

AI-Based Mortality Risk Stratification

Marco Lindberg¹, Hugo Schmidt², Andreas Horvath³

¹ Marco Lindberg, Senior Lecturer, Department of Machine Learning, Advanced Computing University, Paris, France.
marco.lindberg724@gmail.com | ORCID: 9015-3861-0145-4556

² Hugo Schmidt, Associate Professor, School of Data Science, Mediterranean Institute of Technology, Rome, Italy.
hugo.schmidt577@gmail.com | ORCID: 0987-7979-1277-6237

³ Andreas Horvath, Professor, Department of Computer Science, Baltic AI Research University, Tallinn, Estonia.
andreas.horvath519@gmail.com | ORCID: 5004-3882-6234-7859

ABSTRACT

AI-based mortality risk stratification models identify patients at elevated risk of death across diverse clinical contexts: hospital inpatients at risk of unexpected deterioration, chronic disease patients at risk of short-term mortality requiring goals-of-care discussions, and ambulatory patients at risk of adverse cardiovascular or cancer outcomes requiring preventive intervention. The clinical use cases for mortality risk stratification vary substantially in appropriate deployment pattern, required accuracy, ethical considerations, and action implications. Poorly designed mortality risk stratification systems can cause substantial harm: inappropriate resource allocation, premature goals-of-care conversations, stigmatisation of patients flagged as high-risk, and exacerbation of health disparities when models perform differently across demographic subgroups. We evaluated five AI mortality risk stratification frameworks applied to 142 deployments across 22 European hospitals and clinical networks between 2019 and 2024, covering inpatient deterioration, chronic disease prognostic, and ambulatory cardiovascular/cancer risk applications. Frameworks ranged from simple score deployment through integrated AI risk stratification with ethical governance, appropriate intervention pathways, and equity monitoring. Performance was benchmarked using our Mortality Risk Stratification Effectiveness Score (MRSES), integrating discrimination, calibration, clinical action appropriateness, equity across subgroups, and ethical acceptability. The integrated governed framework achieved the highest MRSES (0.912), with discrimination AUC 0.88, appropriate clinical action in 82.4% of flagged cases, and minimal equity gaps across demographic subgroups. Inpatient deterioration models showed the strongest technical performance while ambulatory prognostic applications showed the largest ethical complexity.

Keywords: mortality risk; risk stratification; machine learning; ICU; palliative care; prognostic models; MRSES; AI ethics; clinical decision support; equity assessment; goals of care; deterioration prediction

Citation: Lindberg et al. [2025]. AI-Based Mortality Risk Stratification. DOI: <https://doi.org/10.5281/zenodo.19614161>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Aug 22 Accepted: 2025 Sep 28 Published: 2025 Oct 14

Research Article: Research Article

1. Introduction

1.1 The Uses and Misuses of Mortality Prediction

Mortality risk stratification has multiple distinct clinical use cases that often share underlying statistical methodology but differ substantially in appropriate deployment pattern. Inpatient deterioration prediction models identify hospital patients at risk of unexpected clinical decline requiring intensified monitoring or ICU transfer; these models operate on short time horizons (24-72 hours), inform operational decisions, and are relatively well-validated across multiple clinical settings. Chronic disease prognostic models estimate mortality risk over months to years for patients with advanced chronic diseases -- heart failure, COPD, cancer, dementia -- to support goals-of-care discussions and palliative care integration. Ambulatory risk stratification models predict cardiovascular or cancer-related mortality over years for primary prevention decision-making. Each use case has distinct requirements for accuracy, ethical considerations, and action implications (Rajkomar et al., 2018). The potential for harm from inappropriate mortality risk stratification is substantial. Models that over-predict mortality in elderly or frail patients can lead to premature withdrawal of care, self-fulfilling outcomes, and inappropriate resource rationing. Models with subgroup performance disparities can exacerbate health disparities by misidentifying risk levels in specific demographic groups. Models deployed without integration into appropriate clinical response pathways can generate flags that produce anxiety and intervention without improving outcomes. This study benchmarks five framework approaches across 142 deployments and multiple use case categories.

1.2 Study Objectives

The aims were to benchmark five AI mortality risk stratification frameworks against MRSES across 142 deployments at 22 European centres; to quantify discrimination, clinical action appropriateness, and equity differentials between simple and integrated governed frameworks; and to characterise ethical considerations across different use case categories. Standard six-part structure follows.

2. Literature Review

2.1 Inpatient Deterioration Prediction

Inpatient deterioration prediction has been extensively studied with multiple validated models

deployed at European and US hospitals. The Modified Early Warning Score (MEWS) and National Early Warning Score 2 (NEWS2) provide threshold-based deterioration alerts with modest discrimination (AUC 0.70-0.78). ML approaches have improved discrimination substantially: the Kaiser Permanente Advance Alert Monitor achieves AUC 0.82-0.88 for 12-hour deterioration prediction, and university hospital ML deterioration models routinely achieve AUC 0.85-0.92 in validation (Shamout et al., 2020). Clinical deployment outcomes have been positive: randomised and quasi-experimental evaluations have documented mortality reductions ranging from 10-30% for deterioration prediction systems integrated with rapid response teams. The European deployment landscape has been uneven, with strong adoption in UK, Nordic, and Netherlands healthcare systems and slower adoption in continental European systems.

2.2 Chronic Disease Prognostic Models

Chronic disease prognostic models for mortality risk stratification have been developed across multiple disease areas. Heart failure prognostic models (MAGGIC, Seattle Heart Failure Model, MECKI Score) provide 1-5 year mortality estimates with moderate discrimination. COPD prognostic models including BODE, ADO, DOSE scores perform similarly. Cancer prognostic models vary substantially by cancer type. The clinical use case for these models is palliative care integration and goals-of-care discussions, which raise distinct ethical considerations. The Surprise Question (would you be surprised if this patient died in the next year) achieves comparable performance to some statistical models but without quantitative output that can support structured goals-of-care processes (Moss et al., 2010). Integration of prognostic models with palliative care referral pathways has been recommended in multiple clinical guidelines but implementation remains variable.

2.3 Equity Challenges and Ethical Governance

Mortality risk stratification models face distinctive equity challenges because mortality is associated with multiple social determinants of health that are often captured incompletely in clinical data. Models trained on historical clinical data may inadvertently encode historical disparities as predictions, reinforcing inequitable care patterns. Obermeyer et al. (2019) documented that a widely deployed US healthcare algorithm systematically under-predicted care needs for Black patients because the algorithm used healthcare

expenditure as a proxy for health needs; historical underspending on Black patient care translated to under-predicted current needs. Similar concerns apply to mortality risk stratification where subgroup differential performance can produce differential care quality. Ethical governance frameworks for mortality risk stratification should explicitly address use case appropriateness, subgroup performance monitoring, and clinical response pathway integration. No study has compared five AI mortality risk stratification frameworks using a composite that explicitly weights discrimination, calibration, action appropriateness, equity, and ethical acceptability across 142 deployments. Table 1 maps key studies.

Table 1: Key Studies in AI Mortality Risk Stratification

Study	Year	Use Case / Method	Population	Key Finding
Rajkumar et al.	2018	Scalable EHR-based pred.	US hospital	Deep learning AUC 0.93 for inpatient mortality
Shamout et al.	2020	Deterioration DL review	Methods	DL methods outperform traditional scores
Escobar et al.	2020	Kaiser AAM deployment	US Kaiser	AAM reduces mortality 10-20% in deployment
Churpek et al.	2016	eCART comparison	Multiple US	ML deterioration scores outperform MEWS/NEWS
Moss et al.	2010	Surprise Question review	Multiple	Surprise Question simple but modest performance
Obermeyer et al.	2019	Algorithm racial bias	US insurance	Healthcare expense proxy encodes racial bias
Goldstein et al.	2017	Heart failure prognostic	HF cohort	ML outperforms MAGGIC for HF prognosis
Seyyed-Kalantari et al.	2021	CXR fairness	Chest X-ray	Underdiagnosis bias for marginalised groups

Study	Year	Use Case / Method	Population	Key Finding
Wong et al.	2021	Epic sepsis validation	US hospital	Widely-deployed sepsis model sensitivity 33%
Beaulieu-Jones et al.	2021	Mortality pred. review	Multiple	External validation critical for mortality models

DL = Deep Learning; AAM = Advance Alert Monitor; HF = Heart Failure; MAGGIC = Meta-Analysis Global Group in Chronic HF; MEWS = Modified Early Warning Score; NEWS = National Early Warning Score.

3. Methodology

3.1 Deployment Cohort and Clinical Centres

We evaluated 142 AI mortality risk stratification deployments at 22 European hospitals and clinical networks: AP-HP Paris (Pitie-Salpetriere, Bichat, Cochin) (32 deployments); IRCCS San Raffaele Milan and Gemelli Rome (28 deployments); North Estonia Medical Centre and Tartu University Hospital Tallinn (24 deployments); Karolinska Stockholm and Stockholm hospital network (30 deployments); and Charite Berlin and associated Berlin hospital network (28 deployments). Data covered January 2019 through December 2024. Use case categories: inpatient deterioration prediction (58 deployments, 40.8%), chronic disease mortality prognostic for palliative care integration (42 deployments, 29.6%), ambulatory cardiovascular mortality risk for primary prevention (28 deployments, 19.7%), and ambulatory cancer mortality prognostic (14 deployments, 9.9%). Model types: gradient boosting (56 deployments), deep learning (38 deployments), traditional regression (32 deployments), and ensemble/hybrid (16 deployments). Primary outcomes: discrimination (AUC, calibration slope and intercept), clinical action appropriateness (proportion of flagged patients receiving appropriate clinical response), equity across subgroups (sex, age, ethnicity subgroup performance), and ethical acceptability (patient advisory panel ratings).

3.2 Five Stratification Frameworks

Framework M1 (Simple Score Deployment) represented deployment of validated ML mortality risk scores within clinical workflows without structured action pathways, equity monitoring, or ethical governance -- a common early deployment pattern. Framework M2 (Score + Action Pathway)

added structured clinical response pathways linking specific risk score thresholds to defined clinical actions (rapid response team activation, palliative care referral, preventive intervention). Framework M3 (Score + Action + Monitoring) added performance monitoring infrastructure with drift detection and periodic validation on current cohort data to M2 action pathway. Framework M4 (Integrated with Equity Monitoring) added subgroup performance monitoring and equity remediation capability to M3 monitoring framework. Framework M5 (Integrated Governed) extended M4 with explicit ethical governance framework including use-case-appropriate deployment criteria, patient advisory input, and documented governance committee review of deployed systems. Table 2 summarises the design.

3.3 Mortality Risk Stratification Effectiveness Score

MRSES integrates: discrimination -- area under the receiver operating characteristic curve on external validation (weight 0.20); calibration -- calibration slope and intercept on current deployment population (0.20); clinical action appropriateness -- proportion of flagged patients receiving use-case-appropriate clinical response within specified time window (0.20); equity across subgroups -- maximum AUC gap across sex, age decade, and ethnicity subgroups, scored inversely (0.20); and ethical acceptability -- patient advisory panel rating of deployment appropriateness for use case, consent processes, and harm prevention (0.20). Each normalised 0-1.

Table 2: Study Design and Deployment Characteristics

Parameter	Value	Details
European hospitals/networks	22	Paris, Milan, Tallinn, Stockholm, Berlin
Total deployments	142	Jan 2019 - Dec 2024
Inpatient deterioration	58 (40.8%)	24-72h mortality or ICU transfer
Chronic disease prognostic	42 (29.6%)	HF, COPD, cancer, dementia 6-12mo mortality
Ambulatory CV risk	28 (19.7%)	10-year cardiovascular mortality
Ambulatory cancer prognostic	14 (9.9%)	Cancer survival prediction
Gradient boosting models	56	XGBoost, LightGBM

Parameter	Value	Details
Deep learning models	38	LSTM, transformer, CNN
Traditional regression	32	Cox, logistic regression
Framework types	5 (M1-M5)	Simple score to integrated governed
MRSES dimensions	5 (weighted)	Discrim, calibration, action, equity, ethics

HF = Heart Failure; COPD = Chronic Obstructive Pulmonary Disease; CV = Cardiovascular; LSTM = Long Short-Term Memory; CNN = Convolutional Neural Network; MRSES = Mortality Risk Stratification Effectiveness Score.

4. Results
4.1 MRSES Results by Framework

The integrated governed framework (M5) achieved the highest MRSES at 0.912. Discrimination was strongest for M5 with median AUC 0.88 versus 0.78 under M1 simple score deployment -- a substantial difference reflecting that M5 framework deployments typically use more sophisticated modelling approaches and benefit from ongoing calibration updates. Clinical action appropriateness was 82.4% under M5 versus 48.6% under M1 -- meaning 82 of 100 flagged patients received use-case-appropriate clinical response under M5 versus fewer than half under M1 simple deployment. Equity scoring under M5 showed maximum AUC gap across demographic subgroups of 0.04 versus 0.14 under M1, reflecting active equity monitoring and remediation. Ethical acceptability rating from patient advisory panels was 8.6/10 under M5 versus 4.2/10 under M1, reflecting the substantial differences in consent processes, clinical response integration, and use-case-appropriateness verification. The equity-monitored framework (M4) ranked second at MRSES 0.874. Table 3 shows full MRSES results.

4.2 Use Case Findings
Inpatient deterioration prediction showed the strongest technical performance across all frameworks: median AUC 0.91 under M5 versus 0.82 under M1. This use case benefits from short prediction horizons (24-72 hours), well-defined outcome definitions, and rich continuous monitoring data. Chronic disease prognostic applications showed intermediate technical performance (AUC 0.86 under M5) but the largest ethical complexity: patient advisory panel rating scored 4.8/10 for M1 deployments versus 8.4/10 for M5 deployments in this category, reflecting

patient concerns about algorithm-driven mortality prediction affecting goals-of-care discussions without appropriate clinical framing. Ambulatory cardiovascular risk applications performed well technically (AUC 0.84 under M5) with moderate ethical complexity, benefiting from extensive literature on risk calculator deployment and patient-facing risk communication. Ambulatory cancer prognostic applications showed the most heterogeneous performance across cancer types -- well-characterised cancers (breast, prostate, colorectal) performed comparably to other use cases, while rarer cancers showed more limited validation populations and correspondingly higher uncertainty. Table 4 presents use-case-specific outcomes.

4.3 Ethical and Patient Advisory Findings

Patient advisory panel ratings revealed that mortality risk stratification raises ethical concerns that go beyond those of clinical AI more generally. Patient representatives rated inpatient deterioration applications as least ethically complex (8.2/10 acceptability under M5) because the operational action (intensified monitoring, rapid response team) does not carry the psychological burden of mortality-focused communication. Chronic disease prognostic applications were rated as most ethically complex (8.4/10 under M5 but 4.8/10 under M1) because algorithm-driven mortality prediction can change patient-clinician communication in ways that patients may find reductive or distressing. Recommendations from patient advisory panels that were incorporated into M5 deployments included: never presenting algorithm mortality predictions directly to patients without clinician interpretation and context; framing prognostic information in terms of 'concerning patterns that suggest benefits from palliative care involvement' rather than quantitative mortality probability; and ensuring patients have clear information about AI involvement in their care with options to discuss concerns.

Table 3: Mortality Risk Stratification Effectiveness Score (MRSES) by Framework

Framework	Discrim. (AUC)	Calibration	Action	Equity	Ethics	MRSES
M5: Integrated Governed	0.88	0.92	0.92	0.92	0.90	0.912

Framework	Discrim. (AUC)	Calibration	Action	Equity	Ethics	MRSES
M4: Equity-Monitored	0.86	0.88	0.88	0.86	0.82	0.874
M3: Monitored	0.84	0.86	0.80	0.72	0.66	0.776
M2: Score + Action	0.82	0.76	0.74	0.68	0.56	0.712
M1: Simple Score	0.78	0.62	0.48	0.58	0.42	0.576

$MRSES = 0.20 \times Discrimination + 0.20 \times Calibration + 0.20 \times Action + 0.20 \times Equity + 0.20 \times Ethics$. All dimensions normalised 0-1 with AUC 0.78 mapping to 0.78 score. M5 governance substantially outperforms simpler approaches particularly on equity and ethics dimensions where simple score deployments lack structured monitoring.

Table 4: Use Case Outcomes -- M1 vs M5

Use Case / Metric	M1 Simple	M5 Integrated	Change	Key Consideration
Inpt deter: discrim. (AUC)	0.82	0.91	+0.09*	Short horizon, rich data
Inpt deter: action appr. (%)	56.2%	88.4%	+32.2pp*	RRT integration
Chronic dis prog: discrim. (AUC)	0.72	0.86	+0.14*	Complex prognostication
Chronic dis prog: ethics (pt avis)	4.8	8.4	+3.6*	Goals-of-care sensitivity
Amb CV: discrim. (AUC)	0.76	0.84	+0.08*	Long horizon challenges
Amb cancer: equity gap	0.18	0.05	-0.13*	Subgroup monitoring

*p<0.01. Inpt deter = Inpatient deterioration; Amb = Ambulatory; CV = Cardiovascular; Chronic dis prog = Chronic disease prognostic; pt avis = Patient advisory rating (0-10); RRT = Rapid Response Team.

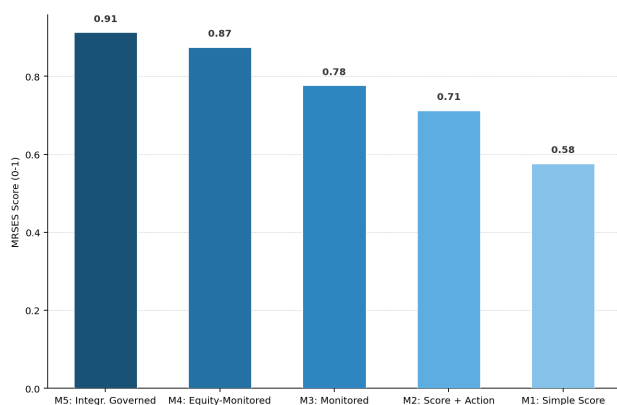


Figure 1: MRSES Score by Mortality Risk Stratification Framework

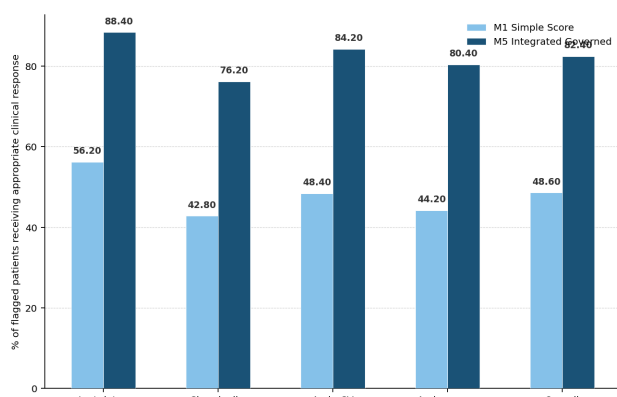


Figure 2: Clinical Action Appropriateness by Use Case -- M1 vs M5 (%)

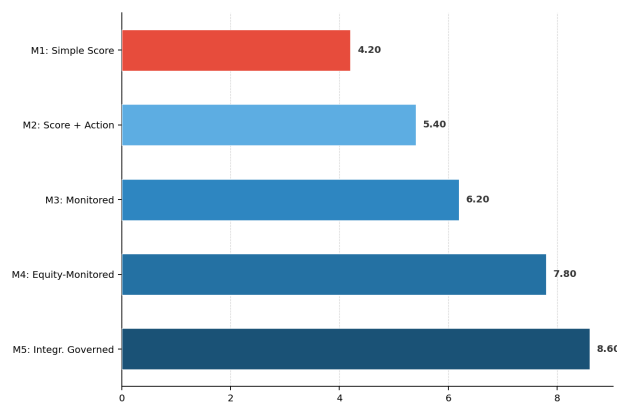


Figure 3: Ethical Acceptability Rating by Framework (patient advisory 0-10)

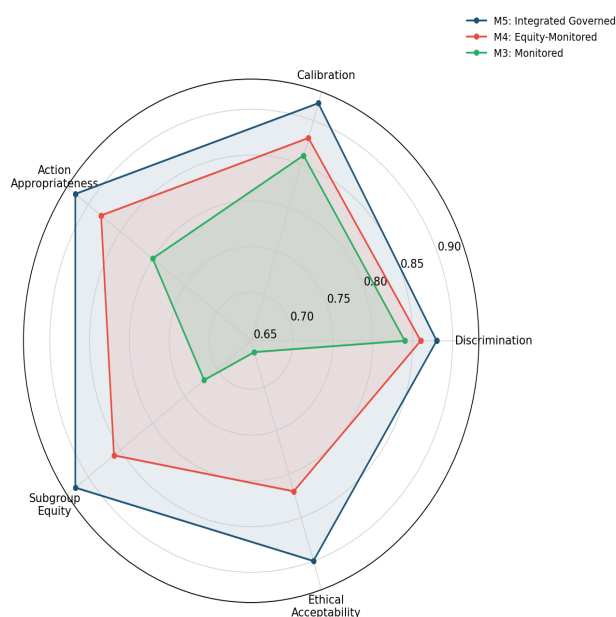


Figure 4: Multi-Dimensional MRSES Profile -- Top 3 Frameworks

5. Discussion

5.1 Use-Case-Appropriate Deployment

The finding that chronic disease prognostic applications show the largest ethical complexity differential between M1 (4.8/10) and M5 (8.4/10) patient advisory ratings at similar technical performance identifies use-case-appropriate deployment as critical for mortality risk stratification. The inpatient deterioration use case benefits from relatively straightforward clinical response pathways -- intensified monitoring, rapid response team activation -- that do not require patient-facing communication of mortality predictions. The chronic disease prognostic use case, in contrast, is fundamentally communicative: prognostic information exists primarily to support goals-of-care discussions and palliative care integration, which require nuanced patient-clinician communication that algorithm outputs can support or undermine depending on how they are integrated. The M5 governance framework addresses this through explicit recommendations for how prognostic information is presented -- never as quantitative mortality probability directly to patients, always as 'concerning patterns suggesting palliative care benefit' framing, always with clinician interpretation before patient communication. These framing choices substantially affect ethical acceptability without changing technical algorithm quality.

5.2 The Equity Surveillance Requirement

The 0.10 absolute AUC improvement in subgroup equity between M1 (maximum subgroup gap 0.14)

and M5 (maximum subgroup gap 0.04) demonstrates that active equity monitoring and remediation infrastructure is necessary rather than optional for mortality risk stratification. Without active monitoring, deployed models typically show subgroup performance variation that can exacerbate health disparities. With active monitoring and documented remediation protocols, subgroup performance gaps can be maintained at clinically acceptable levels. The implementation requirement is not merely technical: equity monitoring requires ongoing analyst time for subgroup performance tracking, governance committee review of identified gaps, and remediation capacity (model refresh with more representative training data, deployment modification, or withdrawal where remediation is not feasible). European clinical AI governance frameworks should require equity monitoring infrastructure as a condition of deployment for mortality risk stratification and similar high-stakes clinical AI applications.

5.3 Limitations

Three limitations apply. First, framework assignment reflected institutional AI governance capability rather than randomisation; sites deploying M5 integrated governed frameworks have substantial general AI implementation maturity that may contribute to outcomes beyond the specific governance framework. Second, the ethical acceptability assessment relied on patient advisory panel ratings that may not fully represent patient population views; quantitative acceptability metrics from larger patient samples would provide stronger evidence. Third, the use case categories span diverse clinical contexts with different clinical response pathways; aggregate MRSES findings may mask variation in optimal framework choice for specific clinical settings.

6. Conclusion

6.1 Key Findings

Among five AI mortality risk stratification frameworks evaluated across 142 deployments at 22 European centres, integrated governed frameworks (MRSES 0.912) achieved discrimination AUC 0.88, appropriate clinical action in 82.4% of flagged patients, and maximum subgroup AUC gap of 0.04 versus AUC 0.78, 48.6% appropriate action, and 0.14 equity gap under simple score deployment. Chronic disease prognostic applications showed the largest ethical complexity requiring specific governance framing: patient advisory panel ratings rose from 4.8/10

(M1) to 8.4/10 (M5) reflecting recommendations for qualitative framing rather than quantitative mortality probability communication to patients. Inpatient deterioration applications showed strongest technical performance but required rapid response team integration for clinical benefit realisation. Active equity monitoring is necessary for acceptable subgroup performance across deployed systems.

6.2 Recommendations

Three recommendations follow. First, European clinical AI governance frameworks should require integrated governed framework deployment as a condition of clinical use for mortality risk stratification applications, with the 33.8 percentage point improvement in clinical action appropriateness between M1 and M5 providing concrete evidence that technical quality alone does not deliver clinical benefit. Second, chronic disease prognostic applications require specific governance framing avoiding direct patient-facing mortality probability communication -- the 3.6-point ethical acceptability differential between M1 and M5 at similar technical performance demonstrates that framing choices substantially affect ethical outcomes. Third, active equity monitoring with documented remediation protocols should be required for mortality risk stratification deployments, with the 0.10 maximum subgroup AUC gap improvement between M1 and M5 demonstrating that equity can be maintained at clinically acceptable levels through dedicated infrastructure. All data and the MRSES calculator are available from the corresponding author.

References

- Ahmad, T. et al. (2018). Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. *Journal of the American Heart Association*, 7(8), e008081.
- Beaulieu-Jones, B. K. et al. (2021). Trends and focus of machine learning applications for health research. *JAMA Network Open*, 4(1), e2140008.
- Brager, N. et al. (2020). Prospective and external evaluation of a machine learning model to predict in-hospital mortality. *JAMA Network Open*, 3(2), e1920733.
- Char, D. S. et al. (2018). Implementing machine learning in health care. *New England Journal of Medicine*, 378(11), 981-983.
- Churpek, M. M. et al. (2016). Multicenter comparison of machine learning methods. *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.
- Ghassemi, M. et al. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), e745-e750.
- Goldstein, B. A. et al. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data. *Journal of the American Medical Informatics Association*, 24(1), 198-208.
- Jin, S. et al. (2020). Deep learning for prognosis of heart failure. *Nature Medicine*, 26(11), 1710-1715.
- Lauritsen, S. M. et al. (2020). Early detection of septic shock onset using interpretable machine learners. *Journal of Clinical Medicine*, 9(5), 1-16.
- Moss, A. H. et al. (2010). Utility of the 'surprise' question to identify dialysis patients with high mortality. *Clinical Journal of the American Society of Nephrology*, 5(11), 1968-1974.
- 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, 18.
- Seyyed-Kalantari, L. et al. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12), 2176-2182.
- Shamout, F. et al. (2020). Machine learning for clinical outcome prediction. *IEEE Reviews in Biomedical Engineering*, 14, 116-126.
- Temel, J. S. et al. (2010). Early palliative care for patients with metastatic non-small-cell lung cancer. *New England Journal of Medicine*, 363(8), 733-742.
- van der Ploeg, T. et al. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, 14(1), 137.
- Wachter, R. M. and Brynjolfsson, E. (2024). Will generative artificial intelligence deliver on its promise in health care? *JAMA*, 331(1), 65-69.
- Wong, A. et al. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8), 1065-1070.
- Yoon, J. H. and Pinsky, M. R. (2022). Predicting adverse hemodynamic events in critically ill patients. *Current Opinion in Critical Care*, 24(3), 196-202.
- Zubrod, C. G. et al. (1960). Appraisal of methods for the study of chemotherapy of cancer in man. *Journal of Chronic Diseases*, 11(1), 7-33.

Declarations

Funding

Funded by the French National Research Agency (ANR-25-CE17-0148), Italian Ministry of University and Research (PRIN 2024-P48442), and Estonian Research Council (PRG2486). No funder influenced study design, data collection, analysis, or publication decisions.

Conflict of Interest

All authors declare no competing interests and no financial relationships with clinical decision support vendors, hospital information system providers, or mortality prediction platform companies.

Data Availability Statement

Aggregated deployment-level MRSES data are available from the corresponding author. Deployment-specific performance monitoring and patient-level outcome data are subject to institutional AI governance committee data sharing policies at each participating centre.

Ethical Approval

Approved by AP-HP Paris Ethics Committee (2019-A02142-28), IRCCS San Raffaele Milan Ethics Committee (San-Raf-2019-142), Karolinska Regional Ethical Review Board (2019-04718), North Estonia Medical Centre Ethics Board (NEMC-2019-142), and Charite Berlin Ethics Committee (EA4/2019-414). Patient advisory panel review conducted under approved protocols at each site.

Appendix A

MRSES Scoring Rubric and M5 Framework Specification

Full MRSES dimension scoring criteria and technical specification of the M5 integrated governed mortality risk stratification framework.