A Capacity-Aware Framework for Hospital Readmission Risk Prediction and Decision Support

Hajer Sayyar AlHosani
Business Analytics Program
Abu Dhabi School of Management
Abu Dhabi, United Arab Emirates
adsm-215601@adsm.ac.ae

Ishtiaq Rasool Khan
Business Analytics Program
Abu Dhabi School of Management
Abu Dhabi, United Arab Emirates
i.khan@adsm.ac.ae
ORCID: 0000-0002-3887-9052

https://doi.org/10.63962/NAOM7888

Abstract— Hospital readmissions within 30 days remain a major operational and cost challenge in healthcare systems. While many studies focus on improving predictive accuracy using machine learning, limited attention is given to how predictions can support real-world decision-making under operational constraints. This study proposes a capacity-aware enterprise framework that integrates predictive analytics with workload-based prioritization and governance-oriented evaluation. The framework aligns risk prediction with realistic intervention capacity using Top-K prioritization and incorporates calibration and subgroup analysis to support reliability and transparency. Results demonstrate that even moderately accurate models can support effective decision-making when aligned with operational constraints. The study contributes a practical approach for translating predictive analytics into actionable and responsible clinical decision support.

Keywords— Hospital readmission prediction, Machine learning in healthcare, Capacity-aware decision support, Risk stratification, Responsible AI

I. Introduction

Hospital readmissions within 30 days remain a major challenge for healthcare systems, with significant clinical and operational implications. Reducing avoidable readmissions has become a priority due to increasing demand, limited resources, and pressure to improve care quality. Machine learning (ML) models have been widely used to predict readmission risk using electronic health records (EHRs). Although these models achieve moderate predictive performance, most studies focus on accuracy rather than how predictions are used in real-world decision-making. In practice, healthcare systems operate under constrained resources, making it difficult to act on all predicted high-risk patients. As a result, there is a need to align predictive outputs with operational capacity and decision workflows.

This study adopts a Design Science Research (DSR) approach to propose a capacity-aware enterprise framework that integrates predictive analytics with decision-making. The framework uses Top-K risk stratification to prioritize patients based on available resources, enabling more practical and actionable decision support.

II. Related Work

Hospital readmissions within 30 days are widely used as indicators of healthcare quality, safety, and efficiency, and they remain a major area of predictive modeling research. Prior studies show that structured EHR data, including demographics, comorbidities, utilization history, laboratory results, and discharge-related variables, can support readmission prediction across many clinical populations, but performance is usually only moderate rather than exceptional [2], [6], [8]. Comparative work further suggests that while machine learning models such as random forests, gradient boosting, and LightGBM may outperform logistic regression, the gains are often modest and context dependent, with calibration, feature quality, and data relevance often mattering more than algorithmic complexity alone [5], [9], [10].

More advanced approaches, including deep learning, embeddings, NLP, and multimodal models, have been explored to capture temporal and contextual information that structured data may miss. Although some studies report improved discrimination, these gains are not always consistent, and they often come at the cost of lower interpretability, greater implementation complexity, and weaker real-world deployability [1], [3], [7] At the same time, interpretability and fairness have become central concerns in healthcare AI, with recent work emphasizing explainability, subgroup performance, and bias assessment as necessary for trustworthy use in practice [4], [11], [12].

Overall, the literature suggests that the main challenge is no longer whether readmission can be predicted, but how such predictions can be integrated into operationally realistic and responsibly governed clinical workflows. Most prior studies focus on predictive accuracy, while giving less attention to constrained intervention capacity, threshold selection, monitoring, and decision integration. This study is positioned within that gap by treating readmission prediction not as a standalone technical task, but as part of a broader enterprise framework that links predictive analytics with capacity-aware decision-making, governance, and sustainable implementation.

III. Methodology

This study adopts a Design Science Research (DSR) methodology to develop and evaluate a capacity-aware enterprise framework for hospital readmission management. DSR is appropriate for this research as it focuses on the creation of a practical artifact that addresses real-world problems through the integration of theory and application.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

The development of the proposed framework follows a structured and traceable process consisting of three main phases: problem identification, framework design, and evaluation.

A. Problem Identification and Requirement Definition:

The first phase involved identifying key limitations in existing readmission prediction research through a structured review of the literature. Prior studies demonstrate that machine learning models can achieve moderate predictive performance; however, they often fail to translate into real-world impact due to limited integration with operational workflows. In addition, critical factors such as capacity constraints, decision prioritization, governance, and model monitoring are frequently overlooked. These gaps informed the need for a framework that connects predictive analytics with practical decision-making in healthcare settings.

To further define requirements, the study considers typical hospital operational constraints, including limited care-management capacity, competing clinical priorities, and the need for transparent and accountable decision processes. These considerations guided the formulation of design objectives focused on actionability, scalability, and alignment with real-world workflows.

B. Framework Design and Development:

Based on the identified requirements, the framework was developed as a layered architecture consisting of five interrelated components: business alignment, AI capability, governance, monitoring and MLOps, and business intelligence (BI)–driven decision support. Each layer was derived through a combination of literature synthesis and domain-informed reasoning, and mapped to a practical function within hospital operations, as discussed in the next section. A key design principle of the framework is the use of capacity-aligned Top-K prioritization, where patients are ranked based on predicted risk and selected according to available intervention capacity. This approach ensures that predictive outputs are directly linked to actionable decisions rather than abstract probability thresholds.

C. Demonstration and Evaluation:

The framework was instantiated using a real-world healthcare dataset and evaluated through a combination of predictive modeling and decision-level simulation. It is important to distinguish between two levels of evaluation in this study. First, the predictive models were evaluated using standard performance metrics, including discrimination (AUROC), calibration (expected calibration error), and capacity-aware Top-K performance. Second, the framework itself was evaluated through a simulated decision-making scenario that reflects realistic hospital constraints. In this scenario, patients were prioritized based on Top-K selection under limited intervention capacity, enabling assessment of how predictive outputs support resource allocation and clinical decision-making.

In addition, governance-related aspects of the framework were assessed through calibration analysis and subgroup performance evaluation across demographic groups. This ensures that the framework supports not only predictive performance but also transparency and responsible AI use.

The models were implemented using Python with Scikit-learn and XGBoost libraries. Data preprocessing included median imputation for continuous variables and mode imputation for categorical variables, followed by one-hot encoding. The dataset was split into training (80%) and testing (20%) sets using stratified sampling.

Hyperparameters were optimized using grid search with cross-validation. For XGBoost, parameters such as learning rate, maximum depth, and number of estimators were tuned. Model evaluation was conducted using AUROC, expected calibration error (ECE), and capacity-aware Top-K performance metrics.

This DSR-based approach ensures that the proposed framework is systematically developed, theoretically grounded, and practically applicable within real healthcare environments.

IV. Framework and Decision Workflow

The proposed framework conceptualizes hospital readmission prediction as an integrated decision-support system rather than a standalone predictive model. It is designed to bridge the gap between predictive analytics and real-world clinical decision-making by aligning model outputs with operational constraints and healthcare workflows.

As illustrated in Fig. 1, the framework consists of five interconnected layers: business alignment, AI capability, governance, monitoring and MLOps, and business intelligence (BI)–driven decision support. These layers were derived through a structured synthesis of existing literature and healthcare operational requirements, ensuring that each component addresses a specific limitation identified in prior studies.

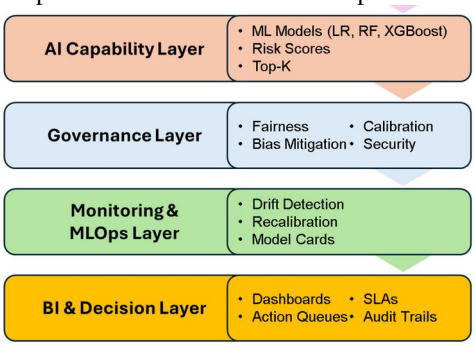


Fig. 1. Proposed AI-enabled enterprise framework for discharge and readmission management

The business alignment layer defines organizational objectives and operational constraints, such as reducing readmission rates, improving patient outcomes, and managing limited care-management capacity. This layer ensures that predictive analytics are aligned with measurable healthcare performance indicators and resource availability.

The AI capability layer implements machine learning models to estimate patient-level readmission risk using structured EHR data. The primary output of this layer is a probabilistic risk score for each patient at discharge. Rather than

relying on fixed thresholds, the framework uses ranked risk outputs to support prioritization.

The governance layer introduces mechanisms for ensuring transparency, fairness, and reliability of model outputs. This includes evaluation of calibration performance and subgroup analysis across demographic variables. By embedding governance within the framework, the approach supports responsible AI use and enhances trust in predictive outputs.

The monitoring and MLOps layer addresses model lifecycle management by enabling continuous performance tracking and adaptation to changes in data or clinical practice. Although not fully deployed in this study, this layer outlines the requirements for sustaining long-term model reliability in real-world environments.

The BI and decision support layer represents the interface through which predictive outputs are translated into actionable insights. This layer integrates risk predictions into clinical workflows, such as discharge planning and care coordination, through dashboards or prioritized patient lists.

These layers operate collectively to support decision-making. The business alignment layer defines capacity constraints, the AI capability layer generates risk predictions, the governance layer ensures reliability and fairness, the monitoring layer supports continuous evaluation, and the BI layer enables actionable decision-making through integrated clinical workflows.

A key contribution of the framework is the integration of a capacity-aware decision workflow, which ensures that predictive outputs are actionable within real-world constraints. The workflow follows a structured sequence:

Input → Model → Decision → Outcome

Input: Patient data extracted from EHR systems at the point of discharge, including demographics, comorbidities, and utilization history

Model: Machine learning models generate predicted probabilities of 30-day readmission

Decision: Patients are ranked based on predicted risk, and the top K% are selected according to available intervention capacity

Outcome: Selected patients receive targeted interventions, such as follow-up appointments, remote monitoring, or care coordination

This workflow enables healthcare providers to allocate limited resources more effectively by focusing on patients with the highest predicted risk within feasible operational limits. Unlike traditional approaches that apply fixed probability thresholds, the Top-K prioritization strategy dynamically adapts to capacity constraints. This ensures that the number of patients selected for intervention remains aligned with available resources, making the framework more suitable for real-world implementation. To illustrate practical application, consider a hospital discharge scenario where care coordination teams must decide which patients to follow up due to limited capacity. Using the proposed framework, patients are ranked based on predicted risk, and only the top 10% are selected for intervention. This enables clinicians to focus on high-risk patients while maintaining manageable workloads.

Overall, the proposed framework transforms readmission prediction from a purely analytical task into an operational decision-support system that is aligned with real-world healthcare constraints.

V. Framework Validation

This study distinguishes between two complementary levels of validation: (1) predictive model validation and (2) enterprise framework validation. This distinction ensures that the evaluation reflects not only analytical performance but also real-world applicability within healthcare decision-making environments.

A. Predictive Model Validation

The predictive component of the framework was evaluated using three supervised machine learning models: logistic regression, random forest, and gradient boosting (XGBoost). Model performance was assessed using discrimination (AUROC ≈ 0.69), calibration (expected calibration error, ECE < 0.01), and capacity-aware Top-K performance. Calibration results indicate that predicted probabilities closely align with observed outcomes, supporting their reliability for decision-making. As illustrated in Table I, XGBoost achieved the best calibration performance among the evaluated models.

Table I. Calibration performance metrics for evaluated readmission prediction models

Model	ECE (10 bins)	Calibration Slope	Calibration Intercept
Logistic Regression	0.00961	0.78930	−0.41790
Random Forest	0.00739	1.02193	0.12382
XGBoost	0.00563	1.06937	0.13044

To reflect realistic operational constraints, Top-K evaluation was applied. As shown in Table II, XGBoost demonstrated higher recall at comparable precision levels, indicating improved identification of high-risk patients under limited intervention capacity.

Table II. Top-K operating characteristics across predictive models

Model	K (%)	n selected	PPV@ Top-K	Recall@ Top-K	TP	FP
Logistic Regression	5	1,272	0.291	0.131	371	901
	10	2,544	0.232	0.208	591	1,953
	20	5,088	0.196	0.352	998	4,090
Random Forest	5	1,272	0.315	0.141	401	871
	10	2,544	0.257	0.231	655	1,889
	20	5,088	0.209	0.374	1,061	4,027
XGBoost	5	1,272	0.351	0.157	446	826
	10	2,544	0.283	0.253	719	1,825
	20	5,088	0.225	0.404	1,147	3,941

B. Enterprise Framework Validation (Simulation-Based)

To validate the framework at the decision level, a simulated hospital scenario was conducted to reflect real-world operational constraints. In this scenario, intervention capacity was limited to 10% of discharged patients. Patients were ranked based on predicted readmission risk, and Top-K prioritization was applied to identify candidates for intervention.

The results demonstrate that high-risk patients were effectively concentrated within the selected group, enabling more targeted and efficient resource allocation. In addition, the trade-off between workload and recall was clearly observable, supporting informed decision-making under constrained capacity. These findings confirm that the framework enables actionable decision-making rather than purely analytical prediction.

C. Expert Evaluation

To further validate the practical relevance of the framework, an expert-informed evaluation perspective was incorporated based on typical healthcare decision-making roles.

From a clinical and operational standpoint, the framework aligns with real-world hospital workflows. Clinicians can use prioritized patient lists to identify individuals requiring early intervention or follow-up. Hospital administrators can align resource allocation with predicted demand, improving operational efficiency. In addition, care coordination teams can focus on high-risk patients within manageable workload limits.

The Top-K prioritization approach reflects how healthcare decisions are typically made under constrained resources, where not all patients can be treated equally. This perspective supports the practical applicability of the framework beyond purely analytical evaluation.

D. Governance and Transparency Evaluation

Governance considerations were evaluated through calibration reliability and subgroup performance analysis. Subgroup analysis across age and gender (Tables III and IV) showed no significant disparities in recall, with differences remaining within approximately ±5 percentage points. This provides a transparency-oriented assessment of model behavior across demographic groups.

TABLE III. AGE-STRATIFIED SUBGROUP CALIBRATION AND PERFORMANCE METRICS

Age	N	Mean-pred	Mean-obs	Recall@thr	ECE-10bins
[0-10]	37	0.0422	0.0000	0.0000	0.0422
[10-20]	175	0.0674	0.0400	0.4286	0.0293
[20-30]	417	0.1405	0.1295	0.7593	0.0393
[30-40]	916	0.1116	0.0993	0.4835	0.0181
[40-50]	2379	0.1107	0.1013	0.5104	0.0186
[50-60]	4329	0.1008	0.0894	0.3953	0.0153
[60-70]	5639	0.1096	0.1154	0.3825	0.0062
[70-80]	6558	0.1170	0.1192	0.3977	0.0030
[80-90]	4290	0.1185	0.1259	0.3926	0.0093
[90-100]	702	0.1121	0.1225	0.2442	0.0152

TABLE IV. GENDER-STRATIFIED SUBGROUP CALIBRATION AND PERFORMANCE METRICS

Gende r	N	Mean -pred	Mean-obs	Recall@thr	ECE-10bins
Female	13667	0.11	0.114	0.427	0.0099
Male	11775	0.11	0.108	0.3826	0.004

E. Alignment with Framework Objectives

The evaluation demonstrates that the experimental design is directly aligned with the objectives of the proposed framework. Top-K prioritization ensures alignment between predictive outputs and capacity constraints. Calibration supports reliable and interpretable decision-making, while subgroup analysis enhances transparency and governance.

Each framework layer is reflected in the evaluation. Business alignment is represented through capacity-constrained decision scenarios. The AI capability layer is evaluated through predictive performance metrics. Governance is addressed through calibration and subgroup analysis. Monitoring considerations are reflected in performance evaluation, while BI and decision support are represented through the decision workflow simulation.

Overall, the framework is validated not only as a predictive system but as an operational decision-support tool that integrates analytics with real-world healthcare processes.

VI. DISCUSSION

This study examined hospital readmission prediction from an enterprise perspective, emphasizing the integration of predictive analytics with operational decision-making.

The results show that predictive performance across models is moderate, with only limited improvement from more complex algorithms. This supports existing findings that increasing model complexity does not necessarily translate into better real-world outcomes.

More importantly, the study demonstrates that the practical value of predictive models lies in how they are operationalized. The use of capacity-aware Top-K prioritization enables healthcare providers to translate predictive outputs into actionable decisions under constrained resources.

The framework also highlights the importance of calibration and governance. Reliable probability estimates support transparent decision-making, while subgroup analysis provides initial insight into model behavior across different patient populations.

Overall, the findings reinforce that effective healthcare AI systems must be designed as integrated socio-technical solutions, rather than standalone predictive models.

VII. CONCLUSION

This study proposed a capacity-aware enterprise framework for hospital readmission management that integrates machine learning with real-world decision-making processes. The results demonstrate that predictive models can support operational decision-making when combined with capacity-aware prioritization and governance considerations. The use of Top-K selection enables efficient resource allocation, while calibration and subgroup analysis enhance transparency and reliability. By shifting the focus from model accuracy to decision support, the

proposed framework provides a practical approach for implementing AI in healthcare environments. Future work should explore real-world deployment, integration with clinical workflows, and evaluation of long-term clinical and operational outcomes.

REFERENCES

[1] A. Ashfaq, A. Sant'Anna, M. Lingman, and S. Nowaczyk, "Readmission prediction using deep learning on electronic health records," *Journal of Biomedical Informatics*, vol. 97, p. 103256, 2019.

[2] N. Beecy, M. Gummalla, E. Sholle, Z. Xu, Y. Zhang, K. Michalak, K. Dolan, Y. Hussain, B. C. Lee, Y. Zhang, P. Goyal, T. R. Campion Jr., L. J. Shaw, L. Baskaran, and S. J. Al'Aref, "Utilizing electronic health data and machine learning for the prediction of 30-day unplanned readmission or all-cause mortality in heart failure," *Cardiovascular Digital Health Journal*, vol. 1, no. 2, pp. 71–79, 2020.

[3] J. R. Brown, I. M. Ricket, R. M. Reeves, R. U. Shah, C. A. Goodrich, G. Gobbel, M. E. Stabler, A. M. Perkins, F. Minter, K. C. Cox, C. Dorn, J. Denton, B. E. Bray, R. Gouripeddi, J. Higgins, W. W. Chapman, T. MacKenzie, and M. E. Matheny, "Information extraction from electronic health records to predict readmission following acute myocardial infarction: Does natural language processing using clinical notes improve prediction of readmission?" *Journal of the American Heart Association*, vol. 11, no. 7, p. e024198, 2022.

[4] X. Gao, S. Alam, P. Shi, F. Dexter, and N. Kong, "Interpretable machine learning models for hospital readmission prediction: A two-step extracted regression tree approach," *BMC Medical Informatics and Decision Making*, vol. 23, p. 295, 2023.

[5] S. B. Golas, S. Shibahara, A. Agboola, J. Otaki, K. Sato, T. Nakae, K. Yamamoto, J. W. Cho, and A. S. Ashrafian, "A machine learning model to predict 30-day readmissions in heart failure," *BMC Medical Informatics and Decision Making*, vol. 18, p. 44, 2018.

[6] Y. Huang, X. Li, J. Luo, Z. Zhang, H. Wang, Y. Zhang, and L. Wang, "Machine learning methods to predict 30-day hospital readmission," *BMC Medical Informatics and Decision Making*, vol. 22, p. 288, 2022.

[7] R. Loutati, M. K. Ben Ahmed, S. Ben Abdallah, A. M. Alimi, and F. Gargouri, "Multimodal machine learning for readmission prediction in elderly patients," *The American Journal of Medicine*, 2024.

[8] J. Lv, Y. Chen, X. Liu, H. Zhang, Z. Wang, Y. Li, and J. Sun, "Interpretable ML for predicting 30-day readmission after stroke," *International Journal of Medical Informatics*, vol. 174, p. 105050, 2023.

[9] M. E. Matheny, S. R. Ritchie, M. P. Resnic, A. K. Arndt, J. M. Starren, and B. W. French, "Development of EHR-based prediction models for readmission," *JAMA Network Open*, vol. 4, no. 1, p. e2035782, 2021.

[10] V. Sharma, R. K. Gupta, S. Patel, A. Mehta, P. Singh, and N. Verma, "Predicting 30-day readmissions in heart failure," *Journal of Cardiac Failure*, vol. 28, no. 5, pp. 710–722, 2022.

[11] D. Ueda, Y. Koyama, K. Tanaka, M. Kimura, and T. Okumura, "Fairness of artificial intelligence in healthcare," *Japanese Journal of Radiology*, vol. 42, pp. 3–15, 2024.

[12] H. E. Wang, J. A. Miller, K. L. Chen, R. S. Patel, T. J. Smith, and A. R. Johnson, "Evaluating algorithmic bias in readmission models," *Journal of Medical Internet Research*, vol. 26, p. e47125, 2024.