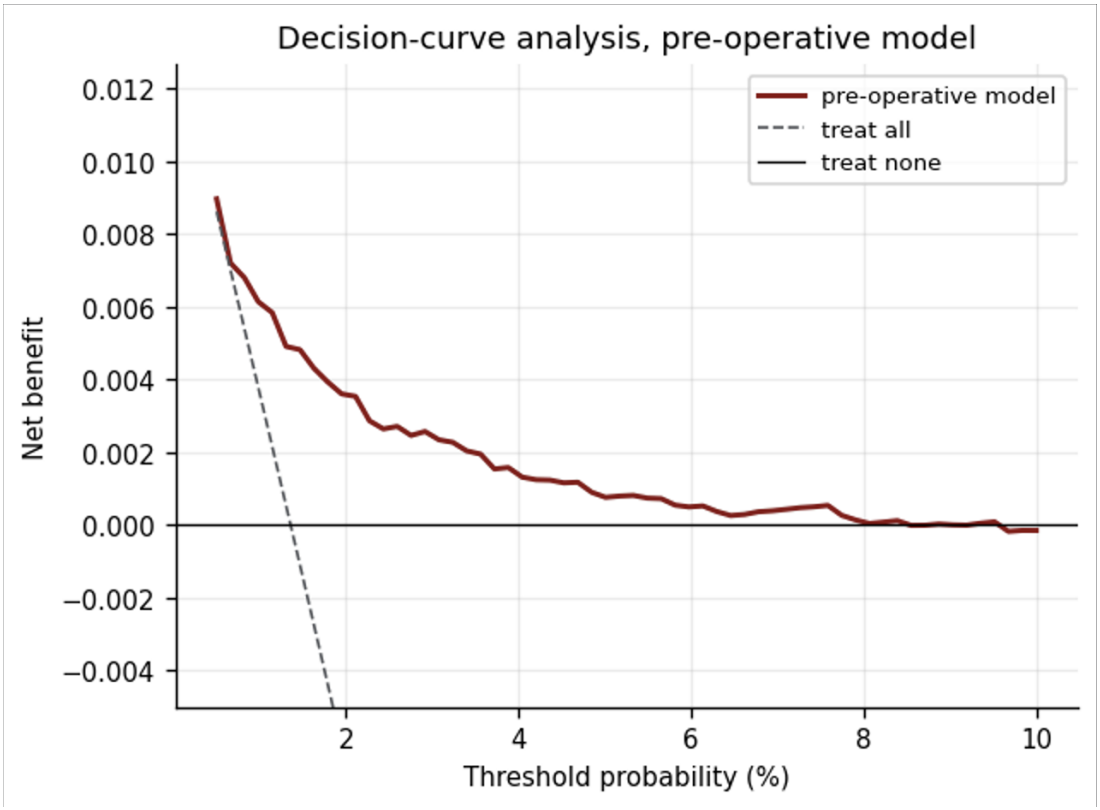

Figure B.2: Decision-curve analysis of the pre-operative model.

B.3 Permutation Test

To confirm the headline is not an artifact of screening many features on few events, the class labels were permuted and the full pipeline re-run many times. Figure B.3 shows the permuted AUROC distribution centered on 0.5, with the observed 0.74 in the far right tail, so the result is not explained by chance fitting.

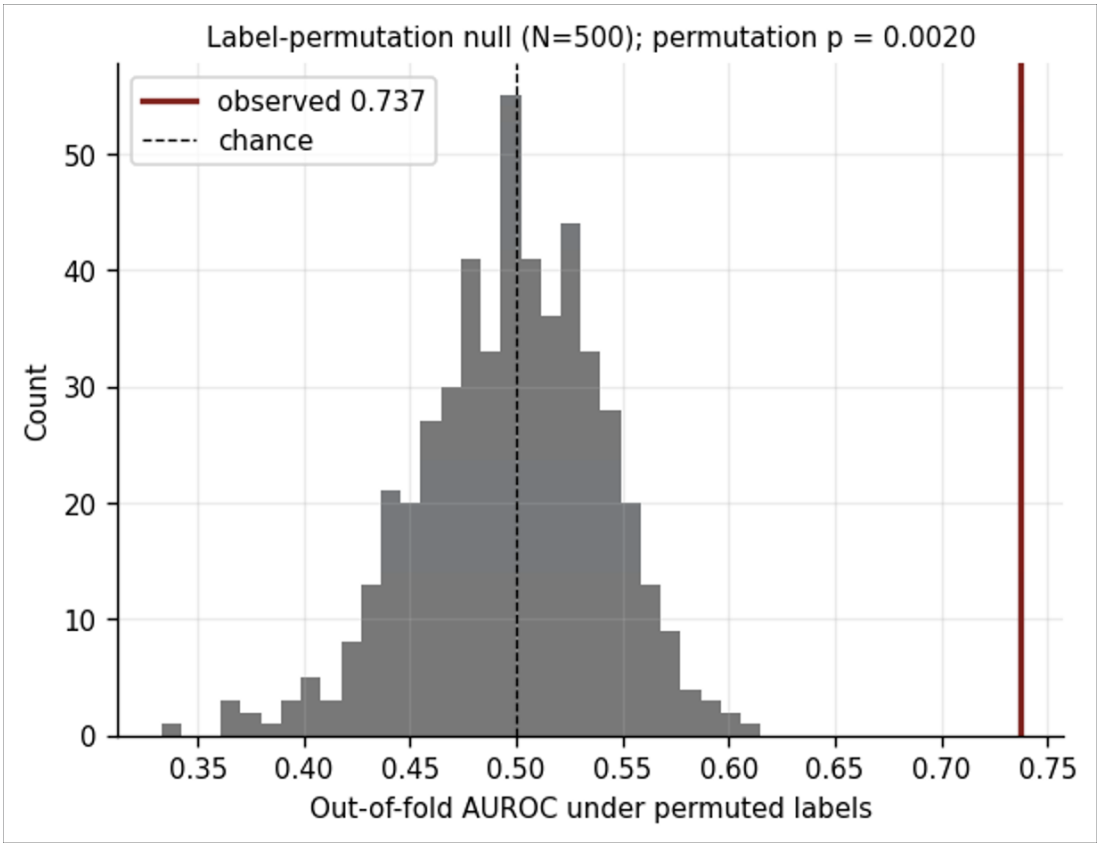

Figure B.3: Label-permutation null for the pre-operative model.

B.4 Model Comparison

Against the penalized logistic reference, lasso, elastic net, gradient boosting, and random forest were each evaluated under repeated cross-validation on the pre-operative cohort. All point estimates fell within noise of the logistic model, with overlapping confidence intervals, and none beat it on a paired bootstrap, so the interpretable logistic model was locked as the reference (Figure B.4).

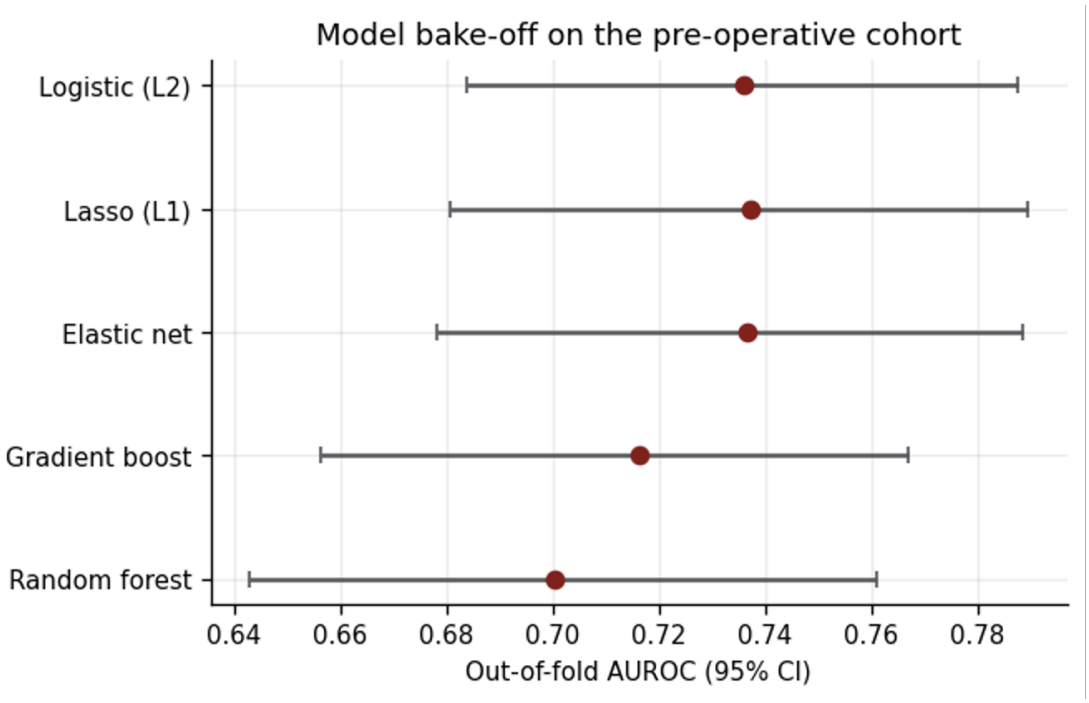

Figure B.4: Model comparison: discrimination of five model families on the pre-operative cohort.

Appendix C

Feature Dictionary

This annex defines every feature family the models use, so the analysis can be reproduced. The numeric parameters set inside the extraction pipeline, the frequency-band edges, the analysis-window length, and the pulse-pressure-variation formula, are noted at the end of this annex.

C.1 Pre-operative covariates (Model 1)

Eleven variables known before the operation: age, body-mass index, left-ventricular ejection fraction, estimated glomerular filtration rate, pre-operative hemoglobin, glycated hemoglobin (HbA1c), frailty score, cardiopulmonary-bypass duration, surgery length, and the circulatory-arrest and selective-antegrade-cerebral-perfusion flags. Race is recorded but excluded by design.

C.2 Early intensive-care structured (Model 2)

The first forty-eight hours of post-operative data, with the hour as the unit of analysis: laboratory values identified by LOINC code, hemoglobin (718-7), hematocrit (4544-3), lactate (32693-4), creatinine (2160-0), and platelet count (777-3); chest-tube output and its hour-on-hour trend; and vital-sign trends.

C.3 Bedside-monitor numeric trends (Model 4)

The low-frequency monitor channels are heart rate (HR), beat-to-beat heart rate (btbHR), respiratory rate (RR), oxygen saturation (SpO_2), perfusion index (Perf), the pleth variability index (PVI), and non-invasive systolic, diastolic and mean blood pressure (NBPs, NBPd, NBPm). Each channel contributes the six summary statistics in Table C.1. Two heart-rate-variability features are added: the root of the NN-interval variance, an RMSSD-type measure of beat-to-beat variability, and pNN50, the proportion of successive beats differing by more than

50 ms. A perfusion-derived pulsus surrogate (`perf_pvi_surrogate`) captures the respiratory swing in the plethysmographic perfusion index, and a coverage term (`cover_h`) records the hours of recording available.

Table C.1: Per-channel summary statistics applied to each numeric-trend channel.

Suffix	Definition
`_mean`	average over the analysis window
`_std`	standard deviation (dispersion) over the window
`_min`	minimum value in the window
`_max`	maximum value in the window
`_last`	final value in the window
`_slope`	linear trend across the window (change per hour)

C.4 Bedside-monitor raw 125 Hz (Model 4)

The raw pressure and plethysmographic channels, arterial (ABP), central-venous (CVP), pulmonary-artery (PAP, where present) and the plethysmogram (Pleth), together with respiration (Resp), yield three kinds of feature. *Time-domain summaries:* the median (`_med`), the per-beat median (`_beat_med`), the standard deviation (`_std`), and the percent variation (`_var_pct`). *Frequency-domain band powers* from the fast Fourier transform: power in the cardiac band (`_cardband`, the frequencies around the heart-rate fundamental) and in the respiratory band (`_respband`, the frequencies around the breathing rate), and their ratio (`_resp_ratio`), which quantifies respiratory modulation of the waveform and is the quantitative form of pulsus paradoxus. The percent variation of the arterial pulse and systolic pressure across the respiratory cycle gives pulse-pressure and systolic-pressure variation. *Channel-presence indicators* (`has_abp`, `has_cvp`, `has_pleth`) record whether each invasive channel exists and are modeled explicitly, so the monitoring-intensity confound cannot masquerade as physiology.

C.5 Pipeline parameters

The following values are fixed in the feature-extraction tool and complete the definitions above: the cardiac frequency band, the respiratory frequency band, the analysis-window length, and the exact pulse-pressure and systolic-pressure variation formula. They are recorded with the code (Annex D) and should be cited there for an exact replication.

Appendix D

Reproducibility and Code

Cohort. The cohort is built from a restricted-access hospital clinical-data warehouse through Google BigQuery, joining the surgical cohort to the procedure and diagnosis tables under the outcome definition of Chapter 3. The funnel and counts are in Table 3.1.

Model and validation. The reference model is an L2-penalized logistic regression with fitted patient-level with one row per patient. Probabilities are left unweighted so they stay calibrated to the true event rate. Discrimination is estimated by repeated stratified cross-validation (five folds, five repeats), with median imputation, feature standardization, and any feature selection performed inside each training fold to avoid leakage. Confidence intervals are obtained by bootstrap, AUROC differences by the DeLong test[1] and a paired bootstrap, and optimism by label permutation. Analyses use Python with scikit-learn, scipy, and pandas; random seeds are fixed for reproducibility.

Attribution and audits. Waveform recordings are matched to patients by the bed-anchored method of Chapter 3; recording dates are realigned by recovering the per-patient jitter offset and verified against the surgery-date check. Routine leakage audits confirm symmetric processing of cases and controls, and the control pool is audited for mislabeled pericardial codes. Reporting follows the TRIPOD+AI guidance[2].

[1] DeLong et al. (1988); see Bibliography.
[2] Collins et al. (2024); see Bibliography.

List of Figures

1.1 The perioperative care pathway and the timing of a tamponade event. 6
1.2 Illustrative bedside waveforms (schematic, not patient data): arterial pressure, central-venous pressure, and the plethysmogram, showing the respiratory swing that tamponade exaggerates as pulsus paradoxus. 7

3.1 Cohort construction funnel. 21

4.1 Pre-operative covariate distributions by class. 23
4.2 Correlation among the pre-operative covariates. 25
4.3 Standardized class difference for the pre-specified tamponade-physiology waveform features; all intervals span zero at the feasibility sample size. 26
4.4 A non-confounder and a confounder: age versus surgery length by class. 27

6.1 ROC curve of the pre-operative model. 34
6.2 What drives the pre-operative model: standardized odds ratios per continuous covariate. 37

B.1 Calibration of the pre-operative model. 57
B.2 Decision-curve analysis of the pre-operative model. 58
B.3 Label-permutation null for the pre-operative model. 59
B.4 Model comparison: discrimination of five model families on the pre-operative cohort. 60

List of Tables

3.1 Cohort attrition, with the count after each filter. 18
3.2 The modeling cohorts. Each sample size follows the data the analysis requires. . 20

4.1 Cohort characteristics by class, median [IQR]; p from the Mann-Whitney test. The final column states where each variable enters the model, so the table can be read as a specification and not only as a description: *1a* = pre-operative model, *1b* = added in the peri-operative model. 22
4.2 Categorical characteristics by class, percent positive; p from the chi-squared test. Tables 4.1 and 4.2 are split by the test each variable requires, rank-based for continuous and chi-squared for binary, not by whether the variable is in the model; the *In model* column carries that information in both. *Tested* means the variable was added to the model and did not improve discrimination (Section 4.2). 24

6.1 Observed against expected event rates by decile of predicted risk, out of fold. Intervals are Wilson 95 percent intervals on the observed rate, which stay valid when a decile contains only one or two events. The expected rate falls inside the observed interval in all ten. 33
6.2 Discrimination and calibration together, out of fold. Slope 1.0 is perfect; above 1 the predictions are too conservative, below 1 too extreme. Scaled Brier above 0 beats always predicting the base rate. Slopes are single-partition estimates and are not used to rank the models. 35
6.3 Discrimination against model size, forcing in the mechanistic variables and screening the rest in-fold. EPV is events per variable; below about ten a model of this kind begins to over-fit. 36
6.4 Each later data source against the pre-operative model. 37

6.5 Discrimination under alternative outcome definitions, all on the same covariates and the same audited control pool. Only n moves. These counts are re-derived from the procedure and diagnosis tables and differ from the funnel in Chapter 3, which was built from a cached cohort file. The large difference is the any-timing severe count, 323 here against 279 there: the cached file assigned severity from diagnosis codes alone, so it missed patients whose only severe evidence is a drainage procedure. The primary count differs by two, from the drainage code list used. What the table establishes is unaffected either way, since discrimination barely moves across a fourfold change in n. 40

6.6 External validation in MIMIC-IV: discrimination of the twenty-one-feature reduced model under three outcome definitions, against its Stanford internal benchmark of 0.634. Only the case definition changes; the control pool is the audited clean-control set throughout. The matched definition reproduces the Stanford primary outcome. Confidence intervals are by bootstrap. 41

B.1 Extended results. AUROC with 95% confidence interval; increment is the paired stacked-minus-base difference. 56

C.1 Per-channel summary statistics applied to each numeric-trend channel. 62