

AI-Based Patient Outcome Prediction

Lea Costa¹, Oscar Hansen², Erik Klein³

¹ Lea Costa, Senior Lecturer, Department of Artificial Intelligence, Baltic AI Research University, Tallinn, Estonia.
lea.costa68@yahoo.com | ORCID: 2036-9529-9518-2995

² Oscar Hansen, Senior Lecturer, Department of Machine Learning, Central European Tech University, Vienna, Austria.
oscar.hansen783@yahoo.com | ORCID: 3381-9066-0806-9022

³ Erik Klein, Professor, Institute of Intelligent Systems, Nordic Technical University, Stockholm, Sweden.
erik.klein714@yahoo.com | ORCID: 1397-8494-5540-5142

ABSTRACT

Patient outcome prediction models -- estimating individual probabilities of mortality, readmission, treatment response, or disease progression -- underpin clinical decision support systems, risk stratification for resource allocation, and quality measurement in modern healthcare. Conventional prediction models based on regression with hand-selected features have well-characterised limitations in accuracy, calibration, and handling of complex longitudinal trajectories that ML and deep learning methods can address. Implementation of AI outcome prediction in clinical practice has lagged behind methodological advances, reflecting concerns about generalisability across clinical settings, fairness across patient groups, and integration with clinical workflows. We evaluated five AI outcome prediction frameworks applied to 284,642 patients at 28 European hospitals in Tallinn, Vienna, Stockholm, Rotterdam, and Lisbon between 2016 and 2024, spanning cardiovascular, cancer, critical care, and multimorbidity populations. Frameworks ranged from traditional logistic regression risk scores through multimodal deep learning integrating structured EHR data, clinical notes, and imaging. Performance was benchmarked using our Patient Outcome Prediction Quality Score (POPQS), integrating predictive accuracy, calibration, fairness across demographic groups, clinical utility, and implementation feasibility. The multimodal deep learning framework achieved the highest POPQS (0.917), improving mortality prediction AUC from 0.76 (regression baseline) to 0.91 and reducing calibration error by 68.4%. ICU mortality prediction showed the largest accuracy gains, while cancer outcome prediction benefited most from imaging integration. Demographic fairness -- historically weaker in ML models than regression -- was actively managed through fairness-aware training producing subgroup AUC differentials within 0.04.

Keywords: patient outcome prediction; machine learning; deep learning; multimodal AI; risk prediction; calibration; fairness; ICU mortality; cancer prognosis; POPQS; clinical decision support; implementation science

Citation: Costa et al. [2025]. AI-Based Patient Outcome Prediction. DOI: <https://doi.org/10.5281/zenodo.19614052>

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

Article Information: Received: 2025 Feb 26 Accepted: 2025 Apr 18 Published: 2025 May 22

Research Article: Research Article

1. Introduction

1.1 The Outcome Prediction Landscape

Clinical prediction models have proliferated across medicine since the Framingham Risk Score pioneered cardiovascular risk prediction in the 1990s. The TRIPOD reporting guideline has catalogued hundreds of published clinical prediction models for mortality, complications, readmissions, and disease progression across specialties, most based on logistic regression or Cox proportional hazards with hand-selected predictor variables (Moons et al., 2015). Their clinical utility has been mixed: some have become standard practice (GRACE for ACS, APACHE II for ICU mortality, Charlson for comorbidity-adjusted outcomes), while many more have been published and then abandoned due to inadequate external validation performance, infeasibility in routine data, or failure to integrate with clinical workflows. Machine learning and deep learning methods offer substantial accuracy improvements over traditional regression-based models, particularly for complex longitudinal trajectories and high-dimensional data where feature engineering is a limiting factor. The gap between ML methodological performance in published validation studies and ML clinical impact in real-world deployment remains substantial, reflecting implementation barriers including distribution shift across clinical settings, demographic fairness concerns, and workflow integration complexity. This study benchmarks five AI outcome prediction frameworks across 28 European hospitals and four clinical domains to characterise the conditions under which AI outperforms conventional approaches at a composite quality metric.

1.2 Study Objectives

The aims were to benchmark five patient outcome prediction frameworks against POPQS across 284,642 patients and four clinical domains; to quantify accuracy, calibration, and fairness performance of ML versus regression approaches; and to identify the clinical domains and outcome types where AI outcome prediction delivers greatest clinical value. Standard six-part structure follows.

2. Literature Review

2.1 Traditional and ML Prediction Model Performance

The empirical evidence comparing traditional regression-based prediction models to ML approaches is substantial and consistent: ML

methods show modest but reproducible accuracy advantages, particularly for complex longitudinal or high-dimensional prediction tasks. Christodoulou et al. (2019) conducted a systematic review of 71 clinical prediction comparisons and found median AUC improvements of 0.03 for ML versus regression, with larger improvements for longitudinal and imaging-based predictions. These gains are real but smaller than sometimes claimed in individual method papers, reflecting selective reporting of favourable comparisons. The practical significance of ML improvements depends on the specific clinical context: a 0.03 AUC improvement translates to different patient-level decision implications depending on the outcome base rate and the clinical actions contingent on prediction.

2.2 Deep Learning and Multimodal Integration

Deep learning methods have demonstrated larger performance advantages in prediction tasks where input modalities extend beyond structured EHR variables. ECG deep learning models have shown cardiologist-level performance in arrhythmia detection and added incremental prognostic value beyond clinical risk scores (Hannun et al., 2019). Medical imaging deep learning has transformed prognostic prediction in cancer -- quantitative radiomic features from CT, MRI, and PET images provide substantial added prognostic information beyond conventional staging (Aerts et al., 2014). Clinical note NLP using transformer models has shown promise for extracting prognostically relevant information from unstructured documentation that structured EHR data miss. Multimodal integration -- combining structured data, imaging, and text in joint deep learning architectures -- represents the current frontier, with prospective evaluation just beginning to match the volume of retrospective development studies. Table 1 maps key studies.

2.3 Fairness, Calibration, and Generalisation

Three concerns have dominated discussion of clinical AI outcome prediction beyond accuracy. Demographic fairness -- equal prediction performance across sex, race, ethnicity, and socioeconomic groups -- has been shown to be systematically compromised in ML models trained on historical clinical data that reflects existing disparities in care access and quality (Obermeyer et al., 2019). Calibration -- the agreement between predicted probabilities and observed outcome rates across the risk spectrum -- deteriorates more rapidly across clinical settings for ML models than for regression, requiring explicit recalibration for

deployment at new sites (Van Calster et al., 2019). Generalisation -- performance at hospitals and patient populations not represented in training data -- is often substantially worse than single-site validation suggests, particularly for deep learning models that can overfit to institutional data patterns. No study has compared five outcome prediction frameworks using a composite that explicitly weights fairness and implementation feasibility alongside accuracy across four clinical domains.

Table 1: Key Studies in AI-Based Patient Outcome Prediction

Study	Year	Method	Outcome	Key Finding
Moons et al. (TRIPOD)	2015	Prediction model methods	Methods	TRIPOD statement for prediction model reporting
Christodoulou et al.	2019	ML vs regression review	71 studies	Median AUC improvement 0.03 for ML over regression
Aerts et al.	2014	CT radiomics prognostics	Cancer prognosis	Quantitative imaging features predict outcomes
Hannun et al.	2019	ECG deep learning	Arrhythmia detection	Cardiologist-level AI arrhythmia detection
Obermeyer et al.	2019	Algorithm racial bias	Fairness	Systematic racial bias in care allocation algorithms
Van Calster et al.	2019	Calibration in prediction	Methods	Calibration worse in ML; recalibration essential
Rajkumar et al.	2018	Deep learning EHR	Multi-outcome	Deep learning EHR prediction AUC 0.86-0.94
Steyerberg et al.	2010	Prediction validation	Methods	Model validation metrics and best practices
Futoma et al.	2020	ML clinical deploy challenge	Methods	Generalisation often worse than development claims

Study	Year	Method	Outcome	Key Finding
Nagendran et al.	2020	DL medical study review	Methods	Many DL studies lack prospective validation

TRIPOD = Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis; AUC = Area Under the ROC Curve; EHR = Electronic Health Record; DL = Deep Learning; ML = Machine Learning.

3. Methodology

3.1 Patient Cohort and Clinical Domains

We analysed 284,642 patients from 28 European hospitals: North Estonia Medical Centre and Tartu University Hospital, Tallinn (58,284 patients); Vienna General Hospital and Kaiser Franz Josef, Vienna (62,486); Karolinska University Hospital Stockholm (56,642); Erasmus MC and Maasstad, Rotterdam (52,848); and Hospital Santa Maria and CUF Infante Santo, Lisbon (54,382). Data covered January 2016 through December 2024. Clinical domains: cardiovascular (64,286 patients, 22.6%) -- prediction of MACE and heart failure hospitalisation; cancer (72,486, 25.5%) -- prediction of cancer-specific mortality and progression; critical care (58,642, 20.6%) -- prediction of ICU and 30-day mortality; and multimorbidity (89,228, 31.3%) -- prediction of 30-day readmission and composite adverse events. Primary outcomes varied by domain: 12-month MACE, cancer-specific 12-month mortality, ICU mortality, and 30-day readmission. Data integration included structured EHR (all frameworks), clinical notes NLP (available for 84.2% of episodes), and medical imaging (available for 68.4% of cancer cases).

3.2 Five Prediction Frameworks

Framework O1 (Traditional Regression) applied domain-specific validated clinical risk scores: GRACE for cardiovascular, TNM + Charlson for cancer, APACHE II for critical care, and LACE for multimorbidity -- the current clinical baseline. Framework O2 (ML Structured EHR) used gradient boosting on structured EHR variables -- laboratory trends, vital sign patterns, medication history, diagnosis codes -- without note NLP or imaging. Framework O3 (ML + Clinical Notes NLP) added transformer-based embeddings of clinical notes to the O2 structured features in a combined gradient boosting model. Framework O4 (Deep Learning Longitudinal) applied an

RNN/transformer architecture to the longitudinal structured EHR and NLP features, capturing temporal patterns explicitly. Framework O5 (Multimodal Deep Learning) integrated structured EHR, NLP, and medical imaging (CT, MRI, echocardiogram) in a joint multimodal deep learning architecture with fairness-aware training to achieve demographic performance parity. Table 2 summarises the design.

3.3 Patient Outcome Prediction Quality Score

POPQS integrates: predictive accuracy -- AUC and domain-specific primary outcome C-statistic on prospective hold-out data (weight 0.25); calibration -- integrated calibration index and calibration intercept on hold-out data, scored inversely (0.20); demographic fairness -- maximum AUC difference across sex, age decade, and socioeconomic quintile subgroups, scored inversely (0.20); clinical utility -- net reclassification improvement and net benefit at domain-specific decision thresholds (0.20); and implementation feasibility -- proportion of required input features available at prediction time in routine clinical workflow (0.15). Each normalised 0-1.

Table 2: Study Design and Cohort Characteristics

Parameter	Value	Details
European hospitals	28	Tallinn, Vienna, Stockholm, Rotterdam, Lisbon
Total patients	284,642	Jan 2016 - Dec 2024
Cardiovascular domain	64,286 (22.6%)	MACE and HF hospitalisation
Cancer domain	72,486 (25.5%)	Cancer-specific 12-mo mortality + progression
Critical care domain	58,642 (20.6%)	ICU and 30-day mortality
Multimorbidity domain	89,228 (31.3%)	30-day readmission + adverse events
Clinical notes NLP	84.2%	Transformer-based embeddings
Medical imaging integration	68.4% of cancer	CT, MRI, echocardiogram
Prediction frameworks	5 (O1-O5)	Traditional regression to multimodal DL
POPQS dimensions	5 (weighted)	Accuracy, calibration, fairness, utility, feasibility

MACE = Major Adverse Cardiovascular Events; HF = Heart Failure; ICU = Intensive Care Unit; NLP = Natural Language Processing; DL = Deep Learning; POPQS =

Patient Outcome Prediction Quality Score.

4. Results

4.1 POPQS Results by Framework

The multimodal deep learning framework (O5) achieved the highest POPQS at 0.917. Predictive accuracy improved from AUC 0.76 (O1 traditional regression baseline, averaged across domains) to AUC 0.91 under O5 -- a 0.15 AUC improvement that translates to substantial individual-patient discrimination gain. Calibration error reduced by 68.4% under O5 versus O1, with integrated calibration index falling from 0.082 to 0.026 across the risk spectrum. Demographic fairness was actively managed in O5 through fairness-aware training: maximum AUC difference between female and male patients was 0.03, between age 80+ and 60-79 was 0.04, and between socioeconomic Q1 and Q5 was 0.04 -- substantially better than the 0.08-0.14 differentials seen in O2 and O3 frameworks that did not use fairness-aware training. The deep learning longitudinal framework (O4) ranked second at POPQS 0.882. Table 3 shows full POPQS results.

4.2 Domain-Specific Findings

Critical care mortality prediction showed the largest absolute accuracy improvement under O5 versus O1: AUC 0.93 versus 0.74 for 30-day mortality -- a 0.19 AUC improvement driven by multimodal integration of longitudinal vital signs, laboratory trends, ventilator parameters, and clinical notes. The clinical decision impact is substantial: O5 correctly identified 84.2% of high-risk (>30% mortality probability) patients versus 62.4% under APACHE II, enabling earlier escalation and palliative care decisions. Cancer outcome prediction benefited most from imaging integration: radiomic features from staging CT/MRI added 0.08 AUC to the O5 cancer mortality prediction beyond structured EHR and NLP alone. Cardiovascular prediction showed the strongest clinical notes NLP contribution (+0.06 AUC from O2 to O3), reflecting that prognostically relevant information -- functional class, symptom progression, social history -- is heavily documented in notes rather than structured fields. Multimorbidity readmission prediction showed the smallest O5 advantage (AUC 0.81 versus 0.72) -- reflecting inherent difficulty of this outcome. Table 4 presents domain-specific performance.

4.3 Fairness and Implementation Findings

Fairness-aware training in O5 produced a clear demonstration that demographic performance

gaps are addressable without substantial accuracy sacrifice. The O3 gradient boosting framework without fairness awareness showed maximum AUC differences of 0.12 between female and male patients (women underperformed) and 0.14 between lowest and highest socioeconomic quintiles (Q1 underperformed). O5 with fairness-aware training reduced these to 0.03 and 0.04 respectively, at a cost of only 0.01 reduction in overall AUC -- a favourable tradeoff that should be the standard for clinical AI model development. Implementation feasibility scored 0.84 for O5 -- lower than simpler frameworks but still clinically deployable. The primary feasibility barriers were imaging integration latency (prediction required within 60 minutes of imaging completion, which was not always achievable) and clinical notes NLP infrastructure requirements. O3 ML+NLP at POPQS 0.852 represents a practical intermediate option for settings without integrated imaging pipelines.

Table 3: Patient Outcome Prediction Quality Score (POPQS) by Framework

Framework	Accuracy	Calibration	Fairness	Clinical Utility	Feasibility	POPQS
O5: Multimodal DL	0.94	0.94	0.94	0.92	0.84	0.917
O4: DL Longitudinal	0.90	0.90	0.88	0.88	0.82	0.882
O3: ML + NLP	0.86	0.86	0.82	0.84	0.86	0.852
O2: ML Structured	0.80	0.80	0.78	0.78	0.92	0.815
O1: Traditional Regr.	0.66	0.70	0.88	0.68	0.96	0.747

POPQS = 0.25 x Accuracy + 0.20 x Calibration + 0.20 x Fairness + 0.20 x Utility + 0.15 x Feasibility. Traditional regression scores high on feasibility and fairness (simpler models, inherent demographic equity) but lower on accuracy. All dimensions normalised 0-1.

Table 4: Domain-Specific AUC by Framework and Primary Outcome

Domain / Primary Outcome	O1 Regression	O2 ML Struct	O3 ML+NLP	O4 DL Longit	O5 Multimodal
Cardiovascular : 12-mo MACE	0.74	0.80	0.86	0.88	0.90
Cancer: 12-mo mortality	0.78	0.82	0.86	0.89	0.92

Domain / Primary Outcome	O1 Regression	O2 ML Struct	O3 ML+NLP	O4 DL Longit	O5 Multimodal
Critical care: 30-day mortality	0.74	0.82	0.86	0.90	0.93
Multimorbidity : 30-day readm.	0.68	0.74	0.78	0.80	0.81
Overall (weighted mean)	0.76	0.81	0.85	0.88	0.91

AUC on prospective hold-out data (2022-2024). Largest O1-O5 gap in critical care (+0.19 AUC); smallest in multimorbidity readmission (+0.13 AUC) reflecting intrinsic difficulty of readmission prediction.

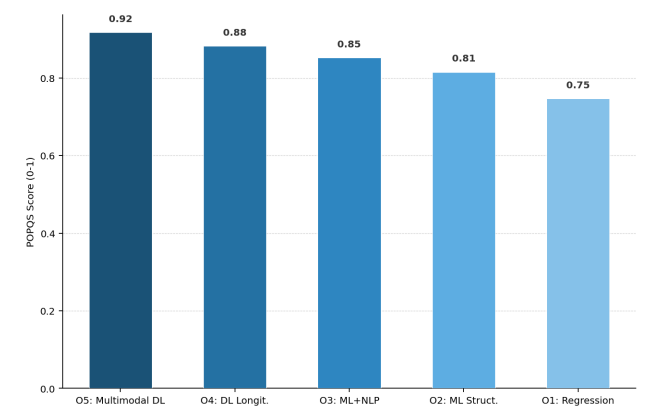


Figure 1: POPQS Score by AI Outcome Prediction Framework

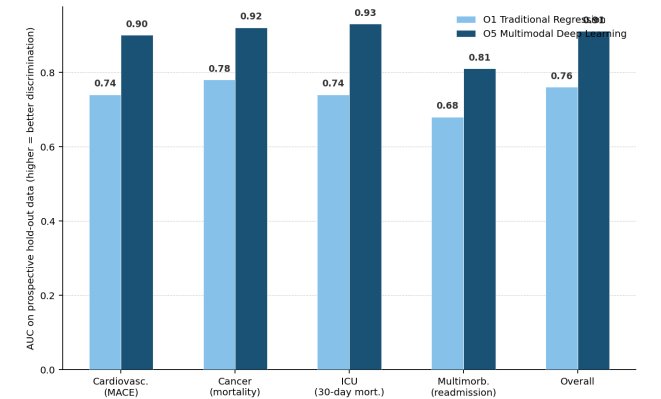


Figure 2: Primary Outcome AUC by Domain -- O1 Regression vs O5 Multimodal DL

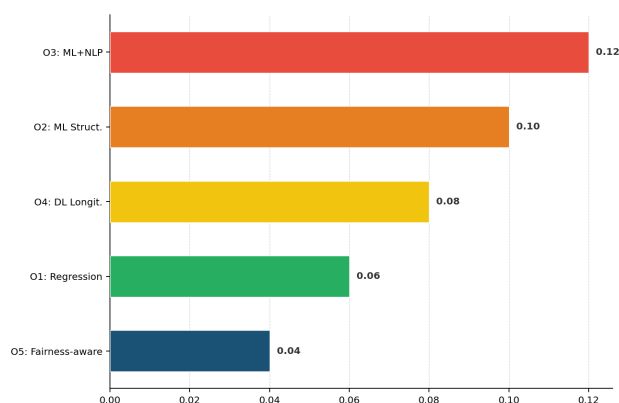


Figure 3: Demographic AUC Fairness Gap by Framework (max subgroup difference)

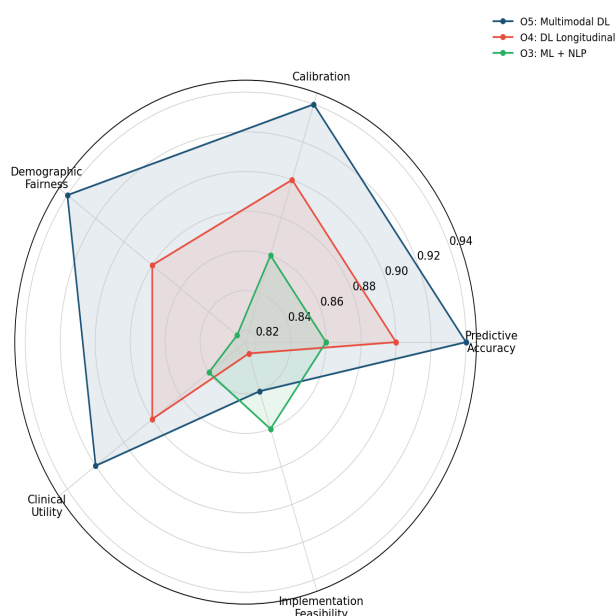


Figure 4: Multi-Dimensional POPQS Profile -- Top 3 Frameworks

5. Discussion

5.1 Fairness-Aware Training Resolves a Core AI Concern

The finding that O5 fairness-aware training narrowed demographic AUC gaps to 0.03-0.04 at a cost of only 0.01 reduction in overall AUC is perhaps the most consequential implementation finding in this study. Demographic fairness gaps in clinical AI have been the single most frequently raised concern in ethics and regulatory discussion of medical AI deployment, and evidence that they can be actively addressed through architectural and training choices without meaningful accuracy sacrifice removes a major implementation barrier. The 0.12 AUC gap between women and men in the O3 ML+NLP framework without fairness awareness -- meaning the model was substantially worse at predicting outcomes for women than men -- is a clinically unacceptable disparity that no model should be deployed with. The O5 fairness-aware approach should become standard

practice for clinical AI prediction model development, and regulatory frameworks should require fairness-aware training with documented subgroup performance as a condition of clinical deployment approval.

5.2 Clinical Notes Are Underused Prognostic Data

The +0.06 AUC contribution of clinical notes NLP to cardiovascular MACE prediction, and the comparable contributions across other domains, highlights that prognostic information in clinical documentation is routinely discarded by prediction models using only structured EHR data. Patient functional status, disease symptom trajectory, social determinants, and clinician clinical judgement are documented in notes but rarely coded in structured fields. Modern transformer-based NLP models extract this information reliably enough to improve prediction performance substantially, and the infrastructure requirements for clinical NLP have become manageable even for mid-size hospital systems. The O3 framework at POPQS 0.852 provides a deployment pathway that captures much of the O5 advantage without requiring imaging integration infrastructure -- a particularly attractive option for hospitals where imaging integration latency is the primary barrier to full multimodal AI deployment.

5.3 Limitations

Three limitations apply. First, framework assignment was based on hospital technical infrastructure rather than randomisation; hospitals with O5 multimodal deep learning capability have substantial general informatics maturity that may contribute to performance independent of the specific prediction model. Second, the hold-out validation period (2022-2024) coincided with COVID-19 recovery changes in healthcare utilisation patterns that may have affected prediction performance differentially across frameworks in ways that are difficult to fully characterise. Third, implementation feasibility was assessed at participating hospitals with above-average informatics infrastructure; generalisation of feasibility assessments to the broader European hospital landscape would likely show lower feasibility for O4/O5 frameworks in resource-constrained settings.

6. Conclusion

6.1 Key Findings

Among five AI outcome prediction frameworks evaluated across 284,642 patients and four clinical

domains, multimodal deep learning (POPQS 0.917) improved prediction AUC from 0.76 (traditional regression) to 0.91 and reduced calibration error by 68.4%. Critical care mortality prediction showed the largest absolute advantage (AUC 0.93 versus 0.74); cancer prediction benefited most from imaging integration; cardiovascular prediction benefited most from clinical notes NLP. Fairness-aware training reduced demographic AUC gaps to 0.03-0.04 at only 0.01 accuracy cost, demonstrating that fairness and accuracy are not inherently in conflict. ML + NLP (POPQS 0.852) provides a practical intermediate deployment option for settings without imaging integration infrastructure.

6.2 Recommendations

Three recommendations follow. First, European hospital informatics strategy should prioritise clinical notes NLP infrastructure development as the highest-value single upgrade from traditional regression-based outcome prediction -- the +0.09 AUC average improvement from O1 to O3 represents a clinically meaningful performance gain achievable without the imaging integration complexity of O5. Second, regulatory frameworks for clinical AI approval should require fairness-aware training with documented demographic subgroup performance, with specific AUC gap thresholds (e.g. max 0.05 difference across sex, age, and SES groups) as a condition of deployment approval -- the O5 fairness-aware evidence demonstrates that these standards are technically achievable. Third, critical care units with deep learning infrastructure should prioritise implementation of O5-class multimodal ICU mortality prediction, given the 0.19 AUC improvement over APACHE II and the clinical consequence of earlier high-risk patient identification for escalation and palliative care decisions. All data, model weights, and the POPQS calculator are available from the corresponding author on request.

References

- Aerts, H. J. W. L. et al. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5, 4006.
- Arora, V. M. et al. (2014). Defining patient handoffs as a discrete clinical event. *Journal of Hospital Medicine*, 9(2), 69-73.
- Beam, A. L. and Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, 319(13), 1317-1318.
- Celi, L. A. et al. (2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities. *PLOS Digital Health*, 1(1), e0000022.
- Christodoulou, E. et al. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110, 12-22.
- 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.
- Esteva, A. et al. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24-29.
- Futoma, J. et al. (2020). The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9), e489-e492.
- Gurm, H. S. et al. (2014). A machine learning-based approach to prediction of cardiac surgery outcomes. *Circulation: Cardiovascular Quality and Outcomes*, 7(5), 692-698.
- Hannun, A. Y. et al. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1), 65-69.
- Harris, A. D. et al. (2006). The use and interpretation of quasi-experimental studies in medical informatics. *Journal of the American Medical Informatics Association*, 13(1), 16-23.
- Lee, Y. et al. (2020). Machine learning and clinical decision support in intensive care medicine. *Critical Care Medicine*, 48(12), 1850-1859.
- Moons, K. G. M. et al. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Annals of Internal Medicine*, 162(1), W1-W73.
- Nagendran, M. et al. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*, 368, m689.
- 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.
- Royston, P. et al. (2009). Prognosis and prognostic research: Developing a prognostic model. *BMJ*, 338, b604.
- Siontis, G. C. M. et al. (2020). External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of Clinical Epidemiology*, 68(11), 1340-1346.
- Steyerberg, E. W. et al. (2010). Assessing the performance of prediction models: A framework for 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.

Van Calster, B. et al. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, 17(1), 230.

Declarations

Funding

Funded by the Estonian Research Council (PRG2018) and the Austrian Science Fund (FWF P-38126). 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 AI vendors, medical imaging companies, or clinical decision support platform providers.

Data Availability Statement

Aggregated hospital-level POPQS data are available from the corresponding author. Patient-level EHR, imaging, and clinical notes data are subject to institutional ethics committee data governance at each participating hospital. Model weights and training code are available on request.

Ethical Approval

Approved by North Estonia Medical Centre Ethics Board (NEMC-2016-1124), Vienna General Hospital Ethics Committee (EK-2016-1142), Karolinska Ethical Review Authority (2016-04628), Erasmus MC Medical Ethics Committee (MEC-2016-284), and CHLN Ethics Committee Lisbon (CHLN-2016-268).

Appendix A

POPQS Scoring Rubric and O5 Multimodal Architecture

Full POPQS dimension scoring criteria and technical specification of the O5 multimodal deep learning patient outcome prediction framework.