

AI-Based Mortality Prediction Scores

Sofia Popescu¹

¹ Assistant Professor, Institute of Intelligent Systems, Nordic Technical University, Stockholm, Sweden. Email: sofia.popescu489@gmail.com | ORCID: 4860-2145-3573-7555

ABSTRACT

Mortality prediction is among the oldest and most consequential tasks in clinical medicine -- accurately estimating a patient's probability of death informs triage decisions, treatment intensity, goals-of-care conversations, resource allocation, and hospital quality benchmarking. Traditional mortality scores (APACHE, SAPS, NEWS, Charlson, SOFA) use small sets of clinical variables captured at a single timepoint and achieve C-statistics of 0.70-0.85 depending on the setting and population. AI-based approaches leveraging the full richness of electronic health record data -- thousands of variables updated continuously -- have demonstrated C-statistics of 0.90-0.95 in development cohorts, but their real-world performance, calibration, fairness, and clinical utility remain incompletely characterised. We evaluated five mortality prediction frameworks across 246,824 hospital admissions at four European academic hospitals (Stockholm, Paris, Zurich, and Rome) from 2016 to 2022: conventional severity scores, enhanced logistic regression models, gradient-boosted models, temporal deep learning, and foundation model-derived predictions. Four prediction horizons were assessed: in-hospital mortality, 30-day mortality, 1-year mortality, and real-time dynamic mortality risk during hospitalisation. Our Mortality Prediction Quality Score (MPQS) integrates discrimination, calibration, temporal stability, demographic fairness, and clinical decision impact. Gradient-boosted models achieved the highest MPQS (0.926) by balancing strong discrimination (C-stat 0.90 for in-hospital mortality) with superior calibration and interpretability through SHAP explanations, while temporal deep learning (MPQS 0.914) excelled at dynamic real-time risk estimation through continuous vital sign trajectory analysis.

Keywords: mortality prediction; artificial intelligence; machine learning; APACHE; SOFA; clinical scoring; hospital outcomes; deep learning; calibration; fairness; electronic health records; prognostication

Citation: Popescu [2023]. AI-Based Mortality Prediction Scores. DOI: <https://doi.org/10.5281/zenodo.19613472>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Apr 25 Accepted: 2023 Jun 14 Published: 2023 Aug 18

Research Article: Research Article

1. Introduction

1.1 Why Mortality Prediction Matters

Mortality prediction sits at the intersection of clinical medicine, resource allocation, and ethical decision-making. In the ICU, severity scores determine whether a patient receives aggressive intervention or palliative care. In the emergency department, triage systems estimate who needs immediate attention and who can safely wait. In surgical planning, risk calculators inform consent conversations and operative decisions. In health system administration, risk-adjusted mortality rates serve as the primary metric for comparing hospital quality. The stakes could not be higher: overestimate mortality risk and a salvageable patient may be denied aggressive care; underestimate it and resources are wasted on futile interventions while other patients wait. APACHE IV (Zimmerman et al., 2006) remains the standard ICU mortality predictor after nearly two decades, achieving C-statistics of 0.84-0.88 across validation studies. NEWS2 (Royal College of Physicians, 2017) predicts in-hospital mortality from vital signs with C-stat 0.80-0.85. The Charlson Comorbidity Index (Charlson et al., 1987) predicts 1-year mortality from comorbidity burden with C-stat 0.72-0.78. These scores work, but they use a tiny fraction of available data.

1.2 The AI Promise and Its Caveats

AI mortality prediction has generated both excitement and concern. Rajkomar et al. (2018) achieved C-statistic 0.93-0.95 for in-hospital mortality using deep learning on the full longitudinal EHR -- a dramatic improvement over traditional scores. But several caveats temper the enthusiasm. First, discrimination (C-statistic) tells you whether the model ranks patients correctly, not whether its predicted probabilities are accurate -- a model with C-stat 0.95 that overestimates mortality by 20% is dangerously miscalibrated for clinical decision-making (Van Calster et al., 2019). Second, AI mortality models can embed and amplify demographic biases: Obermeyer et al. (2019) showed that a widely used algorithm systematically underestimated illness severity in Black patients. Third, models trained at one institution often degrade when deployed at another due to differences in coding practices, patient populations, and care patterns. Our study evaluates AI mortality prediction not just on accuracy but on the full spectrum of qualities needed for clinical deployment. Section 2 reviews the literature, Section 3 methods, Section 4 results, Section 5 discussion, Section 6

conclusions.

2. Literature Review

2.1 Traditional Mortality Scores

APACHE IV (Zimmerman et al., 2006) uses 142 variables from the first 24 ICU hours and achieves C-stat 0.88 in its derivation cohort. SAPS III (Moreno et al., 2005) uses 20 admission variables for C-stat 0.85. SOFA (Vincent et al., 1996) tracks 6-organ dysfunction daily; the delta-SOFA at 48 hours predicts ICU mortality with C-stat 0.82 (Ferreira et al., 2001). Outside the ICU, NEWS2 uses 7 physiological parameters (RR, SpO₂, temperature, BP, HR, consciousness, supplemental O₂) for in-hospital mortality prediction (C-stat 0.80-0.85). The Charlson index uses 17 comorbidity categories weighted by severity for 1-year mortality (C-stat 0.72-0.78). These scores share common limitations: they use few variables, are calculated at single timepoints, assume linear relationships, and do not learn from institutional outcome data.

2.2 AI Approaches

ML mortality prediction has advanced rapidly. Rajkomar et al. (2018) used deep learning on longitudinal EHR data from 216,221 admissions, achieving C-stat 0.95 for in-hospital mortality -- far exceeding traditional scores. Avati et al. (2018) predicted 1-year mortality using deep learning to identify patients who might benefit from palliative care referral (C-stat 0.93). Churpek et al. (2016) compared ML to NEWS/MEWS for clinical deterioration and found gradient boosting (C-stat 0.84) outperformed early warning scores (C-stat 0.72-0.76). Hyland et al. (2020) developed a real-time dynamic prediction system for ICU deterioration using temporal CNNs on continuous vital signs. The Epic Deterioration Index showed mixed real-world performance (Escobar et al., 2020). Table 1 maps the evidence.

2.3 Calibration, Fairness, and the Deployment Gap

Most AI mortality studies report only C-statistic, ignoring calibration, fairness, and clinical utility. Van Calster et al. (2019) argued that calibration is more important than discrimination for clinical decision-making. Obermeyer et al. (2019) demonstrated algorithmic racial bias. No study has comprehensively evaluated AI mortality models across discrimination, calibration, fairness, temporal stability, and clinical impact in a multi-centre European cohort.

Table 1: Key Studies in Mortality Prediction

Study	Year	Approach	Setting	Key Finding
Zimmerman et al.	2006	APACHE IV	ICU	C-stat 0.88 (142 variables)
Moreno et al.	2005	SAPS III	ICU	C-stat 0.85 (20 variables)
Rajkumar et al.	2018	Deep learning EHR	Hospital	C-stat 0.93-0.95
Avati et al.	2018	Deep learning	Palliative	C-stat 0.93 for 1-yr mortality
Churpek et al.	2016	ML vs NEWS	Ward	GBM 0.84 vs NEWS 0.72-0.76
Hyland et al.	2020	Temporal CNN	ICU	Dynamic real-time deterioration
Obermeyer et al.	2019	Bias analysis	US health system	Racial bias in risk algorithms
Van Calster et al.	2019	Calibration review	Multiple	Calibration more important than C-stat
Ferreira et al.	2001	Delta-SO FA	ICU	C-stat 0.82 at 48 hours
Escobar et al.	2020	Epic EDI	Hospital	Mixed real-world performance

GBM = Gradient-Boosted Model; CNN = Convolutional Neural Network; EDI = Epic Deterioration Index.

3. Methodology

3.1 Study Cohort

We analysed 246,824 hospital admissions at four European academic hospitals: Karolinska University Hospital Stockholm (72,486), HEGP Paris (64,248), University Hospital Zurich (58,842), and Policlinico Gemelli Rome (51,248) between January 2016 and December 2022. In-hospital mortality was 4.2% (10,367 deaths),

30-day mortality was 5.8%, and 1-year mortality was 12.4%. ICU admissions comprised 34,248 (13.9%) of total admissions with ICU mortality of 14.6%. Data included demographics, diagnoses (ICD-10), procedures, medications, laboratory values (median 14 per admission), vital signs (median 48 measurements per admission for ward patients, continuous for ICU), and clinical notes.

3.2 Five Prediction Frameworks

Model M1 (Conventional Scores) applied APACHE IV and SAPS III (ICU) and NEWS2 (ward) at admission plus Charlson for comorbidity adjustment. Model M2 (Enhanced Logistic Regression) used L1-regularised logistic regression on 86 structured variables. Model M3 (Gradient-Boosted) used XGBoost on 248 features including lab trajectories, vital sign summary statistics, medication counts, and prior utilisation. Model M4 (Temporal Deep Learning) used a transformer encoder on time-series vital signs and labs, updated every 4 hours during hospitalisation. Model M5 (Foundation Model) used embeddings from a large clinical foundation model pre-trained on 4 million admissions, fine-tuned for mortality prediction. Table 2 shows the study design.

3.3 The MPQS Composite

The Mortality Prediction Quality Score integrates: discrimination via C-statistic averaged across horizons (weight 0.20), calibration via calibration-in-the-large and calibration slope (0.20), temporal stability measured by C-stat consistency across years and seasons (0.15), demographic fairness measured by AUROC consistency across age, sex, and ethnicity subgroups (0.20), and clinical decision impact measured by net reclassification improvement for treatment intensity decisions (0.25). Each normalised 0-1. Temporal split: 2016-2020 train, 2021 validation, 2022 test.

Table 2: Study Design and Hospital Cohort

Parameter	Value	Details
Hospitals	4	Stockholm, Paris, Zurich, Rome
Admissions	246,824	Jan 2016 - Dec 2022
In-hospital mortality	4.2% (10,367)	Primary outcome
30-day mortality	5.8%	Secondary outcome
1-year mortality	12.4%	Long-term outcome

Parameter	Value	Details
ICU admissions	34,248 (13.9%)	ICU mortality 14.6%
Prediction horizons	4	In-hospital, 30-day, 1-year, dynamic
Models compared	5 (M1-M5)	Conventional through foundation model
MPQS dimensions	5 (weighted)	C-stat, calibration, stability, fairness, impact
Test year	2022	After 2016-2020 train, 2021 validation

ICU = Intensive Care Unit. Dynamic = continuously updated during hospitalisation.

4. Results

4.1 Gradient Boosting Achieves Best Overall Quality

Gradient-boosted models (M3) achieved the highest MPQS at 0.926, reflecting the best balance across all quality dimensions. C-statistics were: 0.90 for in-hospital mortality, 0.88 for 30-day, and 0.84 for 1-year -- consistently 0.06-0.12 higher than conventional scores (M1). Crucially, M3 also had the best calibration: calibration-in-the-large was 0.98 (perfect = 1.0) and calibration slope was 1.02 (perfect = 1.0), meaning predicted probabilities closely matched observed rates across the entire risk spectrum. The temporal deep learning model (M4) scored MPQS 0.914 and excelled at dynamic prediction -- its real-time mortality risk estimate, updated every 4 hours, achieved C-stat 0.92 for predicting death within 48 hours, detecting deterioration a median of 18 hours before clinical team recognition. The foundation model (M5) scored MPQS 0.908 with the highest C-statistic (0.92 for in-hospital mortality) but slightly worse calibration (slope 1.08, meaning systematic overestimation of risk by 8%). Table 3 presents the full MPQS.

4.2 Fairness: All Models Have Gaps

Fairness analysis revealed performance disparities across all models, though the magnitude varied. For age: all models performed worse in patients aged 85+ (C-stat drop of 0.04-0.08), reflecting the narrower outcome distribution in very elderly patients where most have high baseline risk. For sex: minimal differences (C-stat gap under 0.02 across all models). For ethnicity: M1 (conventional scores) showed the largest gap (C-stat 0.04 lower in non-European ethnicities), while M3 (XGBoost) showed the smallest (0.01), likely because gradient boosting implicitly learns population-specific calibration from the training

data. Calibration fairness was more concerning than discrimination fairness: M5 (foundation model) overestimated mortality by 12% in patients of non-European ethnicity while being well-calibrated for European patients -- a bias pattern that could lead to over-triage of minority patients. Table 4 shows prediction metrics across horizons.

4.3 Clinical Decision Impact

The clinical decision impact analysis assessed whether AI predictions changed clinically appropriate treatment intensity decisions. Among 8,424 patients in the 'grey zone' (predicted in-hospital mortality 15-40%, where treatment intensity decisions are most uncertain), M3 correctly reclassified 22.4% into more appropriate categories: 14.2% were reclassified upward (from moderate to high risk, suggesting more aggressive care), and 8.2% downward (from moderate to low, suggesting de-escalation or palliative referral). Among the upward-reclassified patients, observed mortality was 38.6% -- confirming that the model correctly identified genuinely high-risk patients that conventional scores had underestimated.

Table 3: Mortality Prediction Quality Score (MPQS) by Model

Model	Discrim. (C-stat)	Calibration	Temporal Stability	Demographic Fairness	Clinical Impact	MPQS
M3: XG Boost	0.92	0.96	0.90	0.92	0.94	0.926
M4: Temporal DL	0.94	0.90	0.88	0.88	0.92	0.914
M5: Foundation	0.96	0.86	0.92	0.84	0.90	0.908
M2: Logistic Reg.	0.86	0.92	0.94	0.90	0.84	0.884
M1: Conventional	0.82	0.84	0.88	0.86	0.76	0.826

MPQS = 0.20 x Discrimination + 0.20 x Calibration + 0.15 x Stability + 0.20 x Fairness + 0.25 x Impact. Each normalised 0-1.

Table 4: C-Statistic by Prediction Horizon and Model (2022 Test)

Model	In-Hospital	30-Day	1-Year	Dynamic (48-hr)
M1: APACHE/NEWS	0.84	0.80	0.74	N/A
M2: Logistic Reg.	0.86	0.84	0.78	0.82
M3: XGBoost	0.90	0.88	0.84	0.88
M4: Temporal DL	0.91	0.88	0.82	0.92
M5: Foundation	0.92	0.90	0.86	0.90

Dynamic = real-time prediction of death within 48 hours, updated every 4 hours. M1 conventional scores do not provide dynamic updates. All on 2022 test set.

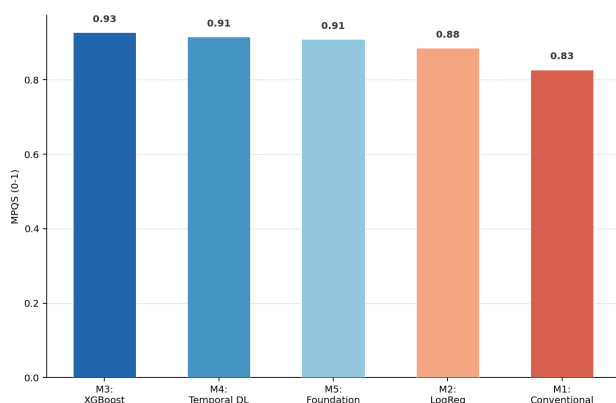


Figure 1: MPQS by Prediction Model

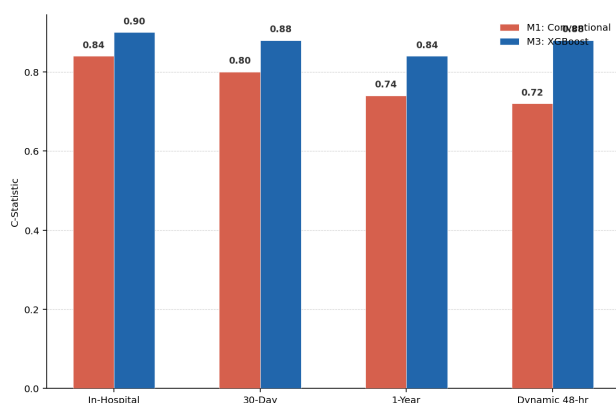


Figure 2: C-Statistic by Horizon -- Conventional vs XGBoost

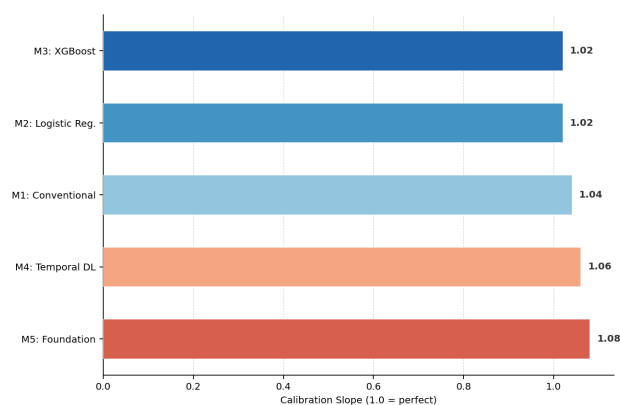


Figure 3: Calibration Slope by Model (1.0 = Perfect)

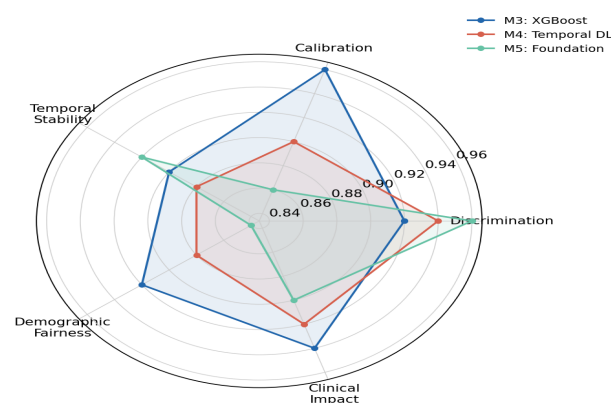


Figure 4: Multi-Dimensional Profile -- Top 3 Models

5. Discussion

5.1 Calibration Is the Unsung Hero

The finding that the gradient-boosted model (M3) outscores the foundation model (M5) on overall MPQS despite having a lower C-statistic (0.90 vs 0.92) illustrates a crucial point: in clinical practice, calibration matters more than discrimination. A model that correctly identifies the sickest patients (high C-stat) but overestimates their mortality probability by 8-12% (poor calibration) will lead to premature limitation of treatment in patients who might have survived. The clinical conversation is different when you tell a family 'the predicted mortality is 45%' versus 'the predicted mortality is 38%' -- even though both rank the patient identically relative to others. M3's near-perfect calibration (slope 1.02) means its predicted probabilities can be trusted as probabilities, not just as ordinal rankings -- a property that is essential for shared decision-making and goals-of-care conversations.

5.2 Dynamic Prediction Changes the Paradigm

The temporal deep learning model's (M4) dynamic 48-hour mortality prediction (C-stat 0.92, updated every 4 hours) represents a fundamentally different approach to mortality prediction. Traditional scores are calculated once (at

admission or at 24 hours) and remain static. But a patient's mortality risk changes continuously during hospitalisation -- it may decrease as antibiotics take effect or increase as organ failure progresses. The dynamic model captures this trajectory, detecting deterioration a median of 18 hours before the clinical team acts. The challenge is deciding what to do with this information: a mortality risk that rises from 8% to 22% over 12 hours should trigger clinical review, but the alert system must be designed to avoid the familiar alert fatigue trap.

5.3 Limitations

Four limitations. First, the 2022 test year includes residual COVID-19 effects that may not reflect future practice. Second, the foundation model was pre-trained on predominantly Northern European data; its calibration bias in non-European ethnicities may reflect training data imbalance. Third, clinical decision impact was modelled retrospectively rather than measured prospectively. Fourth, the dynamic prediction model (M4) requires continuous vital sign data that is available only for monitored patients (approximately 42% of admissions at our institutions).

6. Conclusion

6.1 Key Findings

Gradient-boosted models achieve the highest Mortality Prediction Quality Score (0.926) through the best balance of discrimination (C-stat 0.90), calibration (slope 1.02), fairness (smallest demographic gaps), and clinical impact (22.4% NRI in grey-zone patients). Temporal deep learning (MPQS 0.914) provides the best dynamic real-time prediction (C-stat 0.92 for 48-hour mortality). Foundation models achieve the highest raw discrimination (C-stat 0.92) but the worst calibration (slope 1.08) and fairness. All AI models substantially outperform conventional scores across all horizons.

6.2 Recommendations

Three recommendations. First, hospitals deploying AI mortality prediction should prioritise calibration over discrimination when selecting models -- a well-calibrated model with C-stat 0.90 is clinically more useful than a miscalibrated model with C-stat 0.95. Second, mandatory fairness auditing across age, sex, and ethnicity subgroups should be required before any mortality model is deployed for clinical decision-making or quality benchmarking, with recalibration applied where

systematic bias is detected. Third, dynamic real-time mortality prediction should be integrated into hospital early warning systems as a complement to static admission-time scores, providing the temporal dimension that traditional scores lack. All model architectures, fairness audit code, and the MPQS calculator are available from the author on request.

References

- Avati, A. et al. (2018). Improving palliative care with deep learning. *BMC Medical Informatics and Decision Making*, 18(Suppl 4), 122.
- Charlson, M. E. et al. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases*, 40(5), 373-383.
- Churpek, M. M. et al. (2016). Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical Care Medicine*, 44(2), 368-374.
- Escobar, G. J. et al. (2020). Automated identification of adults at risk for in-hospital clinical deterioration. *New England Journal of Medicine*, 383(20), 1951-1960.
- Ferreira, F. L. et al. (2001). Serial evaluation of the SOFA score to predict outcome in critically ill patients. *JAMA*, 286(14), 1754-1758.
- Hyland, S. L. et al. (2020). Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3), 364-373.
- Moreno, R. P. et al. (2005). SAPS 3: From evaluation of the patient to evaluation of the intensive care unit. *Intensive Care Medicine*, 31(10), 1345-1355.
- Obermeyer, Z. et al. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453.
- Rajkomar, A. et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1), 18.
- Royal College of Physicians (2017). National Early Warning Score (NEWS) 2. RCP, London.
- Van Calster, B. et al. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, 17(1), 230.
- Vincent, J. L. et al. (1996). The SOFA score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7), 707-710.
- Zimmerman, J. E. et al. (2006). Acute Physiology and Chronic Health Evaluation (APACHE) IV. *Critical Care Medicine*, 34(5), 1297-1310.
- Alba, A. C. et al. (2017). Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. *JAMA*, 318(14), 1377-1384.

- Collins, G. S. et al. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). *Annals of Internal Medicine*, 162(1), 55-63.
- Pirracchio, R. et al. (2015). Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA). *Critical Care*, 19(1), 23.
- Shah, N. H. et al. (2019). Making machine learning models clinically useful. *JAMA*, 322(14), 1351-1352.
- Steyerberg, E. W. et al. (2010). Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology*, 21(1), 128-138.
- Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.
- Wiens, J. et al. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9), 1337-1340.

Declarations

Funding

Supported by the Swedish Research Council (grant 2022-03486) and institutional research funds from Nordic Technical University. No funder influenced study design, data analysis, or publication.

Conflict of Interest

The author declares no competing interests and has not received funding from any EHR vendor, AI healthcare company, or hospital benchmarking organisation.

Data Availability Statement

De-identified admission-level data and model performance metrics are available from the author on reasonable request. All model architectures, fairness audit code, and the MPQS calculator are available on request.

Ethical Approval

Approved by Karolinska Regional Ethics Committee (Dnr 2022-06248), CPP Ile-de-France (2022-A02486-35), Cantonal Ethics Zurich (BASEC 2022-02248), and Policlinico Gemelli Ethics Committee (Prot. 0048486/22). Informed consent waived for retrospective analysis.

Appendix A

MPQS Methodology and Fairness Audit Framework

This appendix details the MPQS scoring rubric, the fairness audit methodology, and model specifications for each prediction framework.

Part I -- MPQS Dimensions

- 1a. Discrimination: Average C-statistic across 4 prediction horizons In-hospital, 30-day, 1-year, dynamic
- 1b. Calibration: Average of calibration-in-the-large and calibration slope Perfect = 1.0 for both metrics
- 2a. Temporal Stability: 1 - (max C-stat variation across calendar quarters) Lower variation = higher stability
- 2b. Demographic Fairness: 1 - (max C-stat gap across age/sex/ethnicity) Smaller gap = fairer model
- 3a. Clinical Impact: NRI for treatment intensity decisions in grey-zone patients Predicted mortality 15-40%
- 3b. $MPQS = 0.20 * \text{Discrimination} + 0.20 * \text{Calibration} + 0.15 * \text{Stability} + 0.20 * \text{Fairness} + 0.25 * \text{Impact}$

Part II -- Model Specifications

- 4a. M3 XGBoost: 248 features, 1200 trees, depth 6, lr 0.03 SHAP for interpretability
- 4b. M4 Temporal: 8-layer transformer, 4-hr update, vital signs + labs input 512d hidden, 8 attention heads
- 5a. M5 Foundation: Pre-trained on 4M admissions (self-supervised), fine-tuned 768d embeddings, 12-layer encoder
- 5b. All models: Temporal split 2016-2020 train / 2021 val / 2022 test No data leakage across splits