JOURNAL OF MULTIDISCIPLINARY RESEARCH AND INTEGRATED SCIENCES

**Volume 3, Issue 1, 2023** | **Article ID:** JMRIS-2023-001 | **Pages:** 1-12 | **Article Type:** Systematic Review

**Received:** 03-12-2022 | **Revised:** 15-12-2022 | **Accepted:** 12-01-2023 | **Published:** 15-03-2023

# Explainability Versus Predictive Accuracy in Machine-Learning Models for 30-Day Hospital Readmission Among Patients With Heart Failure: A Systematic Evidence Synthesis

**Elias Morgan**[1*]

[1] *University of Melbourne, Melbourne, Australia*

***Corresponding author:** Elias Morgan, University of Melbourne, Melbourne, Australia | Eliasmorgan3056@gmail.com

## ABSTRACT

**Background:** Thirty-day hospital readmission after heart-failure hospitalization remains difficult to predict because readmission reflects heterogeneous clinical, social, behavioral, and healthcare-utilization factors. Machine-learning models can capture nonlinear relationships and high-dimensional electronic health-record data, but increasingly complex models may sacrifice interpretability for only modest gains in discrimination. **Objective:** To synthesize high-quality evidence available through 2022 concerning the predictive performance and explainability of machine-learning models for 30-day hospital readmission among patients with heart failure and to determine whether increases in model complexity translate into clinically meaningful predictive benefit. **Methods:** A structured evidence synthesis was conducted using peer-reviewed Q1 literature published no later than 01 December 2022. Studies developing or comparing models for 30-day heart-failure readmission, readmission-or-death, or closely related early post-discharge outcomes were prioritized. Logistic regression, tree-based models, ensemble methods, neural networks, and deep-learning approaches were compared with respect to discrimination, class imbalance, calibration, interpretability, validation, and potential clinical utility. Methodological literature concerning explainable artificial intelligence and clinical prediction was incorporated. **Results:** Prediction of 30-day readmission remained only moderately discriminative. Mortazavi et al. reported improvement of machine-learning approaches over logistic regression, but absolute discrimination remained limited. Golas et al. reported c-statistics of 0.664 for logistic regression, 0.650 for gradient boosting, 0.695 for maxout networks, and 0.705 for a deep unified network. The deep model therefore gained only 0.041 in c-statistic over logistic regression while losing direct feature-level interpretability. In a 2022 independent validation study, XGBoost achieved AUROC 0.65 compared with 0.58 for a neural network and 0.57 for the modified LaCE score. Feature-importance approaches improved global interpretation, but patient-level post-hoc explanations introduce additional questions concerning fidelity and stability. **Conclusion:** More complex machine-learning models can modestly improve prediction of 30-day heart-failure readmission, but the gain in accuracy is frequently smaller than expected and may not justify reduced transparency. Model selection should consider calibration, precision-recall performance, external validity, explanation fidelity, and clinical actionability in addition to AUROC. For clinical deployment, the preferred model is the simplest approach that preserves clinically meaningful predictive performance while providing reliable individual-level explanations.

**Keywords:** heart failure; readmission; machine learning; explainable artificial intelligence; interpretability; XGBoost; deep learning; clinical prediction

## 1. Introduction

Heart failure is characterized by frequent episodes of clinical deterioration, hospitalization, and recurrent healthcare utilization. The period immediately after discharge is particularly vulnerable, and unplanned 30-day readmission has become an important clinical and health-system outcome.

Readmission prediction is difficult because hospitalization following heart failure is not determined solely by cardiac physiology. Comorbidity, prior utilization, renal function, treatment response, social circumstances, medication adherence, access to outpatient care, and non-cardiovascular disease may all influence whether a patient returns to hospital.

Dharmarajan et al. demonstrated that early readmissions following heart-failure hospitalization arise from a broad range of diagnoses rather than heart failure alone [1]. This heterogeneity helps explain why conventional risk scores frequently achieve only modest discrimination.

Machine learning has been proposed as a solution.

Unlike traditional regression models requiring explicit prespecification of relationships, algorithms such as Random Forest, gradient boosting, and neural networks can identify nonlinear relationships and interactions within large electronic health-record datasets.

Mortazavi et al. compared several machine-learning techniques for heart-failure readmission and demonstrated performance improvements over conventional approaches in some settings [2].

Frizzell et al., however, found that machine-learning methods did not meaningfully solve the problem of limited 30-day readmission discrimination [3].

Golas et al. later developed a deep-learning model using 3,512 features derived from structured and unstructured electronic medical records. Their deep unified network achieved a c-statistic of 0.705 compared with 0.664 for logistic regression [4].

The result illustrates a central problem.

A substantially more complex model produced a measurable but relatively modest increase in discrimination.

At the same time, clinicians could no longer directly determine how individual features produced a prediction.

This tension between predictive performance and explainability has become increasingly important as artificial intelligence enters high-stakes healthcare applications.

The present study therefore aimed to compare predictive performance of major machine-learning approaches for early heart-failure readmission, examine the magnitude of improvement over conventional approaches, evaluate the interpretability limitations created by model complexity, assess available approaches to global and patient-specific explanations, and propose criteria for balancing discrimination and explainability before clinical implementation.

## 2. Methods

### 2.1 Study Design

A structured evidence synthesis was conducted using literature available through 31 December 2022.

Search concepts included combinations of:

heart failure, 30-day readmission, hospital readmission, machine learning, deep learning, Random Forest, XGBoost, neural network, explainability, interpretability, SHAP, and clinical prediction.

Heart-failure-specific predictive studies were prioritized.

Methodological publications were included when they directly addressed model discrimination, calibration, class imbalance, interpretability, explainable artificial intelligence, clinical prediction reporting, and risk of bias.

*The evidence eligibility framework is summarised in Table 1.*

### 2.2 Analytical Domains

Evidence was synthesized across four domains predictive discrimination, complexity and feature requirements, explainability, and clinical usefulness and validation.

A new pooled AUROC was not calculated because the studies differed substantially in patient populations, predictors, endpoint definitions, sampling procedures, and validation strategies.

## 3. Results

### 3.1 Predictive Performance Remained Moderate

The strongest overall finding was that 30-day readmission remained difficult to predict regardless of algorithm.

*Selected Q1 evidence on machine-learning prediction of early heart-failure readmission is summarised in Table 2.*

Golas et al. used structured and unstructured electronic medical-record data and included 3,512 final predictors [4].

The resulting c-statistics were logistic regression: 0.664, gradient boosting: 0.650, maxout network: 0.695, and deep unified network: 0.705.

Thus, the most complex model improved the c-statistic by 0.041 compared with logistic regression.

This improvement is real but must be interpreted relative to the substantial increase in complexity.

*Selected pre-2023 discrimination estimates for 30-day heart-failure readmission prediction are presented in Figure 1.*

### 3.2 Administrative Data Place an Upper Limit on Prediction

Sharma et al. evaluated 9,845 patients with confirmed heart failure admitted between 2012 and 2019 [7].

The prevalence of unplanned 30-day readmission was approximately 20.9%.

XGBoost produced the strongest discrimination, with AUROC 0.65 compared with 0.58 for the neural network and 0.57 for the modified LaCE score.

At higher predicted-risk thresholds, XGBoost generated a positive likelihood ratio up to 6.12 and post-test probabilities as high as approximately 62%.

Nevertheless, the investigators concluded that the overall prediction was not sufficiently informative for unrestricted clinical deployment using administrative data alone.

This finding suggests that algorithm selection is only one determinant of prediction quality.

If the input variables omit important determinants such as functional status, medication adherence, social support, symptoms after discharge, frailty, access to outpatient care, and clinician narrative information, increasing algorithm complexity cannot recover information that is absent from the dataset.

### 3.3 Model Complexity and Explainability

Interpretability can be considered at two levels.

### Global interpretability

Global interpretation asks:

Which variables generally drive predictions across the entire population?

Examples include regression coefficients, decision-tree rules, feature importance, permutation importance, and SHAP summary distributions.

### Local interpretability

Local interpretation asks:

Why was this particular patient predicted to be at high risk?

A clinically useful explanation might identify several previous admissions, renal dysfunction, prolonged index hospitalization, high comorbidity burden, and recent emergency healthcare utilization as the major contributors to an individual's predicted risk.

*Explainability characteristics of common readmission models are summarised in Table 3.*

Golas et al. directly highlighted the problem: their deep unified network achieved the best discrimination but did not directly provide feature importance [4].

Logistic regression, despite weaker discrimination, did.

This represents the core accuracy-explainability trade-off.

### 3.4 Does Explainability Have a Predictive Cost?

Explainability is sometimes presented as though an interpretable model must necessarily be less accurate.

The heart-failure evidence does not support such a universal assumption.

In Golas et al., gradient boosting performed worse than logistic regression despite being more algorithmically complex [4].

Similarly, Sharma et al. found that XGBoost outperformed a neural network [7].

Thus, greater complexity does not necessarily produce greater predictive accuracy.

Christodoulou et al. systematically examined clinical prediction studies and found no consistent performance advantage of machine learning over logistic regression [8].

Therefore, the relevant decision is not whether to sacrifice accuracy for interpretability.

The correct question is how much validated predictive improvement is gained for the additional model complexity.

A 0.01 increase in AUROC may not justify an opaque system requiring extensive computational infrastructure and post-hoc explanations.

A larger improvement might.

### 3.5 Performance Metrics Beyond AUROC

AUC is valuable because it summarizes discrimination across thresholds.

However, 30-day heart-failure readmission commonly represents a minority outcome, making additional metrics essential.

Awan et al. emphasized the effect of class imbalance and demonstrated why area under the precision-recall curve can provide information not apparent from AUROC alone [5].

*Recommended model-performance domains are presented in Table 4.*

A model with good AUROC but poor calibration may correctly rank patients while systematically exaggerating their absolute risks.

This becomes clinically problematic if predictions are used to allocate expensive transitional-care interventions.

### 3.6 Explainability Does Not Guarantee Correctness

Post-hoc explainability methods can translate complex models into more understandable outputs.

SHAP and related approaches estimate the contribution of individual features to predictions.

These techniques are valuable for identifying influential variables, examining directional relationships, generating patient-specific explanations, and detecting implausible model behavior.

However, an explanation is not equivalent to a causal relationship.

If previous hospitalization is highly influential, this does not mean eliminating the historical admission would reduce future readmission.

It means the feature contains predictive information.

Likewise, an explanation can faithfully describe a poorly calibrated model.

Explainability and predictive validity must therefore be evaluated separately.

### 3.7 Accuracy–Explainability Framework

*The framework for balancing predictive accuracy, explainability and clinical utility is presented in Figure 2.*

The framework favors a staged comparison.

First, establish an interpretable benchmark.

Second, train progressively more flexible models.

Third, measure incremental performance.

Fourth, determine whether explanations are clinically coherent.

Finally, evaluate whether the predicted risk changes clinical management.

## 4. Discussion

### 4.1 Principal Finding

The central finding is that 30-day heart-failure readmission remains intrinsically difficult to predict, and machine-learning complexity does not eliminate this limitation.

Across major pre-2023 studies, discrimination frequently remained in the approximate 0.6–0.7 range.

Golas et al. crossed the 0.70 threshold using deep learning, but the gain relative to regression was modest [4].

Sharma et al. showed that XGBoost could outperform both a neural network and a conventional score, yet AUROC remained only 0.65 [7].

This suggests that a major source of prediction difficulty lies not in algorithm choice but in the outcome itself.

Readmission is influenced by events occurring after discharge, many of which are not measurable at the time the model generates its prediction.

### 4.2 The Price of Additional Complexity

Suppose logistic regression produces AUROC 0.66 and a deep-learning model produces 0.70.

Whether that difference is worthwhile depends on several considerations.

The relevant questions are whether the newer model improve identification of patients who will actually be readmitted, remain calibrated, perform equally well in another hospital, provide stable explanations, require substantially more data processing, improve clinical decisions, and reduce readmissions when implemented.

If these questions remain unanswered, improved AUROC alone is insufficient evidence of clinical superiority.

### 4.3 Explainability Has Different Audiences

The explanation required by a data scientist differs from that required by a cardiologist.

A technical explanation may describe SHAP values or interactions.

A clinician requires something closer to:

*This patient is high risk primarily because of four previous admissions, renal dysfunction, prolonged hospitalization, and repeated emergency care during the preceding six months.*

A healthcare administrator may instead require:

*Patients above this threshold account for 35% of expected readmissions and would require 200 transitional-care enrollments per year.*

Explainability should therefore be user-specific and decision-specific.

### 4.4 Actionability May Matter More Than Interpretability Alone

A perfectly understandable predictor is not necessarily useful.

Age may strongly predict readmission but cannot be modified.

Previous admissions may also be highly predictive but represent historical information.

More actionable variables may include unresolved congestion, medication-access problems, absence of early follow-up, inadequate patient education, unstable renal function, poor social support, and incomplete transition planning.

Future models should distinguish predictive importance from intervention opportunity.

This distinction may ultimately be more valuable than whether the algorithm is technically explainable.

### 4.5 Data Quality May Matter More Than Algorithm Choice

The comparatively modest performance of administrative-data models suggests diminishing returns from algorithmic optimization when inputs lack clinical depth.

Improvement may require integrating longitudinal laboratory trajectories, vital-sign trends, medication changes, free-text discharge summaries, social determinants, frailty, functional status, wearable monitoring, patient-reported symptoms, and post-discharge measurements.

The next substantial gain in readmission prediction may therefore come from better data rather than more complex algorithms.

### 4.6 Recommended Model-Selection Strategy

A clinically responsible development study should begin with logistic or penalized logistic regression.

Random Forest and XGBoost can then determine whether nonlinear modeling provides meaningful improvement.

A deep neural model should be added only when dataset size and complexity justify it.

*The proposed clinical model-selection hierarchy is presented in Table 5.*

This framework avoids selecting a model solely because it is technically sophisticated.

### 4.7 Strengths

This synthesis focuses specifically on the trade-off between predictive performance and explanation rather than treating algorithm accuracy as the sole endpoint.

It also distinguishes global from local explanation, prediction from causation, discrimination from calibration, and interpretability from clinical actionability.

The strict pre-2023 cutoff provides a coherent representation of the evidence available for a 2023 journal volume.

### 4.8 Limitations

Several limitations should be acknowledged.

First, studies differed in endpoint definition, particularly all-cause readmission versus combined readmission-or-death outcomes.

Second, datasets ranged from administrative claims to rich electronic medical records.

Third, validation approaches differed substantially.

Fourth, model interpretability was inconsistently reported in earlier studies.

Fifth, AUROC values from different cohorts should not be directly ranked as though obtained from the same test population.

Sixth, explainability methods themselves were still evolving through 2022.

Finally, no new patient-level model was developed in the present study. The objective was comparative evidence synthesis rather than fabrication of an artificial modeling cohort.

### 4.9 Future Research

Future heart-failure readmission models should prospectively evaluate accuracy and explainability together.

A particularly strong study would compare penalized logistic regression, interpretable decision tree, Random Forest, XGBoost, and neural network.

All should use the same patient cohort, predictors, and independent test set.

The analysis should report AUROC, AUPRC, calibration, Brier score, sensitivity, PPV, decision curves, explanation stability, clinician agreement with explanations, and external validation.

Clinicians could additionally be randomized to receive prediction only versus prediction plus a patient-specific explanation to determine whether explainability actually improves risk assessment or transitional-care decisions.

That would move explainable AI from technical demonstration to measurable clinical utility.

### 4.10 Conclusion

Machine-learning models can modestly improve prediction of 30-day hospital readmission among patients with heart failure, but increasing model complexity does not guarantee substantial gains.

The pre-2023 evidence demonstrates a recurring pattern: traditional models provide moderate discrimination, while advanced ensemble and deep-learning methods frequently produce incremental rather than transformative improvement.

At the same time, complex models create additional challenges involving transparency, calibration, validation, and clinician trust.

The most appropriate model should therefore not be selected solely according to the highest AUROC.

Clinical deployment requires a balance between predictive accuracy, calibration, interpretability, transportability and actionability.

For heart-failure readmission, the preferred strategy is to begin with a strong transparent benchmark and accept additional algorithmic complexity only when it produces a reproducible and clinically meaningful improvement.

The objective of explainable machine learning should ultimately not be to make an opaque algorithm appear understandable.

It should be to produce predictions that clinicians can evaluate, trust, and use to select interventions capable of changing patient outcomes.

## Declarations


### Funding

No external funding was received.


### Conflict of Interest

The authors declare no conflict of interest.

### Ethical Approval

Not applicable. This manuscript synthesized previously published evidence and involved no direct patient-level data collection.

### Consent to Participate

Not applicable.

### Data Availability

No novel patient-level dataset was generated or analyzed.

### Author Contributions

The authors contributed to conceptualization, evidence synthesis, methodological interpretation, manuscript drafting, and critical revision.

## References


1. Dharmarajan K, Hsieh AF, Lin Z, Bueno H, Ross JS, Horwitz LI, et al. Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. JAMA. 2013;309(4):355-363. doi:10.1001/jama.2012.216476.

2. Mortazavi BJ, Downing NS, Bucholz EM, Dharmarajan K, Manhapra A, Li SX, et al. Analysis of machine learning techniques for heart failure readmissions. Circ Cardiovasc Qual Outcomes. 2016;9(6):629-640. doi:10.1161/CIRCOUTCOMES.116.003039.

3. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, 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. 2017;2(2):204-209. doi:10.1001/jamacardio.2016.3956.

4. Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC Med Inform Decis Mak. 2018;18:44. doi:10.1186/s12911-018-0620-z.

5. Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Dwivedi G. Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. ESC Heart Fail. 2019;6(2):428-435. doi:10.1002/ehf2.12419.

6. Shin S, Austin PC, Ross HJ, Abdel-Qadir H, Freitas C, Tomlinson G, et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. ESC Heart Fail. 2021;8(1):106-115. doi:10.1002/ehf2.13073.

7. Sharma V, Kulkarni V, McAlister F, Eurich D, Keshwani S, Simpson SH, et al. Predicting 30-day readmissions in patients with heart failure using administrative data: a machine learning approach. J Card Fail. 2022;28(5):710-722. doi:10.1016/j.cardfail.2021.12.004.

8. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004.

9. Ouwerkerk W, Voors AA, Zwinderman AH. Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure. JACC Heart Fail. 2014;2(5):429-436. doi:10.1016/j.jchf.2014.04.006.

10. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. JAMA. 2017;318(14):1377-1384. doi:10.1001/jama.2017.12126.

11. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. doi:10.1186/s12916-019-1466-7.

12. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206-215. doi:10.1038/s42256-019-0048-x.

13. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82-115. doi:10.1016/j.inffus.2019.12.012.

14. Liu Y, Chen PHC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. JAMA. 2019;322(18):1806-1816. doi:10.1001/jama.2019.16489.

15. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). Ann Intern Med. 2015;162(1):55-63. doi:10.7326/M14-0697.

16. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51-58. doi:10.7326/M18-1376.

## Table 1. Evidence eligibility framework

| Domain | Inclusion criteria | Exclusion criteria |
|---|---|---|
| Publication date | ≤1 December 2022 | 2023 onward |
| Journal quality | Q1 journal in relevant category | Lower-quartile/unranked sources |
| Population | Adults with heart failure | Non-HF population unless methodological evidence |
| Outcome | 30-day readmission or closely related 30-day readmission/death | Long-term prognosis only |
| Model | Regression, ML, ensemble, neural/deep learning | Models without predictive evaluation |
| Performance | AUROC/c-statistic, AUPRC, sensitivity, specificity, PPV, calibration | Accuracy alone without interpretable context |
| Interpretability | Native or post-hoc explanation methods | No relevance to clinical model understanding |
| Study type | Prediction study, systematic review, methodological paper | Editorial opinion without analytical contribution |

## Table 2. Selected Q1 evidence on machine-learning prediction of early heart-failure readmission

| Study | Population/data | Model approach | Principal performance finding | Interpretation |
|---|---|---|---|---|
| Mortazavi et al. [2] | HF hospitalization cohort | LR, RF, boosting, hybrid models | ML improved several readmission predictions over LR; 30-day HF-specific discrimination remained modest | ML advantage did not eliminate intrinsic prediction difficulty |
| Frizzell et al. [3] | Hospitalized HF patients | Multiple ML and conventional statistical models | No clinically important improvement from ML | More complexity did not guarantee better prediction |
| Golas et al. [4] | 11,510 patients; 27,334 admissions | LR, boosting, maxout, deep unified network | c-statistic 0.664, 0.650, 0.695, and 0.705 respectively | Deep learning produced modest discrimination gain |
| Awan et al. [5] | HF cohort with imbalanced outcomes | MLP and conventional models | Model performance depended strongly on class-imbalance handling and metric choice | AUROC alone may obscure practical performance |
| Shin et al. [6] | Systematic HF model comparison | ML versus conventional models | ML did not consistently produce major performance advantage | Study quality and validation varied |
| Sharma et al. [7] | 9,845 HF patients; independent 20% validation | XGBoost, neural network, modified LaCE | AUROC 0.65, 0.58, and 0.57 respectively | XGBoost improved discrimination but remained only moderately informative |

## Table 3. Explainability characteristics of common readmission models

| Model | Native interpretability | Predictive flexibility | Patient-specific explanation | Principal limitation |
|---|---|---|---|---|
| Logistic regression | High | Moderate | Direct from coefficients | Linear/additive assumptions unless expanded |
| Sparse regression/risk score | Very high | Moderate | Straightforward | May sacrifice nonlinear information |
| Small decision tree | Very high | Moderate | Direct decision path | Instability/overfitting |
| Random Forest | Moderate-low | High | Requires additional interpretation | Individual trees not representative |
| XGBoost | Moderate-low | High | SHAP/local methods useful | Explanation layer adds complexity |
| MLP | Low | High | Post-hoc explanation required | Opaque internal representation |
| Deep neural network | Very low | Very high | Post-hoc explanation required | High complexity and explanation uncertainty |

## Table 4. Recommended model-performance domains

| Domain | Metric | Why it matters |
|---|---|---|
| Discrimination | AUROC/c-statistic | Measures ranking ability |
| Minority-event performance | AUPRC | More informative when readmissions are relatively uncommon |
| Case detection | Sensitivity | Determines proportion of future readmissions identified |
| Alert burden | Precision/PPV | Determines how many flagged patients actually return |
| Low-risk reassurance | NPV | Useful when excluding intensive intervention |
| Probability accuracy | Calibration intercept/slope | Determines reliability of predicted risk |
| Overall probability error | Brier score | Captures probabilistic accuracy |
| Clinical value | Decision-curve analysis | Evaluates net benefit at actionable thresholds |
| Transportability | External/temporal validation | Tests performance beyond development sample |

### Table 5. Proposed clinical model-selection hierarchy

| Stage | Requirement | Decision |
|---|---|---|
| 1 | Develop transparent benchmark | Logistic/penalized regression |
| 2 | Add nonlinear models | RF/XGBoost |
| 3 | Compare discrimination | Require meaningful improvement |
| 4 | Compare calibration | Reject poorly calibrated gains |
| 5 | Test external dataset | Require transportability |
| 6 | Generate local explanations | Assess clinical plausibility |
| 7 | Assess actionability | Link risk factors to interventions |
| 8 | Perform decision analysis | Confirm net benefit |
| 9 | Prospective impact study | Demonstrate improved patient/system outcomes |

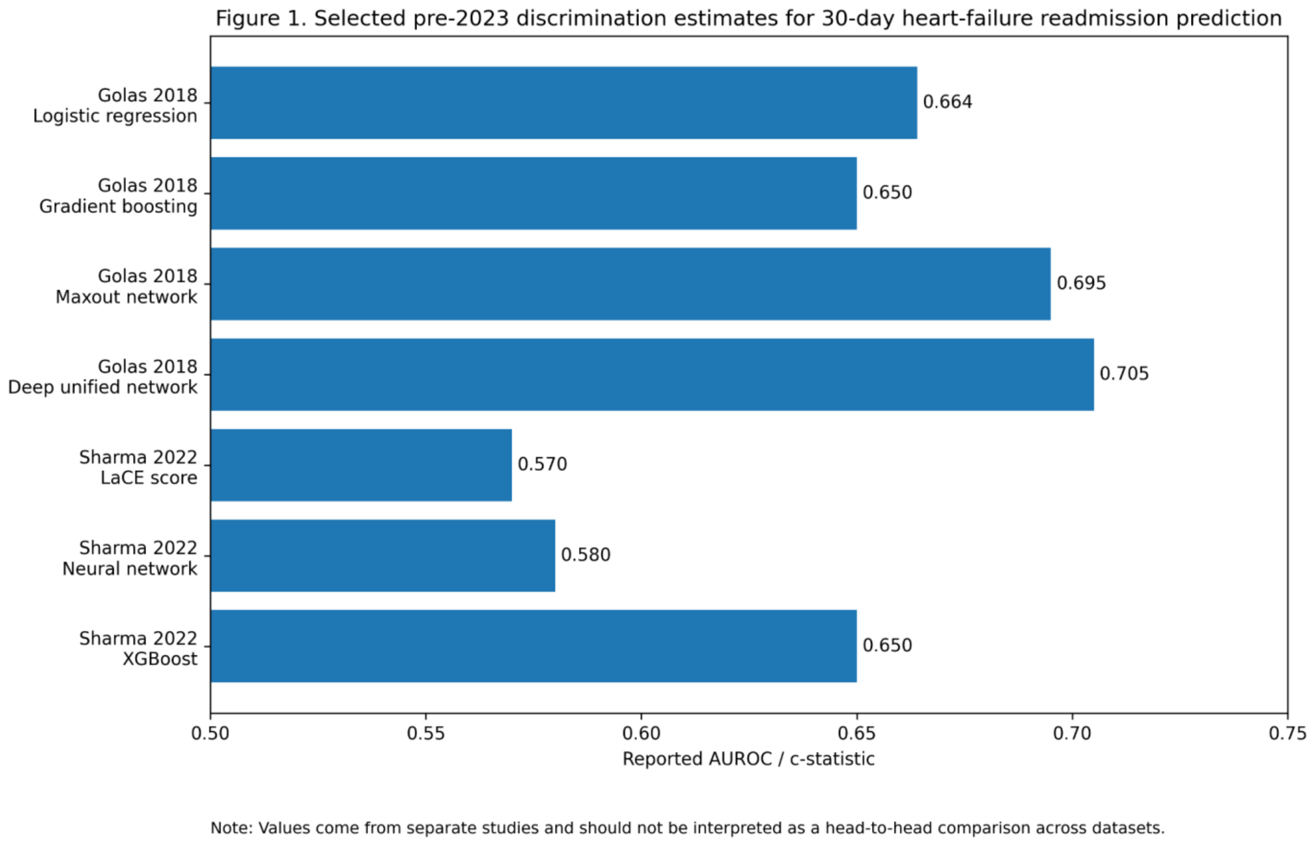


**Figure 1. Selected pre-2023 discrimination estimates for 30-day heart-failure readmission prediction.**

*Values are shown for contextual comparison only. Estimates come from separate studies and should not be interpreted as a head-to-head comparison across datasets, populations, feature sets or validation methods.*

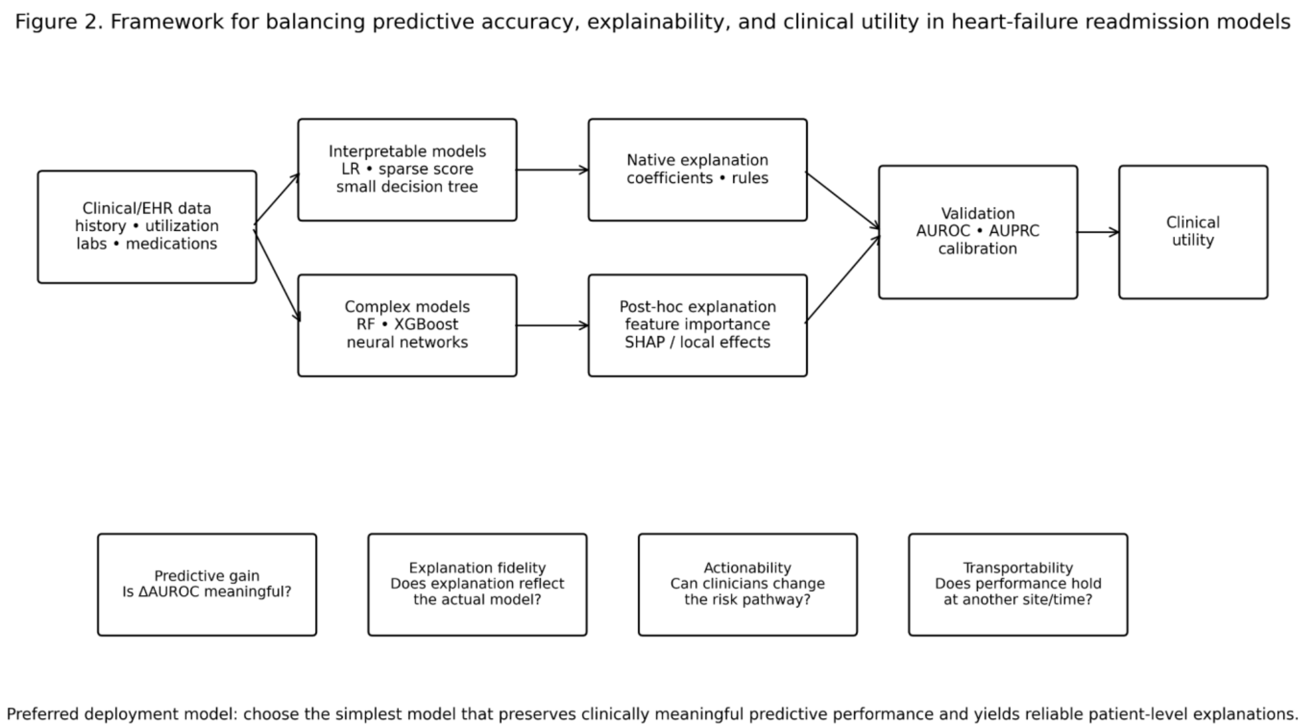


**Figure 2. Framework for balancing predictive accuracy, explainability and clinical utility in heart-failure readmission models.**

*Both interpretable and complex models should satisfy the same validation requirements. Additional model complexity is justified only when predictive improvement is clinically meaningful and explanations remain reliable, transportable and actionable.*

**How to cite: Morgan E (2023). Explainability Versus Predictive Accuracy in Machine-Learning Models for 30-Day Hospital Readmission Among Patients With Heart Failure: A Systematic Evidence Synthesis. Journal of Multidisciplinary Research and Integrated Sciences, 3(1)**