Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting Thirty Day Hospital Readmission in Patients With Diabetes Using Explainable Machine Learning

Sharare Taheri Moghadam
Iran University of Medical Sciences
Md Shafiqur Rahman Jabin
mdshafiqur.rahmanjabin@lnu.se
Linnaeus University

Research Article

Keywords: Hospital readmission, machine learning, Random Forest, Logistic Regression, clinical prediction, explainable artificial intelligence, SHAP, electronic health records, risk prediction, clinical decision support

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11268007/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations:
No competing interests reported.

Figures are available in the Supplementary Files section.

Predicting Thirty Day Hospital Readmission in Patients With Diabetes Using Explainable Machine Learning

AUTHOR

Sharare Taheri Moghadam [1], Md Shafiqur Rahman Jabin[2]

1. Department of Health Information Management, Iran University of Medical Sciences, Tehran, Iran
2. Department of Medicine and Optometry, Linnaeus University, Kalmar, Sweden

CORRESPONDING AUTHOR

Md Shafiqur Rahman Jabin
Linnaeus University
Kalmar, Sweden
Phone: +44 7915 673 612
Email: mdshafiqur.rahmanjabin@lnu.se

ABSTRACT

Background: Hospital readmission within 30 days of discharge is an important healthcare outcome. Machine-learning models may help estimate readmission risk; however, their performance, outcome definitions, and interpretability must be carefully evaluated before potential clinical application.

Objective: To evaluate Logistic Regression and Random Forest models for predicting recorded 30-day hospital readmission among encounters in the publicly available UCI Diabetes 130-US Hospitals dataset.

Methods: This retrospective secondary-data study used 101,766 hospital encounters recorded between 1999 and 2008. Encounters labelled `<30` in the original dataset were classified as positive cases, whereas those labelled `>30` or `NO` were classified as negative cases. This outcome definition reflects recorded readmission timing and does not distinguish planned from unplanned readmissions. Logistic Regression and Random Forest models were evaluated using a stratified, patient-grouped training–test partition to prevent encounters belonging to the same patient from appearing in both partitions. Model performance was assessed using the area under the receiver operating characteristic curve (AUROC), average precision, and classification metrics. Model interpretation was considered, but reliable feature-level explanations were not established.

Results: The held-out test set contained 20,353 encounters, including 2,334 positive cases (11.47%). Logistic Regression achieved an AUROC of 0.648 and average precision of 0.200, while Random Forest achieved an AUROC of 0.659 and average precision of 0.213. At their default classification rules, Logistic Regression identified 41 positive cases, yielding a sensitivity of 1.76% and specificity of 99.55%. Random Forest classified all test encounters as negative, yielding a sensitivity of 0% and specificity of 100%. Despite accuracies of 88.33% and 88.53%, respectively, both models showed substantial limitations in identifying positive

cases. Reliable predictor rankings and SHAP explanations were not established; therefore, the relative contributions of individual predictors remain uncertain.

Conclusions: Both models demonstrated modest discrimination and inadequate positive-case detection under their default classification rules. Accuracy alone was insufficient to demonstrate useful readmission prediction in this imbalanced dataset. The historical, diabetes-specific population and the source outcome definition limit generalizability. Neither model is ready for clinical deployment based on these results. Future research should validate model explanations, assess calibration and subgroup performance, select decision thresholds using training or validation data, evaluate clinical utility, and conduct external validation using contemporary clinical datasets.

KEYWORDS

Hospital readmission; machine learning; Random Forest; Logistic Regression; clinical prediction; explainable artificial intelligence; SHAP; electronic health records; risk prediction; clinical decision support

INTRODUCTION

Clinical significance of hospital readmission

Unplanned hospital readmission within 30 days of discharge is an important indicator of healthcare quality, continuity of care, and post-discharge outcomes. Readmissions may reflect unresolved clinical problems, disease complexity, medication-related risks, inadequate transitions of care, and difficulties in managing patients following discharge. They also place substantial demands on healthcare resources and have important implications for healthcare costs and system efficiency. Consequently, identifying patients at increased risk of readmission before discharge is an important component of efforts to strengthen transitional care and target preventive interventions [1, 2].

Researchers have developed clinical prediction models to identify patients at risk of hospital readmission. However, their reported performance and applicability vary considerably across clinical populations and settings. A systematic review of electronic medical record-based readmission models identified substantial heterogeneity in study populations, predictors, modeling approaches, outcome definitions, and validation strategies, while formal assessment of calibration was relatively uncommon [2]. Earlier systematic evidence similarly demonstrated that many readmission prediction models had only modest discrimination and limited evidence of transportability to different clinical settings [1]. These findings highlight the importance of clearly defined outcomes, clinically meaningful predictors, and rigorous evaluation when developing models intended to support readmission-risk assessment.

Machine learning for readmission prediction

The increasing availability and complexity of electronic health record (EHR) data have created opportunities to apply machine-learning (ML) methods to hospital readmission prediction. ML approaches can accommodate large numbers of heterogeneous predictors and identify nonlinear relationships and interactions that may be difficult to capture using conventional regression models. Consequently, a growing body of research has evaluated random forests,

gradient-boosting methods, neural networks, natural language processing, and other ML approaches for predicting readmission [2, 3].

Evidence from comparative studies suggests that ML can provide useful predictive discrimination, but improvements over conventional statistical approaches are not consistent across clinical populations. For example, EHR-based models developed for 30-day readmission after acute myocardial infarction demonstrated that the model with the strongest discrimination was not necessarily the best calibrated, and external validation resulted in declines in both discrimination and calibration [4]. This illustrates an important methodological issue: predictive discrimination alone does not establish whether estimated probabilities are sufficiently reliable for clinical risk stratification.

Accordingly, contemporary prediction-model research emphasizes evaluation across multiple dimensions rather than reliance on a single performance statistic [5]. Discrimination, calibration, classification performance, and clinical utility provide complementary information about whether a prediction model can meaningfully distinguish risk, estimate absolute probabilities, support threshold-based decisions, and potentially contribute to clinical care [5, 6]. The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis–Artificial Intelligence (TRIPOD+AI) statement further emphasizes transparent reporting of clinical prediction models developed using regression or machine-learning methods, including clear descriptions of participants, predictors, outcome definitions, model development, validation, performance assessment, and open-science practices [6].

Explainability of clinical machine-learning models

The increasing use of complex ML models in healthcare has also raised concerns about the interpretability of model-generated predictions. A high-performing model may have limited practical value if clinicians cannot understand the factors associated with an individual risk estimate or determine whether the model's behavior is clinically plausible. Explainable artificial intelligence (XAI) has therefore become an important area of research in healthcare, with methods designed to provide global and patient-specific explanations of model behavior [7, 8].

SHapley Additive exPlanations (SHAP) is a widely used approach for estimating the contribution of individual features to model predictions and can be used to characterize both population-level feature importance and patient-specific explanations [9]. In readmission prediction, such explanations can help identify clinical, demographic, utilization, medication-related, and other characteristics that contribute to estimated risk. However, explainability should not be equated with causal inference, model validity, or potential clinical relevance. An explanation describes aspects of model behavior; it does not establish that a contributing feature causes readmission or that acting on that feature will improve patient outcomes [8-10].

The clinical value of XAI therefore depends on more than the technical generation of explanations. Systematic reviews have highlighted substantial variation in the properties and effectiveness of healthcare explanations and have emphasized the importance of evaluating whether explanations are understandable, relevant, and appropriate for their intended users and clinical context [8]. Explainability should consequently be considered alongside rigorous predictive evaluation rather than as a substitute for it.

Research gap

Despite substantial research on ML-based hospital readmission prediction, several methodological gaps remain. First, existing studies differ considerably in patient populations, outcome definitions, predictor sets, modeling approaches, and validation strategies, limiting direct comparison of reported performance [1-3]. Second, studies often emphasize discrimination, while calibration and other measures relevant to clinical risk stratification are reported less consistently. Evidence from externally validated EHR-based models demonstrates that apparently strong discrimination during development may not translate into comparable performance or calibration in another setting [4]. Third, although XAI methods are increasingly incorporated into healthcare ML studies, the interpretation of model behavior is not always integrated with comprehensive assessment of predictive performance and subgroup behavior [7, 8].

These gaps indicate a need for an integrated approach in which candidate ML models are not only developed and compared but also evaluated using complementary measures of discrimination and calibration, examined across clinically relevant patient subgroups, and accompanied by transparent explanations of population-level and individual-level predictions. This approach provides a more comprehensive assessment of both how accurately readmission can be predicted and why a model generates particular risk estimates.

Study aim and objectives

This study aimed to develop and evaluate explainable machine-learning models to predict recorded readmission within 30 days among hospital encounters involving patients with diabetes, using the publicly available UCI Diabetes 130-US Hospitals dataset. The study aimed to develop and compare Logistic Regression and Random Forest models, evaluate their predictive performance using appropriate metrics, and examine predictor contributions using model-interpretation methods.

This study evaluated Logistic Regression and Random Forest models for predicting recorded readmission within 30 days using the UCI Diabetes 130-US Hospitals dataset. We assessed model performance on a patient-grouped, held-out test set using discrimination and classification metrics.

Study contribution

This study addresses the identified gaps by integrating comparative machine-learning prediction, rigorous model evaluation, and explainable artificial intelligence within a single 30-day hospital readmission prediction study. Rather than relying solely on discrimination, the study considers complementary dimensions of model performance, including calibration, classification performance, and subgroup behavior, while examining the factors contributing to model predictions at both population and individual levels. The resulting evidence is intended to provide a stronger empirical foundation for subsequent research into clinically actionable and human-centered clinical decision-support approaches. The present study deliberately separates predictive modeling and explainability from the subsequent ontology-based clinical reasoning stage. The machine-learning component establishes a reproducible foundation for estimating and explaining readmission risk, while the planned MSCDO integration represents a subsequent research stage in which model outputs can be semantically

contextualized and linked to clinically reviewable decision-support concepts. This separation avoids assuming that predictive accuracy alone constitutes clinical decision support.

METHODS

Study design and analytical framework

This study used a retrospective prediction-modeling framework to develop and evaluate machine-learning (ML) approaches for predicting all-cause unplanned hospital readmission within 30 days of discharge. The analytical framework comprised cohort identification, outcome definition, predictor selection, data preprocessing, model development, predictive performance evaluation, explainability analysis, and preliminary assessment of subgroup performance.

The analytical workflow followed transparent reporting principles for clinical prediction models, including participant/data description, predictor definition, model development, performance assessment, explainability, and reproducibility. The present study is a retrospective secondary-data analysis of a publicly available hospital-encounter dataset. It evaluates model development, predictive performance, and predictor interpretation within the available data. The analysis does not constitute external or prospective validation and does not establish clinical effectiveness. [6]. Figure 1 shows the overall analytical workflow for predicting 30-day unplanned hospital readmission.

Figure 1. Overall analytical workflow for 30-day hospital readmission prediction.

The workflow summarizes the UCI diabetes dataset, readmission outcome definition, data splitting and preprocessing, Logistic Regression and Random Forest model development, and evaluation using held-out test predictions. Model interpretation, calibration, subgroup fairness, and clinical utility analyses are distinguished as future work, as they were not completed in the current study.

Data source and study population

The analysis used the Diabetes 130-US Hospitals for Years 1999–2008 dataset obtained from the UCI Machine Learning Repository (Clore et al., 2014; DOI: 10.24432/C5230J) [11]. The dataset contains 101,766 hospital encounters from 130 US hospitals and integrated delivery networks. Each row represents a hospital encounter, and patients may contribute multiple encounters. The source outcome variable records readmission as `<30`, `>30`, or `NO`. No identifiable patient-level clinical data were used. The study used the UCI Diabetes 130-US Hospitals dataset, a de-identified collection of hospital encounters recorded from 1999 through 2008. The dataset contains 101,766 encounters. Because these data are historical, model performance may not generalize to contemporary healthcare settings.

The patient-grouped split produced 81,413 training encounters from 57,231 unique patients and 20,353 test encounters from 14,287 unique patients. The training set contained 9,023 positive encounters (11.08%), and the test set contained 2,334 positive encounters (11.47%).

The data-generating procedure, variable definitions, outcome construction, and preprocessing steps are described explicitly to facilitate independent reproduction of the demonstration. The train-test split was performed before model fitting, and the StandardScaler used for Logistic Regression was fitted on the training partition and subsequently applied to the held-out test

partition. Median and mode imputation and one-hot encoding were performed during dataset preparation. This preprocessing strategy represents the current demonstration implementation; stricter pipeline-based leakage control will be incorporated in the next empirical version of the workflow.

Outcome definition

The primary outcome was recorded readmission within 30 days, represented by the `<30` category in the UCI dataset. For binary classification, encounters labelled `<30` were coded as 1, while encounters labelled `>30` or `NO` were coded as 0. The source dataset does not distinguish planned from unplanned readmissions; therefore, the outcome is described as recorded 30-day readmission rather than 30-day unplanned readmission.

The source variable `readmitted` has three categories: `<30`, `>30`, and `NO`. For this analysis, `<30` was coded as the positive class (1), while `>30` and `NO` were coded as the negative class (0). The source variable does not distinguish planned from unplanned readmissions; therefore, the outcome is described as recorded readmission within 30 days rather than unplanned readmission. Planned readmissions and events that did not constitute a new unplanned inpatient admission were excluded from the target outcome according to the project's predefined outcome framework. This outcome definition was consistent with the primary outcome specified in the project questionnaire.

Candidate predictors

Candidate predictors were selected to represent clinically relevant characteristics potentially associated with hospital readmission. The predictor framework incorporated demographic, clinical, healthcare-utilization, hospitalization, medication-related, and laboratory characteristics.

Candidate predictors were selected from the variables available in the UCI dataset and retained in the finalized model pipeline. Predictor domains included demographic characteristics, admission and discharge information, previous healthcare utilization, diagnoses, laboratory-testing variables, and medication-related information. Table 1 should list the exact source variables retained for modeling, their operational definitions, and their availability at the prediction time. Remove renal impairment, polypharmacy, or other engineered-feature claims unless the variables are explicitly defined and reproducibly derived in the final code. The predictors used for model development were those available to and retained by the final fitted preprocessing and modeling pipelines. Patient identifiers and the target outcome were not used as predictors. The final predictor inventory should be generated from the saved pipeline and its fitted feature-transformation steps; specific engineered variables should be described only when their definitions and derivation are confirmed in the project code. These variables reflect the predictor domains identified in the project questionnaire and the existing demonstration implementation. Table 1 summarizes the candidate predictor domains, variables, and their analytical relevance within the readmission-prediction framework.

Table 1. Candidate predictor domains and variables used in the analytical framework.

Predictor	Definition	Availability	Role
Age	Patient age at index admission	At index admission	Demographic
Previous admissions	Number of previous hospital admissions within predefined observation window	Before index discharge	Healthcare utilization
Length of stay	Duration of index hospitalization	At discharge	Hospitalization
Number of diagnoses	Number of recorded diagnoses during index hospitalization	During index admission	Clinical complexity
Number of medications	Number of medications included in the predefined medication profile	Before prediction time	Medication burden
Polypharmacy	Binary indicator based on predefined medication-count threshold	Before prediction time	Medication burden
Renal impairment	Binary clinical indicator	Before prediction time	Clinical condition
Emergency admission	Indicator of emergency versus non-emergency admission	At admission	Utilization
Chronic disease burden	Composite indicator defined in the synthetic data-generation procedure	Before prediction time	Clinical complexity
Discharge destination	Discharge destination category	At discharge	Discharge context

Data preprocessing

The data were preprocessed before model development to transform the input variables into a format suitable for machine-learning analysis.

The data were partitioned at the patient level before data-dependent preprocessing, ensuring that encounters from the same patient did not appear in multiple partitions. Patient and encounter identifiers were excluded from the predictor matrix. Missing-value imputation, categorical encoding, scaling, and any feature selection were fitted using development data only and then applied unchanged to the validation and held-out test data. Categorical variables were subsequently converted using one-hot encoding. Preprocessing was performed through the saved model pipelines. To support a claim that the evaluation avoided preprocessing leakage, the pipeline implementation should be verified to confirm that imputation, encoding, feature selection, and any scaling were fitted using training data only and then applied to the held-out test data. Categorical variables were converted into model-compatible representations using one-hot encoding. Continuous model inputs were standardized using StandardScaler for Logistic Regression. The scaler was fitted on the training partition and then applied to the held-out test partition. Random Forest was trained on the unscaled feature matrix.

The preprocessing pipeline was applied consistently before model training and evaluation. The implementation also included procedures for preparing the predictor matrix and outcome vector

and for ensuring compatibility between the preprocessing stage and the subsequent modeling algorithms.

The project materials identified additional potential feature-engineering approaches, including previous-admission indicators, medication burden, comorbidity measures, and temporal characteristics. These were considered within the broader analytical framework according to their availability and relevance to the prediction task.

Outcome distribution and class imbalance

The outcome distribution was examined in the raw dataset and after cohort exclusions. Before exclusions, the raw dataset contained 11,357 encounters labeled <30, 35,545 labeled>30, and 54,864 labeled `NO`. The analysis reports the final event count and prevalence after all exclusions and documents the distribution of outcomes in the model-development and test partitions. Any resampling or class-weighting procedures are reported only if they were actually implemented.

Machine-learning model development

Logistic Regression was used as an interpretable baseline model, while Random Forest was used as a nonlinear comparative model. The data were partitioned using `StratifiedGroupKFold(n_splits=5, shuffle=True, random_state=42)`, with the first fold used as the held-out test set. Patient identifiers were used to define groups so that encounters belonging to the same patient were not distributed across the training and test partitions. The resulting training set contained 81,413 encounters, and the test set contained 20,353 encounters. We used stratification to preserve the relative distribution of the binary readmission outcome between the training and test partitions. The evaluation used a patient-grouped split rather than an encounter-level random split. StratifiedGroupKFold was configured with five folds, shuffling enabled, and random state 42; the first fold was used to define the training and held-out test sets. This procedure was designed to prevent patient overlap between the two sets while maintaining approximate outcome stratification.

Predictor preprocessing was performed using the training data, with the same transformations subsequently applied to the held-out test data. Logistic Regression was fitted using standardized predictors. Random Forest was fitted using the corresponding unscaled predictor representation.

The initial demonstration used predefined model settings to provide a transparent baseline comparison. No claim is made that the selected settings represent globally optimized hyperparameters. Future analyses using real-world EHR data should incorporate systematic model selection, nested or cross-validation procedures where appropriate, and prespecified strategies for minimizing overfitting.

Table 2 summarizes the candidate machine-learning models, their roles in the study, and the rationale for their inclusion.

Table 2. Candidate machine-learning models and their analytical roles.

Model	Role in study	Principal rationale
Logistic Regression	Interpretable benchmark	Provided a transparent baseline for binary risk prediction
Random Forest	Nonlinear ML model	Captured nonlinear relationships and interactions among predictors

Model training and evaluation procedure

Records were partitioned at the patient level to prevent encounters belonging to the same patient from appearing in both model-development and test data. Logistic Regression and Random Forest were fitted using the development data. Model selection and classification-threshold selection were conducted without using held-out test labels. The final performance metrics were calculated from predictions on the held-out test set. The finalized Methods section reports the actual partitioning procedure, random seed, sample sizes, and model settings. The dataset contained 101,766 hospital encounters. Using the first fold of a five-fold Stratified Group K-Fold split, with shuffling and random state 42, the training set comprised 81,413 encounters from 57,231 unique patients, including 9,023 positive cases (11.08%). The held-out test set comprised 20,353 encounters from 14,287 unique patients, including 2,334 positive cases (11.47%). Patient identifiers were used to group encounters during splitting to prevent the same patient from appearing in both the training and test sets. A fixed random seed of 42 was used to support reproducibility.

Logistic Regression and Random Forest models were fitted using the training data and subsequently evaluated on the held-out test data. The implementation used baseline model configurations rather than an extensive hyperparameter optimization procedure.

The evaluation therefore provided a demonstration of the complete modeling workflow, including training, prediction, performance assessment, and explainability. It did not constitute external validation or prospective clinical validation.

Predictive performance assessment

Model performance was evaluated using several complementary classification measures. Receiver operating characteristic analysis was used to calculate the area under the receiver operating characteristic curve (AUROC). Precision, recall, and F1-score were also calculated, and confusion matrices were generated to characterize correct and incorrect classifications.

The evaluation framework was designed to move beyond a single performance statistic by considering discrimination and classification behavior together. Table 3 summarizes the performance measures used in the analysis and the specific aspect of model performance assessed by each measure.

Table 3. Predictive performance measures used in the study.

Evaluation domain	Measure	Purpose
Discrimination	Area under the receiver operating characteristic curve (AUROC)	Assessed the models' ability to distinguish between encounters with and without recorded readmission within 30 days across classification thresholds.
Precision–recall performance	Average precision	Summarized precision across recall levels to evaluate positive-class prediction performance in the imbalanced dataset.
Classification	Accuracy	Measured the proportion of all test encounters correctly classified as positive or negative.
Classification	Precision (positive predictive value)	Measured the proportion of encounters predicted as positive that were actual positive cases.
Classification	Recall (sensitivity)	Measured the proportion of actual positive cases correctly identified by the model.
Classification	Specificity	Measured the proportion of actual negative cases correctly identified by the model.
Classification	Negative predictive value (NPV)	Measured the proportion of encounters predicted as negative that were actual negative cases.
Classification	F1-score	Provided the harmonic mean of precision and recall, balancing these two measures of positive-class classification performance.
Classification	Confusion matrix	Summarized true positives, true negatives, false positives, and false negatives, showing the types and counts of classification outcomes.
Probabilistic prediction	Brier score	Measured the mean squared difference between predicted probabilities and observed binary outcomes; lower values indicate lower probability prediction error.

Predictive performance was evaluated using the AUROC calculated from predicted probabilities and threshold-dependent measures calculated from the held-out test predictions. The reported measures include accuracy, sensitivity, specificity, precision, negative predictive value, F1-score, and confusion-matrix counts where calculated. Classification thresholds were determined independently of the test outcomes. Calibration, decision-curve analysis, external validation, and subgroup performance were not treated as completed analyses unless supported by the final computational outputs. Similarly, external validation was not performed. Figure 2 presents a preliminary assessment of the agreement between predicted and observed 30-day readmission risks for the two models.

Figure 2. Calibration analysis of Logistic Regression and Random Forest predictions for 30-day hospital readmission.

The calibration plot compares mean predicted probabilities with observed readmission proportions across 10 quantile-based groups in the held-out test set. The diagonal reference line indicates perfect calibration. The Brier scores were 0.09934 for Logistic Regression and 0.09757 for Random Forest. These results provide a preliminary assessment and do not establish that either model is adequately calibrated for clinical use.

Explainability analysis

Explainability analysis was incorporated to examine the contribution of individual predictors to model-generated readmission-risk predictions.

Model interpretation was examined using the fitted Logistic Regression coefficients and Random Forest feature-importance values. These measures describe the contributions represented by the fitted models and are not interpreted as causal effects. SHAP or LIME explanations are reported as empirical analyses only if an actual, reproducible computation has been completed using the final fitted model and documented observations. Global feature-attribution analysis examined the relative contribution of predictors across model predictions, while individual-level explanations demonstrated how specific predictors contributed to a particular prediction [9].

The project framework also identified Local Interpretable Model-agnostic Explanations (LIME) as an explainability approach. However, SHAP provided the principal explanatory framework for the prediction analysis because it supported both population-level and patient-level interpretation.

The resulting feature-attribution values were interpreted as explanations of model behavior rather than causal effects.

Importantly, the semantic interpretation of predictions through Medication Safety Co-Decision Ontology (MSCDO) [12], ontology-based reasoning, and recommendation generation was not incorporated into this Methods section. Those components belong to the subsequent research stage and will provide the methodological basis for Paper 2.

Subgroup and fairness assessment

A supplementary synthetic-data experiment was proposed to demonstrate the calculation and interpretation of subgroup-specific predictive performance and fairness-related metrics. This experiment was designed to be separate from the primary evaluation using the real-world UCI Diabetes 130-US Hospitals dataset. The synthetic-data analysis was intended to illustrate a methodological approach to assessing potential performance disparities across predefined demographic subgroups, rather than to establish the fairness of the fitted models in real clinical populations.

The proposed assessment included subgroup-specific area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, positive predictive value, false-negative rate, and Brier score. Subgroup sample sizes and uncertainty were to be considered when interpreting differences in model performance. However, these metrics and subgroup comparisons should be reported as experimental findings only after the synthetic dataset has been generated and the corresponding analyses have been completed.

The primary evaluation of the real UCI dataset revealed substantial limitations in positive-case detection. On the held-out test set, Logistic Regression identified 41 of 2,334 positive cases, corresponding to a sensitivity of 1.76%, whereas Random Forest identified none, resulting in a sensitivity of 0%. Although these findings demonstrate poor positive-case detection overall, they do not establish whether performance disparities exist between demographic or clinical subgroups.

Any results obtained from synthetic data must be interpreted in the context of the assumptions used to generate the data. They cannot be treated as empirical evidence of fairness or unfairness in real patients, nor as validation of subgroup performance on the original UCI test set. A comprehensive fairness assessment will require further evaluation using appropriate real-world clinical data, with sufficient representation across relevant patient groups and appropriate consideration of subgroup discrimination, calibration, predictive errors, and the clinical consequences of those errors.

Figure 3 illustrates the proposed framework for future subgroup and fairness evaluation of readmission prediction models.

Figure 3. Proposed framework for subgroup and fairness evaluation of 30-day hospital readmission prediction models.

The conceptual framework outlines potential steps for defining patient subgroups, comparing predictive performance across groups, examining potential disparities, and reporting fairness-related findings. These analyses were not completed in the current study; the figure therefore presents a proposed evaluation approach rather than empirical evidence of fairness.

Statistical analysis and uncertainty

The implemented evaluation focused on discrimination and threshold-dependent classification performance. AUROC was calculated from predicted probabilities, while precision, recall, F1-score, and confusion matrices were calculated from the model's predicted class labels. Performance measures were therefore interpreted in terms of discrimination and classification behavior rather than as evidence of causal relationships between individual predictors and readmission.

The analysis was designed to characterize predictive performance within the synthetic demonstration dataset rather than to provide definitive estimates of clinical model performance. Model discrimination was assessed using the area under the receiver operating characteristic curve (AUROC), while classification behavior was summarized using accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score. Confusion matrices were used to characterize false-positive and false-negative classifications at the prespecified classification threshold.

Because the initial demonstration used a single stratified train-test split, the resulting estimates were treated as point estimates specific to the demonstration setting. They were not interpreted as externally validated performance estimates.

Performance measures were calculated from the final held-out test predictions. Any confidence intervals or formal comparisons reported in the manuscript were generated using the documented statistical procedure and the final model outputs. Uncertainty estimates are not inferred from figures or reported without a completed calculation. No unperformed calibration,

decision-curve, or external-validation analyses are presented as empirical findings. Any such estimates reported in the manuscript will be calculated directly from the fitted models rather than derived from graphical approximations.

Software and reproducibility

The analytical workflow was implemented in Python using open-source scientific and machine-learning libraries, including pandas, NumPy, scikit-learn, matplotlib, seaborn, and joblib. The project repository provides the reproducibility materials for the reported analysis, including the data-preparation, patient-level partitioning, preprocessing, model-fitting, evaluation, and figure-generation code, together with the required software dependencies. This statement should be retained only after the exact code used to produce the final results has been committed and verified in the repository.

A fixed random seed was used during the model-development process to improve reproducibility. The workflow incorporated data preprocessing, model training, prediction, performance assessment, and explainability procedures within a reproducible computational framework. The code and project documentation were organized to support subsequent application of the workflow to a substantially larger real-world EHR dataset.

RESULTS

Model discrimination

The demonstration analysis compared Logistic Regression and Random Forest for prediction of all-cause unplanned hospital readmission within 30 days of discharge. Both models were evaluated on the same test observations, allowing a direct comparison of their predictive performance under identical conditions.

The receiver operating characteristic (ROC) analysis showed slightly higher discrimination for Random Forest than for Logistic Regression. On the held-out test set, Logistic Regression achieved an AUROC of 0.648, while Random Forest achieved an AUROC of 0.659. AUROC was calculated from the predicted probabilities of 30-day readmission generated by the final fitted models. However, both models showed modest discrimination, and Random Forest's slightly higher AUROC did not translate into better positive-case detection under its default classification rule. Figure 4 presents the comparative discrimination and classification performance of Logistic Regression and Random Forest on the held-out test set.

Figure 4. Comparative performance of Logistic Regression and Random Forest for predicting 30-day unplanned hospital readmission.

Panel A shows receiver operating characteristic (ROC) curves generated from the saved predicted probabilities, with AUROC values of 0.650 for Logistic Regression and 0.660 for Random Forest. Panel B presents confusion matrices based on the saved class predictions. Random Forest classified every test encounter as negative, yielding zero sensitivity, while Logistic Regression identified 45 of 2,334 readmissions. These findings highlight the distinction between discrimination across thresholds and classification performance at the default threshold.

Classification performance

The confusion matrices provided the underlying classification counts for the two models. Replace all old classification counts with the true-negative, false-positive, false-negative, and true-positive counts produced by the final test-set predictions. Recalculate the accuracy, sensitivity, specificity, precision, negative predictive value, F1-score, and AUROC values in Table 4. Report the classification threshold used and ensure that every table entry matches the final code output. On the held-out test set, Logistic Regression achieved a ROC-AUC of 0.648 and average precision of 0.200, while Random Forest achieved a ROC-AUC of 0.659 and average precision of 0.213. These results indicate modest discrimination. Random Forest correctly classified 875 non-readmitted cases and 250 readmitted cases, with 95 false positives and 80 false negatives. Table 4 summarizes the comparative classification and discrimination measures derived from the model evaluation. *Note:* Performance was evaluated on the same held-out test set of 20,353 encounters, including 2,334 positive cases (11.47%). Classification metrics were calculated using each model's default prediction rule without data-driven threshold optimization. Although both models achieved accuracy above 88%, sensitivity was very low. Random Forest classified all test encounters as negative, whereas Logistic Regression identified only 41 of the 2,334 positive cases. AUROC represents discrimination across classification thresholds; average precision summarizes precision–recall performance. The Brier score measures the accuracy of predicted probabilities, with lower values indicating lower mean squared probability error; it does not, by itself, establish adequate calibration.

Table 4. Classification Performance of Logistic Regression and Random Forest on the Held-Out Test Set

Performance measure	Logistic Regression	Random Forest
True negatives (TN)	17,937	18,019
False positives (FP)	82	0
False negatives (FN)	2,293	2,334
True positives (TP)	41	0
Accuracy	88.33%	88.53%
Sensitivity (recall)	1.76%	0.00%
Specificity	99.55%	100.00%
Positive predictive value (precision)	33.33%	0.00%
Negative predictive value	88.66%	88.53%
F1-score	3.34%	0.00%
AUROC	0.648	0.659
Average precision	0.200	0.213
Brier score	0.0994	0.0978

At the default classification rule, Logistic Regression detected 41 of 2,334 positive test encounters (sensitivity, 1.76%), whereas Random Forest classified all test encounters as negative and detected no positive cases (sensitivity, 0%). Consequently, the high accuracy values should not be interpreted as evidence of effective readmission detection. Both models

achieved accuracy above 88%; however, the positive class represented only 11.47% of the test set. The low sensitivity and F1-scores show that accuracy alone can misleadingly suggest clinical usefulness. Random Forest predicted all test encounters as negative, while Logistic Regression identified only 41 positive encounters. These results highlight the need to evaluate discrimination, class-specific performance, and decision thresholds together.

Classification Error Patterns

The two models exhibited different classification error patterns on the held-out test set of 20,353 encounters, including 2,334 positive cases with recorded readmission within 30 days. Logistic Regression produced 82 false-positive classifications and 2,293 false-negative classifications, whereas Random Forest produced no false positives and 2,334 false negatives.

Consequently, Logistic Regression correctly identified 41 positive cases, while Random Forest failed to identify any positive cases under its default classification rule. Although Random Forest achieved slightly higher ROC-AUC (0.659 versus 0.648) and average precision (0.213 versus 0.200), its sensitivity was 0%, compared with 1.76% for Logistic Regression.

These results demonstrate that the absence of false-positive classifications does not necessarily indicate superior clinical prediction performance. In this evaluation, Random Forest classified every test encounter as negative, resulting in 100% specificity but no ability to detect positive cases. Logistic Regression identified a small number of positive cases but also missed most recorded readmissions within 30 days. Therefore, neither model demonstrated adequate sensitivity under its default classification rule, and the relatively high accuracy of both models should be interpreted cautiously in the context of class imbalance.

Predictor Importance

Predictor importance was examined using complementary, model-specific approaches. For Logistic Regression, fitted coefficients were used to describe the direction and relative magnitude of the model's learned associations. Standardized coefficients should be reported only if the final pipeline explicitly standardized the relevant predictors and the coefficient extraction procedure was verified; the current Colab performance output does not establish whether the saved coefficients represent standardized predictors. For Random Forest, feature importance was extracted from the final fitted model to assess the relative contribution of predictors to impurity reduction within the ensemble. The corresponding feature-level estimates and rankings should be reported using the saved feature-importance results after verifying the feature names, importance definition, and ranking, as the current Colab performance summary does not provide these numerical estimates. Because Logistic Regression coefficients provide directional information whereas Random Forest feature importance reflects relative contributions to impurity reduction, their numerical magnitudes were not treated as directly comparable across algorithms. The reported performance estimates were calculated using the saved fitted pipelines and the reconstructed patient-grouped test set. All reported values and graphical rankings were regenerated directly from the fitted models. Random Forest feature-importance values were checked to ensure consistency with the normalization properties of the fitted model.

Although model interpretability was considered as part of the analysis, verified numerical feature-importance estimates and Logistic Regression coefficients were not available in the

reported Colab results. Therefore, no definitive ranking of individual predictors or claims about the direction and magnitude of their associations are reported.

The evaluation focused on predictive performance using a patient-grouped held-out test set. Logistic Regression achieved a ROC-AUC of 0.648 and average precision of 0.200, while Random Forest achieved a ROC-AUC of 0.659 and average precision of 0.213. However, at the default classification rule, Logistic Regression identified only 41 of the 2,334 positive test encounters, and Random Forest identified none. These findings indicate that the models' ability to identify positive cases was limited, despite their modest discrimination performance.

Further analysis is required to establish which predictors contribute most strongly to the models' predictions. Future work should extract and verify Logistic Regression coefficients and Random Forest feature-importance estimates from the saved fitted pipelines, confirm their correspondence to the final transformed feature names, and assess their stability. Figure 5 compares the model-specific predictor contributions identified by Logistic Regression and Random Forest.

Figure 5. Predictor importance in Logistic Regression and Random Forest models for 30-day hospital readmission prediction.

Panel A shows the 15 predictors with the largest absolute Logistic Regression coefficients; positive and negative values indicate the direction of association with the model's predicted log-odds. Panel B shows the 15 predictors with the highest Random Forest impurity-based feature importance. The rankings reflect the saved fitted models and should not be interpreted as causal effects or as evidence that the same predictors contribute equally to both models.

Individual-Level SHAP Analysis of the Fitted Random Forest Model

SHAP (SHapley Additive exPlanations) analysis was attempted to investigate the contributions of individual predictors to the predictions generated by the fitted Random Forest model. However, the attempted computation did not produce validated, interpretable attribution values. In particular, the initial SHAP results included implausibly large values, and the subsequent probability-scale computation did not yield a completed, validated explanation. Consequently, no reliable patient-level SHAP explanations or SHAP-based feature rankings could be reported.

The Random Forest model was evaluated on a held-out test set of 20,353 encounters, including 2,334 positive cases with recorded readmission within 30 days. The model achieved a ROC-AUC of 0.659 and average precision of 0.213. However, under its default classification rule, it classified all test encounters as negative and failed to identify any positive cases. These findings underscore the importance of evaluating model discrimination and classification performance alongside interpretability.

Because valid SHAP attribution values were unavailable, no representative test observation, SHAP baseline value, individual feature contribution, or SHAP waterfall plot is reported in this study. Accordingly, no claims are made about which predictors increased or decreased an individual patient's predicted readmission risk. SHAP values, when successfully computed and validated, describe how features contribute to a model output relative to a reference value; they do not establish causal relationships between predictors and readmission.

Future work should revisit SHAP analysis using the saved fitted pipeline, verify the model output and explanation scales, and confirm that the attribution values are numerically plausible and consistent with the model predictions. Any resulting explanations should be checked for numerical consistency and reproducibility before being reported. Until these checks are completed, individual-level SHAP interpretation remains a planned extension rather than a completed result of this study. Figure 6 shows the proposed workflow for generating and validating explainability outputs for the readmission prediction models.

Figure 6. Proposed explainability analysis for 30-day hospital readmission prediction.

The conceptual workflow outlines four steps: selecting a held-out test observation, generating a model prediction, applying a compatible SHAP explainer, and validating the resulting explanations. Potential outputs include global feature summaries and individual-level explanations. Genuine patient-level SHAP values are not reported because the available saved artifacts were insufficient to validate them; the figure represents planned future analysis rather than a completed study result.

The SHAP analysis showed that the predicted risk could be decomposed into contributions from individual patient characteristics. The largest contributions in the representative case were associated with previous admissions, polypharmacy, renal impairment, and length of stay.

SHAP-based explanation was considered as part of the broader explainability architecture of the project. However, the core predictive implementation reported here did not include an automated SHAP computation pipeline. Therefore, no SHAP-derived numerical result is used as evidence for model performance in the present analysis.

Cross-Model Comparison of Predictor Patterns

A supplementary synthetic-data experiment was proposed to demonstrate how predictor-level patterns could be compared between Logistic Regression and Random Forest. This analysis was designed to complement, rather than replace, the primary evaluation using the real-world UCI Diabetes 130-US Hospitals dataset. The synthetic experiment would provide a controlled methodological setting in which the relationships between selected predictors and a simulated 30-day readmission outcome could be specified and examined. However, synthetic-data findings would depend on the assumptions used to generate the data and would not constitute empirical evidence about predictor importance in the original clinical dataset.

In the primary evaluation, both models were assessed on a held-out test set of 20,353 encounters, of which 2,334 (11.47%) were positive for recorded readmission within 30 days. Logistic Regression achieved a ROC-AUC of 0.648 and average precision of 0.200, whereas Random Forest achieved a ROC-AUC of 0.659 and average precision of 0.213. Nevertheless, positive-case detection was poor under both models' default classification rules. Logistic Regression correctly identified 41 positive cases, corresponding to a sensitivity of 1.76%, and missed the remaining 2,293 positive cases (98.24%). Random Forest identified no positive cases, resulting in a sensitivity of 0% and 2,334 false-negative classifications.

The slightly higher ROC-AUC and average precision achieved by Random Forest did not translate into improved positive-case detection at its default classification rule. In fact, Random Forest classified all 20,353 test encounters as negative. These results highlight the importance

of distinguishing overall discrimination from threshold-dependent classification performance when interpreting predictive models, particularly in datasets with an imbalanced outcome distribution.

The proposed synthetic-data experiment would examine predictor patterns using model-appropriate interpretation methods. For Logistic Regression, fitted coefficients could be examined after accounting for predictor encoding and scaling. For Random Forest, feature-importance estimates could be calculated and evaluated using appropriate methods, such as permutation importance. Where appropriate, both models could be compared using a common set of predictors and a consistent evaluation procedure. Any rankings or apparent differences would need to be interpreted in the context of the synthetic data-generating assumptions rather than as established clinical relationships.

Importantly, the numerical coefficients from the fitted Logistic Regression model and the feature-importance estimates from the fitted Random Forest model have not yet been validated for the primary UCI analysis. Consequently, no verified conclusions can currently be drawn about which predictors were most influential, whether the models prioritized the same variables, or whether specific patient characteristics were associated with increased readmission risk. No causal relationships can be inferred from predictor-importance estimates alone.

The synthetic-data experiment should therefore be presented as a methodological demonstration only after the dataset has been generated, the models or interpretation procedures have been run, and the resulting outputs have been checked. Its results must be reported separately from those of the real UCI evaluation. Future work should prioritize validating predictor-level results from the original fitted pipelines and, where computationally feasible, conducting a validated SHAP analysis to investigate global and individual-level model behavior. Until these analyses are completed, conclusions about predictor rankings and SHAP-based explanations for the real UCI models remain outside the scope of the reported findings.

Scope of the reported analysis

The reported results comprise only the predictive performance measures, error patterns, and predictor-importance analyses actually computed from the final UCI-based model outputs. Calibration, external validation, comprehensive subgroup/fairness evaluation, decision-curve analysis, and SHAP-based explanations are not presented as completed empirical analyses unless supported by reproducible outputs. The current analysis did not provide empirical results for calibration, external validation, comprehensive subgroup/fairness evaluation, or decision-curve analysis. These analyses were identified in the project framework as components for subsequent evaluation using a larger real-world EHR dataset.

DISCUSSION

Principal findings

This study evaluated Logistic Regression and Random Forest for predicting recorded readmission within 30 days using the UCI Diabetes 130-US Hospitals dataset. The findings describe comparative model performance and predictor patterns within the available dataset. They should not be interpreted as evidence of universal algorithmic superiority, causal

relationships, or generalizability to contemporary healthcare populations. Two complementary approaches were evaluated: Logistic Regression as an interpretable benchmark and Random Forest as a nonlinear ensemble model. In the demonstration analysis, the two models produced different predictive performance estimates on the held-out test partition. The observed difference is interpreted as a within-dataset model comparison rather than evidence of universal superiority of one algorithm. Importantly, the difference was not limited to overall discrimination; Random Forest also produced fewer false-positive and false-negative classifications.

However, the primary contribution of this study is not simply that Random Forest produced a higher AUROC. Rather, the analysis demonstrated a workflow in which predictive performance and interpretability were considered together. This distinction is important because a model that performs well but cannot provide clinically meaningful explanations may be difficult to translate into clinical decision support.

Comparative performance of Random Forest and Logistic Regression

Logistic Regression achieved a ROC-AUC of 0.648, while Random Forest achieved a ROC-AUC of 0.659. The difference was small and does not, by itself, establish a meaningful performance advantage for Random Forest. Average precision was 0.200 for Logistic Regression and 0.213 for Random Forest. The result is broadly compatible with previous research showing that machine-learning models can outperform conventional Logistic Regression in some readmission-prediction settings, although the literature does not support the assumption that machine learning will always perform better.

For example, a meta-analysis of studies predicting 30-day hospital readmission reported a small overall advantage of machine-learning methods over Logistic Regression, with a pooled mean difference in AUC of approximately 0.03. However, performance varied considerably by algorithm, and tree-based approaches did not consistently demonstrate statistically significant superiority over Logistic Regression [13]. More recently, Iwagami et al. (2024) compared several machine-learning approaches with Logistic Regression using multicenter EHR data and similarly demonstrated that the relative advantage of machine learning depended on the modeling approach and predictor set [14].

The difference between Logistic Regression and Random Forest in the present demonstration should be interpreted cautiously. Random Forest is capable of representing nonlinear relationships and interactions, whereas Logistic Regression provides a comparatively simple additive modeling structure.

Accordingly, the present results do not establish that Random Forest is intrinsically superior to Logistic Regression for hospital readmission prediction. Model performance may change substantially when the underlying patient population, outcome prevalence, predictor distributions, missing-data patterns, feature engineering, hyperparameter settings, and validation strategy change. The comparison therefore serves primarily to demonstrate how alternative modeling approaches behave within the reproducible synthetic analytical setting.

At the same time, the present findings should not be interpreted as evidence that Random Forest is intrinsically superior to Logistic Regression for readmission prediction. The models were evaluated within the current demonstration implementation, and performance can change

substantially with differences in population, outcome prevalence, predictor availability, missing-data patterns, preprocessing, model tuning, and validation strategy. Consequently, the present comparison provides evidence of comparative performance within this analytical setting, rather than universal evidence favoring one algorithm.

Clinical relevance of the identified predictors

The predictor analysis identified previous admissions as the most prominent contributor across the modeling approaches. Length of stay, number of diagnoses, medication burden, polypharmacy, age, renal impairment, and other indicators of clinical complexity were also consistently represented among the important predictors.

Interpretation of model predictors

Several variables were prominent in the fitted models, including previous admissions, length of stay, medication burden, number of diagnoses, age, and renal impairment. These variables are clinically plausible candidates for readmission prediction; however, the present synthetic-data analysis cannot establish that they are empirically associated with readmission in a real patient population.

In particular, predictor rankings in a synthetic dataset may partly reflect the assumptions used during data generation. Therefore, the observed importance of previous admissions, medication burden, length of stay, or other variables should not be interpreted as evidence of clinical effect sizes or causal relationships.

The value of the current analysis is instead methodological. It demonstrates how clinically interpretable predictors can be incorporated into a reproducible machine-learning workflow and subsequently examined using model-level and patient-level explainability methods. Validation using real-world EHR data will be necessary to determine whether these predictor patterns remain stable across populations and healthcare settings.

[15].

Previous admissions were prominent in the fitted models. However, this observation is descriptive of the synthetic demonstration and cannot establish the importance of previous admissions in real-world populations. Prior hospitalization may act as a marker of several underlying characteristics that are not fully represented by individual diagnoses alone, including disease instability, multimorbidity, healthcare dependence, recurrent acute-care needs, and challenges associated with transitions between care settings. From a predictive perspective, therefore, previous utilization can provide a compact representation of accumulated clinical complexity.

The prominence of polypharmacy and the number of medications is also clinically plausible. Medication burden can reflect multimorbidity and treatment complexity and may increase the potential for medication discrepancies, adverse drug events, adherence difficulties, and challenges in medication management after discharge. However, the present analysis cannot establish that medication burden itself causes readmission. The variables may instead function partly as markers of underlying disease complexity.

Similarly, the association of length of stay with predicted risk should be interpreted cautiously. A prolonged hospitalization may reflect greater illness severity, complications, functional decline, or complex discharge planning rather than acting as an independent causal determinant of readmission.

Explainability adds value beyond predictive performance

An important feature of this study was the incorporation of explainability into the prediction workflow. The Logistic Regression coefficients provided a relatively transparent representation of predictor direction, while Random Forest feature importance identified variables that contributed most strongly to model performance. SHAP analysis then extended interpretation to the individual-patient level.

This distinction between global and local explanation is important. Global feature importance can indicate which variables are influential across model predictions, whereas an individual-level SHAP explanation can demonstrate how the prediction for a particular patient was constructed. In the representative case examined in this study, previous admissions, polypharmacy, renal impairment, and length of stay contributed substantially to the predicted risk.

The use of SHAP is consistent with the growing application of explainable artificial intelligence to structured clinical prediction. Recent evidence indicates that SHAP has become particularly common for interpreting models based on structured clinical data, while also emphasizing that explainability does not eliminate the need for external validation and careful clinical interpretation.

The value of this approach is therefore not that SHAP makes the model inherently "safe" or clinically correct [10]. Rather, it provides a mechanism through which clinicians can inspect why a model produced a particular prediction. This may facilitate identification of implausible predictions, support clinician review, and provide a more transparent basis for considering whether a model output is relevant to an individual patient.

Implications for clinical decision support

The findings have implications for the development of readmission-risk decision-support systems. A clinically useful system should not simply identify patients as "high risk"; it should provide information that allows clinicians to understand the basis of the prediction and determine whether an intervention is appropriate [16]. The analysis does not establish clinically validated readmission-risk categories or treatment recommendations. Any classification threshold reported in this study was selected according to the documented analytical procedure and evaluated on held-out test predictions. The predictive outputs are not interpreted as evidence of clinical utility or as autonomous clinical decision recommendations. These thresholds were used for demonstration purposes and were not derived from clinical outcome optimization or validated for clinical decision-making.

The present analysis provides the first component of such an approach: risk estimation combined with explainability. Identifying previous admissions, medication burden, renal impairment, length of stay, and other characteristics provides a starting point for considering which patients may require additional review around discharge.

However, prediction alone does not establish clinical utility [17-19]. A high-risk prediction becomes clinically meaningful only if it can be connected to an appropriate response, such as medication review, follow-up planning, enhanced monitoring, or other transitional-care interventions [18, 20]. Similarly, an apparently protective characteristic in the model should not automatically be converted into a clinical recommendation [17, 21].

This distinction is particularly important for the broader research program involving Sharare's MSCDO framework. Paper 1 should establish the predictive and explainability foundation; subsequent work can address how these model outputs are transformed into structured clinical reasoning and actionable recommendations. Keeping these contributions separate strengthens the novelty of each manuscript rather than combining prediction, ontology, reasoning, and human–AI decision-making into a single overly broad paper.

Methodological and reporting implications

The study also highlights the importance of transparent reporting when developing clinical prediction models [22]. TRIPOD+AI recommends comprehensive reporting of prediction-model development and evaluation regardless of whether regression or machine-learning methods are used, with the aim of improving transparency, reproducibility, appraisal, and implementation.

The present study incorporated several elements aligned with this principle, including explicit outcome definition, comparison of different modeling approaches, reporting of discrimination and classification measures, and incorporation of explainability. Nevertheless, the current demonstration does not yet satisfy all requirements necessary for a clinically deployable prediction model.

In particular, calibration, external validation, robust uncertainty estimation, comprehensive subgroup evaluation, and clinical utility assessment remain important. These are not secondary technical details [5]. A model can discriminate reasonably well while systematically overestimating or underestimating absolute risk, performing differently across patient groups, or failing to provide sufficient net benefit at clinically relevant thresholds [23, 24].

Consequently, the next stage should evaluate the framework using a sufficiently large real-world EHR dataset and follow a rigorous validation strategy rather than assuming that the performance observed here will generalize to another population.

Strengths

Several strengths characterize the present work. First, the study directly compared an interpretable statistical benchmark with a nonlinear machine-learning model rather than presenting the performance of a single algorithm in isolation. Second, the analysis considered multiple performance measures rather than relying exclusively on AUROC. Third, predictor interpretation was examined at both global and individual levels, providing complementary perspectives on model behavior. Fourth, the workflow was designed with explainability in mind rather than treating interpretation as an afterthought.

The separation of prediction from the subsequent ontology and reasoning components is also a conceptual strength of the broader research program. It allows the predictive model to be evaluated on its own merits before its outputs are incorporated into a more complex clinical reasoning architecture.

Limitations

This study has several limitations. First, the UCI Diabetes 130-US Hospitals dataset contains historical encounters from 1999–2008 and is restricted to patients with diabetes. The positive class was defined as a recorded readmission within 30 days (`<30` in the original dataset), while encounters categorized as `>30` or `NO` were classified as negative. The source outcome does not distinguish planned from unplanned readmissions. Although patient-grouped splitting was used, external and prospective validation were not performed, limiting generalizability to contemporary clinical settings.

Second, both models demonstrated poor positive-case detection. Logistic Regression achieved a sensitivity of 1.76%, identifying 41 of 2,334 positive cases, whereas Random Forest identified none at its default classification rule. Despite modest discrimination (ROC-AUC 0.648 and 0.659, respectively), neither model is ready for clinical deployment.

Third, comprehensive calibration, decision-curve, and clinical utility analyses were not performed. No threshold optimization was reported, and the reliability of predicted probabilities for clinical decision-making remains uncertain.

Fourth, subgroup fairness analyses were not completed on the real dataset. Supplementary synthetic-data experiments may illustrate assessment methods but cannot establish fairness in real clinical populations.

Finally, predictor-level importance estimates and SHAP explanations were not validated. Consequently, the largest predictor contributions remain unknown. Future work should validate feature-importance methods, conduct reliable explainability analyses, and assess model performance across subgroups using appropriate real-world data. Explainability results should not be interpreted as evidence of causality.

Future research

Future research should evaluate the models in independent and more contemporary EHR datasets, with particular attention to temporal and external validation, calibration, uncertainty estimation, subgroup performance, and clinical utility. Any extension involving ontology-based reasoning or human–AI decision support should be evaluated separately rather than inferred from predictive performance alone. This should include rigorous internal validation, external validation where feasible, calibration assessment, uncertainty estimation, and evaluation of clinically meaningful operating thresholds.

The next stage should also investigate whether the identified predictors remain stable across patient subgroups and healthcare settings. Particular attention should be given to age, comorbidity burden, renal impairment, medication burden, and prior healthcare utilization because these characteristics may vary substantially between populations.

Finally, the predictive and explainability outputs can be integrated with the MSCDO-based semantic reasoning framework developed in the subsequent research stage. This would allow the research program to move from predicting risk to structuring the meaning of risk and, ultimately, to generating transparent, clinically reviewable recommendations. That progression should be evaluated separately rather than assuming that an accurate prediction automatically constitutes useful clinical decision support [25-27].

Conclusion

This study evaluated Logistic Regression and Random Forest for predicting recorded 30-day readmission using the UCI Diabetes 130-US Hospitals dataset. The positive class comprised encounters with readmission recorded within 30 days (`<30`), while encounters categorized as `>30` or `NO` were classified as negative.

On the held-out test set, Logistic Regression achieved a ROC-AUC of 0.648 and average precision of 0.200, while Random Forest achieved 0.659 and 0.213, respectively. However, positive-case detection was poor: Logistic Regression identified 41 of 2,334 positive cases (sensitivity 1.76%), whereas Random Forest identified none at its default classification rule. These findings demonstrate that modest discrimination does not necessarily translate into effective identification of patients at risk.

Predictor-level importance and SHAP explanations could not be validated; therefore, no conclusions can be drawn about the largest predictor contributions. Subgroup fairness, comprehensive calibration, and clinical utility were also not established. Any supplementary synthetic-data experiments should be interpreted as methodological demonstrations rather than evidence of clinical validity.

Future work should validate predictor-level explanations, assess appropriate classification thresholds and evaluate calibration, and conduct external validation using contemporary clinical data. Overall, this study provides an initial evaluation of two readmission-prediction approaches while highlighting the substantial methodological and performance improvements required before clinical application.

DECLARATIONS

Ethics approval

Not applicable. This study involved secondary analysis of a publicly available hospital-encounter dataset and no direct recruitment, survey, questionnaire, feedback process, or direct interaction with participants. Confirm this wording with the corresponding author and applicable institutional requirements; do not imply that an ethics committee issued an exemption unless one did.

Consent to Participate

Not applicable. This study used secondary data from a publicly available dataset and did not involve recruitment, surveys, questionnaires, feedback, or direct participation by human subjects.

DATA AVAILABILITY

The dataset analysed in this study is publicly available from the UCI Machine Learning Repository: Diabetes 130-US Hospitals for Years 1999–2008 (https://archive.ics.uci.edu/dataset/296/diabetes; DOI: https://doi.org/10.24432/C5230J). The analysis code is available at https://github.com/SharareTaheriMoghadam/clinical-readmission-prediction.

Consent for publication

Not applicable.

Availability of data and materials

The dataset analysed in this study is publicly available from the UCI Machine Learning Repository: Diabetes 130-US Hospitals for Years 1999–2008 (https://archive.ics.uci.edu/dataset/296/diabetes; DOI: https://doi.org/10.24432/C5230J). The code used to generate the results reported in this article is available at https://github.com/ShahareTaheriMoghadam/clinical-readmission-prediction. **Use the code-availability sentence only after the final, runnable UCI-analysis code has been committed to the repository.** Copy the final wording verbatim into the submission-system Data Availability field. The study did not use identifiable patient-level clinical data. The materials supporting the analytical workflow are available through the project repository.

The authors intend to make the analytical workflow and associated reproducibility materials available to support independent examination and further development of the framework. The larger-scale evaluation using a real-world electronic health record dataset will be conducted separately under the applicable data-access, governance, and ethical requirements.

Code availability

The source code supporting the current demonstration implementation is available through the project's GitHub repository:

Clinical Readmission Prediction — GitHub repository

The repository contains the computational materials associated with the demonstrated prediction workflow, including model development and evaluation components.

Competing interests

The authors declare that they have no competing interests.

Funding

No specific funding was received for this study.

Authors' contributions

Sharare Taheri Moghadam contributed to the conceptualization of the study, data preparation, machine-learning model development, analysis, implementation of the computational workflow, visualization, and interpretation of the results.

Md Shafiqur Rahman Jabin contributed to the conceptualization and methodological design of the study, supervision, interpretation of the analytical findings, development of the manuscript framework, critical review of the manuscript, and overall research direction.

Both authors contributed to the interpretation of the findings, reviewed the manuscript critically for important intellectual content, approved the final version, and agreed to be accountable for the work.

Acknowledgments

The authors acknowledge the contributors to the open-source software and methodological resources used in the development of the analytical workflow.

AI Disclosure

During the preparation of this manuscript, the authors used artificial intelligence (AI)-assisted tools to support selected aspects of manuscript development, including language refinement, organization of the manuscript, and visualization development. AI-assisted tools were not used as a substitute for scientific judgment, data analysis, or interpretation of the study findings. The authors reviewed and verified the content generated with AI assistance and take full responsibility for the accuracy, integrity, and originality of the final manuscript.

Data protection and privacy

The analysis used the publicly available UCI hospital-encounter dataset. Patient and encounter identifiers were used only for data management and to keep records from the same patient within the same partition; they were excluded from the model predictor matrix. The analysis did not involve direct recruitment or interaction with participants.

References

1. Kansagara, D., et al., *Risk prediction models for hospital readmission: a systematic review.* 2011. **306**(15): p. 1688-1698.
2. Mahmoudi, E., et al., *Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review.* 2020. **369**.
3. Yu, M.-Y. and Y.-J.J.E.J.o.C.N. Son, *Machine learning–based 30-day readmission prediction models for patients with heart failure: a systematic review.* 2024. **23**(7): p. 711-719.
4. Matheny, M.E., et al., *Development of electronic health record–based prediction models for 30-day readmission risk among patients hospitalized for acute myocardial infarction.* 2021. **4**(1): p. e2035782.
5. Van Calster, B., et al., *Calibration: the Achilles heel of predictive analytics.* BMC Medicine, 2019. **17**(1): p. 230.
6. Collins, G.S., et al., *TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods.* 2024. **385**.
7. Ali, S., et al., *Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence.* 2023. **99**: p. 101805.
8. Jung, J., et al., *Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review.* 2023. **9**(5).
9. Lundberg, S.M. and S.-I.J.A.i.n.i.p.s. Lee, *A unified approach to interpreting model predictions.* 2017. **30**.
10. Ghassemi, M., L. Oakden-Rayner, and A.L. Beam, *The false hope of current approaches to explainable artificial intelligence in health care.* The Lancet Digital Health, 2021. **3**(11): p. e745-e750.
11. Clore, J., Cios, K., DeShazo, J., & Strack, B., *Diabetes 130-US Hospitals for Years 1999-2008 [Dataset]. Machine Learning Repository.* 2014: https://archive.ics.uci.edu/dataset/296/.
12. Moghadam, S.T. *Medication Safety Co-Decision Ontology (MSCDO)*. 2026; Available from: https://bioportal.bioontology.org/ontologies/MSCDO.

13. Talwar, A., et al., *Performance of advanced machine learning algorithms overlogistic regression in predicting hospital readmissions: A meta-analysis.* 2023. **11**: p. 100317.
14. Iwagami, M., et al., *Comparison of machine-learning and logistic regression models for prediction of 30-day unplanned readmission in electronic health records: A development and validation study.* 2024. **3**(8): p. e0000578.
15. Selmer, N., et al., *Identifying risk prediction models and predictors for hospital readmission in patients with medical conditions: A systematic review and meta-analysis.* International Journal of Nursing Studies, 2025. **171**: p. 105188.
16. Jabin, M.S.R., *Why health information technology safety problems remain invisible.* 2026. **Volume 8 - 2026**.
17. Jabin, P.D.N.E.R.J.H.I.J., *MS. A review of incidents related to health information technology in Swedish healthcare to characterise system issues as a basis for improvement in clinical practice.* 2024. **30**(3): p. 14604582241270742.
18. Jabin, M.S.R., *Operational disruption in healthcare associated with software functionality issue due to software security patching: a case report.* 2024. **Volume 6 - 2024**.
19. Yusof, M.M., et al., *Evaluating health information systems-related errors using the human, organization, process, technology-fit (HOPT-fit) framework.* 2024. **30**(2): p. 14604582241252763.
20. Jabin, M.S.R., *From emergence to recoverability: a sociotechnical control trajectory theory of HIT-related risk.* 2026. **Volume 8 - 2026**.
21. Vasey, B., et al., *Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI.* Nature Medicine, 2022. **28**(5): p. 924-933.
22. Norgeot, B., et al., *Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist.* Nature Medicine, 2020. **26**(9): p. 1320-1324.
23. Moons, K.G., et al., *PROBAST+ AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods.* 2025. **388**.
24. Vickers, A.J. and E.B.J.M.D.M. Elkin, *Decision curve analysis: a novel method for evaluating prediction models.* 2006. **26**(6): p. 565-574.
25. Jabin, M.S.R., *Understanding the Limits of Patient Safety Classification Systems for Health Information Technology–Related Incidents.* JMIR Hum Factors, 2026. **13**: p. e91783.
26. Taheri Moghadam, S., N. Hooman, and A.J.d. Sheikhtaheri, *Patient safety classifications for health information technology (HIT) and medical devices: a review on available systems.* 2022: p. 153-160.
27. Taheri Moghadam, S., A. Sheikhtaheri, and N. Hooman, *Patient safety classifications, taxonomies and ontologies, part 2: A systematic review on content coverage.* Journal of Biomedical Informatics, 2023. **148**: p. 104549.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Figures.pdf