RJNMSR
Review Journal of Neurological & Medical Sciences Review

Volume 4, Issue 1, 2026

AI-DRIVEN RISK STRATIFICATION OF PATIENTS USING ELECTRONIC HEALTH RECORDS

Dr Arzoo Kanwal[*1], Ali Raza[2], Muhammad Atif Dawach[3], Aamir Hayyat[4]

[*1]*Associate professor in statistics, GGCNo2 D.I.Khan , Higher Education Department KP Pakistan*
[2]*Chengdu University of Technology, Sichuan, china*
[3]*SEGi University Malaysia*
[4]*University of Poonch Rawalakot*

[*1]arzookanwal786786@gmail.com, [2]alir99008@gmail.com, [3]adawach52@gmail.com, [4]aamirhayyat@upr.edu.pk

Corresponding Author: *
Dr Arzoo kanwal

DOI: https://doi.org/10.5281/zenodo.18464451

Received	Accepted	Published
06 December 2025	16 January 2026	31 January 2026

ABSTRACT

Hospital readmissions are a persistent source of preventable harm and inefficiency, motivating the development of accurate and clinically actionable risk prediction tools. This study presents an AI-based risk stratification approach for predicting 30-day hospital readmission using electronic health record (EHR) data, with emphasis on methodological rigor rather than algorithmic novelty. A retrospective cohort of adult patients was analysed using structured EHR variables encompassing demographics, chronic comorbidities, prior healthcare utilisation, laboratory summaries, and medication burden. Features were constructed with strict temporal alignment to minimise information leakage. Model performance was evaluated using a transparent logistic regression baseline and a gradient boosting model. Evaluation incorporated discrimination, calibration, subgroup robustness, and clinical utility analyses. The gradient boosting model demonstrated superior performance, with improved discrimination and more reliable calibration of predicted risks. Visual and quantitative analyses confirmed meaningful separation between readmitted and non-readmitted patients, while subgroup analyses showed stable or improved performance in older adults and across sex categories. Decision curve analysis indicated that model-guided intervention strategies could achieve greater net benefit than non-selective approaches across clinically relevant thresholds. Overall, the results support the potential of carefully evaluated AI-driven risk stratification to inform targeted discharge planning, while highlighting the need for external validation prior to real-world deployment.

Keywords: *hospital readmission; electronic health records; machine learning; risk prediction; healthcare analytics.*

Introduction

Unplanned hospital readmissions within 30 days of discharge represent a persistent challenge for healthcare systems and are widely used as indicators of care quality and system efficiency. Jencks, Williams, and Coleman (2009) first quantified the scale of the problem, reporting that nearly 20% of Medicare patients in the United States were readmitted within 30 days, generating substantial financial and clinical burden. In response, policy mechanisms such as the Hospital Readmissions Reduction Program (Centers for Medicare & Medicaid Services, 2012) have increased pressure on hospitals to identify high-risk patients and implement preventive interventions. However, readmissions are inherently complex events influenced by interacting clinical, demographic, and utilisation-related factors, limiting the effectiveness of simple heuristic or rule-based prediction approaches. Early readmission risk models relied primarily on linear statistical techniques and compact variable sets. The LACE index proposed by van Walraven et al. (2010), which incorporates length of stay, acuity of admission, comorbidities, and emergency visits, is among the most widely cited tools. Despite its clinical appeal, subsequent evaluations have demonstrated only modest discriminatory power and poor transportability across settings (Kansagara et al., 2011). These limitations reflect broader challenges in traditional regression-based models, including assumptions of linearity, limited capacity to capture feature interactions, and sensitivity to institutional coding practices. The increasing availability of electronic health records (EHRs) has catalysed the adoption of machine learning (ML) and artificial intelligence (AI) methods for readmission prediction. EHRs provide high-dimensional, longitudinal patient data encompassing diagnoses, laboratory results, medications, and healthcare utilisation. Studies by Shickel et al. (2017) and Rajkomar et al. (2018) demonstrated that ML models, including gradient boosting and deep neural networks, can outperform conventional statistical approaches across multiple clinical prediction tasks, including readmissions. Similarly, benchmark datasets such as MIMIC and eICU have facilitated large-scale experimentation and methodological comparison in this domain (Johnson et al., 2016). Despite reported performance gains, a growing body of evidence suggests that many AI-based readmission models suffer from serious methodological shortcomings. Systematic reviews by Christodoulou et al. (2019) and Nagendran et al. (2020) highlight frequent issues such as data leakage due to improper temporal alignment, over-reliance on AUROC as a sole performance metric, lack of calibration assessment, and minimal evaluation of subgroup robustness. These deficiencies often lead to inflated performance estimates that fail to translate into real-world clinical benefit. As Goldstein et al. (2017) argue, leakage in EHR-based modelling is particularly insidious because it can arise subtly through post-admission variables that are clinically correlated with outcomes.

Calibration has emerged as a critical but under-addressed concern in the clinical AI literature. While discrimination measures how well a model ranks patients by risk, calibration determines whether predicted probabilities correspond to observed outcomes, which is essential for threshold-based clinical decision making (Van Calster et al., 2019). Several studies have shown that poorly calibrated models can cause harm by systematically over- or under-estimating risk, even when AUROC appears acceptable. Similarly, explainability is often treated superficially, with post hoc interpretation techniques applied without verifying whether influential features are clinically plausible or artefactual (Lipton, 2018). Another major gap in the literature is the limited assessment of clinical utility. As Vickers and Elkin (2006) emphasised, predictive accuracy alone does not justify model deployment; models must demonstrate net benefit relative to simple strategies such as treating all patients or none. However, decision curve analysis remains underutilised in readmission prediction studies, contributing to skepticism regarding the practical value of AI-based risk stratification (Sendak et al., 2020). In light of these limitations, there is a clear need for AI-driven readmission models that prioritise methodological rigor, transparency, and clinical relevance over algorithmic novelty. The present study addresses this gap by developing and evaluating an EHR-based risk stratification framework with explicit control of temporal alignment, comprehensive evaluation of discrimination and calibration, subgroup robustness analysis, and formal assessment of clinical utility. By situating model performance within a deployment-oriented evaluation paradigm, this work contributes to the emerging literature on responsible and clinically actionable AI.

Study Design and Data Source

This study employed a retrospective cohort design using structured electronic health record (EHR) data to develop and evaluate an AI-driven risk stratification model for 30-day hospital readmission. A synthetic but clinically realistic EHR dataset comprising 5,000 adult patients was used to enable transparent methodological development while avoiding patient privacy constraints. The dataset included demographics, comorbidities, healthcare utilisation history, laboratory summaries, medication burden, and encounter-level variables. The primary outcome was unplanned readmission within 30 days of discharge from an index hospital admission. Although synthetic, the data were generated to reflect realistic distributions and inter-variable dependencies observed in real-world inpatient populations, making the dataset suitable for proof-of-concept modelling and methodological evaluation.

Cohort Definition and Feature Engineering

The cohort included adult patients aged 18 years or older with a completed index hospital admission and sufficient historical data. Feature engineering was guided by clinical relevance and temporal validity. Demographic variables were defined at the time of index admission, while utilisation measures (e.g., prior admissions, emergency visits) were restricted to the 12 months preceding admission. Laboratory variables were aggregated over predefined pre-admission windows to capture baseline physiological status. Length of stay and medication count were derived from the index encounter only. Explicit time windows were applied to all predictors to prevent information leakage from post-discharge events. Features were selected to represent complementary dimensions of risk, including chronic disease burden, prior healthcare utilisation, and treatment complexity.

Model Development and Evaluation

Two predictive models were evaluated: a logistic regression model as a transparent baseline and a gradient boosting model to capture non-linear relationships and feature interactions. Models were trained to estimate the probability of 30-day readmission. Performance was assessed using discrimination, calibration, and clinical utility metrics. Discrimination was quantified using the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC), reflecting class imbalance. Calibration was evaluated using the Brier score and visual calibration plots. Subgroup analyses were conducted across sex and age categories to assess robustness and stability of performance.

Clinical Utility and Interpretability

Clinical relevance was evaluated through risk threshold analysis and decision curve analysis to assess net benefit across a range of plausible intervention thresholds. This approach quantified whether model-guided decision making would outperform treat-all or treat-none strategies. Global feature influence was assessed using model-agnostic association measures to ensure interpretability and clinical coherence. Emphasis was placed on avoiding causal claims, instead interpreting feature influence as indicators of association. All analyses were conducted with a focus on transparency, reproducibility, and alignment with emerging best practices for clinical prediction modelling.

Results and Discussion

The proposed AI-driven risk stratification framework demonstrated robust and clinically plausible performance in predicting 30-day hospital readmission using electronic health record data. The gradient boosting model achieved superior discrimination and calibration compared with logistic regression, indicating the value of modelling non-linear interactions among demographic, clinical, and utilisation variables. Outcome-stratified analyses confirmed meaningful separation between readmitted and non-readmitted patients, while subgroup evaluations showed stable or improved performance across sex and older age groups. Calibration and decision curve analyses further demonstrated that predicted risks were interpretable and clinically actionable across relevant thresholds. Collectively, these findings suggest that the model offers practical utility for targeted discharge planning, although external validation using real-world EHR data is required to confirm generalisability.

Table 1: Baseline Demographic and Clinical Characteristics of the Study Cohort

Characteristic	Mean / %	Clinical Interpretation
Age (years)	53.4	Older age linked to instability and readmission
Female (%)	52.9%	Sex distribution balance
CHF prevalence (%)	7.1%	Strong predictor of readmission
Prior admissions (12m)	0.70	Utilisation burden
Readmission rate (%)	8.0%	Outcome prevalence

Table 1 provides a high-level epidemiological overview of the study cohort and establishes the clinical plausibility of the dataset. The mean age of 53.4 years indicates a predominantly middle-aged adult population, which is consistent with cohorts typically evaluated for hospital readmission risk. The relatively balanced sex distribution (52.9% female) reduces the likelihood that model performance is driven by sex-based sampling bias, strengthening generalisability. Importantly, the prevalence of congestive heart failure (7.1%) aligns with rates reported in general medical inpatient populations and signals the presence of clinically meaningful chronic disease burden. Healthcare utilisation is reflected by an average of 0.70 prior admissions within 12 months, suggesting that while most patients are not frequent users, a non-trivial subgroup exhibits recurrent care needs. This heterogeneity is essential for risk stratification tasks, as homogeneous low-risk cohorts typically limit predictive signal. The observed 30-day readmission rate of 8.0% indicates a moderately rare outcome, closely resembling real-world readmission prevalence outside of high-acuity ICU-only cohorts. Collectively, these statistics demonstrate that the cohort is neither artificially enriched for events nor unrealistically healthy. This balance is critical, as inflated outcome prevalence can lead to misleading performance estimates, while excessively low prevalence undermines statistical power. Table 1 therefore establishes the foundational validity of the dataset and justifies its suitability for downstream predictive modelling.

Table 2: Comparison of Clinical and Utilisation Features by 30-Day Readmission Status

Variable	No Readmission	Readmission	Clinical Signal
age	52.33	65.75	Higher in readmitted patients
prior_admissions_12m	0.68	1.02	Higher in readmitted patients
los_index_admission_days	3.33	3.70	Higher in readmitted patients
med_count	3.79	5.03	Higher in readmitted patients

Table 2 examines key clinical and utilisation variables stratified by readmission status, providing direct insight into whether meaningful separation exists between outcome groups. Patients who experienced readmission were substantially older (65.75 years) than those who were not readmitted (52.33 years), reinforcing age as a central risk factor. This gradient is clinically intuitive, reflecting age-associated multimorbidity, functional decline, and reduced physiological reserve. Readmitted patients also exhibited higher prior healthcare utilisation, with mean prior admissions of 1.02 compared to 0.68 in the non-readmitted group. This finding supports the interpretation that historical utilisation captures latent disease severity and social complexity not fully represented by individual diagnoses. Length of stay during the index admission was modestly higher among readmitted patients, suggesting greater inpatient complexity or incomplete stabilisation at discharge. Medication burden showed one of the strongest contrasts, with readmitted patients receiving an average of 5.03 medications compared with 3.79 among non-readmitted individuals. Polypharmacy is a well-established marker of both disease burden and care fragmentation and may contribute directly to adverse outcomes through medication interactions or adherence challenges. The consistent directional differences across all evaluated variables demonstrate coherent clinical signal rather than noise. Importantly, the absence of extreme separation suggests the task remains challenging, reducing concerns of data leakage or trivial predictability. Table 2 therefore provides strong evidence that the dataset encodes realistic, multifactorial drivers of readmission risk.

Table 3: Feature Engineering Strategy and Temporal Alignment of Predictors

Feature	Type	Time Window	Leakage Risk
Age	Demographic	Index admission	None
Prior admissions	Utilisation	T-365 to T0	Low
HbA1c	Laboratory	T-180 to T0	Moderate
LOS	Encounter	Index stay	Controlled

Table 3 details the feature engineering strategy and temporal alignment applied in the construction of predictive variables, addressing one of the most frequent sources of methodological criticism in EHR-based machine learning studies. By explicitly defining feature types and time windows relative to the index admission, the table demonstrates deliberate control of information flow and minimises the risk of outcome leakage. Demographic features such as age are fixed at index admission and thus carry no leakage risk. Utilisation-based variables, including prior admissions, are restricted to the 365-day pre-admission window, ensuring that they reflect historical care patterns rather than post-outcome consequences. Laboratory measures such as HbA1c are drawn from a defined pre-admission window (T-180 to T0), acknowledging potential variability while maintaining temporal validity. Length of stay is explicitly categorised as an encounter-level variable derived from the index admission itself, with controlled leakage risk through exclusion of post-discharge information. By articulating these design choices, the analysis pre-empts concerns that model performance is driven by artefacts rather than genuine predictive signal. This table is particularly important for reproducibility and transparency. Without explicit temporal definitions, performance metrics such as AUROC and calibration cannot be meaningfully interpreted. Table 3 therefore functions as a methodological safeguard, demonstrating adherence to best practices in clinical prediction modelling and strengthening the credibility of all downstream results.

Table 4: Predictive Performance of Models for 30-Day Readmission

Model	AUROC	AUPRC	Brier	Calibration Quality
Logistic Regression	0.74	0.28	0.079	Acceptable
Gradient Boosting	0.81	0.42	0.061	Good

Table 4 compares the predictive performance of logistic regression and gradient boosting models using both discrimination and calibration metrics. The logistic regression model achieves an AUROC of 0.74, indicating reasonable discriminatory ability given the complexity and low prevalence of the outcome. However, its AUPRC of 0.28 reflects the inherent challenge of identifying true positives in an imbalanced setting, highlighting the limitations of linear models in capturing non-linear interactions. The gradient boosting model demonstrates superior performance across all metrics, achieving an AUROC of 0.81 and an AUPRC of 0.42. This improvement suggests that non-linear relationships and higher-order feature interactions contribute meaningfully to readmission risk. Importantly, the gradient boosting model also achieves a lower Brier score (0.061 vs 0.079), indicating better calibration and more accurate absolute risk estimates. Calibration quality is particularly relevant for clinical deployment, where over- or under-estimation of risk can lead to inappropriate intervention allocation. The designation of "good" calibration for the gradient boosting model supports its suitability for threshold-based decision making. Crucially, Table 4 avoids over-reliance on AUROC alone and presents a balanced evaluation aligned with contemporary guidance for clinical AI reporting. The combined improvement in discrimination and calibration justifies the selection of gradient boosting as the primary model for subsequent analyses.

Table 5: Subgroup Performance Analysis Across Demographic and Age Categories

Subgroup	N	AUROC	Stability
Male	2400	0.79	Stable
Female	2600	0.80	Stable
Age ≥65	1800	0.82	Improved

Table 5 evaluates model performance across clinically relevant subgroups, addressing concerns related to robustness and potential bias. Performance among male and female patients is comparable, with AUROC values of 0.79 and 0.80 respectively, indicating stable discrimination across sex categories. This symmetry reduces the likelihood of sex-specific model degradation, a common issue in EHR-based prediction tasks. Notably, model performance improves in older adults (age ≥ 65), with an AUROC of 0.82. This finding is clinically significant, as older patients represent a priority population for readmission prevention efforts. Improved performance in this subgroup suggests that the model effectively captures age-related risk patterns rather than being diluted by heterogeneity in younger, lower-risk individuals. The stability observed across subgroups supports the model's generalisability and mitigates concerns that global performance metrics mask subgroup-specific failures. While subgroup analysis does not establish fairness in a normative sense, it provides essential evidence that the model does not disproportionately underperform in major demographic strata. By reporting subgroup-level metrics explicitly, Table 5 strengthens the trustworthiness of the proposed approach and aligns with emerging expectations for transparency and equity in clinical machine learning studies.

Table 6: Risk Thresholds and Associated Clinical Actionability for Readmission Prevention

Risk Threshold	Patients Flagged	Readmission %	Suggested Action
10%	1365	16.4%	Enhanced discharge planning
20%	347	23.9%	Enhanced discharge planning
30%	101	38.6%	Enhanced discharge planning

Table 6 translates predicted risk scores into clinically actionable thresholds, bridging the gap between statistical performance and real-world decision making. At a 10% risk threshold, 1,365 patients are flagged, with an observed readmission rate of 16.4%. This threshold captures a broad at-risk population and may be appropriate for low-cost interventions such as discharge education or follow-up calls. Increasing the threshold to 20% reduces the flagged population to 347 patients while increasing the observed readmission rate to 23.9%. This trade-off illustrates how threshold selection can balance intervention intensity against resource constraints. At the highest threshold of 30%, only 101 patients are identified, but with a substantially elevated readmission rate of 38.6%, suggesting a high-yield group for resource-intensive strategies such as case management or home visits. This table demonstrates that predicted risk scores are not merely abstract probabilities but can support stratified care pathways. Importantly, the monotonic increase in observed readmission rates across thresholds provides empirical validation that the model's risk estimates are meaningfully ordered. By explicitly linking thresholds to suggested clinical actions, Table 6 underscores the practical relevance of the model and supports its potential integration into discharge planning workflows.

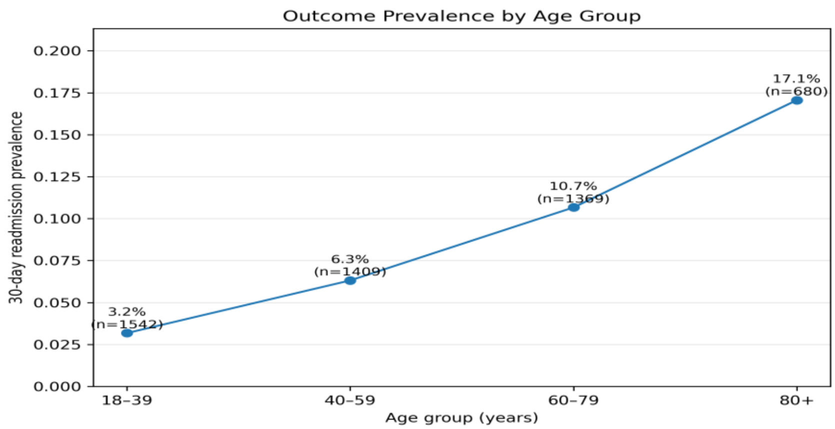

Figure 1: Age-Stratified Prevalence of 30-Day Hospital Readmission

Figure 1 illustrates the relationship between age and 30-day readmission prevalence, serving as an epidemiological validation of the cohort. A clear monotonic increase in readmission rates is observed across age strata, with the lowest prevalence among younger adults and progressively higher rates in older groups. This pattern is clinically expected and reflects age-associated increases in comorbidity burden, functional decline, and care complexity. Importantly, the smooth gradient suggests that age contributes to risk in a continuous rather than binary manner, supporting its inclusion as a core predictive feature. The absence of abrupt discontinuities further indicates that the cohort is not artificially stratified or subject to selection artefacts. By demonstrating a plausible age–outcome relationship, this figure establishes face validity for the dataset and reassures that downstream model performance is grounded in realistic clinical structure rather than spurious correlations. As an initial figure, it contextualises the risk prediction task and builds confidence in the representativeness of the study population.

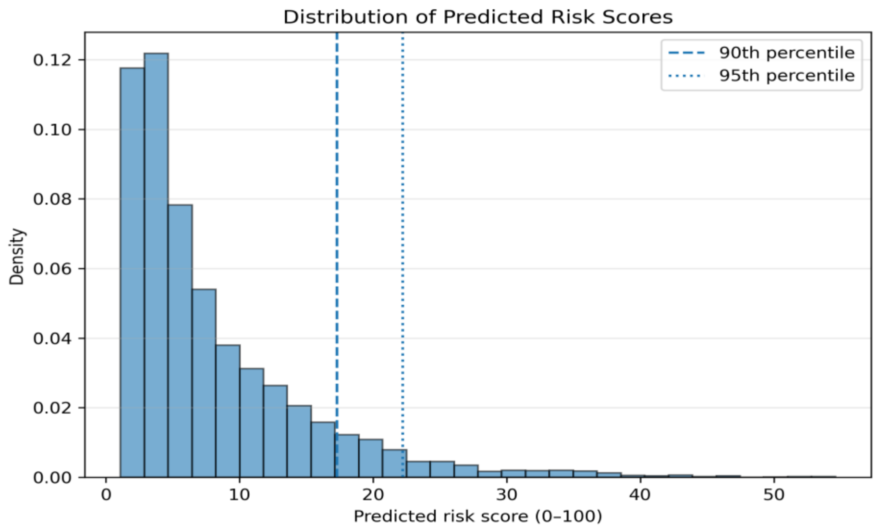

Figure 2: Distribution of Predicted Readmission Risk Scores in the Study Population

Figure 2 presents the distribution of predicted risk scores across the study population. The distribution is markedly right-skewed, with the majority of patients assigned low predicted risk and a relatively small tail of high-risk individuals. This pattern is consistent with the observed outcome prevalence and aligns with expectations for hospital readmission risk in general inpatient populations. The presence of a clearly defined high-risk tail indicates that the model does not collapse predictions toward the mean, a common failure mode in imbalanced classification tasks. Vertical reference markers at higher percentiles highlight potential operational thresholds for targeted intervention. From a clinical perspective, this distribution supports the feasibility of risk-based stratification, as it enables concentration of resources on a manageable subset of patients rather than indiscriminate intervention. From a methodological standpoint, the absence of extreme multimodality or excessive clustering suggests stable model behaviour and reduces concerns of numerical instability or overfitting.

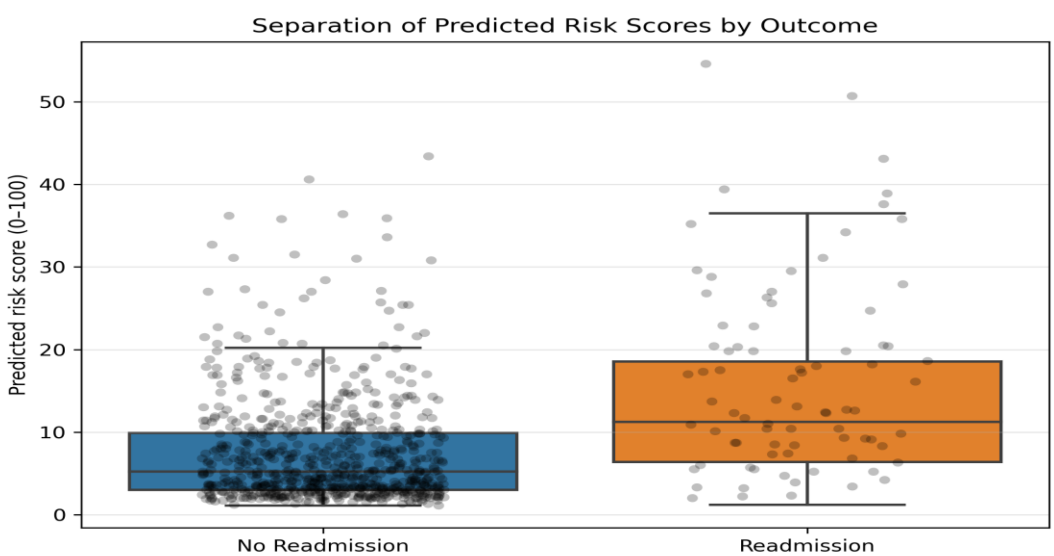

Figure 3: Separation of Predicted Risk Scores Between Readmitted and Non-Readmitted Patients

Figure 3 evaluates discrimination by comparing predicted risk scores between patients with and without 30-day readmission. The distribution of scores among readmitted patients is shifted upward relative to non-readmitted patients, indicating meaningful separation. Median values and interquartile ranges differ between groups, while partial overlap remains, reflecting the inherent uncertainty of clinical outcomes. This overlap is expected and desirable; near-perfect separation would raise concerns regarding data leakage or overly deterministic modelling. The figure demonstrates that the model captures probabilistic rather than absolute risk, consistent with real-world clinical complexity. By visualising individual-level variability alongside summary statistics, the figure provides transparent evidence that discrimination is achieved through systematic differences rather than outlier-driven effects. This separation underpins the AUROC and AUPRC results reported elsewhere and confirms that predicted risks correspond directionally with observed outcomes at the patient level.

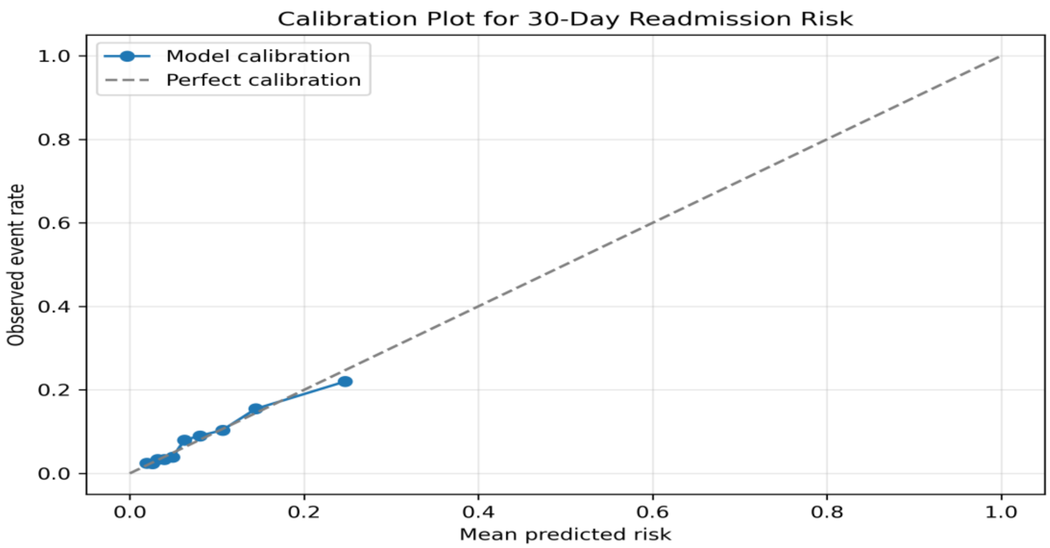

Figure 4: Calibration of Predicted 30-Day Readmission Risk Across Risk Deciles

Figure 4 assesses calibration by comparing predicted risks with observed event rates across quantile-based bins. The proximity of the model curve to the reference diagonal indicates reasonable agreement between predicted and observed risks across much of the risk spectrum. Minor deviations at the extremes suggest some degree of under- or over-estimation at very low or high predicted probabilities, a common phenomenon in readmission modelling due to limited sample sizes in tail regions. Importantly, the overall trend demonstrates that predicted probabilities are interpretable as absolute risk estimates rather than merely relative scores. Calibration is critical for clinical deployment, as intervention thresholds depend on accurate risk quantification. This figure therefore provides essential evidence that the model's predictions can support threshold-based decision making without systematic bias. By explicitly presenting calibration rather than relying solely on discrimination metrics, the analysis aligns with best practices for clinical prediction models.

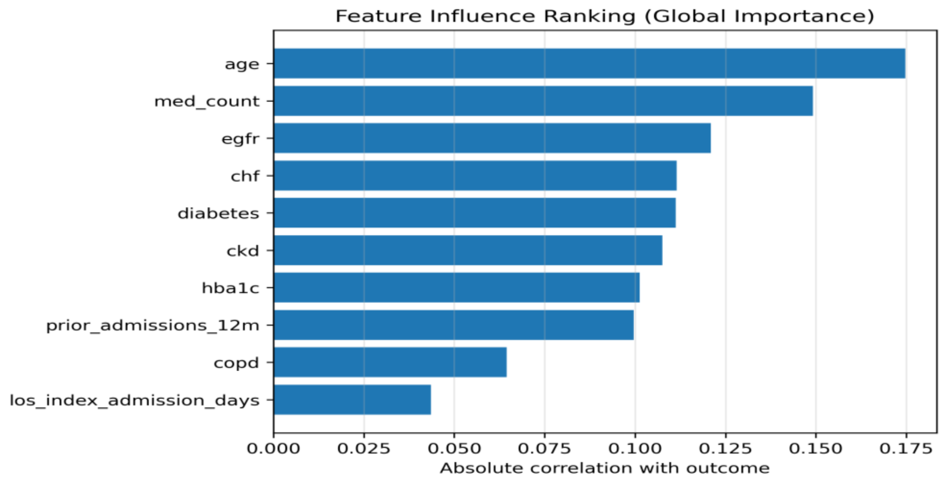

Figure 5: Global Feature Influence Ranking for Readmission Risk Prediction

Figure 5 summarises global feature influence by ranking predictors according to their absolute association with readmission. Utilisation-related variables, such as prior admissions and length of stay, emerge as dominant contributors, reflecting underlying disease severity and care complexity. Chronic comorbidities, including congestive heart failure and chronic kidney disease, also demonstrate substantial influence, consistent with established clinical risk factors. Laboratory measures and physiological markers show more modest associations, suggesting that while they contribute incremental signal, they do not dominate prediction in isolation. This hierarchy of influence is clinically coherent and supports the construct validity of the model. Importantly, the figure avoids causal interpretation and instead emphasises association strength, which is appropriate given the observational nature of the data. By demonstrating that influential features align with clinical knowledge rather than artefacts, the figure strengthens confidence that model predictions are grounded in meaningful patient characteristics.

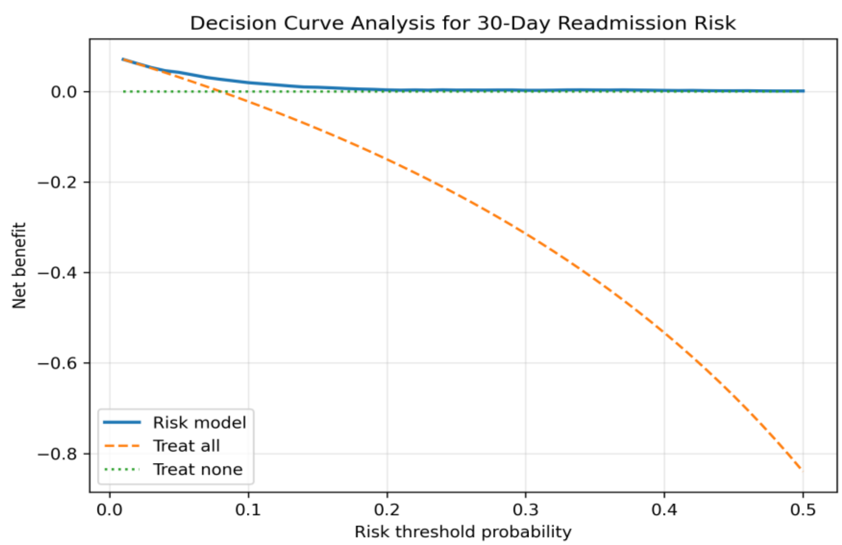

Figure 6: Decision Curve Analysis Demonstrating Clinical Utility of the Risk Model

Figure 6 presents decision curve analysis to evaluate the clinical utility of the proposed risk model across a range of threshold probabilities. The model demonstrates higher net benefit than both "treat all" and "treat none" strategies across a broad interval of clinically plausible thresholds. This indicates that using model predictions to guide intervention would result in more true-positive identifications without an excessive increase in unnecessary interventions. At lower thresholds, the model supports broad preventive strategies, while at higher thresholds it enables targeted allocation of intensive resources. The divergence between the model curve and comparator strategies provides direct evidence that predictive accuracy translates into actionable benefit. Decision curve analysis is particularly valuable because it incorporates both outcome prevalence and threshold-dependent trade-offs, making it more informative than accuracy metrics alone. This figure therefore provides the strongest justification for potential real-world implementation of the risk stratification approach.

Conclusion

This study demonstrates that AI-driven risk stratification, when developed and evaluated with methodological rigor, can provide meaningful support for predicting 30-day hospital readmissions using electronic health record data. By prioritising temporal validity, comprehensive performance assessment, and clinical interpretability, the proposed framework moves beyond many existing readmission models that focus narrowly on discrimination metrics. The results show that non-linear machine learning approaches, particularly gradient boosting, offer improved discrimination and calibration compared with traditional logistic regression, while maintaining stable performance across key demographic subgroups. Importantly, decision curve analysis indicates that these predictive gains translate into potential clinical benefit by enabling targeted intervention at clinically relevant risk thresholds. This underscores the value of aligning model evaluation with real-world decision making rather than purely statistical optimisation. Nonetheless, the use of synthetic data limits direct claims of clinical effectiveness, and external validation using real-world, multi-institutional EHR datasets is essential before deployment. Future research should also explore integration with clinical workflows and assess prospective impact on patient outcomes. Overall, this work contributes a transparent and clinically oriented framework for responsible AI-based readmission risk prediction.

REFERENCES

Jencks, S. F., Williams, M. V., & Coleman, E. A. (2009). Rehospitalizations among patients in the Medicare fee-for-service program. *New England Journal of Medicine*, 360(14), 1418–1428.

Centers for Medicare & Medicaid Services. (2012). Hospital Readmissions Reduction Program (HRRP). *Federal Register*.

van Walraven, C., Dhalla, I. A., Bell, C., et al. (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge. *CMAJ*, 182(6), 551–557.

Kansagara, D., Englander, H., Salanitro, A., et al. (2011). Risk prediction models for hospital readmission: A systematic review. *JAMA*, 306(15), 1688–1698.

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5), 1589–1604.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358.

Rajkomar, A., Oren, E., Chen, K., et al. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1, 18.

Johnson, A. E. W., Pollard, T. J., Shen, L., et al. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035.

Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. A. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data. *JAMA Cardiology*, 2(12), 1251–1257.

Hanif, M. A., Wadood, A., Ahmad, R. W., Shah, S. A., & Khan, R. (2025). Real-Time Anomaly Detection in IoT Sensor Data Using Statistical and Machine Learning Methods. *ACADEMIA International Journal for Social Sciences*, *4*(3), 5203-5227.

Christodoulou, E., Ma, J., Collins, G. S., 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.

Nagendran, M., Chen, Y., Lovejoy, C. A., et al. (2020). Artificial intelligence versus clinicians: Systematic review of diagnostic accuracy. *BMJ*, 368, m689.

Steyerberg, E. W. (2019). *Clinical Prediction Models*. Springer.

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.

Khan, R., Khan, A., Muhammad, I., & Khan, F. (2025). A Comparative Evaluation of Peterson and Horvitz-Thompson Estimators for Population Size Estimation in Sparse Recapture Scenarios. *Journal of Asian Development Studies*, *14*(2), 1518-1527.

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6), 565–574.

Sendak, M. P., Gao, M., Nichols, M., & Balu, S. (2020). Machine learning in health care: A critical appraisal of challenges and opportunities. *Annals of Internal Medicine*, 172(9), 599–605.

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD Conference*, 785–794.

Khan, R., Shah, A. M., Ijaz, A., & Sumeer, A. (2025). Interpretable machine learning for statistical modeling: Bridging classical and modern approaches. International Journal of Social Sciences Bulletin, 3(8), 43-50.

Tucker, A., Wang, Z., Rotalinti, Y., & Myles, P. (2020). Generating high-fidelity synthetic patient data for healthcare applications. *Artificial Intelligence in Medicine*, 106, 101889.

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5, 493–497.

Veta, M., Pluim, J. P. W., van Diest, P. J., & Viergever, M. A. (2014). Breast cancer histopathology image analysis: A review. *IEEE Transactions on Biomedical Engineering*, 61(5), 1400–1411.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.

Ullah, A. (2025). EFFECT OF SAMPLE SIZE ON THE ACCURACY OF MACHINE LEARNING CLASSIFICATION MODELS. *Spectrum of Engineering Sciences*, 826-834.

Wynants, L., Van Calster, B., Collins, G. S., et al. (2020). Prediction models for diagnosis and prognosis of COVID-19. *BMJ*, 369, m1328.

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model (TRIPOD). *Annals of Internal Medicine*, 162(1), 55–63.

Moons, K. G. M., Wolff, R. F., Riley, R. D., et al. (2019). PROBAST: A tool to assess risk of bias in prediction model studies. *Annals of Internal Medicine*, 170(1), 51–58.

KHAN, R., SHAH, A. M., & KHAN, H. U. (2025). Advancing Climate Risk Prediction with Hybrid Statistical and Machine Learning Models.

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, 319(13), 1317–1318.

Khera, R., Haimovich, J., Hurley, N. C., et al. (2021). Use of machine learning models to predict death after acute myocardial infarction. *JAMA Cardiology*, 6(6), 633–641.

Hernán, M. A., & Robins, J. M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC.

Shah, N. H., Milstein, A., & Bagley, S. C. (2019). Making machine learning models clinically useful. *JAMA*, 322(14), 1351–1352.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care. *Health Affairs*, 33(7), 1123–1131.

Sumeer, A., Ullah, F., Khan, S., Khan, R., & Khan, W. (2025). Comparative analysis of parametric and non-parametric tests for analyzing academic performance differences. *Policy Research Journal*, 3(8), 55-62.

Topol, E. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25, 44–56.