

Beyond AUROC: External Validation, Calibration, and Explainability of Machine Learning Models for 30-Day Hospital Readmission Using Electronic Health Records—A Rapid Systematic Review

Md Eftakhar Ahamed

eftakharahamed52@gmail.com

Chittagong University of Engineering and Technology

Md. Istiakul Alam

Chittagong University of Engineering & Technology

Mohammad Shamsul Arefin

Chittagong University of Engineering & Technology

Systematic Review

Keywords: PRISMA 2020, Rapid Systematic Review, Hospital Readmission, Machine Learning, Electronic Health Records, External Validation, Calibration, Explainable AI

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11417231/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Beyond AUROC: External Validation, Calibration, and Explainability of Machine Learning Models for 30-Day Hospital Readmission Using Electronic Health Records—A Rapid Systematic Review

Md Eftakhar Ahamed^{1*}, Md. Istiakul Alam¹ and
Mohammad Shamsul Arefin¹

^{1*}Chittagong University of Engineering and Technology, Chittagong,
Bangladesh.

*Corresponding author(s). E-mail(s): eftakharahamed52@gmail.com;
Contributing authors: sagaristiak827@gmail.com; sarefincse@gmail.com;

Abstract

Studies evaluating machine learning (ML) algorithms predicting 30-day hospital readmission using EHRs have relied on AUROC as a sole criterion. The AUROC score provides no insight into whether an ML model transfers to new hospitals or periods, whether the probabilities provided are calibrated, or whether the explanation offered can be trusted. This study was conducted as a rapid systematic review using streamlined methods informed by Cochrane rapid review guidance and reported according to the applicable PRISMA 2020 reporting items. It examined the extent to which EHR-based machine learning studies predicting 30-day hospital readmission reported external or temporal validation, calibration, explainability, clinical utility, and subgroup fairness. PubMed and Europe PMC databases were searched for studies conducted between January 2020 and October 2026, resulting in the screening of 1,644 records in two phases and inclusion of 156 studies coded based on their abstracts. Out of 62 studies providing an AUROC score for readmission as the only outcome, the median highest AUROC score was 0.76 (interquartile range 0.69–0.85). Only 19 studies (12.2%) included external (10) or temporal (10; one both) validation, and AUROC decreased in four of five studies that evaluated both settings. Calibration was mentioned in 24 abstracts (15.4%) but only one study used a metric specific to calibration; decision curve analysis appeared in 6 (3.8%) and fairness analysis in 7 (4.5%), and 112 studies (71.8%) did not report any of these four evaluations. Sixty abstracts (38.5%) reported explainability, mainly via SHAP, but no one evaluated the

explanations. From 2020-2023 to 2024-2026, SHAP usage increased from 5.8% to 29.9% ($p < 0.001$), while external or temporal validation remained unchanged at 11.6% and 12.6% ($p = 1.0$). The rate of adoption of post-hoc explanation has been much higher than the rate of adopting these evaluations for deciding about the safety of a model. We propose a minimal “beyond AUROC” evaluation set for readmission prediction models.

Keywords: PRISMA 2020, Rapid Systematic Review, Hospital Readmission, Machine Learning, Electronic Health Records, External Validation, Calibration, Explainable AI

1 Introduction

A 30-day unplanned readmission after discharge is a very common benchmark of the quality of care and one of the biggest cost drivers in hospitals. In the US, there are financial incentives attached to the Hospital Readmissions Reduction Program [1] and in other places, it is a quality metric tracked by health organizations. Hence, it makes sense for hospitals to find out which patients are at higher risk and allocate scarce transitional-care services to them.

Traditional scores for readmission predict poorly. In 2011, a systematic review of 26 different scoring systems concluded that the majority exhibited poor discrimination [2]. With the increased use of EHRs and ML techniques, we could hope for better performance, as EHRs have laboratory results, drugs, vital signs and clinical notes that are not considered by administrative scores, and ML methods can model complex interaction between multiple predictors. This hope is less supported by data than often thought. According to a systematic review from 2020 of 41 studies using EHRs, the ML models had an average C statistic of 0.74 vs. 0.71 for regression models, which was not significantly different; all studies used internal validation and just 15 models were calibrated [3]. Also, a scoping review from 2021 of 43 ML studies found a median AUROC of 0.68 [4].

Like primary studies, these reviews tend to evaluate models primarily using AUROC (which is a measure of model ability to rank-order the patients’ risks). AUROC says nothing about three key properties that determine whether a model will be practical in terms of safety and utility: *transportability* to another hospital/period/patient cohort; *calibration* (i.e., agreement between predicted risks and observed readmissions rates) [7]; and *explainability* (clinicians’ ability to understand why a certain patient was assigned a risk score and the quality of these explanations) [8]. In clinical prediction studies, machine learning does not outperform logistic regression in terms of AUROC in low-bias studies, and 79% of studies do not report calibration [9]. The new TRIPOD+AI reporting guideline calls for reporting discrimination, calibration, explainability, and fairness [10], and the PROBAST+AI risk-of-bias tool calls for evaluating these dimensions as well [11]. No recent review has looked into the reporting of these dimensions among ML readmission prediction studies based on EHR data.

The main contributions made by this rapid systematic review are as follows:

1. A structured, PRISMA 2020-informed evidence synthesis of 156 EHR-based machine learning studies on 30-day hospital readmission published between 2020 and 2026.
2. A quantification of how often these studies report external or temporal validation, calibration, explainability, clinical utility and subgroup fairness, and how discrimination changes in a new setting.
3. A summary of model families, public datasets and clinical populations studied, with reporting broken down by population.
4. A tested comparison of reporting between 2020–2023 and 2024–2026.
5. A minimum “beyond AUROC” reporting set for readmission models, aligned with TRIPOD+AI and PROBAST+AI.
6. Practical guidance for researchers, hospitals and editors on evaluating readmission models before deployment.

2 Background

Readmission prediction is one of the most frequent use cases of ML on EHR data. This part outlines the clinical question, reviews the process of constructing ML models to solve it, and discusses the evaluation properties analyzed in this work (Fig. 1).

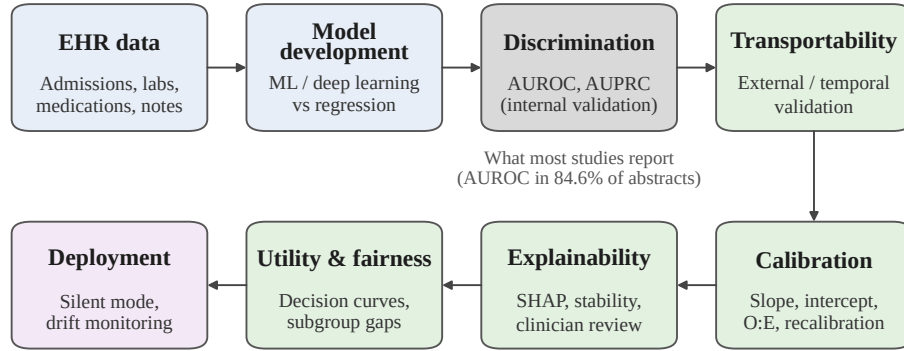
2.1 Hospital Readmission and Risk Prediction

Readmission is defined as re-admittance into the hospital after a given time period post discharge, usually 30 days. Some readmissions are planned (staged surgical operations), and thus not used for quality improvement purposes, while others are unplanned and partially preventable by improved discharge planning, follow-ups, and medication management. Risk prediction models can be applied to identify cases suitable for intervention. Traditional risk scores like LACE and HOSPITAL employ a few variables available at discharge time and are easy to calculate, but have limited discriminative ability [2].

2.2 Machine Learning for Readmission Prediction

The models in ML include several times more variables than what is used in LR models, including diagnoses, procedures, labs, drugs, vitals, prior utilisation and free text. The common ML model classes include tree ensemble (random forest, gradient boosted trees like XGBoost and LightGBM), neural networks, recurrent neural networks for longitudinal data, and, most recently, transformer language models on discharge summaries. Penalised logistic regression (such as LASSO) still makes a good benchmark [9]. The development and validation datasets come from the same institution and are randomly split.

Evaluating a readmission prediction model: AUROC and beyond



"Beyond AUROC" evaluations (green boxes): each reported in only 3.8%–15.4% of abstracts;
 112 of 156 studies (71.8%) reported none of validation, calibration, decision curves or fairness.
 Explainability was claimed in 38.5% of abstracts, but no abstract reported evaluating the explanations.

Fig. 1 Evaluation of a readmission prediction model: AUROC and the “beyond AUROC” properties examined in this review.

2.3 Evaluation Beyond AUROC

2.3.1 Discrimination

AUROC is the likelihood that a patient who was readmitted gets assigned a higher predicted probability of readmission than a patient who was not readmitted. It is robust to prevalence of outcomes and to transformations of predicted probabilities that preserve monotonicity, so two models having the same AUROC can have vastly different absolute probabilities. Readmission is uncommon (usually 10–20% of discharges), so the area under the precision–recall curve (AUPRC) is usually more useful in understanding the number of flagged patients who will be readmitted.

2.3.2 Transportability

Internal validation, using methods such as random split, cross-validation, or bootstrapping, measures performance for future patients at the same hospital during the same time frame. Internal validation does not measure performance for situations where there are changes to the patient population mix, coding schemes, documentation procedures, discharge procedures, and community services. External validation involves testing the model using data from another hospital or health care delivery system, while temporal validation involves testing the model using data collected after the period used to develop the model.

2.3.3 Calibration

Calibration refers to the consistency of risk predictions and actual risks. It is normally quantified using the calibration slope (which should ideally be 1) and intercept (ideally 0) from a calibration plot or O/E ratio [7]. The Hosmer-Lemeshow test has low power and the Brier score confounds discrimination and calibration; therefore, neither is adequate for calibration alone. Calibration problems may frequently be solved by recalibration if there is sufficient local data [12].

2.3.4 Explainability

Post hoc feature explanation techniques, for example, SHapley Additive exPlanations (SHAP) [13], explain predictions by assigning importance values to input features. They explain the model and not the factors leading to re-admission. They also tend to be unstable to bootstrapping [8].

2.3.5 Clinical Utility

The decision curve analysis (DCA) calculates the net benefit of taking an action based on model predictions at different threshold probabilities, considering both true and false positives [14]. It determines if applying the model is advantageous to intervention for all or none of the patients.

2.3.6 Fairness

Fairness assessment is done on discrimination, calibration, or errors across different demographic groups of gender, age, race, ethnicity, insurance status, or deprivation. The readmission category depends to an extent on access to care, meaning that the algorithm could learn bias from the outcome [15, 16].

3 Methodology

This study was performed as a rapid systematic review using simplified methodology, which was based on the guidelines set forth by the Cochrane Rapid Reviews Methods Group [17]. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guideline [18] was followed during the reporting process. Methodological adaptations included searching through two free-to-access databases, following the pre-defined keywords-based screening criteria, performing one-reviewer-only screening and extracting the data mainly from titles and abstracts. No full-text screening and risk-of-bias assessments were performed systematically. Such adaptations limited the comprehensiveness of the review and were taken into account while analyzing the results.

The following is the methodology that was used in this review:

1. Defining the review protocol
2. Formulating the research questions
3. Searching the information sources
4. Screening and eligibility assessment

5. Reporting appraisal
6. Data extraction and coding
7. Descriptive synthesis and statistical comparison

3.1 Review Protocol

Prior to conducting the search, the aims of the review, the sources of information, inclusion/exclusion criteria, stage-1 screening criteria, and coding criteria were determined. The review protocol was not registered. The aim of this rapid systematic review was to explore the primary literature concerning development and/or validation of machine learning prediction models for 30-day hospital readmissions from inpatients in adults using EHRs.

3.2 Research Questions

To achieve the objectives of this review, the following research questions (RQs) were formulated.

RQ1. How often are models externally or temporally validated, and how does discrimination change in a new setting?

RQ2. How often, and how, is calibration reported?

RQ3. Which explainability methods are used, and are the explanations evaluated?

RQ4. Are clinical utility and subgroup fairness assessed?

RQ5. Has reporting of these properties changed between 2020–2023 and 2024–2026?

3.3 Information Sources

Two freely accessible electronic databases were searched on 9–10 October 2026:

1. PubMed/MEDLINE
2. Europe PMC (which also indexes preprint servers)

The inclusion criteria for studies published between 1 January 2020 and the search date applied; the most recent date of a PubMed entry amongst the exported papers is 9 October 2026. The year of eligibility for each paper depended on the year of publication stated by each database such that preprints were considered as being in the same year. The first year was selected following the review of EHR-based readmission models in 2020 [3].

3.4 Search Strategy

The search combined free-text terms and controlled vocabulary (MeSH) with Boolean operators, as illustrated in Fig. 2. The search terms were separated into three concept blocks:

1. Readmission
2. Machine learning
3. Prediction or risk modelling

The terms in each block were combined with OR and the blocks were combined with AND. The search details, record counts and the concept blocks are provided in Supplementary Table S3; the exact query strings were not saved. The records from both databases were merged and de-duplicated by DOI, PMID and normalized titles.

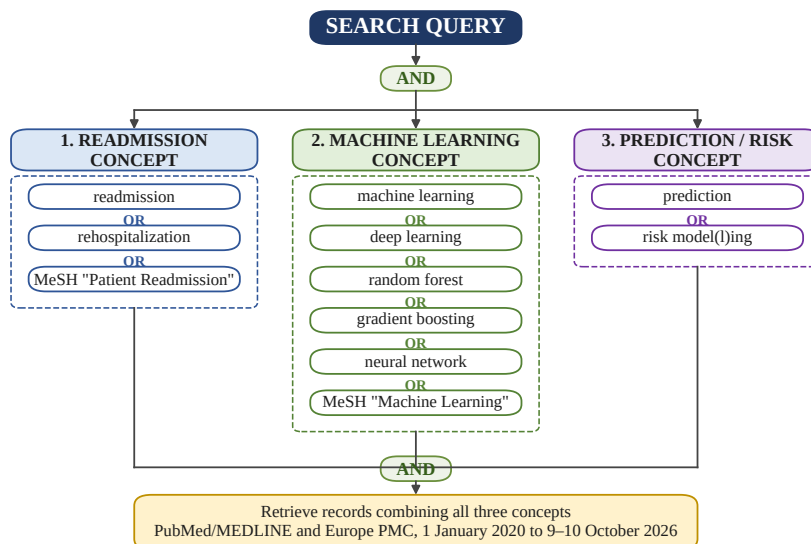


Fig. 2 Search query framework.

3.5 Eligibility Criteria

Eligibility criteria were defined before screening to ensure consistency and reduce selection bias.

Inclusion Criteria

Studies that satisfied all of the following criteria were included:

- Adult hospital inpatients, any clinical specialty.
- Developed or validated at least one ML model (tree ensembles, neural networks, support vector machines, other non-parametric learners or language models).
- Predicted all-cause or unplanned readmission within 30 days of discharge.
- Used hospital EHR data, clinical data warehouses or EHR-derived research databases (e.g., MIMIC, Diabetes 130-US Hospitals).
- Primary prediction studies published from 2020 to 2026 as journal articles or stand-alone preprints.

Exclusion Criteria

A study was excluded if it met any of the following criteria:

- Paediatric, psychiatric-only or ICU-readmission-only cohorts.

- Regression-only models, including penalised regression (e.g., LASSO), which was treated as a comparator, not as ML.
- Other prediction horizons, composite outcomes, or abstracts not stating the 30-day horizon.
- Claims, administrative or quality-registry data (e.g., NSQIP, MBSAQIP, Nationwide Readmissions Database).
- Association or implementation studies; reviews, protocols, conference abstracts, and preprints of included articles.

Studies using ICU-derived databases such as MIMIC were included when the abstract described readmission to hospital within 30 days of discharge rather than readmission to the ICU, and readmission defined by an emergency-department admission (the usual route for unplanned readmission) was accepted as unplanned readmission.

3.6 Study Selection

The study selection process followed the four phases of the PRISMA 2020 framework:

1. Identification
2. Screening
3. Eligibility assessment
4. Inclusion

Screening involved two rounds, both based on titles and abstracts. Stage 1 consisted of pre-defined keyword-based exclusion rules identifying those articles that lacked information regarding the readmission outcome, prediction purpose, machine learning (ML) model, 30-day period, use of claims or registry data, and primary research report. To verify the accuracy of the rules, a sample of 100 records excluded at stage 1 was hand-checked: none of these were eligible, one had a wrong exclusion reason, and nine of these were excluded because of a lack of information about the readmission period. Hand-checking of exclusions specifically for missing ML terms and incorrect publication type resulted in the identification of two additional eligible articles [19, 20] that were included in stage 2 of the review process. In stage 2, all of the remaining articles were hand-screened based on the inclusion criteria and one reason per exclusion was provided.

The flow of the selection process is illustrated in Fig. 3. The searches produced **1,644** records (PubMed 1,247; Europe PMC 397). Excluding **359** duplicates, **1,285** records were screened. The stage-1 criteria rejected 1,006 records; however, after the review stage 1 filtered out **1,004**, and passed **281**. Of those stage-1 exclusions, 447 did not have a 30-day outcome, 177 were reviews, 169 lacked a prediction aim, 63 had claims or registry data, 51 were not primary studies, 50 did not use an ML model, and 47 for other reasons (16 without readmission outcome, 10 paediatric studies, 8 with ICU readmission only, 7 without an abstract, 4 psychiatric studies only, 2 not within the date range). Stage 2 rejected **125** studies based on ineligible data source (51), design (33) or outcome (31), duplicated and/or abstract only papers (8), and population (2). There were altogether **156** studies included, 12 of which were preprint only studies.

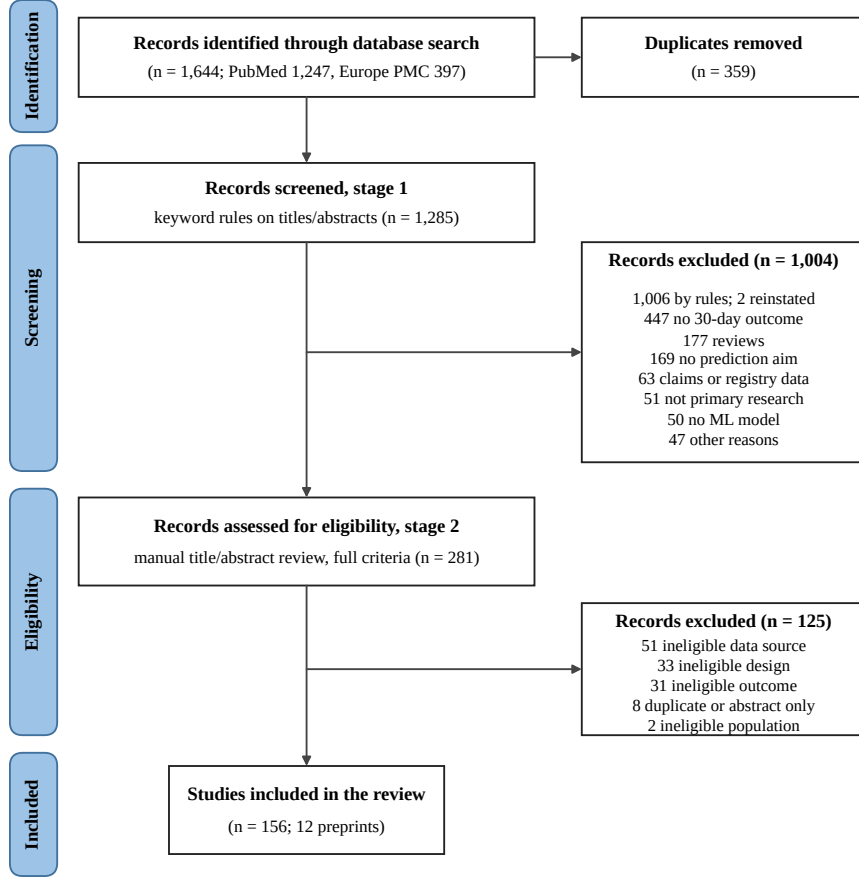


Fig. 3 PRISMA 2020 flow diagram of the study selection process.

3.7 Reporting Appraisal

A formal domain-level risk-of-bias assessment (e.g., PROBAST+AI [11]) was not feasible from abstracts. Instead, each included study was appraised for whether its abstract reported the evaluations that TRIPOD+AI [10] and PROBAST+AI [11] treat as essential beyond discrimination. The appraisal items (RA) were:

- RA1.** Is the model validated externally or temporally?
- RA2.** Is calibration assessed?
- RA3.** Is clinical utility assessed with decision curve analysis?
- RA4.** Is performance compared across subgroups (fairness)?
- RA5.** If explanations are presented, are they evaluated (e.g., stability, clinician review)?

Each item was scored as shown in Table 1. The RA1–RA4 scores were summed to give a “beyond AUROC” score from 0 to 4 for each study; RA5 was reported separately because it applies only to studies that present explanations. Every positive validation, decision-curve and fairness code was checked against the supporting abstract sentence.

Table 1 Reporting appraisal scoring criteria

Score	Description
1	Reported in the abstract
0	Not reported in the abstract

3.8 Data Extraction and Coding

Full texts were not retrieved systematically and were not used for coding; coding was based on titles and abstracts. The open-access full texts of five studies [21–25] were read only to illustrate selected findings. From each included abstract the following items were coded, using a form adapted from CHARMS [26]:

- **Study and data:** publication year; clinical population (assigned from the title to one of nine groups); multi-site development data; public benchmark data (MIMIC, Diabetes 130-US Hospitals); clinical text or natural language processing (NLP); prediction before discharge.
- **Modelling:** model families named; deep learning; comparison with regression or a conventional score; class-imbalance handling; missing-data handling.
- **Discrimination:** whether AUROC was reported; the highest AUROC, recorded only for studies whose sole outcome was readmission; AUPRC or precision–recall.
- **Validation (RQ1):** external (a different hospital or health system) or temporal (a later period).
- **Calibration (RQ2):** any calibration assessment; separately, the Brier score and calibration-specific measures (slope, intercept, Hosmer–Lemeshow test, O:E ratio).
- **Explainability (RQ3):** SHAP; any explainability method; any evaluation of explanations.
- **Utility and fairness (RQ4):** decision curve analysis; subgroup or fairness analysis.

3.9 Synthesis and Statistical Analysis

Frequency tables, percentages, medians, and IQRs were used for describing the included studies. In regards to RQ5, publications from 2020–2023 were compared with those published in 2024–2026, and also separated by publications prior to and after TRIPOD+AI [10], using Fisher’s exact test. AUROC distribution was compared with the Mann-Whitney U test, while frequency of validation among multi-site studies and single-site studies with Fisher’s exact test. The tests were two-sided with $\alpha = 0.05$, without correction for multiple testing and should be considered as exploratory analyses. The analyses were conducted using Python (SciPy). Coded data and exclusion criteria are provided in Supplementary Tables S1 and S2.

4 Results

4.1 Characteristics of the Included Studies

The number of selected articles increased from 14 articles in 2020 to 35 in 2026 at the time of search (Fig. 4a). There were seventy articles (44.9%) which developed models for all-cause readmission among general inpatients, while others were focused on surgical or orthopaedic (26), heart failure (14), diabetes (13), cardiovascular (9), respiratory (9), stroke (6), cancer (6) and liver or kidney diseases (3) (Table 8).

The characteristics of both models and datasets have been presented in Table 2. More than half of the models (78, 50.0%) have been benchmarked against machine learning and regression or traditional scores. Nearly a third of the models (56, 35.9%) have made use of deep learning. Only 32 (20.5%) have created their models based on data from multiple hospitals, while 20 (12.8%) have used public benchmark datasets and 12 (7.7%) have predicted readmissions even before discharge.

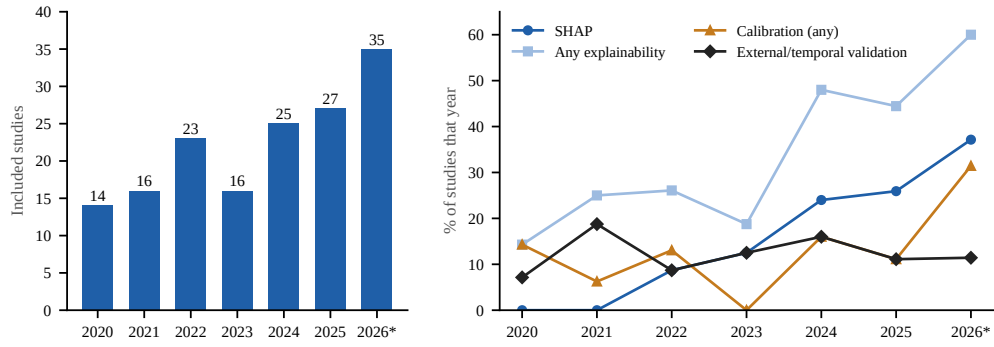


Fig. 4 (a) Included studies by publication year. (b) Share of studies each year reporting SHAP, any explainability, any calibration assessment and external or temporal validation. *2026 up to the search date.

Table 2 Model and data characteristics of the 156 included studies (coded from abstracts)

Characteristic	<i>n</i>	%
Compared with regression or a conventional score	78	50.0
Used deep learning	56	35.9
Multi-site development data	32	20.5
Used clinical text / NLP	31	19.9
Reported AUPRC or precision–recall	19	12.2
Handled class imbalance (e.g., SMOTE)	13	8.3
Predicted before discharge (e.g., at admission)	12	7.7
Used MIMIC	12	7.7
Used Diabetes 130-US Hospitals dataset	8	5.1
Reported missing-data handling	5	3.2

4.2 Machine Learning Models

The list of models cited in the titles or abstracts is shown in Table 3. Tree ensemble was the most popular type of machine learning models mentioned: 70 articles mentioned random forests or decision trees, 67 used gradient boosting. Neural networks and multilayer perceptron were mentioned in 33 articles, support vector machine in 20 articles and transformers or other language models in 17 articles, usually applied to discharge summaries and clinical notes. Recurrent (5) and convolutional (3) neural networks were less popular due to the mostly tabular nature of EHR data. 59 abstracts used logistic or penalised regression as a control.

Table 3 Model families named in the 156 included abstracts (not mutually exclusive)

Model family	Studies	Advantages	Weaknesses
Random Forest / Decision Trees	70	Can handle tabular data with interactions; strong default settings	Poorly calibrated probabilities; big ensembles difficult to explain
Gradient Boosting (XGBoost, LightGBM, CatBoost)	67	Excellent performance on EHR tabular data; can handle missing data	Susceptible to overfitting on small samples; tend to be overconfident
Logistic / penalised regression (comparator)	59	Transparent; usually well calibrated; strong baseline	Needs explicit interaction terms
Neural network / MLP	33	Flexible non-linear modelling	Needs large samples; limited transparency
Support vector machine	20	Effective in high-dimensional spaces	No native probabilities; sensitive to scaling
Transformer / language model	17	Uses discharge summaries and notes directly	High computational cost; sensitive to documentation drift
Recurrent network (RNN, LSTM, GRU)	5	Models sequences of encounters	Prolonged training; needs longitudinal data
Convolutional network (CNN)	3	Learns local patterns in codes or signals	Rarely suited to tabular EHR data

4.3 Publicly Available Datasets

A total of twenty studies (12.8%) were based on one of two publicly available EHR datasets (Table 4). These were 12 studies that relied on MIMIC (four of MIMIC-III and seven of MIMIC-IV, version was not indicated for one) [27–29] consisting of ICU and ED admissions from a single academic medical center in Boston and eight studies utilizing the Diabetes 130-US Hospitals dataset [30, 31], consisting of ten years of diabetic encounters in 130 US hospitals. Both of these are well-established benchmarking resources, yet neither is amenable to external validation as traditionally understood, and the latter lacks lab values and notes.

Table 4 Public datasets used by the included studies

Dataset	Source	Setting and period	Size	Studies
MIMIC-III [27]	PhysioNet	ICU admissions, Beth Israel Deaconess Medical Center, Boston, 2001–2012	46,520 patients; 58,976 admissions	4
MIMIC-IV [28, 29]	PhysioNet	ICU and emergency-department admissions, Beth Israel Deaconess Medical Center, 2008–2022	364,627 patients; 546,028 hospitalisations (v3.1)	7 ¹
Diabetes 130-US Hospitals [30, 31]	UCI Machine Learning Repository	Diabetic inpatient encounters, 130 US hospitals and delivery networks, 1999–2008	101,766 encounters; 47 features; label <30, >30 days or no readmission	8

¹One further MIMIC study did not state the version.

4.4 Discrimination

Of 156 abstracts, 132 (84.6%) provided the AUROC (Fig. 5). Of the 62 studies that provided an AUROC for readmission as their only outcome measure, the median maximum AUROC was 0.76 (IQR 0.69-0.85; range 0.56-0.99; Fig. 6). The median is optimistic since the maximum value was chosen and, in general, it was calculated internally. There were five studies providing an AUROC of 0.90 or higher, and all of them had not been externally validated.

There were no differences in discrimination found between studies using deep learning and non-deep-learning methods (median 0.76 in both cases, $p = 0.49$), using clinical text or not ($p = 0.70$), and multi-center studies versus single-center studies ($p = 0.62$). Six studies using public benchmark datasets showed worse AUROC values compared to other studies (median 0.69 versus 0.78, $p = 0.022$), but again, the sample size is rather limited. In the few comparisons of ML and regression models applied to large datasets, differences were minimal: in 38 Japanese hospitals, gradient boosting achieved 0.764 compared to 0.755 for LASSO regression [32], while in Australia’s general medicine population, LASSO and logistic regression performed equal or better than random forest, XGBoost, and a neural network [21].

4.5 External and Temporal Validation (RQ1)

Out of all 156 studies, 19 (12.2%) included external validation in an additional setting – 10 through external validation at another hospital or healthcare system and 10 through temporal validation at a different time, of which one study included both types of external validation. The remaining 137 (87.8%) included internal validation only.

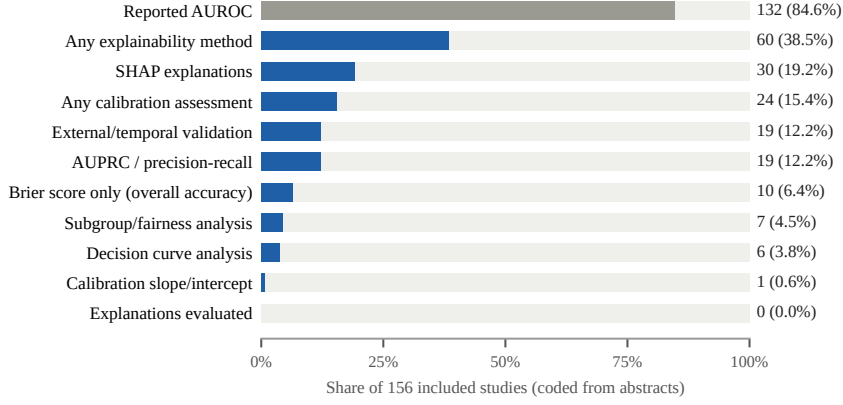


Fig. 5 Share of the 156 included studies whose abstract reports each item.

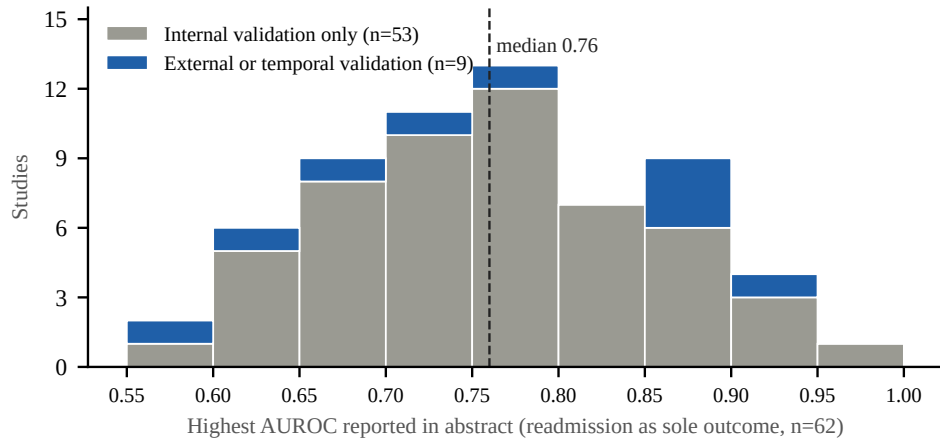


Fig. 6 Distribution of the highest AUROC reported in 62 abstracts whose sole outcome was readmission, by whether the study reported external or temporal validation. The plotted value may be an internal estimate even in validated studies.

Validation occurred more often in studies with multi-center development datasets (9 out of 32, or 28.1%) than in single-center studies (10 out of 124, or 8.1%; $p=0.005$). However, a multi-center model development is not external validation; only 23 out of 32 multi-center studies included external validation.

Of these 19, nine provided results that could be used for Table 5; the other ten did not separate development and new setting performance. Five studies have provided results for AUROC in both settings, and discrimination has decreased in all of them, from 0.686-0.704 to 0.558-0.655 for acute myocardial infarction [33], from 0.841 to 0.783 in different Korean hospitals [34], by little bit (0.853 to 0.838) in cirrhosis [35], and from 0.78 to 0.67 over two years for a discharge summary language model as documentation changes and recovers only after calibration [36]. The fifth model has kept its performance in the temporal validation (0.866 to 0.85-0.88) [20] being surgical

and predicting readmission through the emergency department. Of those reporting only new setting results, the best prospective multicentre study has given only 0.59 AUROC which is only a bit better than HOSPITAL (0.54) and LACE (0.52) [37], and one domain generalisation model has been stable for three years in an unseen rural hospital, then declined in the fourth year [19].

Table 5 Overview of studies reporting performance at a new site or in a later period

Ref.	Population and setting	Validation	Development AUROC	New-setting AUROC
[33]	Acute myocardial infarction, US EHR	External	0.686–0.704	0.558–0.655
[34]	All-cause, Korea (common data model)	External	0.841	0.783
[38]	Pituitary surgery	External	–	0.76
[35]	Cirrhosis, two centres	External	0.853	0.838
[37]	Multimorbid patients, multicentre	External	–	0.52–0.59
[19]	Diabetes, unseen rural site	External + temporal	–	F1 0.79 ¹
[32]	All-cause, 38 hospitals	Temporal	–	0.764
[20]	Surgical, ED-route readmission	Temporal	0.866	0.85–0.88
[36]	Cardiac surgery, language model	Temporal	0.78	0.67

– = not reported.

¹Proposed model at the unseen site (baseline F1 0.73); AUROC not reported.

4.6 Calibration (RQ2)

Abstracts that addressed any kind of calibration assessment were 24 (15.4%). Abstracts that indicated a numerical value in calibration assessment were 11 (7.1%), however, ten of these used only Brier Score, indicating overall accuracy and not calibration, whereas one abstract used slopes in calibration. Five abstracts used both validation in new settings and some kind of calibration assessment, but none of the 19 validated abstracts indicated a measure of calibration in new settings.

The calibration-specific metrics revealed the issues that AUROC did not uncover in the whole texts considered. At an Argentinean hospital, calibration slope values for logistic regression, random forest, and LightGBM were 0.75, 1.13, and 1.76 [22], respectively, while in Australia, XGBoost could not pass the Hosmer-Lemeshow test without using Platt scaling [21]. The process of recalibration itself requires more data than small hospitals can have: two rural hospitals had 3,187 and 905 patients versus 16,461 and 15,299 needed to recalibrate a readmission model [12].

4.7 Explainability (RQ3)

A total of sixty (38.5%) abstracts mentioned some form of explainability while 30 (19.2%) used SHAP whereas others used feature/permutation importance, attention weights, or coefficients/nomograms. There was no mention of any explanation evaluation such as robustness and review by clinicians (RA5). In 30 SHAP papers, there were five papers that validated their model in new settings, six papers that included any calibration evaluation and one paper that included any fairness analysis. The explanation was shown in the form of lists of important variables such as C-reactive protein, eGFR and natriuretic peptides for elderly patients with diabetes and heart failure [23] or diagnosis (18.4%) and nursing assessments (11.2%) at eight Swiss hospitals [39].

4.8 Clinical Utility and Fairness (RQ4)

Six studies (3.8%) applied decision curve analysis, and seven (4.5%), all of which appeared in 2024 and afterward, used subgroup analysis. After hip replacement surgery, worse model performance was observed among Hispanic/Latinx people, women and diabetic patients [40]. In heart failure, more sophisticated social determinants of health features increased AUROC (0.671 vs 0.632), and decreased inequality in predictions between different races (equalized odds ratio 0.437 vs 0.329, where 1 corresponds to equal results) [41]. One paper reduced gender and insurance biases through post-processing of results [42].

Table 6 shows the distribution of the “beyond AUROC” appraisal score (RA1–RA4). In total, 112 studies (71.8%) reported none of the four evaluations, 34 (21.8%) reported one, 8 (5.1%) reported two and 2 (1.3%) reported three. No study reported all four.

Table 6 Distribution of the “beyond AUROC” appraisal score (RA1–RA4)

Score (max = 4)	Studies, n	%
0	112	71.8
1	34	21.8
2	8	5.1
3	2	1.3
4	0	0.0

4.9 Trends Over Time (RQ5)

Comparison is done in Table 7. From 2020–2023 (69 papers) to 2024–2026 (87 papers), there was an increase in the use of SHAP (from 5.8% to 29.9%, $p < 0.001$), any claims for explainability (21.7% to 51.7%, $p < 0.001$), calibration (8.7% to 20.7%, $p = 0.046$), decision curve analysis (0% to 6.9%, $p = 0.034$) and fairness analysis (0% to 8.0%, $p = 0.018$). Reporting of Brier score or any other measure of calibration (5.8% vs 8.0%,

$p = 0.76$) and external/temporal validation (11.6% vs 12.6%, $p = 1.0$) did not differ (Fig. 7). Use of deep learning, clinical text, multi-center data, or comparison with regression methods did not change either. Annual trend showed that the percentage of papers employing SHAP went up from 0 in 2020–2021 to 37% in 2026 while the percentage of papers describing validation in a new setting varied between 7% and 19% (Fig. 4b).

Table 7 Reporting in studies published in 2020–2023 and 2024–2026

Item	2020–2023 ($n = 69$), n (%)	2024–2026 ($n = 87$), n (%)	p^1
Any explainability claim	15 (21.7)	45 (51.7)	< 0.001
SHAP	4 (5.8)	26 (29.9)	< 0.001
Any calibration assessment	6 (8.7)	18 (20.7)	0.046
Brier score or calibration-specific measure	4 (5.8)	7 (8.0)	0.76
Decision curve analysis	0 (0.0)	6 (6.9)	0.034
Subgroup or fairness analysis	0 (0.0)	7 (8.0)	0.018
External or temporal validation	8 (11.6)	11 (12.6)	1.00
Deep learning	21 (30.4)	35 (40.2)	0.24
Clinical text / NLP	14 (20.3)	17 (19.5)	1.00
Multi-site development data	16 (23.2)	16 (18.4)	0.55
Compared with regression or a conventional score	34 (49.3)	44 (50.6)	1.00

¹Fisher’s exact test, two-sided, unadjusted.

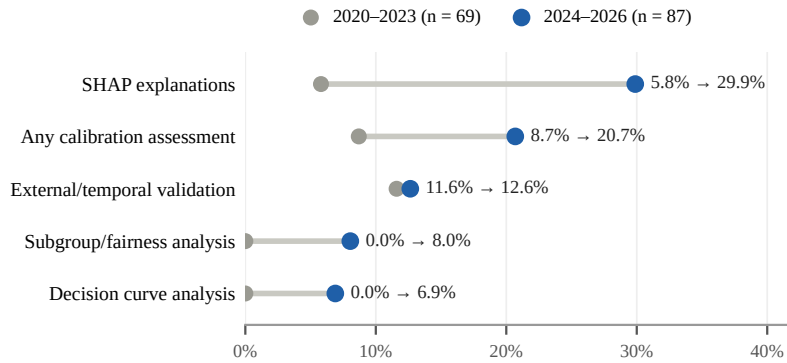


Fig. 7 Change in reporting between 2020–2023 and 2024–2026 (Fisher’s exact test: SHAP $p < 0.001$; validation $p = 1.0$).

4.10 Findings by Clinical Population

Table 8 summarises reporting by clinical population. Groups other than general inpatients contributed fewer than 15 studies each, and most contributed fewer than ten AUROC values, so the comparisons below are descriptive.

Table 8 Reporting by clinical population. AUROC is the highest value reported in abstracts whose sole outcome was readmission

Population	Studies	AUROC		Validated ¹	Calibration ²	SHAP
	<i>n</i>	<i>n</i>	Median	<i>n</i>	<i>n</i>	<i>n</i>
General / all-cause	70	22	0.76	9	8	8
Surgical / orthopaedic	26	13	0.86	2	6	6
Heart failure	14	7	0.81	1	2	5
Diabetes	13	7	0.69	2	2	3
Other cardiovascular	9	4	0.72	1	5	3
Respiratory	9	3	0.69	0	1	0
Stroke	6	2	0.83	2	0	2
Cancer	6	3	0.71	1	0	1
Liver / kidney	3	1	0.85	1	0	2
Total	156	62	0.76	19	24	30

¹External or temporal validation.

²Any calibration assessment, including Brier score only.

4.10.1 General Inpatients

The next-largest category was all-cause readmission in general medicine or mixed inpatients (70 studies; median AUROC 0.76 based on 22). The category also included the vast majority of validated models (9 out of 19), including temporal validation in 38 Japanese hospitals [32] and external validation of ML in Korean hospitals using the same data model [34]. This indicates that a model developed in one hospital should not be considered transferable: discrimination dropped when transferring models between hospitals [34], and in Australia, penalised regression was equal or better than ML [21].

4.10.2 Surgical and Orthopaedic Patients

The surgical and orthopaedic cohort categories (26 papers) had the highest AUROC median (0.86 for 13), yet only two papers had external validation in a new context, temporally [20, 36]. Some of the high values were achieved in single-centre cohorts. This category produced some of the most interesting results related to fairness: predictive models performed poorly on Hispanic/Latinx patients, women, and those with diabetes [40] following hip arthroplasty and total joint replacement [42].

4.10.3 Heart Failure and Other Cardiovascular Conditions

The median AUROCs for heart failure (14) and other cardiovascular diseases (9) were 0.81 and 0.72, respectively. SHAP was used in heart failure models (5 out of 14), and Matheny et al.’s study on acute myocardial infarction had the biggest decrease in discrimination during external validation [33]. Predictors from social determinants of health increased the discrimination and fairness of heart failure models [41], while an explainable gradient boosting model provided equivalent performance to a deep LSTM in heart failure readmissions [25].

4.10.4 Diabetes

Out of the 13 studies on diabetes, five (median AUROC = 0.69) relied on the Diabetes 130 US hospitals database. Two of these were validated using a different data set, one being a domain generalisation approach which was tested in a novel rural hospital setting over four years [19]. For elderly type 2 diabetic patients with heart failure, the factors which played an important role in the explanation included the levels of C-reactive protein, eGFR, and natriuretic peptides [23].

4.10.5 Other Populations

Respiratory diseases (9), stroke (6), cancer (6), and liver/renal diseases (3) added up to 24 studies. No respiratory studies mentioned validation in an external setting, calibration, and SHAP. For decompensated cirrhosis, one prospective study from two centers indicated only marginal decrease in discrimination at external validation (0.853 to 0.838) [35].

5 Discussion

There have been many advances made in ML readmission literature; however, reported assessment has not kept up. Over 88% of the abstracts only describe internally validated results, over 72% do not provide any of the assessments needed to determine the model’s transportability, calibration, usefulness, or fairness, and the community has embraced SHAP far sooner than external validation, quantitative calibration, or fairness assessments. Previous full-text reviews had reached the same conclusion regarding external validation and calibration in general [3, 4] and heart failure [5, 6] readmission literature.

5.1 Why AUROC Is Not Enough

Readmission programmes target absolute risk. A transitional-care programme that is able to handle 200 patients per month should be aware of how many of its high-risk patients return, which is a calibration issue [7]. Readmission rates vary across hospitals and over time, such that a predictive model can maintain its rankings while the predictions themselves get less accurate. Discrimination was constant in one acute kidney injury predictive model based on a Veterans Affairs population, while calibration of both the regression model and machine learning approach fell [44]. This was not a

readmission prediction model but the problem, a change in case mix and in incidence of the event, applies just as well to readmission prediction models, as did the need for recalibration in the language-based readmission model reviewed here [36]. Even gains in discrimination are fragile; the benefit of machine learning falls away when the risk of bias in studies is low [9], and in heart failure cases machine learning was no better than regression (C statistics 0.59-0.62) [45].

5.2 Transportability

Readmission is dependent on local case mix, discharge practices, and community services such that models perform poorly elsewhere. Four out of five of the studies presented in Table 5 that reported AUROC in both settings lost discrimination, and institution-specific models have therefore been recommended [46]. Our search did not identify at least one eligible external-validation study in which XGBoost performed worse at 0.707 compared to logistic regression at 0.753 after falling from 0.912 at an external hospital [24], indicating low search sensitivity and similar findings. A vendor-derived readmission score, the method used for which is unknown, achieved an AUROC of 0.692 in Switzerland, which was no better than LACE+ (AUROC = 0.703) [47]. Multi-site development data being associated with increased validation attempts indicates that data availability, and not just reporting trends, restricts validation. Leave-one-site-out and temporal validation are cheap means of testing transportability, and deployed models require ongoing monitoring and recalibration [12, 19, 36].

5.3 Explainability and Fairness

SHAP charts identify the driving inputs, not patient rehospitalization reasons or prevention measures, and post hoc reasoning can be fragile and misleading [8]. Given SHAP popularity without evaluation and its poor correlation with validation and calibration, it can be assumed that explainability is simply a presentation aspect rather than a validated one. It is possible to maintain transparency without losing accuracy, as an explainable gradient-boosting model had similar performance on the task of 100-day heart failure readmission compared with a deep LSTM model, although most features were non-actionable [25]. Fairness study started in 2024, and the seven subgroup analyses showed true disparities. Readmission scores also tend to have poorer calibration among Black patients and those of lower income [16]. Since rehospitalization is partially an access issue, there exists a danger of label bias as seen in cost-based algorithms [15], and fairness can degrade in production [43].

5.4 Recommendations

Table 9 gives a minimum reporting set consistent with the TRIPOD+AI [10] and PROBAST+AI [11] reporting guidelines. Since the reader sees the abstract first, these important elements need to be included in the abstract. The most important addition for the researcher is a more thorough evaluation, not a more complicated algorithm. For the hospital evaluating a readmission model from a publication or software company, the minimum is a local silent-mode evaluation of discrimination and calibration before

clinical predictions are made. For the editor or reviewer, this set gives an easy way to tell if a readmission model has been well evaluated.

Table 9 Minimum “beyond AUROC” reporting set for readmission models

Dimension	Minimum	Preferred
Discrimination	AUROC with 95% CI on held-out data; AUPRC; outcome prevalence	Training and test AUROC side by side
Validation	Temporal or leave-one-site-out	Independent external health system
Calibration	Plot, slope and intercept	O:E ratio per site; recalibration plan
Explainability	Method named, with limitations	Stability across resamples and sites; clinician review
Utility	Decision curve analysis	Patients flagged per readmission caught
Fairness	Discrimination and calibration by sex, age, race, deprivation	Error-rate gaps over time; mitigation
Deployment	Prediction time point; leakage safeguards; missing data	Silent-mode prospective evaluation; drift monitoring

5.5 Limitations

The rapid systematic review was not preregistered and the screening and coding process was performed by a single reviewer without independent dual screening, while the audit of stage-1 exclusions and recovery of two missing articles minimize, but do not completely solve this problem. The search was performed in two freely accessible databases, and there is at least one article that should be retrieved [24], besides papers indexed in Embase, Scopus, IEEE Xplore or Web of Science databases, and papers which do not report the 30-day time horizon in their abstracts. The coding was done based on abstracts, where the methodological details are less informative than in full texts, and the percentages represent the estimates of what was done and cannot be compared with results of full-text reviews. Subpopulation subgroups were chosen from the title, and the family of models was determined based on keywords of abstracts; the comparisons of AUROC between subpopulation subgroups are based on few cases. The sample size, number of events per variable and outcome prevalence were not coded, while the risk-of-bias assessment was not performed. MIMIC studies use ICU patients with different readmission criteria.

6 Conclusion

For this review, 156 articles that predicted 30-day readmission using ML on EHR data from 2020 to 2026 were analyzed. AUROC was mainly validated internally (median =

0.76). Only 12% had validation in another location or a different time period, where performance was lower. Calibration, decision curves and fairness metrics were done by less than 1 in 6 studies, more than 7 in 10 had done none of these. No abstracts had evaluation of their explanations. The use of SHAP increased five-fold, but not validation.

More future gains will come from improved assessment rather than through more sophisticated algorithms. Validation, either temporal or external, calibration and recalibration, decision curves, subgroup performance and assessed explanations should be considered the minimum standards before multi-site prospective validation in discharge processes. The minimum set of reporting criteria suggested here can serve as a useful checklist for researchers, hospitals, and editors when deciding about the clinical usefulness of a hospital readmission prediction model.

References

- [1] Centers for Medicare & Medicaid Services, “Hospital Readmissions Reduction Program (HRRP),” 2026. [Online]. Available: <https://www.cms.gov/medicare/payment/prospective-payment-systems/acute-inpatient-pps/hospital-readmissions-reduction-program-hrrp>
- [2] D. Kansagara *et al.*, “Risk prediction models for hospital readmission: A systematic review,” *JAMA*, vol. 306, no. 15, pp. 1688–1698, 2011.
- [3] E. Mahmoudi *et al.*, “Use of electronic medical records in development and validation of risk prediction models of hospital readmission: Systematic review,” *BMJ*, vol. 369, Art. no. m958, 2020.
- [4] Y. Huang, A. Talwar, S. Chatterjee, and R. R. Aparasu, “Application of machine learning in predicting hospital readmissions: A scoping review of the literature,” *BMC Med. Res. Methodol.*, vol. 21, Art. no. 96, 2021.
- [5] S. Shin *et al.*, “Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality,” *ESC Heart Fail.*, vol. 8, no. 1, pp. 106–115, 2021.
- [6] M.-Y. Yu and Y.-J. Son, “Machine learning-based 30-day readmission prediction models for patients with heart failure: A systematic review,” *Eur. J. Cardiovasc. Nurs.*, vol. 23, no. 7, pp. 711–719, 2024, doi: 10.1093/eurjcn/zvae031.
- [7] B. Van Calster *et al.*, “Calibration: The Achilles heel of predictive analytics,” *BMC Med.*, vol. 17, Art. no. 230, 2019.
- [8] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, “The false hope of current approaches to explainable artificial intelligence in health care,” *Lancet Digit. Health*, vol. 3, no. 11, pp. e745–e750, 2021.
- [9] E. Christodoulou *et al.*, “A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models,” *J. Clin. Epidemiol.*, vol. 110, pp. 12–22, 2019.
- [10] G. S. Collins *et al.*, “TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods,” *BMJ*, vol. 385, Art. no. e078378, 2024.
- [11] K. G. M. Moons *et al.*, “PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods,” *BMJ*, vol. 388, Art. no. e082505, 2025.
- [12] K. E. Brown, B. A. Malin, and S. E. Davis, “Community medical centers struggle to produce well-calibrated clinical prediction models: Data augmentation can help,” *J. Biomed. Inform.*, vol. 177, Art. no. 105012, 2026.

- [13] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4765–4774.
- [14] A. J. Vickers and E. B. Elkin, “Decision curve analysis: A novel method for evaluating prediction models,” *Med. Decis. Making*, vol. 26, no. 6, pp. 565–574, 2006.
- [15] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science*, vol. 366, no. 6464, pp. 447–453, 2019.
- [16] H. E. Wang, J. P. Weiner, S. Saria, and H. Kharrazi, “Evaluating algorithmic bias in 30-day hospital readmission models: Retrospective analysis,” *J. Med. Internet Res.*, vol. 26, Art. no. e47125, 2024.
- [17] C. Garritty *et al.*, “Cochrane Rapid Reviews Methods Group offers evidence-informed guidance to conduct rapid reviews,” *J. Clin. Epidemiol.*, vol. 130, pp. 13–22, 2021.
- [18] M. J. Page *et al.*, “The PRISMA 2020 statement: An updated guideline for reporting systematic reviews,” *BMJ*, vol. 372, Art. no. n71, 2021.
- [19] A. A. Hai, M. G. Weiner, A. Livshits, *et al.*, “Domain generalization for enhanced predictions of hospital readmission on unseen domains among patients with diabetes,” *Artif. Intell. Med.*, vol. 158, Art. no. 103010, 2024.
- [20] V. V. Mišić, E. Gabel, I. Hofer, *et al.*, “Machine learning prediction of postoperative emergency department hospital readmission,” *Anesthesiology*, vol. 132, no. 5, pp. 968–980, 2020.
- [21] Y. Sharma *et al.*, “Performance of machine learning models in predicting 30-day general medicine readmissions compared to traditional approaches in Australian hospital setting,” *Healthcare*, vol. 13, no. 11, Art. no. 1223, 2025.
- [22] D. Halac, C. Cocucci, and S. Camerlingo, “Predictive machine learning model for 30-day hospital readmissions in a tertiary healthcare setting,” *Bioinform. Adv.*, vol. 5, no. 1, Art. no. vbaf121, 2025.
- [23] L. Wang *et al.*, “Machine learning-based prediction model for 30-day readmission risk in elderly patients with type 2 diabetes mellitus and heart failure: A retrospective cohort study with SHAP interpretability analysis,” *Front. Cardiovasc. Med.*, vol. 12, Art. no. 1673159, 2025.
- [24] Y. He *et al.*, “Development and validation of a risk prediction model for 30-day readmission in elderly type 2 diabetes patients complicated with heart failure: A multicenter, retrospective study,” *Front. Endocrinol.*, vol. 16, Art. no. 1534516, 2025.

- [25] A. Soliman, B. Agvall, K. Etmiani, O. Hamed, and M. Lingman, “The price of explainability in machine learning models for 100-day readmission prediction in heart failure,” *J. Med. Internet Res.*, vol. 25, Art. no. e46934, 2023.
- [26] K. G. M. Moons *et al.*, “Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The CHARMS checklist,” *PLoS Med.*, vol. 11, no. 10, Art. no. e1001744, 2014.
- [27] A. E. W. Johnson, T. J. Pollard, L. Shen, *et al.*, “MIMIC-III, a freely accessible critical care database,” *Sci. Data*, vol. 3, Art. no. 160035, 2016.
- [28] A. E. W. Johnson, L. Bulgarelli, L. Shen, *et al.*, “MIMIC-IV, a freely accessible electronic health record dataset,” *Sci. Data*, vol. 10, Art. no. 1, 2023.
- [29] PhysioNet, “MIMIC-IV (version 3.1),” 2024. [Online]. Available: <https://physionet.org/content/mimiciv/3.1/>
- [30] B. Strack, J. P. DeShazo, C. Gennings, *et al.*, “Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records,” *BioMed Res. Int.*, vol. 2014, Art. no. 781670, 2014.
- [31] UCI Machine Learning Repository, “Diabetes 130-US Hospitals for Years 1999–2008.” [Online]. Available: <https://archive.ics.uci.edu/dataset/296/diabetes-130-us-hospitals-for-years-1999-2008>
- [32] M. Iwagami *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,” *PLOS Digit. Health*, vol. 3, no. 8, Art. no. e0000578, 2024.
- [33] M. E. Matheny *et al.*, “Development of electronic health record-based prediction models for 30-day readmission risk among patients hospitalized for acute myocardial infarction,” *JAMA Netw. Open*, vol. 4, no. 1, Art. no. e2035782, 2021.
- [34] B. Ryu, S. Yoo, S. Kim, *et al.*, “Development of prediction models for unplanned hospital readmission within 30 days based on common data model: A feasibility study,” *Methods Inf. Med.*, vol. 60, no. S2, pp. e65–e75, 2021.
- [35] F. Yang *et al.*, “Construction and validation of a machine learning-based model predicting early readmission in patients with decompensated cirrhosis: A prospective two-center cohort study,” *BioData Min.*, vol. 18, no. 1, Art. no. 63, 2025.
- [36] D. A. Alber *et al.*, “A cardiac surgery-specific artificial intelligence model to automate risk stratification for 30-day readmissions using discharge summaries,” *Ann. Thorac. Surg. Short Rep.*, vol. 4, no. 3, pp. 1076–1080, 2026.

- [37] A. Pikatza-Huerga *et al.*, “Machine learning using PROFUND components for 30-day readmission prediction in multimorbid patients: A prospective multicentre study,” *Sci. Rep.*, vol. 16, 2026, doi: 10.1038/s41598-026-51343-y.
- [38] B. T. Crabb *et al.*, “Machine learning-based analysis and prediction of unplanned 30-day readmissions after pituitary adenoma resection: A multi-institutional retrospective study with external validation,” *Neurosurgery*, vol. 91, no. 2, pp. 263–271, 2022.
- [39] M. M. Havranek *et al.*, “Machine learning predictions of unplanned readmissions using electronic medical records: Predictor importance across medical and surgical patient populations,” *PLoS ONE*, vol. 20, no. 9, Art. no. e0331263, 2025.
- [40] M. M. Raza *et al.*, “Underperformance of machine learning algorithms predicting extended lengths of stay and readmission in underrepresented patient cohorts after primary total hip arthroplasty,” *J. Arthroplasty*, vol. 41, no. 9S1, pp. S30–S35, 2026.
- [41] C. Fensore *et al.*, “Beyond composite indices: Comprehensive social determinants improve heart failure readmission prediction,” *J. Am. Heart Assoc.*, vol. 15, no. 7, Art. no. e043735, 2026.
- [42] V. Digumarthi *et al.*, “Preoperative prediction model for risk of readmission after total joint replacement surgery: A random forest approach leveraging NLP and unfairness mitigation,” *J. Orthop. Surg. Res.*, vol. 19, no. 1, Art. no. 287, 2024.
- [43] S. E. Davis *et al.*, “Emerging algorithmic bias: Fairness drift as the next dimension of model maintenance and sustainability,” *J. Am. Med. Inform. Assoc.*, vol. 32, no. 5, pp. 845–854, 2025.
- [44] S. E. Davis, T. A. Lasko, G. Chen, E. D. Siew, and M. E. Matheny, “Calibration drift in regression and machine learning models for acute kidney injury,” *J. Am. Med. Inform. Assoc.*, vol. 24, no. 6, pp. 1052–1061, 2017.
- [45] J. D. Frizzell *et al.*, “Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: Comparison of machine learning and other statistical approaches,” *JAMA Cardiol.*, vol. 2, no. 2, pp. 204–209, 2017.
- [46] S. Yu *et al.*, “Predicting readmission risk with institution-specific prediction models,” *Artif. Intell. Med.*, vol. 65, no. 2, pp. 89–96, 2015.
- [47] A. B. Hwang, G. Schuepfer, M. Pietrini, and S. Boes, “External validation of EPIC’s Risk of Unplanned Readmission model, the LACE+ index and SQLape as predictors of unplanned hospital readmissions,” *PLoS ONE*, vol. 16, no. 11, Art. no. e0258338, 2021.