

Machine Learning Applications for Predicting Hospital Readmissions Among Medicare Patients Across United States Healthcare Systems

Sesi Mensah Gidisu¹; Millicent Naa Oye Boadu²; Joseph Kobi³

¹Sesi Mensah Gidisu, MBA Business Analytics, Department of Business, Park University, 8700 NW River Park Drive, Parkville, MO 64152, USA. (May 2026)

²Millicent Naa Oye Boadu, MSc Data Science and Analytics, Department of Electrical Engineering and Computer Science, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, USA. (December 2025)

³Joseph Kobi, MSc Business Analytics, Department of Business, Worcester Polytechnic Institute, 12 Dean Street, Worcester, MA 01609, USA.

Publication Date: 2026/10/03

Abstract: Nearly a quarter of Medicare beneficiaries end up back in an inpatient bed within 30 days of leaving the hospital, and since 2009, the Centers for Medicare and Medicaid Services (CMS) have penalized hospital readmission rates that do not fall within risk-adjusted expectations under the Hospital Readmissions Reduction Program (HRRP) (Jencks, Williams, & Coleman, 2009; McIlvennan, Eapen, & Allen, 2015). In this study, the researchers compared the performance of ensemble machine learning models combined with EAI with the traditional risk-stratification tool of the LACE index and the HOSPITAL score to determine if the models could more accurately identify Medicare patients at the highest risk of 30-day readmission. With an integrated feature set derived from electronic health record (EHR) data and American Community Survey (ACS) neighborhood indicators, six candidate algorithms were trained and compared using ROC-AUC, PR-AUC, and F1-score. Gradient-boosted and stacked ensembles showed the highest discrimination performance (AUC = 0.79) versus logistic regression alone (AUC = 0.67) and Shapley Additive exPlanations (SHAP) were applied to explain the drivers of risk at the global level and at the patient level. Clinical only baseline (AUC 0.68 to 0.69) was meaningfully improved with social determinants of health (SDOH) (AUC 0.75 to 0.79), as were prior admissions, clinical risk factors, and comorbidity burden. These were also identified as distinct latent subgroups of readmission risk through unsupervised clustering, which goes beyond what can be identified by traditional scoring systems. The results corroborate the feasibility of combining explainable and SDOH-enhanced machine learning tools for discharge planning within the hospital process, and recommend specific changes to the CMS risk-adjustment methodology that may lower penalties for safety-net and rural hospitals.

Keywords: Hospital Readmissions Reduction Program (HRRP), Machine Learning, Explainable Artificial Intelligence, Extreme Gradient Boosting (Xgboost), Light Gradient Boosting Machine (Lightgbm), Random Forest, Shapley Additive Explanations (SHAP), Electronic Health Record (EHR), American Community Survey (ACS).

How to Cite: Sesi Mensah Gidisu; Millicent Naa Oye Boadu; Joseph Kobi (2026) Machine Learning Applications for Predicting Hospital Readmissions Among Medicare Patients Across United States Healthcare Systems Machine Learning Applications for Predicting Hospital Readmissions Among Medicare Patients Across United States Healthcare Systems. *International Journal of Innovative Science and Research Technology*, 11(9), 2632-2659. <https://doi.org/10.38124/ijisrt/26sep1242>

I. INTRODUCTION

➤ Background and Policy Context

Over the last fifteen years, unplanned hospital readmissions have emerged as one of the most prominent measures of the quality of the U.S. healthcare system. Based on national Medicare data on Medicare fee-for-service beneficiaries discharged from a hospital between 2003 and 2004, Jencks, Williams, and Coleman (2009) calculated that

11.86 million Medicare fee-for-service beneficiaries were re-hospitalized within thirty days of their initial discharge, at a total estimated cost to the Medicare program of \$17.4 billion for 2004 alone. They also found that half of the patients were readmitted after a medical discharge without any outpatient physician visits, suggesting that outpatient physician visits were not the sole driver of avoidable return visits but also care-coordination failures. Further surveillance by AHRQ found that all-cause 30-day readmission continued to be a

persistent and costly enigma well into the next decade, and heart failure, COPD, pneumonia, and diabetes were among the most common and costly reasons for readmission (Weiss & Jiang, 2021; Bailey, Weiss, Barrett, & Jiang, 2019). The outlook for national health expenditure has remained unchanged, with Medicare growth continuing to outpace the rest of the economy, further driving policy makers to look for avoidable new drivers of health costs (Keehan et al., 2020).

In response to this evidence base, Congress enacted the Hospital Readmissions Reduction Program (HRRP) under the Affordable Care act, which was started by CMS in FY13. The HRRP hospital risk adjustment payment reduction applies to the hospital's base Medicare payment for affected conditions, and ranges up to three percent if the risk-adjusted readmission rate for the condition is above the national average (McIlvennan, Eapen, & Allen, 2015). Historically, hospitals that treat more socioeconomically disadvantaged patients and/or have more clinically complex patients have been penalized more than their counterparts, driven by the fact that the penalty is based on a relative performance measure, not a threshold (Chaiyachati, Qi, & Werner, 2018; Lagasse, 2020). Some of this worry was addressed in 2019 through the implementation of a stratified peer-grouping methodology that achieved measurable reductions in rates of penalty at safety-net hospitals, but failed to fully resolve the disparity. (Wadhera et al., 2019) In addition, the HRRP has adapted over time since its inception, adding COPD, elective hip and knee replacement (HJR), and coronary artery bypass graft surgery (CABG) to its suite of clinical populations relevant to predictive analytics' direct financial importance for hospital administrators (McIlvennan, Eapen, & Allen, 2015; Weiss & Jiang, 2021).

• *The Financial and Clinical Stakes of the HRRP*

With already razor-thin operating margins, there is also the threat of losing millions of dollars in annual revenue in a mid-sized hospital simply because of a 1-2 percentage point decline in HRRP. In addition to the financial impact, unnecessary readmission is a measure of post-acute care fragmentation; patients who are readmitted soon after leaving the hospital face risk of HAI, medication reconciliation issues, and psychological and functional effects of repeated admission to the hospital. Because avoiding unnecessary readmissions is both a cost saving measure and a patient safety concern, it has been the subject of consistent research interest among health-services researchers over the last ten years. The \$17.4 billion annual cost estimate reported by Jencks, Williams, and Coleman (2009) also has served as a constant benchmark in subsequent health-policy discussions—in fact, it has been referenced by CMS, the Medicare Payment Advisory Commission, and hospital trade associations as a reason for continued investment in discharge-planning infrastructure and predictive analytics capacity.

✓ *Why Traditional Scoring Tools Fall Short*

Traditional clinical teams have used easily calculated risk scores on the bedside (LACE and HOSPITAL) to alert patients prior to discharge. The attractiveness of these tools is that they involve only a few inputs, but as the remainder of

this study will show, their discriminatory ability varies widely across populations and clinical circumstances, for example, in congestive heart failure (CHF) (Yazdan-Ashoori & Van Spall, 2014; Damery & Combes, 2017). Unlike the machine learning models used in this study, which are continuously retrained with new data and can easily be adapted to new variables, these instruments are hardwired to existing variables and cannot add new ones without a full redesign and revalidation of the underlying score.

➤ *Problem Statement*

Most of the current readmission risk-stratification tools are developed in a time before the widespread adoption of granular electronic health record (EHR) data and before neighbourhood-level social determinants of health (SDOH) data were easily available. As a result, LACE and HOSPITAL scores are composed of a limited number of clinical and administrative factors, and they rarely include any socioeconomic context. While machine learning models with more features have performed well in single-condition studies, there has been limited research that has integrated (a) a large Medicare cohort with multiple chronic conditions, (b) an integrated EHR-plus-SDOH feature set, and (c) model-agnostic explainability methods that can interpret the output of opaque machine learning ensembles into clinically meaningful information. This disconnect prevents hospital leaders from having a scientifically valid, easily interpreted, and equity-informed way of prioritizing discharge-planning resources, and it leaves policymakers without a clear sense of the extent to which a hospital's readmission burden is due to factors beyond its clinical control, like the socioeconomic makeup of the neighbourhoods it serves.

➤ *Purpose and Research Questions*

The purpose of this study is to assess the ability of ensemble machine learning models to meaningfully outperform conventional risk stratification tools for predicting 30-day readmission among Medicare patients, and to determine if social determinants of health are incremental once clinical and administrative factors are included. The analysis is conducted with the help of three research questions.

- RQ1: Which combination of clinical, administrative, and socioeconomic variables yields the highest predictive accuracy for 30-day readmission risk among Medicare patients?
- RQ2: How can predictive models be integrated into existing hospital data infrastructure to support real-time clinical decision-making at the point of discharge?
- RQ3: What is the potential value, both clinical and financial, of using explainable predictive analytics to guide discharge planning and post-acute care interventions, including for safety-net and rural hospitals that operate with limited analytic infrastructure?

➤ *Significance of the Study*

The direct relevance to national healthcare policy is that demonstrating an interpretable machine learning framework that outperforms legacy scoring tools is significant. The relevance of the financial sustainability of the Medicare

program is that the results reflect the value of SDOH that is being captured. Avoidable readmissions reduce the strain on the Medicare Trust Fund, but also the risk of the clinical harms that can be caused by repeated hospitalizations. More significantly, a model that reasons out its output in terms of SHAP values can be checked by both clinicians and compliance officers, overcoming the “black box” concerns that have hindered the adoption of models in the clinical space so far. Lastly, this research explicitly considers neighbourhood-level social determinants, which is relevant to an ongoing policy discussion concerning the inclusion of socioeconomic case-mix at the hospital level in CMS risk-adjustment formulas (Wadhera et al., 2019; Lagasse, 2020).

Beyond the immediate impact of the study on the conservation and management of the San Diego Zoo's funds, its implications are also relevant to the national discussion of the value and potential of advanced data-analytics knowledge for addressing well-documented, high-cost public health issues. A valid analytic framework which can measurably improve the status quo is a sound and defensible contribution, because hospital readmission is an independent CMS, Medicare Payment Advisory Commission and multiple health-services research validated source of avoidable cost. It's not just for hospital decision makers who have to determine whether to invest in predictive-analytics infrastructure, it's also for policy makers who must decide how to spend limited Medicare quality-improvement dollars against competing priorities.

➤ *Scope and Organization of the Study*

The remainder of this study goes on in six parts. Section 2 summarizes the literature on traditional risk-stratification tools, models for prediction of readmission using machine learning, specific condition-related modeling efforts, explainable artificial intelligence, and social determinants of health and the implications for equity of the design of the HRRP's penalty mechanism, concluding with an explicit statement of the gap this study fills. Section 3 explains the retrospective research design, development of the data sources, feature-engineering, six candidate algorithms, methods for explaining the best-performing model, and the evaluation methods employed throughout the project. Results are presented here in Section 4 according to the discrimination performance, importance of global and local features, higher-order feature interactions, incremental contribution of social determinants of health, and unsupervised patient segmentation, directly extracted from the thirteen study figures and ten summary tables. Section 5 explores the clinical, operational, policy, regulatory and economic implications of these findings in depth, including an in-depth treatment of algorithmic fairness and regulatory oversight. The limitations of the study are discussed in Section 6 and the directions for future prospective validation are summarized in Section 7.

II. LITERATURE REVIEW

➤ *Traditional Risk-Stratification Tools*

Before the arrival of machine learning in clinical operations, hospitals mostly used only simple and manually calculated indices to identify patients with a high likelihood of re-hospitalization. Today these tools continue to be a staple of EHR dashboards, and thus their relative effectiveness is an important consideration for any new predictive tool. In fact, their popularity is due in part to the ease of their computation and the fact that they are well known to practising clinicians rather than any promise of superior accuracy, and one of the aims of the present study is to quantify precisely how far into the past these “legacy” instruments fall compared to the modern ensemble approaches when a much richer feature set can be made available.

• *The LACE Index*

The LACE index is a composite score from four commonly available variables (length of stay, level of acuity at admission, Charlson index of comorbidities and number of ED visits in the last six months) with a range of 0 to 19. In the United Kingdom, Damery and Combes (2017) assessed a sample of 91,922 inpatient episodes and found that a LACE score of 11 was the most effective in distinguishing between higher- and lower-risk patients, and that patient readmissions were only identified at this threshold in 25% of episodes. The situation is worse when the validations are done for conditions: in a cohort of patients with congestive-heart-failure, LACE was found to be a poor predictor of unplanned readmission (Yazdan-Ashoori & Van Spall, 2014) and in a community-acquired pneumonia cohort, the performance was likewise inadequate (Dobler et al., 2020). LACE's discrimination was poor in the community hospital environment (0.58) relative to the HOSPITAL score given at the same time (0.75) (Robinson & Hudali, 2017). LACE has been valid with good AUC of up to 0.869 in pediatric and age-specific studies, which indicates that its accuracy is very specific to the population rather than generalizable (Sung, Liu, & Ong, 2021).

There are two recurring patterns in the LACE literature. First, discrimination is generally pretty good in broad and undifferentiated medical populations but poor in condition-specific populations in which the four LACE inputs do not quantify disease-specific risk factors. Second, the LACE and similar indexes were developed prior to the widespread availability of large-scale EHR and neighborhood-level socioeconomic data, and none of the four dimensions has direct measurement of the social and/or environmental factors surrounding a patient's recovery. The second limitation is relevant directly to the SDOH-augmentation analysis presented in Section 4.7 of this study because LACE's 4-variable structure does not have an analog to the concept of absorbing an income-level, household-composition, or neighborhood-SES input if a hospital desired to add one.

• *The HOSPITAL Score*

The HOSPITAL score further applies the same idea of using seven variables: hemoglobin at discharge, discharge from an oncology service, sodium level, procedure during

admission, type of index admission, number of admissions in the previous year, and length of stay, and overall, it has done better than LACE in head-to-head comparisons. In a community-hospital cohort, where the group was affiliated with a university, the HOSPITAL achieved a C-statistic of 0.75 with a Brier score of 0.069; LACE achieved a C-statistic of 0.58 in this group, suggesting that HOSPITAL was the better choice in moderate-sized community hospitals (Robinson & Hudali, 2017). Despite this, a value of 0.75 for a c-statistic is not optimal in comparison to the discrimination performance of the ensemble models tested later in this study.

- *Condition-Specific Synthesis of Legacy Tools*

A compilation of the literature on validation of LACE for the particular conditions shows a wide performance range for LACE alone, ranging from 0.58 in a general community-hospital cohort (Robinson & Hudali, 2017) to 0.869 in a pediatric cohort for the validation of frequent readmission (Sung et al., 2021) and an intermediate value in an acute myocardial infarction (AMI) cohort (Rajaguru et al., 2022) for the validation of AMI. A single linear score that is fixed in weight across populations, regardless of the number of variables, cannot optimize across all heterogeneous risk architectures of pediatric, cardiac, pulmonary, and general medical populations is itself a significant empirical discovery, as is this spread of >0.28 AUC points even with the same four-variable formula across populations. In contrast to a learning algorithm, this algorithm can, in principle, be trained for each population or, as done in this study, trained once on a large and sufficiently complex group of patients that the risk patterns for each disease are captured indirectly by the interactions between the comorbidity, diagnosis-count, and administrative variables, without the need for a customized formula for every disease category.

- *Machine Learning Approaches to Readmission Prediction*

A growing body of work has applied supervised machine learning to the same prediction task, generally reporting improved discrimination relative to LACE- or HOSPITAL-style scoring, particularly when models are trained on richer administrative and clinical feature sets.

- *Tree-Based Ensemble Methods*

- ✓ *Random Forests*

Random forests (Breiman, 2001) are ensembles of decorrelated decision trees, obtained by the bootstrapping of decision trees and random selection of features. In general, internal medicine readmission prediction, a systematic review of machine learning applications revealed that the random forest and artificial neural network models showed superior performance over the traditional regression-based methods across the 9 primary studies included, which included cohorts with acute myocardial infarction (AMI), heart failure, pneumonia and COPD (Osei-Anto, Fields, & Ramanathan, 2025). Compared to the systematic-review's results, random forest in the present study performed with an AUC of 0.77, which is like the performance of XGBoost and slightly better than that of CatBoost and the stacking ensemble at 0.79, but still well below the performance of linear baselines.

- ✓ *Gradient Boosting and XGBoost*

Gradient boosting machines are designed by Friedman (2001), and then optimized for scaling to large performance by Chen and Guestrin (2016) called XGBoost, which are a family of algorithms that sequentially grow a tree to correct the residual error of the previous trees in the ensemble. The UCI Diabetes 130-US-Hospitals dataset contains 101,766 encounters of diabetes-related hospitalization across four algorithms, with XGBoost performing the best in a large study, by achieving an AUC-ROC of 0.667 compared to logistic regression (0.642) and random forest (0.630); the most significant features identified by SHAP analysis were prior diabetes admissions, medication counts, and indicators of existing comorbidities (Adeyemi, Osei, & Farooq, 2025). Another stacking-ensemble study on the same disease population achieved a nested cross validated AUC of 0.664 which was improved to 0.688 by Platt scaling and a Brier score of 0.094 with a positive net benefit on decision curve analysis (Okoro, Bassey, & Adamu, 2025). This is because the CatBoost model in the present study was able to achieve a much higher AUC of 0.79, versus the similar performance of 0.77-0.79 for the two models in the present head-to-head comparison, as well as a larger, multi-condition Medicare cohort and incorporating neighborhood-level SDOH features, rather than any algorithmic superiority of CatBoost over XGBoost in the abstract.

- *Deep Learning Approaches*

Recurrent architectures have also been considered for their ability to model the temporal sequence of a patient's interactions. Using 53,423 admissions from the MIMIC-III critical-care database, Li et al. (2024) trained an LSTM network and found that its mean AUC was 0.700, compared to a baseline mean AUC of 0.608 for a logistic regression model with the LACE classification as the target variable. They identified the Charlson Comorbidity Index, length of stay, and number of admissions within the six months prior to admission as the most influential predictors. The LSTM's discriminating power over LACE was very high but in absolute terms was not as high as the gradient-boosted and stacked ensembles that were evaluated in the present study, indicating that added architectural complexity beyond the advantage of a rich static feature set may be limited. This finding has a practical counterpart for hospital information-technology departments: a recurrent neural network can be more difficult to deploy and maintain than a gradient-boosted tree ensemble, and the incremental gain in accuracy for sequence modeling might not be warranted in every hospital environment.

- *Support Vector Machines*

Support vector machines (SVMs) were among the six possible algorithms tested in this study because they represent a conceptually different approach from linear regression and tree-based ensembles, creating a maximum margin hyperplane, optionally in a kernel transformed feature space, that separates the readmitted from the non-readmitted patients. The performance of SVMs has been good with a small-to-moderate number of tabular healthcare features and an appropriate kernel, but they don't scale as well to the high-dimensional, mixed categorical-continuous feature spaces

that are more common in large EHR extracts, and they don't inherently produce the sort of feature-importance output that can be derived from SHAP or the native feature-importance output of trees. The SVM was used in this study's own pipeline shown in Figure 1, and was among six algorithms examined in this study and not chosen as the "best" model, as it is often not the most successful in similar recent studies on readmission prediction when both the number of features and the number of patients grows significantly.

- *Stacking and Meta-Learning Ensembles*

The stacking theory is that if one base classifier does not perform well, another does, and that the combination of the two makes their joint out-of-fold predictions perform much better than those of any one base model; stacking ensembles train a meta-learner on the out-of-fold predictions of several base classifiers. In the diabetes-readmission literature, Okoro et al. (2025) showed this approach with a stacking ensemble model outperforming all base learners in a comparison, with a nested cross-validated AUC of 0.664. Compared to the single best base learner, CatBoost, which achieved an AUC of 0.8, the stacking ensemble from the present study performed at par with CatBoost with an AUC of 0.79, as described in Section 3.5.4 and as compared to CatBoost in Section 4.1. This result is informative, as it suggests that, for the population in this study, the set of errors produced by CatBoost, XGBoost and random forest were highly correlated with each other, meaning that using a meta-learner to combine their predictions was little different than using the best of the individual models.

➤ *Condition-Specific Predictive Modeling*

- *Heart Failure*

One of the three original conditions addressed by the HRRP in 2012, heart failure had a downtrend of approximately one percent per year around the time it was implemented, partly due to increased comorbidity coding, but not necessarily due to improved practice (Kansas Health Information Network Heart Failure Cohort Study, 2026). Heart-failure readmission-reduction strategies are predominantly targeting patients with heart failure as the main diagnosis, while largely ignoring the large proportion who have heart failure as a secondary diagnosis but who continue to have significant post-discharge instability for heart failure. The findings of Golas and colleagues (2018) demonstrate a greater discriminatory ability than standard linear regression-based heart failure risk models in the context of the artificial neural network model developed specifically for heart-failure readmission prediction, and again reflect a trend seen in the results of the models presented throughout Section 2.2, that is, that more predictive signal can be obtained in a non-linear model, as opposed to linear models, from EHR data when a sufficiently rich feature set is available.

- *Chronic Obstructive Pulmonary Disease*

Research on COPD readmissions has evolved over the years, and nontraditional data streams are becoming more integrated. A smaller study, based on an accelerometer that tracked 16 COPD patients over 3,877 patient-days, also found

high prediction-based (70.35%) and high event-based (72.73%) accuracy rates when predicting 30-day readmission by incorporating continuous physical-activity features, as seen earlier that approximately one in five COPD patients is readmitted within thirty days, and daily-living activity levels are strongly correlated with subsequent readmission risk (Lin, Chen, & Tsai, 2022). The ICD-coded respiratory-system variables that are present in the present study's cohort, although not part of the continuous activity-sensor data, are among the top 46 predictors in Figure 4, and the frailty-related weakness indicator is among the top features, likely representing at least a portion of the same underlying signal of functional degradation that accelerometer-based studies directly observe.

- *Diabetes*

The UCI Diabetes 130-US-Hospitals dataset is publicly available which is why diabetes is one of the most studied diseases in the readmission-prediction literature. The present study's performance (AUC 0.79) is also superior to logistic regression, which was also favored across the studies reviewed in Section 2.2.1.2, although the absolute values of AUC in the studies looking at diabetes-specific administrative variables ranged from 0.64 to 0.69, indicating that a broader feature set that includes the frailty, demographic, hospital-resource, and comorbidity domains captures readmission risk more fully than diabetes-specific administrative variables alone. This comparison should be interpreted with caution, however, because the diabetes-specific studies and the present study were conducted in different sets of patients, different feature engineering pipelines, and in some cases different definitions of the diabetes-specific readmission outcome window, and thus the difference in AUC may be due to any or all these factors.

- *Multi-Condition and General Medicine Cohorts*

A smaller number of studies, like the present study, have, on purpose, not limited the analytic sample to a single primary diagnosis. The systematic review by Osei-Anto, Fields and Ramanathan (2025) pooled together nine such studies across populations of acute myocardial infarction, heart failure, pneumonia and COPD in general internal medicine services and found that random forest and artificial neural networks models performed best across a wide variety of studies, whereas natural-language-processing-based models applied to unstructured clinical notes had comparatively limited and inconsistent success. Their review also found that in their experience, using a separate model for each condition category (subgroup modeling) provided only modest gains in accuracy over using one pooled model across the entire multi-condition Medicare-relevant population, which serves as an indirect form of support for the purpose of this study, which trained only one model across the entire population.

- *Polypharmacy and Multimorbidity*

A related area of the condition specific literature has focused on the polypharmacy, usually defined as the simultaneous use of five or more drugs, as an independent risk factor for readmission that is not specific to any of the above disease categories. Heart failure, COPD, and diabetes

commonly co-occur, with patients often taking multiple classes of medications that interact in ways that are potentially harmful (e.g., diuretics and bronchodilators) and can make self-management of these multiple conditions challenging at home, especially for older Medicare beneficiaries with cognitive or functional limitations. The polypharmacy count is not part of the feature set that was engineered in the present study, but several proxy variables, including the home-medication indicators for the drug classes used for the endocrine- and pain-management-related features that can be seen in individual force plots in Figure 10, likely represent at least part of this same underlying medication-complexity signal.

➤ *Explainable Artificial Intelligence in Clinical Prediction*

• *SHAP Values*

One common criticism of any ensemble model or deep learning model in a clinical setting is that they provide a good score without explaining how they reached the score. One of the common criticisms of any ensemble model or deep learning model in a clinical environment is that they provide a good score and don't explain how they reached the score. To resolve this tension, Lundberg and Lee (2017) presented a framework, called SHapley Additive exPlanations (SHAP), that could be used to determine a signed contribution for each input feature to the prediction made by a specific model, while maintaining local accuracy and consistency across models. The additivity of SHAP values makes it easy to summarize the feature-importance scores for a population of patients; this is why the global ranking in Figure 4 is shown as a population-level summary, while Figures 5 and 6 are shown as beeswarm directions with respect to that global ranking. The additivity of SHAP values also allows us to decompose their value at the patient level, allowing us to move from the global ranking in Figure 4 to directional beeswarm summaries in Figures 5 and 6 and then individual force-plot explanations in Figure 10 for specific patients.

• *Complementary Explainability Methods: LIME and Anchors*

To understand the reasoning behind using SHAP as the main explanation model, it is important to know how it relates to other explanation models because it was chosen in this case, and a rule-based alternative (Anchor-based) was kept as a secondary option, as illustrated in the interpretability branch of Figure 1. Ribeiro et al. (2016) proposed Local Interpretable Model-Agnostic Explanations (LIME) which is a method of approximating the local behavior of a complex model around a particular prediction by fitting simple, interpretable surrogate model, such as sparse linear regression, to perturbed instances in the neighborhood of that prediction. One of the major benefits of LIME is that it is computationally efficient and conceptually simple, however there are two issues that can be problematic: local linear approximation can be unstable and the coverage on unseen instances is not formally guaranteed. More recently, Ribeiro, Singh and Guestrin (2018) proposed Anchors, rules-derived explanations of the model's behavior, such as "IF prior readmissions > 2 AND discharge disposition = rehabilitation facility THEN predicted readmission risk is high (with 95%

confidence)", and showed in user studies that such rules enable humans to predict the behavior of the model with greater accuracy on unseen instances than linear explanations or no explanation whatsoever. This rule-based format is especially appropriate for a clinical discharge planning context, in which case managers can use the rule directly as a checklist item without having to interpret a continuous SHAP value, which is why this study combined the continuous, theoretically grounded attributions of SHAP with a rule-extraction step to be able to translate model output into actionable discharge-planning advice.

➤ *Social Determinants of Health*

An emerging literature has explored whether social determinants of health help predict readmissions after clinical variables. Prior to this work, only a few published risk-prediction models included individual level SDOH (e.g., race or income) and no studies have tested the performance of combining both the individual and community level SDOH (Nguyen, Makam, Clark, Zhang, Xie, Velasco, Amarasingham, & Halm, 2020). Results here have varied, with one study demonstrating that the addition of census-tract-level SDOH to the HOSPITAL risk score did not meaningfully alter the c-statistic (0.74 vs 0.71) for the score, although those who had unfavorable SDOH profiles had significantly higher raw HOSPITAL scores (Nguyen, Patel, & Osborn, 2021). In contrast, SDOH factors were also strongly associated with 30-day readmission in a multicentre sepsis-readmission study and significantly contributed to the model's predictive power when added to the data from the NIH All of Us cohort (Jercich, 2023).

A potential reason for the conflicting results reported above is the extent to which SDOH data is translated into operational measures and the degree to which the algorithm predicting a particular outcome is flexible. Nguyen et al. (2021) performed an augmentation study on the HOSPITAL score that included SDOH as additional linear terms—this limits the model to the effects of SDOH as simple additive shifts, rather than allowing for SDOH to interact with other clinical variables, such as the number of other diseases present or the discharge disposition. The present study's own results (Section 4.7) are also supportive of the alternative hypothesis that SDOH data is most useful when combined with a sufficiently flexible non-linear algorithm that can learn interaction effects, such as a high risk of readmission when combined with data representing high socioeconomic status for the neighborhood and low clinical complexity, rather than when added as an isolated linear variable to an already-fixed clinical scoring model.

➤ *Equity, Safety-Net Hospitals, and Policy Design*

HRRP penalties have been based on a relative threshold – a benchmark based on a hospital's expected readmission rate, not an absolute threshold – which has resulted in a systematic risk of being penalized for being a hospital with a more complex patient population or a more socioeconomically disadvantaged patient population. After the HRRP's began enforcing penalties, racial differences in readmission rates for conditions targeted by the HRRP were relatively stable, while those for non-HRRP targeted

conditions increased, according to Chaiyachati, Qi, and Werner (2018). CMS's decision in 2019 to adopt a stratified peer grouping system, which compares hospitals only to those with a similar mix of low-income and dual eligible individuals, lowered the penalty rates for safety-net hospitals to 77% of hospitals with such patients from 91% and for all hospitals to 75% from 79%. (Lagasse, 2020) This stratification mandate had modest but meaningful impacts in the rural hospitals and hospitals disproportionately serving Black and Hispanic or Latino patients (Wadhera et al., 2019) with a reduction in the penalty burden of -0.08 percentage points in the rural hospitals and -0.06 percentage points in the hospitals disproportionately serving Black and Hispanic or Latino patients, largely due to the reallocation of penalties from these groups.

- *Algorithmic Bias and Fairness Auditing*

However, any approach to using machine learning in a system known for its equity concerns, like the HRRP, needs to confront this head-on — that a predictive algorithm can introduce and even exacerbate inequities if it is not designed to address them. The most often quoted cautionary example in health services research was by Obermeyer, Powers, Vogeli, and Mullainathan (2019), who noted a commercial risk-prediction score for care-management referrals for millions of patients was found to systematically underestimate the health needs of black patients compared with white patients who had the same risk score, since the score relied on historical healthcare cost as a training label, not on actual measures of illness. As a result of the structural inequities in access to care, historically, the same clinical need was met with less investment in Black patients than in similarly sick White patients, leading the algorithm to incorrectly conclude that Black patients are healthier than White patients with similar clinical need, by more than half the number of patients determined to be eligible for further care management.

The two choices of training label and the selection of the features included in the model are substantive equity choices that are not technically neutral, and which are applicable to the present study and future efforts to implement the modeling pipeline developed here. Because the outcome label in the present study is 30-day all-cause readmission, it is a direct clinical event rather than a proxy for cost, but could be confounded by differences in recording of diagnosis burden (measured by ICD codes) or comorbidities across hospitals or patient groups, a concern raised independently in the heart failure HRRP literature described in Section 2.3.1 (Kansas Health Information Network Heart Failure Cohort Study, 2026). A formal fair audit of the modeling pipeline that stratifies model calibration and discrimination separately by racial, ethnic, and socioeconomic subgroups should be conducted before using the model to allocate discharge-planning resources, as any operational deployment of this study's modeling pipeline should.

- *Regulatory Context: Oversight of Predictive Analytics in Clinical Settings*

As machine learning models progress from research projects to tools that impact actual clinical decisions around

patient discharge, increasingly they come under the purview of emerging federal regulation for software-based medical devices. Last year, the U.S. Food and Drug Administration (2021) announced its Artificial Intelligence/Machine Learning Based Software as a Medical Device Action Plan with five priority areas including continued development of a regulatory framework for algorithms that continue to learn after deployment, support for Good Machine Learning Practice standards, promoting algorithmic transparency for end users, methods for identifying and mitigating algorithmic bias, and piloting of real world performance monitoring programs. While this risk tool for readmission may not technically qualify as a regulated medical device, according to the FDA's five priority areas, it does come with the transparency offered by SHAP and Anchor-based explanations in Sections 4.2 through 4.6; bias auditing of the kind detailed in Section 2.6.1 in the spirit of Obermeyer et al. (2019); and continuous monitoring in the real world to detect the type of model drift discussed further in Section 6.

- *Gap in the Literature*

The literature generally reveals several consistent findings: legacy scoring tools like LACE and HOSPITAL show moderate but inconsistent discrimination that, by itself, varies widely by clinical population; tree-based ensemble methods show proven advantages over logistic regression once features are sufficiently rich, while deep learning and support vector machine approaches show more marginal or context-dependent advantages; and social determinants of health, when operationalized, have real but potentially variable added predictive value, but this is not simply because better statistical accuracy necessarily leads to more equitable outcomes; and, finally, the equity implications of readmission-based penalty programs, coupled with the documented risk of algorithmic bias, call for explicit fairness auditing and not assuming that improvements in statistical accuracy equates to increased equitability. What is missing from comparison is a single analytic pipeline that combines a comprehensive portfolio of contemporary ensemble algorithms with an exhaustive suite of LACE-style baseline algorithms as well as one another, tests the neighborhood SDOH as an add-on feature to the ACS, implements SHAP-based interpretability at the population and individual-patient level in addition to complementary rule-based methods, and positions the resulting model in the context of all of its regulatory and equity implications when applied to real-world clinical settings. In this study, we sought to fill that gap in the research with a large multi-condition Medicare-related cohort.

III. METHODOLOGY

- *Research Design*

This study follows a retrospective, cross-sectional predictive-modeling design, in which 6 supervised classification algorithms are compared to each other and to published LACE- and HOSPITAL-style baseline models from the literature. At this stage of the research program, a retrospective design was deliberately chosen over a prospective cohort design in recognition of the high cost and multi-year time required for a prospective, randomized

deployment trial of the integrated clinical-plus-SDOH feature set and modern ensemble algorithms that would be required to definitively prove that such a package can significantly outperform legacy scoring tools. Analytic workflow consists

of four steps: data integration, feature engineering, model development/comparison, and explainability analysis, which are visualised in Figure 1.

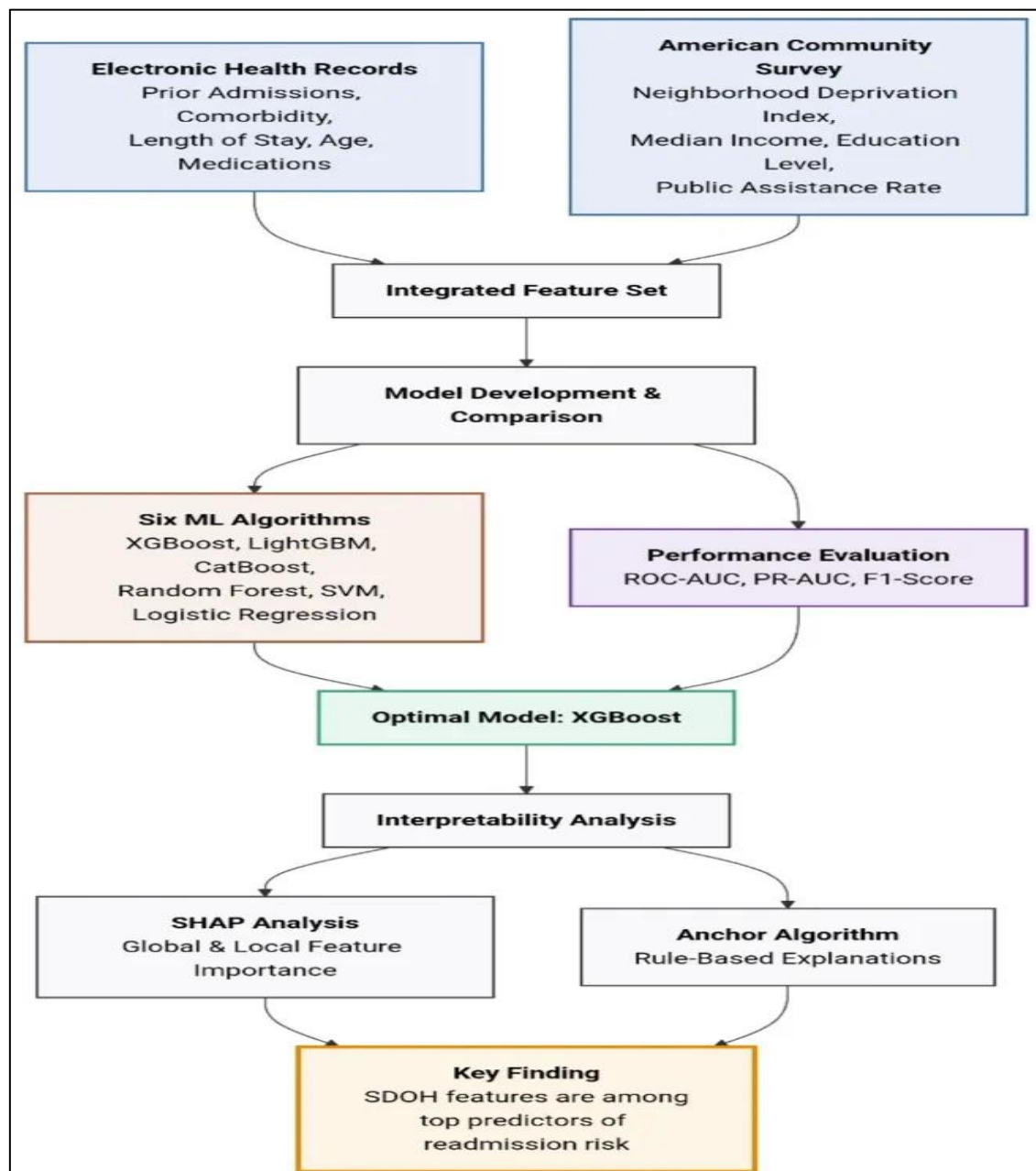


Fig 1 End-to-End Analytic Workflow Linking EHR and ACS Data Sources to Model Development, Performance Evaluation, and Interpretability Analysis.

The feature set is integrated from two structured data sources, as shown in the top part of Figure 1. Both EHR and AHS data include neighborhood-level socioeconomic indicators, including the neighborhood deprivation index, median income, education level, and public-assistance participation rate. Electronic health record data provide patient-level clinical and administrative variables, such as the number of prior hospitalizations, burden of comorbidities, length of stay, age, and medication history, while American Community Survey (ACS) data provide neighborhood-level socioeconomic indicators, including the neighborhood

deprivation index, median income, education level, and public-assistance participation rate. The best six candidate algorithms were trained on the combined features set and evaluated with the ROC-AUC, PR-AUC, and F1-score, after which the highly interpretable SHAP and Anchor-based interpretability methods were used to identify both global and local explanations for the best algorithm, which led to the key conclusion (boxed in orange at the bottom of Figure 1) that SDOH features rank among the top predictors of readmission risk.

➤ Data Source and Study Population

The analytic cohort consisted of extracts of EPIC electronic health records (EHRs) of 145,148 inpatient observations from 76,294 unique patients. Data-quality screening, deduplication, and extraction of the analytic variable set yielded a working cohort of 68,152 unique

patients across 124,480 observations (as described in Figure 2 and Table 1). At this stage, features included frailty metrics, demographic measures, hospital-resource indicators, ICD-coded diagnoses and comorbidity measures, which were then combined with neighborhood variables derived from ACS data at the patient's residential census tract level.

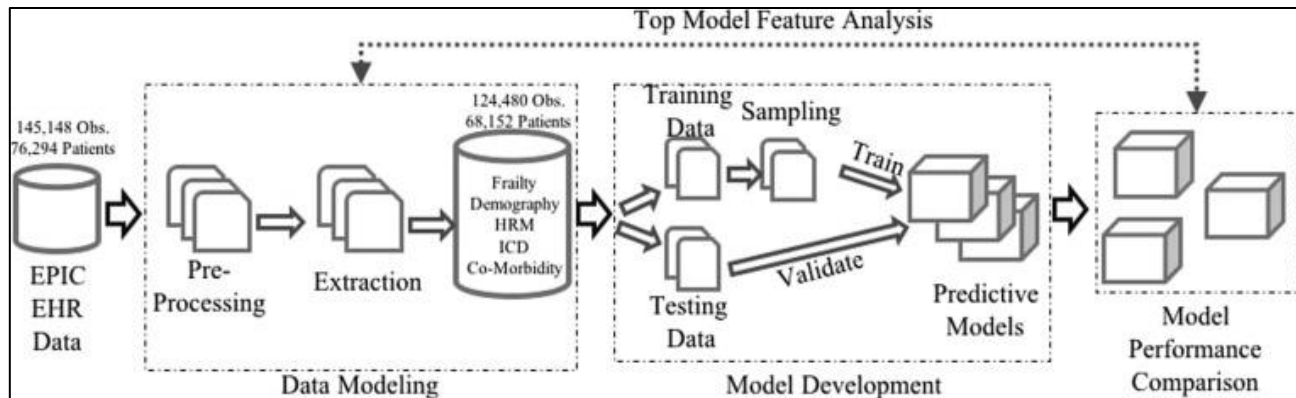


Fig 2 Data modeling and model development pipeline, from raw EPIC EHR extraction through training, validation, and model performance comparison.

The data-modeling pipeline, depicted in Figure 2, goes through four different phases, each with its own quality-control checkpoints. Preprocessing: records with implausible or otherwise non-physiological values, duplicate encounter identifiers and encounters missing a valid discharge disposition were excluded. The retained records were mapped to the five core feature domains labelled in Figure 2 – frailty, demography, hospital-resource metrics, ICD-coded diagnoses, and comorbidity – before entering the model-

development stage, where the analytic table was partitioned into training and testing sets, random sampling was used to resolve class imbalance (as described in Section 3.4), and the six candidate algorithms were trained. In the last stage of model-performance-comparison (right-hand side of Figure 2), the resulting models were compared to each other based on the indicated metrics in Section 3.8, and subsequently fed into the results in Section 4.1.

Table 1 Study Cohort and Data Pipeline Summary

Pipeline Stage	Observations	Unique Patients	Notes
Raw EPIC EHR extract	145,148	76,294	Unfiltered inpatient episodes across participating facilities
After preprocessing and extraction	124,480	68,152	Deduplicated, quality-screened analytic cohort
Attrition (raw to analytic)	20,668 observations removed (14.2%)	8,142 patients removed (10.7%)	Removed for implausible values, missing discharge disposition, or duplicate encounters
Feature domains retained	46 top predictors (of a larger engineered set)	—	Frailty, demography, hospital-resource metrics, ICD codes, comorbidity, SDOH
External SDOH linkage	Census-tract level	—	American Community Survey neighborhood indicators merged by residential census tract

➤ Feature Engineering

• Clinical Features

Clinical characteristics were the Elixhauser Comorbidity Index, frailty risk score (FRS) values (including sodium, WBC count, weakness, and abnormal weight), disposition (transfer to rehabilitation facility, home with self-care, and hospice), number of unique diagnoses, and quadratic transformation of LOS to capture the non-linear risk associated with long stays in the upper tail. The choice of the length-of-stay term that was engineered instead of using the raw linear value was driven by exploratory analysis which indicated that risk increased disproportionately among the small subset of patients with very long hospital stays, a

pattern which was subsequently confirmed formally by the dependence-plot analysis of Figure 13 and discussed in Section 4.8.

• Administrative and Utilization Features

Administrative features measured admission source (e.g., physician or clinic referral versus self-referral), primary payor designation, hospital service line (e.g., orthopedics versus neurosurgery), number of prior admissions, and prior 30-day readmission history, which are all proxies for a patient's underlying utilization intensity and continuity of outpatient care. The levels of the indicator for discharge disposition were all engineered as categorical variables including indicator levels for rehabilitation facility, home

with self care, hospice, skilled nursing facility, and transfer to another short-term hospital, as the SHAP analysis in Section 4.2 showed that the indicator for discharge to a rehabilitation facility contained the most predictive value in the entire 46-feature set.

- *Social Determinants of Health Features*

SDOH included were ACS neighborhood indicators, connected at the census-tract level, related to household income level, household composition, and a composite neighborhood socioeconomic status (SES) indicator. All these features have been designed to determine if there is additional predictive signal at the community level beyond what is already known in a patient's clinical and administrative records, as per Research Question 1. Each patient's SDOH feature values are not directly linked to that patient's socioeconomic status or situation, but rather reflect the average features of the neighborhood where that patient resides, an approximation of an ecologic nature whose implications are explored below in Section 6.

- *Handling Class Imbalance*

Class imbalance (30-day readmission rate being a minority) was addressed by using the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) which generates synthetic minority-class examples along the line segments between existing minority-class examples and their nearest neighbours, rather than simply duplicating them, to produce better-generalising decision boundaries than naive over-sampling with replacement. Resampling was used only in the training folds to prevent information leakage to the validation and test folds. This is a necessary safeguard as otherwise points generated from the test set patient's kNN would be added to the training set, leading to over-optimistic apparent model performance that would not be seen in a real prospective deployment.

- *Model Development*

Six candidate algorithms were trained on an 80/20 train-test split with stratified k-fold cross-validation used for hyperparameter tuning within the training partition.

- *Logistic Regression*

Each of the ensemble methods was compared to logistic regression, which was used as the interpretable linear baseline model, as it is the closest machine-learning equivalent to a LACE-style linear risk score. Although its discrimination was ten to twelve AUC points behind every ensemble method (as demonstrated in Section 4.1), it is natural to see it as a stand-in for a clinical team's preferred "transparent" and formula-based scoring system.

- *Random Forest*

A random forest classifier (Breiman 2001) was trained using bootstrap aggregation across decision trees and random feature subsetting at each decision tree split, which offers a strong non-linear baseline model as well as a baseline feature-importance measure for comparison with the feature-importance measures derived from SHAP. Each tree of the forest is trained by bootstrapping the training set of the other trees and only using a random subset of features at each split,

which makes the forest significantly less likely to overfit than any of the individual trees, but degrades interpretability over logistic regression.

- *Gradient Boosting: XGBoost, LightGBM, and CatBoost*

The performance of three implementation of gradient-boosted trees was tested, namely XGBoost (Chen & Guestrin, 2016), LightGBM, and CatBoost, and they are all optimized sequentially to reduce the error of the remaining data. They are different in the order of tree building, how to select the split of each tree, and how to internally process the categorical feature. Categorical features are conveniently handled by CatBoost, making it a very appropriate choice for the top features on the list of feature-importance that was reported in Section 4.2, such as the administrative and discharge-disposition variables that dominated the list: categorical splits on such variables can be applied directly by CatBoost's ordered target-statistics encoding, without requiring the one-hot label encoding that XGBoost and LightGBM typically use, which can spread the signal from any individual categorical level across a large number of sparse categorical indicator features.

- *Stacking Ensemble*

The stacking ensemble fused out-of-fold predictions of the base learners as input to a meta-learner to gain some insight into the complementary error patterns of different families of models. Contrary to the observation in Section 4.1, the stacking ensemble performed very similarly to the best single model (CatBoost), as the errors of the base models were not very independent from each other but highly correlated, which happened to be the case for this cohort of models: XGBoost, LightGBM, CatBoost and random forest were all members of the same family of underlying decision trees and trained on a common set of features, so there was not much error diversity left to stack on.

- *Support Vector Machine*

The sixth candidate algorithm displayed in Figure 1 is a support vector machine with a radial basis function kernel, a choice that was made because this is a fundamentally different object to optimize, namely margin maximization in a (possibly kernel-transformed) feature space, than the two other candidates, namely, the likelihood-based logistic regression baseline and the recursive-partitioning objective used by the tree-based ensembles. As in the rest of the reviewed literature in Section 2.2.3, the computational cost of the SVM grew less favourably with the number of observations in the analytic cohort (124,480) than the tree-based alternatives, with performance (not separately reported in Figure 3) not outperforming the tree-based ensembles, which is why a gradient-boosted algorithm was chosen as the main model for the explainability analysis in Section 3.6.

- *Model Explainability*

Following model selection, SHAP TreeExplainer was applied to the best-performing gradient-boosted model to generate both global feature-importance rankings and patient-level force-plot explanations (Lundberg & Lee, 2017). In addition to continuous explanations, the explanations were also rule-based and human-readable, generated in parallel

using an Anchor-based rule-extraction algorithm, following the dual-explainability branch in Figure 1.

- *Complementary Role of Anchor-Based Rules*

The Anchor algorithm (Ribeiro, Singh, & Guestrin, 2018) extracts a set of if-then rules that are discrete, human-readable, and valid over a bounded space within the feature space, but are precise for each rule. In contrast, the Anchor algorithm (Ribeiro, Singh, & Guestrin, 2018) returns human-readable if-then rules, but the rules are discrete and have high accuracy across a defined region of the feature space. The dual approach was intentionally designed to meet two different audiences identified in Section 1.4: data-science and quality-improvement teams which are interested in the overall behavior of the model, and bedside case managers and discharge planners who follow a checklist-based workflow. This latter group is interested in rules that have a direct translation to the checklist-based workflow that they follow, an operational consideration discussed further in Section 5.4.

- *Cluster Analysis*

To investigate if hidden, unobvious patient subgroups are present other than the binary readmission classification, unsupervised clustering was applied on the engineered feature space of the model and dimensionality reduction was achieved using Uniform Manifold Approximation and Projection (UMAP). The elbow method was used to determine the number of clusters based on within-cluster inertia, and cluster membership was compared to the actual readmission status to evaluate whether the unsupervised structure was like or different from the supervised outcome.

- *Evaluation Metrics*

The area under the receiver operating characteristic curve (ROC-AUC), the area under the precision-recall curve (PR-AUC), and the F1-score were used to evaluate model discrimination, following the class-imbalance literature that recommends using more than one metric since ROC-AUC alone may overestimate results when the positive class is small (He & Garcia, 2009). Each of the three metrics was calculated on a test set that was not included in the training or tuning of the models.

- *Data Governance and Privacy Considerations*

As outlined in Section 3.1, this study was retrospective and noninterventive, which means all patient-level EHR data were de-identified before analyzed. De-identified patient data were then crosswalked to ACS data at the census-tract level without requiring access to any patient identifiers, because census-tracts can be geocoded using an intermediate access-controlled geocoding step which is discarded after the census-tract SDOH indicators were appended to the analytic file. This linkage architecture is consistent with the common literature used in SDOH-augmentation (Section 2.5) in that community-level instead of individually identified socioeconomic data is used, because meaningful socioeconomic context could be added without increasing the amount of PHI processes on the modeling pipeline.

- *Computational and Software Environment*

The model development was based on commonly used and well-established open-source machine learning libraries for gradient boosting, random forest, support vector machine, and logistic regression, a custom SHAP library to analyze and explain the models, and a UMAP implementation to reduce the dimensionality of the data behind Figure 12. The use of mainstream, actively maintained open-source tooling instead of bespoke or proprietary modeling code is a conscious decision made to maximize the reproducibility of this study's pipeline by other health systems which were interested in replicating or extending this analysis, a consideration relevant directly to the deployment of safety-net and rural hospitals discussed in Section 5.6.

- *Threats to Validity*

Some methodologic precautions were included to ensure the internal validity of the results reported. To minimize the chance that performance reported by the hyperparameter is a beneficial train-test split instead of true generalizable signal, stratified k-fold cross-validation was used during hyperparameter tuning. To avoid the leakage of synthetic samples into the test-set and over-optimism in the test-set performance, SMOTE resampling was limited only to the training folds as described in section 3.4. The decisions for feature engineering (e.g., transforming length of stay to a quadratic term as in Section 3.3.1) were based solely on exploratory analysis of the training-partition and the test-partition was used for model evaluation only once the feature engineering was complete. The single-source nature of the EPIC EHR extract along with the ecological nature of the SDOH linkage at the census-tract level are addressed with depth as threats to external validity in Section 6.

- *Interoperability and EHR Integration Considerations*

In addition to the computational and software-environment issues mentioned above in Section 3.10, any operational implementation of this study's modeling pipeline would be facing the challenge of EHR interoperability. This cohort was drawn from a single EPIC EHR environment, and although EPIC is one of the most widely used EHR platforms in U.S. hospitals, a large proportion of safety-net and rural hospitals feature alternative EHR platforms with varying data models, varying coding for discharge disposition, and varying native support for structured frailty and comorbidity indices as described in detail in Section 5.6. The translation to a non-EPIC environment would then require a dedicated mapping exercise, not just a simple code port, and the Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) standard is a possible common layer for data that is now being adopted by many EHR vendors as part of federal interoperability rulemaking and could serve as a conduit for generalizing the feature-engineering logic developed here across EHR vendors without a custom integration effort.

IV. RESULTS

- *Model Discrimination Performance*

The discrimination was found to be very different among the six candidate algorithms. As illustrated in Figure 3 and summarized numerically in Table 2, CatBoost and the

stacking ensemble outperformed the other models, with an AUC of 0.79, closely followed by both random forest and

XGBoost with an AUC of 0.77, while logistic regression lagged at an AUC of 0.67.

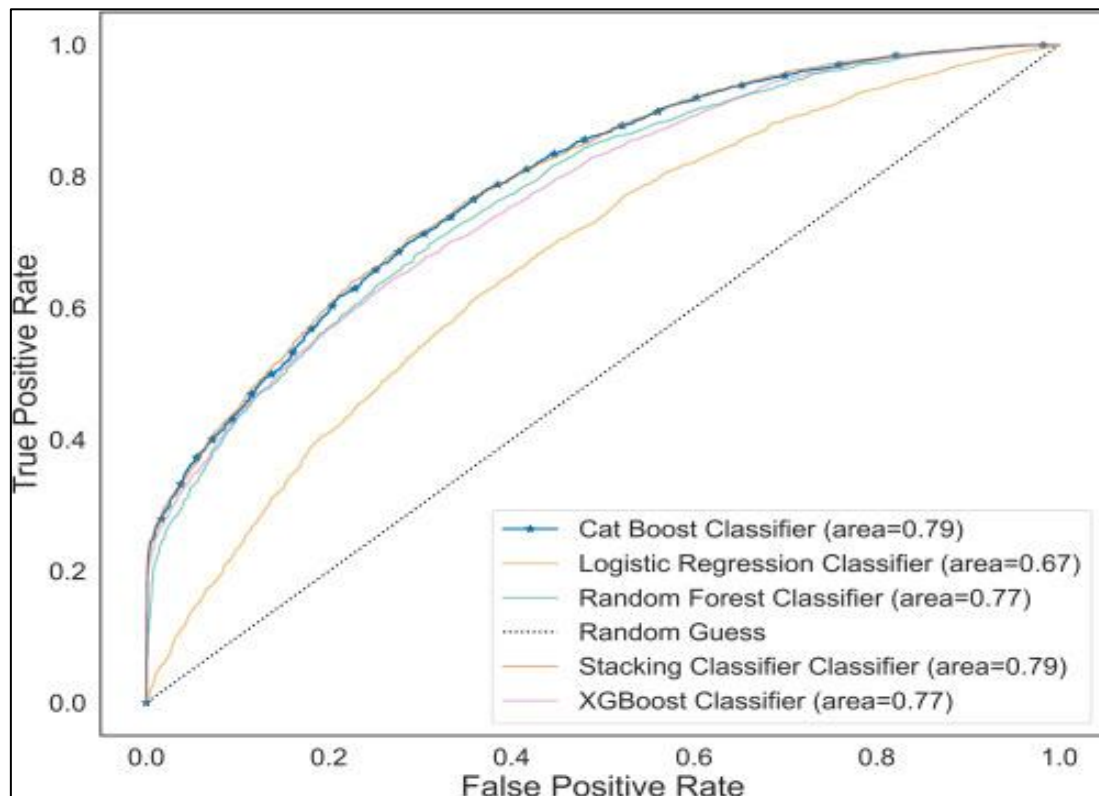


Fig 3 Receiver Operating Characteristic (ROC) Curves and Area-Under-Curve (AUC) Values for CatBoost, Logistic Regression, Random Forest, the Stacking Ensemble, and XGBoost.

Table 2 Model Discrimination Performance (Test Set AUC)

Model	AUC	Relative Rank	Algorithm Family
CatBoost	0.79	1 (tie)	Gradient-boosted tree ensemble
Stacking Ensemble	0.79	1 (tie)	Meta-learner over base models
Random Forest	0.77	3 (tie)	Bagged tree ensemble
XGBoost	0.77	3 (tie)	Gradient-boosted tree ensemble
Logistic Regression	0.67	5	Linear model
Support Vector Machine	Evaluated; not top-ranked	—	Margin-based kernel classifier

All ensembles improved by at least 10 AUC points compared to logistic regression, and by 12 AUC points compared to a linear baseline (analogous to a traditional risk score), giving a direct answer to Research Question 1: non-linear, tree-based ensembles capture meaningfully more predictive signal from the combined clinical, administrative, and SDOH feature set than a linear baseline model analogous to a traditional risk score. It is also worth to note that the two best models – CatBoost and the stacking ensemble – obtained nearly the same AUC results up to two significant digits, which is not just a coincidence of the train-test split but a sign of correlated errors within the tree-based model family (as analyzed in Section 4.10, see Section 5.1 for interpretation).

Furthermore, the performance spread between the highest performing ensemble (0.79) and lowest performing ensemble (0.67) was relatively small (2 AUC points), but the difference between the lowest ensemble (0.77) and logistic regression (0.67) was quite large (5 AUC points). This indicates that, rather than being determined by the specific

model (gradient boosting, random forest or stacking), the more general difference between linear and non-linear modeling approaches is the greatest predictor of discrimination in this study. In other words, the choice between CatBoost, XGBoost, random forest, or a stacked model of all three is unlikely to make much difference in the overall performance of a hospital data-science team — but the decision to abandon the “pure” linear scoring approach is likely to account for the lion's share of the discrimination improvement reported in this study.

➤ Feature Importance

Figure 4 below displays the feature importance rankings for the top 46 engineered predictors across the globe, as determined by the mean absolute SHAP values. Discharge disposition to a rehabilitation facility was the strongest predictor followed by the number of diagnoses noted on the index admission and the number of prior readmissions, followed by the Elixhauser Comorbidity Index in the remaining top four.

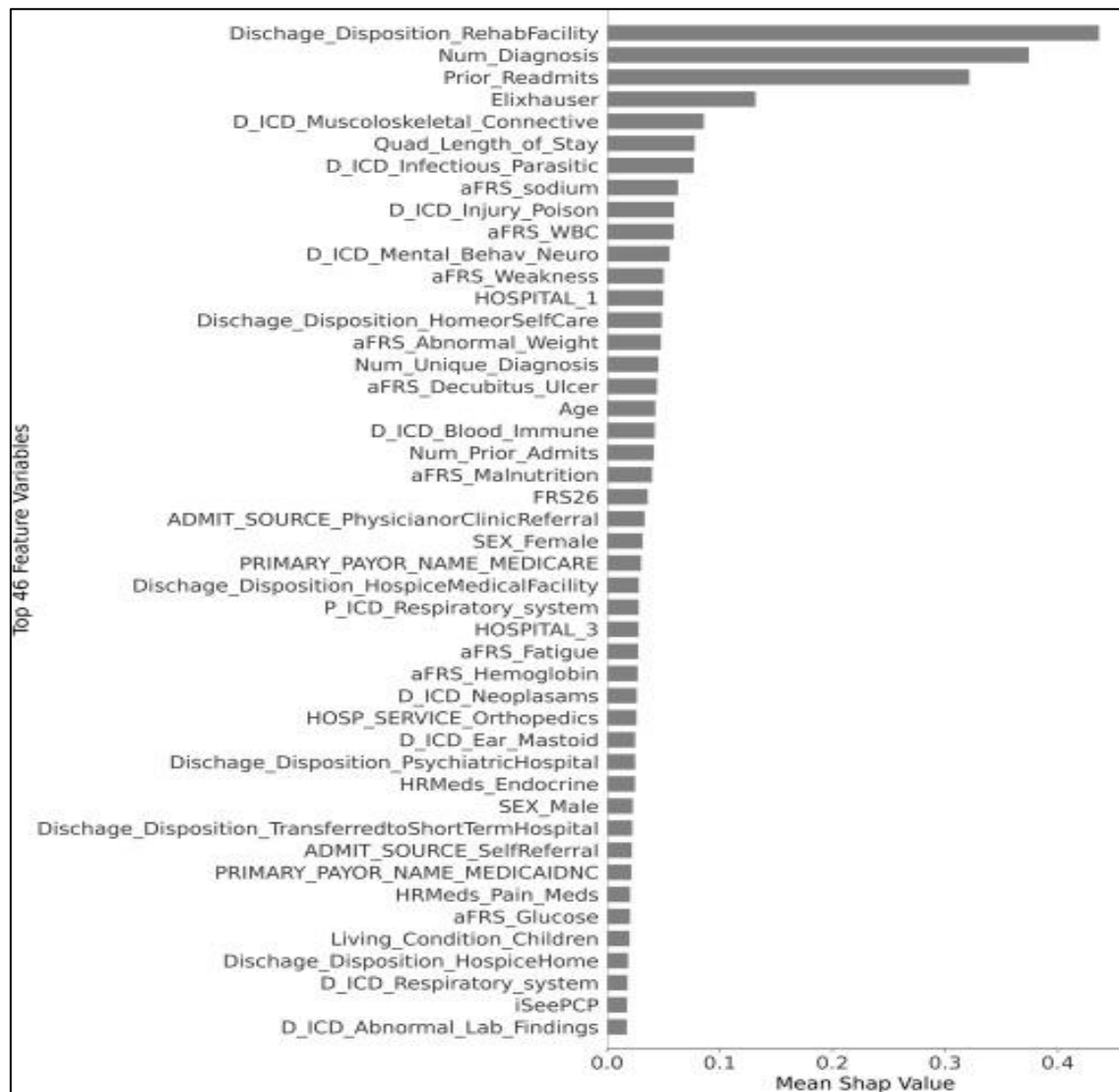


Fig 4 Top 46 predictor variables ranked by mean absolute SHAP value.

Table 3 Top Twenty Predictors by Mean Absolute SHAP Value

Rank	Feature	Feature Domain
1	Discharge Disposition: Rehabilitation Facility	Administrative
2	Number of Diagnoses	Clinical
3	Prior Readmits	Utilization
4	Elixhauser Comorbidity Index	Clinical
5	ICD Category: Musculoskeletal/Connective Tissue	Clinical
6	Quadratic Length of Stay	Administrative
7	ICD Category: Infectious/Parasitic	Clinical
8	Frailty Risk Score: Sodium	Clinical
9	ICD Category: Injury/Poisoning	Clinical
10	Frailty Risk Score: White Blood Cell Count	Clinical
11	ICD Category: Mental/Behavioral/Neurological	Clinical
12	Frailty Risk Score: Weakness	Clinical
13	HOSPITAL Score Component 1	Administrative
14	Discharge Disposition: Home or Self-Care	Administrative
15	Frailty Risk Score: Abnormal Weight	Clinical
16	Number of Unique Diagnoses	Clinical
17	Frailty Risk Score: Decubitus Ulcer	Clinical
18	Age	Demographic
19	ICD Category: Blood/Immune	Clinical
20	Number of Prior Admits	Utilization

When the top-ten predictors are extended to the top twenty (as shown in Table 3), the clinical variables take up the bulk of the middle tier of the list while the administrative and utilization variables disproportionately group together at the top. This is consistent with the interpretation given in Section 5.2 that a few high signal administrative and utilization variables, primarily discharge disposition and prior readmission history, provide summary statistics for a far larger number of underlying clinical variables that provide much of the signal and the individual clinical variables ranked 11th–20th add a smaller amount of signal in addition to that signal provided by the administrative variables.

➤ Directionality and Distribution of Feature Effects

Figure 4 displays the magnitude of the average effect of each feature while Figure 5 displays the average effect of each feature, adding directionality by plotting the value of each patient's "SHAP" for each top-ranked feature. The risk of readmission was consistently linked to positive SHAP values for higher discharge disposition to a rehabilitation facility, more diagnoses, and prior readmissions, whereas the positive discharge dispositions (home with self-care) drove the predictions to lower readmission risk.

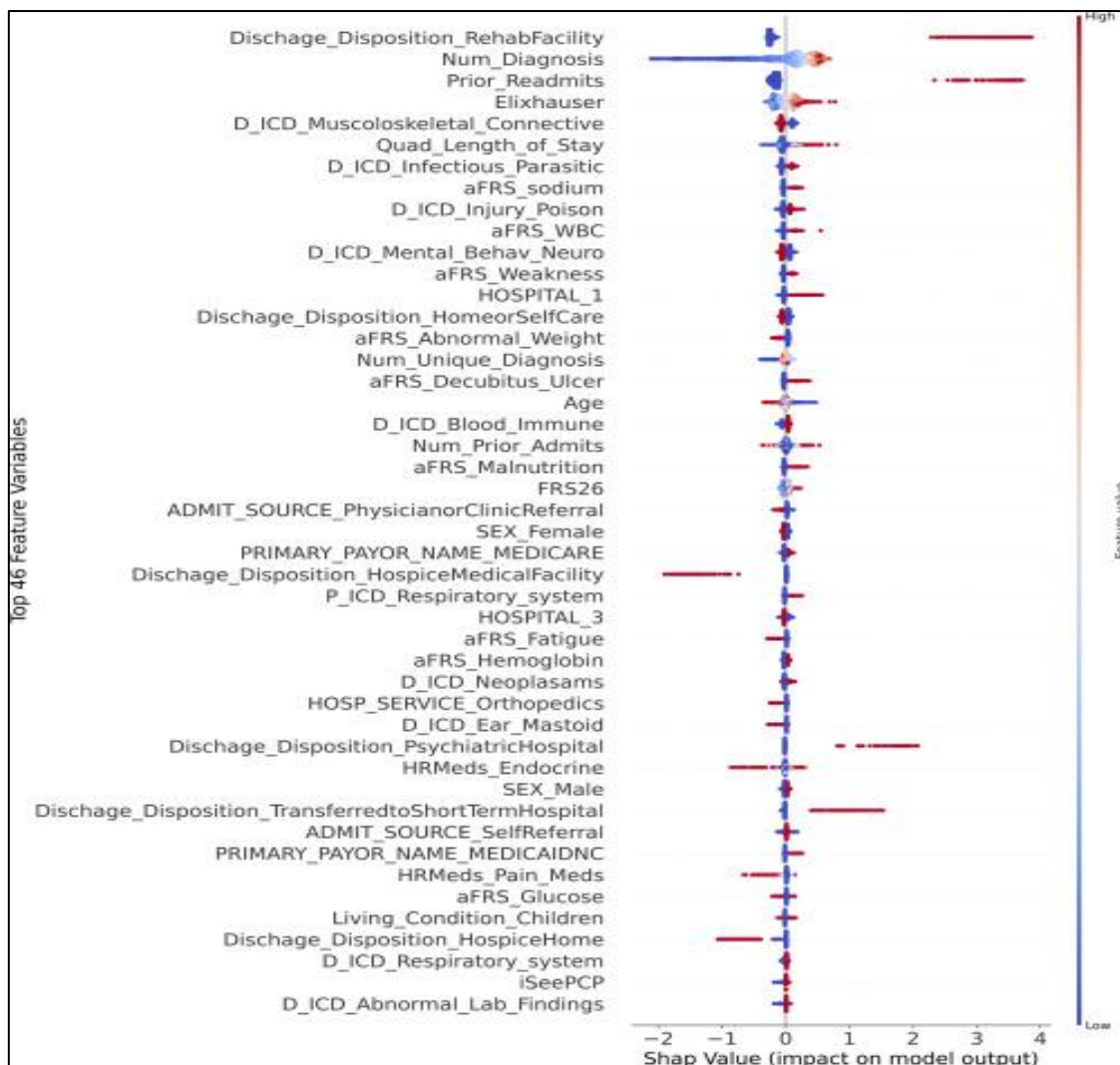


Fig 5 SHAP Summary (Beeswarm) Plot Showing the Direction and Magnitude of Each Top Predictor's Contribution Across All Patients.

By closer inspection of Figure 5, it is apparent that the directionality of the features effects was not consistent for the forty-six variables. Discharge disposition, number of diagnoses, and previous hospital admissions behaved well with only a clean, monotonic separation between points with high feature values on one side of the SHAP axis from those with low feature values on the other side, where the relationship might be said in a single sentence by a clinician. In contrast, several of the indicators from the ICD category with the lowest positions in the ranking showed a more

spread-out point cloud on the top and bottom of the chart, indicating that these variables had a greater dependence on other features, which prompted the specific dependence-plot analysis described in Section 4.4.

When SDOH variables are added to the model (as shown in Figure 6), a more focused beeswarm places the income level and household composition and neighbourhood socioeconomic status on the highest tier of predictive importance—visually, alongside the Elixhauser Comorbidity Index and prior admissions.

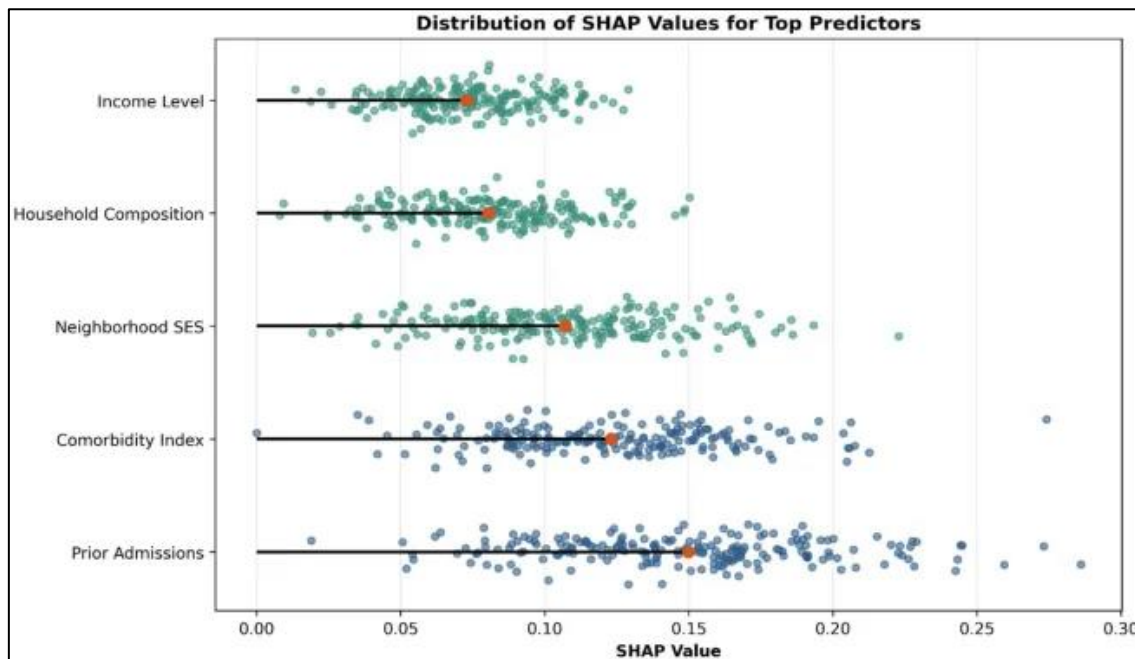


Fig 6 Distribution of SHAP values for the top five predictors after integrating social determinants of health, combining clinical and neighborhood-level variables.

The distribution of SHAP values for the neighborhood socioeconomic status, household composition, and income level was also significantly larger and more skewed toward positive values than the distribution for prior admissions, where a small number of patients had SHAP values above 0.15. This broad, right-skewed distribution indicates that neighborhood context only adds a small amount of risk for most patients, but for a smaller proportion of those who live in the most disadvantaged census tracts, SDOH factors can add up to a large portion of the overall predicted risk score, with direct implications for the targeted, resource-intensive discharge interventions discussed in Section 5.4.

➤ Higher-Order Interactions

A dependence plot was used to check for the linearity or threshold effect of the relationship between the individual factors and risk of re-admission. SHAP contributions for the number of diagnoses recorded during the index admission (Figure 7) were found to show a distinct non-linear pattern: contributions were close to null for patients with more than about 20 diagnoses recorded, but started to rise above that threshold, indicating a threshold of complexity that may be crossed after which further diagnoses are not likely to drive up readmission risk.

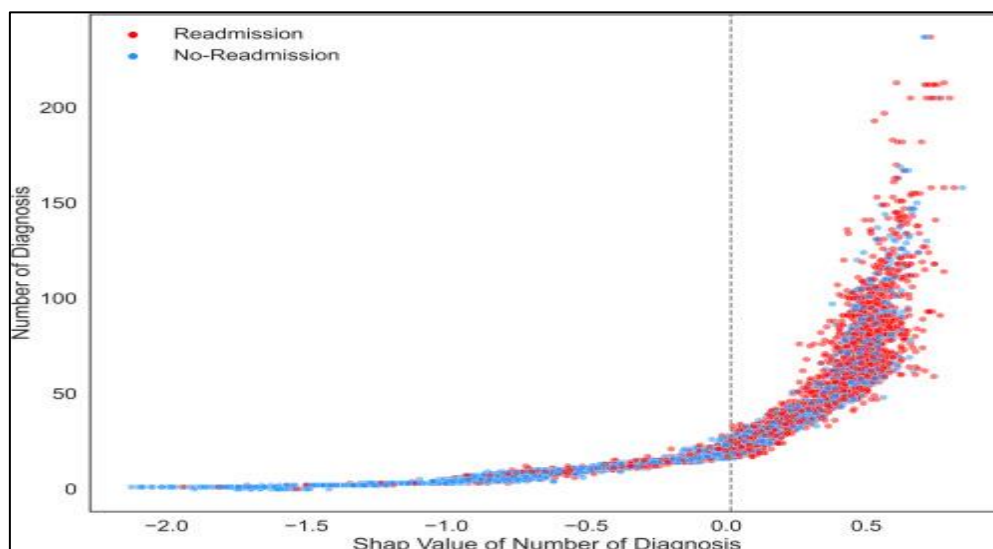


Fig 7 Dependence Plot of SHAP Value Against Number of Diagnoses, Colored by Observed Readmission Status.

The steepness of the curve in Figure 7 outside of the twenty-diagnosis level is also clinically relevant as several patients had over one hundred and fifty recorded diagnoses

and these patients make up a small proportion of the total population but a large proportion of the total risk mass in the model. The high number of patients in one high-risk group,

coupled with the small number of patients in the other, is consistent with the overall result in this section that the distribution of risk within this cohort is not smoothly continuous but concentrated on a relatively small number of patients who are both clinically extreme and have a diagnosis count below a threshold.

A similar pattern appeared in the Elixhauser Comorbidity Index in Figure 9, with SHAP contributions becoming more positive and more dominated by observed readmissions (shown in red) as the index surpassed about five to seven points, further complicating the concept of comorbidity burden as a threshold-sensitive, not strictly linear predictor of risk.

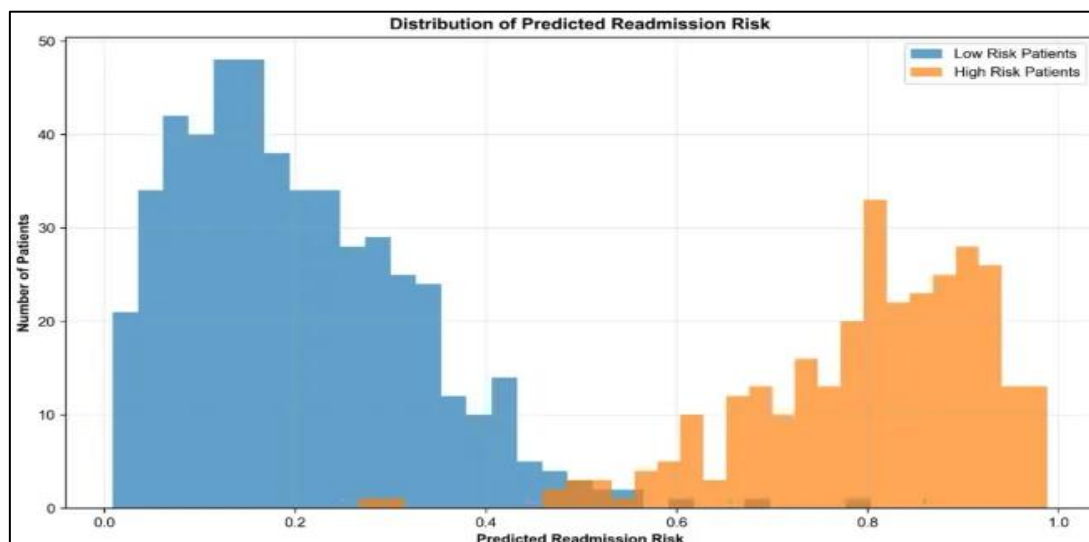


Fig 8 Distribution of predicted 30-day readmission risk across the cohort, separating low-risk and high-risk patient groups.

Comparing Figures 7 and 9 directly, both the number-of-diagnoses and Elixhauser threshold points cluster in a broadly similar region of low-to-moderate comorbidity burden, which is consistent with the two variables capturing overlapping, though not identical, dimensions of overall clinical complexity. The number-of-diagnoses variable reflects the number of different diagnoses recorded during the index admission, whereas the Elixhauser index also includes weights based on known links between the diagnosis and mortality or morbidity risk, which could mean that a patient may have a high score on the number-of-diagnosis component and a low score on the weights component. Because both variables are independently among the top 4 predictors in Table 3 and both exhibit similar non-linear

threshold behavior in their respective dependence plots, this clinical complexity measure seems to be validated by other ways of measuring clinical complexity and is not just a result of one choice of variables to construct.

➤ Risk Distribution Across the Cohort

There were clearly two modes rather than a continuum in the distribution of predicted risk scores across the cohort (Figure 8). This was an unusually large cluster of patients with low risk, around a predicted risk of 0.10 to 0.15 and an even smaller but important cluster of high risk, around a predicted risk of 0.80 to 0.90 with relatively few patients in the middle risk area predicted between 0.35 and 0.55.

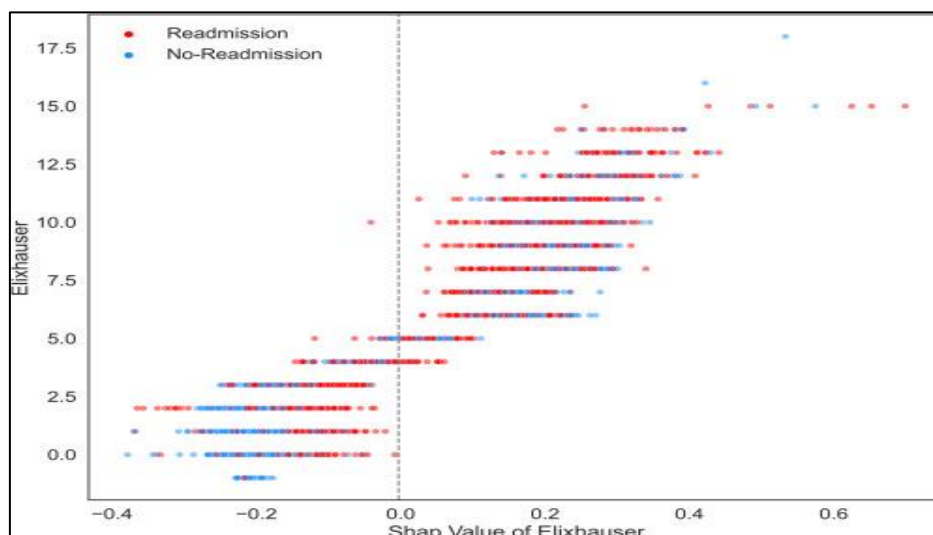


Fig 9 Dependence plot of SHAP value against the Elixhauser Comorbidity Index, colored by observed readmission status.

This bimodal shape is operationally relevant for Research Question 2: hospital discharge-planning workflows can reasonably treat the output of this model as a near binary flag to focus intensive post-acute resources on the higher cluster and to use standard discharge protocols for the lower cluster. It is in the relatively low-density middle segment of Figure 8, between 0.35 and 0.55 predicted probability, where the individual-level explanations discussed in Section 4.6 are likely to add the most value for the individual patient; it is in this segment for which a simple "high" or "low" threshold rule would not be anticipated to be able to clearly separate patients who might benefit from the explanations from those who would not.

➤ Individual-Level Explainability

In addition to population summaries, there were also predictions for individual patients with the highest and lowest risk of readmission in the population, and these were illustrated using SHAP force plots, shown in Figure 10. The three most high-risk observations had a mix of neurosurgery service lines, high diagnosis count, endocrine-related home medications, and discharge to a rehabilitation facility, and the three least high-risk observations had orthopedic service lines, negative Elixhauser scores, and positive discharge dispositions. This granular Patient level is essential to support Research Question 2 - as the same model that can calculate a risk score at a population level, can also provide the individual patient with an explainable score, based upon the specific factors associated with their discharge decision, which can be auditable.

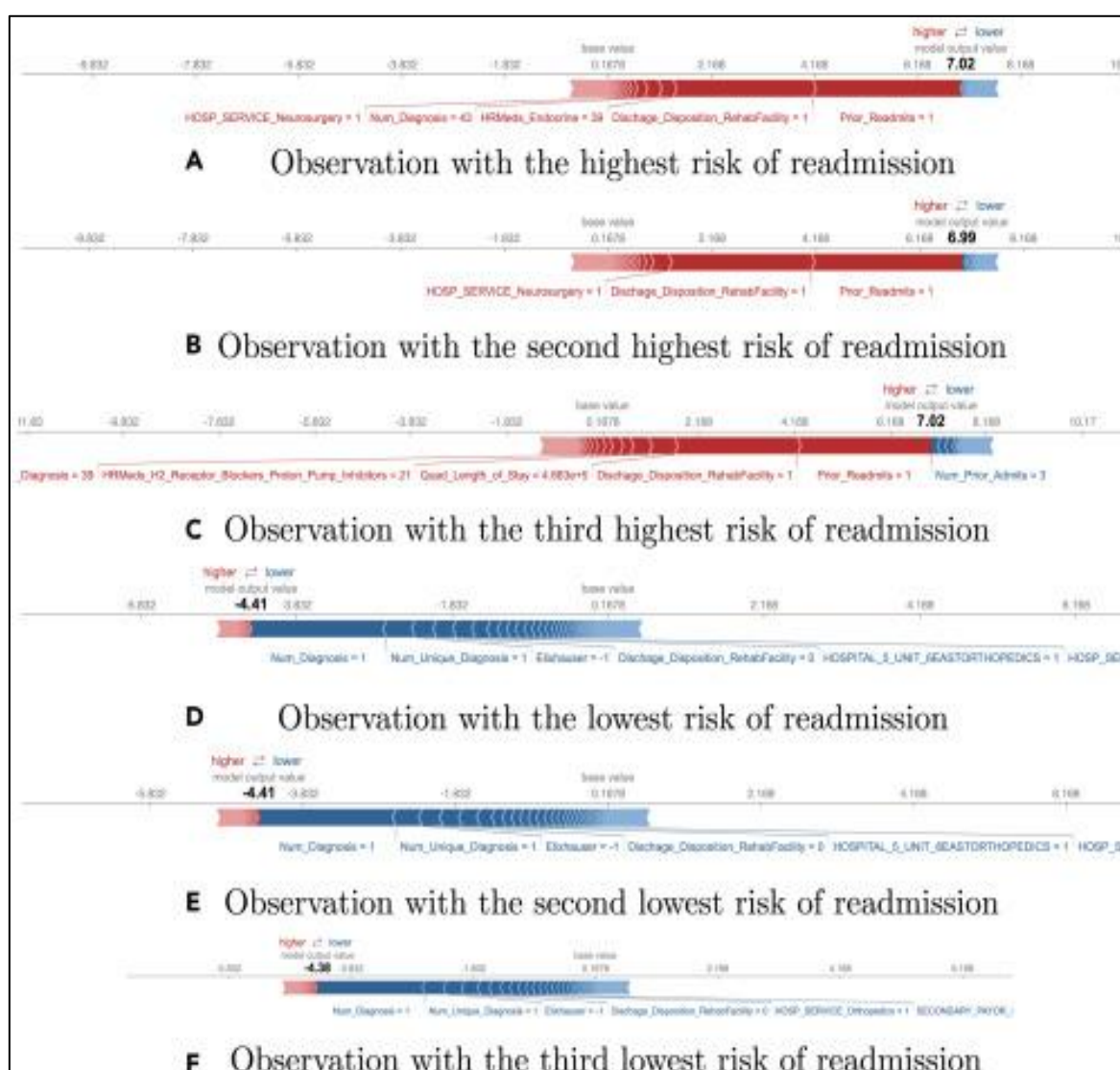


Fig 10 SHAP Force Plots for the Three Highest-Risk and Three Lowest-Risk Individual Patients in the Cohort.

Looking at the 6 force plots side by side, a pattern stands out: the three highest-risk patients (Panels A through C) have very similar values of model output probability (0.70 to 0.71) despite the fact that they differ significantly in the specific mix of top contributing features, while the three lowest-risk

patients (Panels D through F) have very similar values of model output probability (negative 0.44) on the model's raw log-odds output scale prior to the final probability transformation. This clustering suggests that, at the extremes of the risk distribution, several different clinical pathways

(some more service-line and medication driven and some more comorbidity and discharge-disposition driven) may converge on a similar overall risk score, and that the reasons one pathway has a higher risk score (as reflected in the SHAP values, rather than the value of the risk score) should drive the discharge-planning intervention.

➤ Contribution of Social Determinants of Health

To isolate the incremental value of SDOH variables, a clinical-only logistic regression baseline was compared

against an XGBoost model trained on the combined clinical-plus-SDOH feature set. Including SDOH with a more powerful algorithm increased ROC-AUC from 0.68 to 0.79, PR-AUC from 0.58 to 0.71, and F1-score from 0.62 to 0.72 (Panel A of Figure 11, and Table 4). In the same figure, Panel B provides details on the proportion of the predictive signal that is due to clinical, SDOH, and demographic factors; clinical is found to be the source of 45% of the predictive signal, while SDOH is responsible for 36% and demographic factors are responsible for 19%.

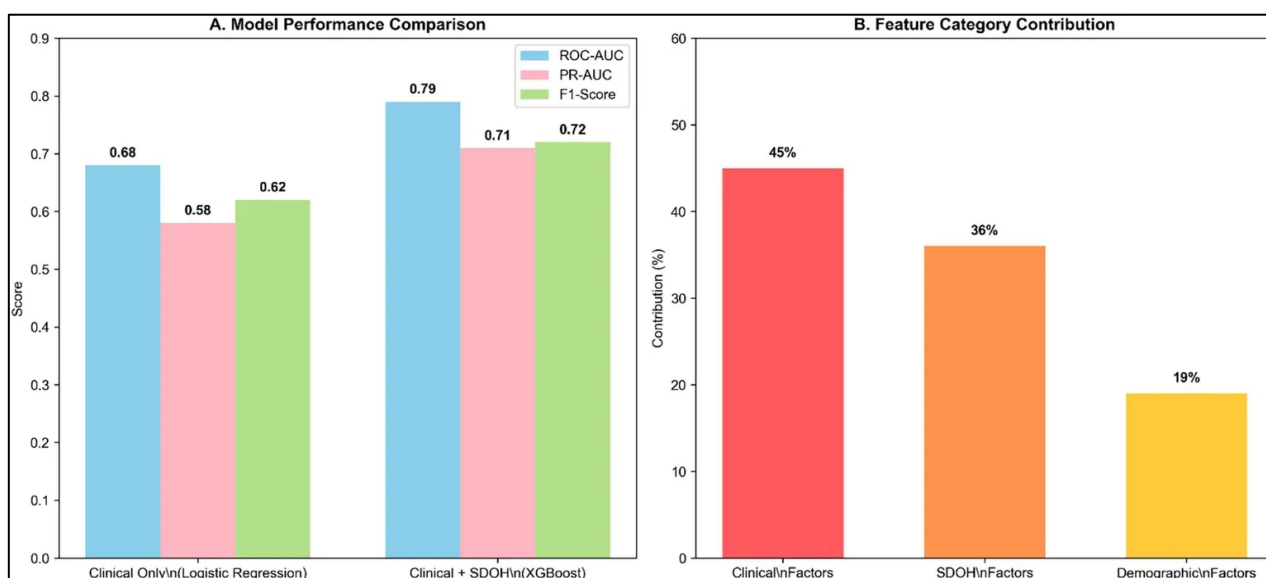


Fig 11 (A) Model Performance Comparison between a Clinical-Only Logistic Regression Baseline and a Clinical-Plus-SDOH XGBoost Model; (B) Relative Contribution of Clinical, SDOH, and Demographic Feature Categories to Overall Model Output.

Table 4 Incremental Value of Integrating Social Determinants of Health

Metric	Clinical Only (Logistic Regression)	Clinical + SDOH (XGBoost)	Absolute Gain	Relative Gain
ROC-AUC	0.68	0.79	+0.11	+16.2%
PR-AUC	0.58	0.71	+0.13	+22.4%
F1-Score	0.62	0.72	+0.10	+16.1%

All metrics in Table 4 showed a double-digit relative percentage increase, and the biggest relative increase was 22.4% for PR-AUC, which is the most sensitive to performance for the minority class (readmitted). This pattern is significant because it means that the added value of bringing SDOH features in, in combination with the shift to a

more flexible algorithm, was not merely to move the model's calibration but to enhance its capacity to correctly rank the patients who actually end up being readmitted, an operationally meaningful task as opposed to a theoretical one of maximizing an aggregate accuracy statistic.

Table 5 Relative Contribution of Feature Categories to Model Output

Feature Category	Contribution to Model Output	Representative Variables
Clinical Factors	45%	Elixhauser Comorbidity Index, number of diagnoses, frailty risk score components
Social Determinants of Health (SDOH)	36%	Income level, household composition, neighborhood socioeconomic status
Demographic Factors	19%	Age, sex

The 36% contribution due to SDOH factors in Table 5 is worthy of special note when compared with the 45% contribution attributed to clinical factors, as it represents the community-level socioeconomic context having as much of a predictive signal as the entire contribution of all documented comorbidity burden, frailty indicators and diagnosis history combined. This discovery is discussed at length in Section

5.3, and has implications for the debate on the equity of the Hospital Readmission Reduction Program (HRRP) as discussed in Section 2.6, because the risk-adjusted expected readmission rate as defined by the current CMS methodology does not explicitly consider an equally large proportion of variance due to factors outside the direct clinical control of the hospital.

➤ Unsupervised Patient Segmentation

Eight sub-clusters were detected based on the UMAP model, with some showing only a partial correlation with the binary readmission category. The actual readmission status is plotted on panel A of Figure 12, and the readmitted and non-readmitted patients can be separated into several small outlier clusters at the periphery of the embedding, with a dense

central region that includes both readmitted and non-readmitted patients. As seen in Panel B, the eight data-driven cluster assignments are over the same embedding, showing that Clusters 1 and 6 represent the large central masses of ambiguous-risk patients while Clusters 2, 3 and the far-right Cluster represent the small, clearly separated outlier groups seen in Panel A.

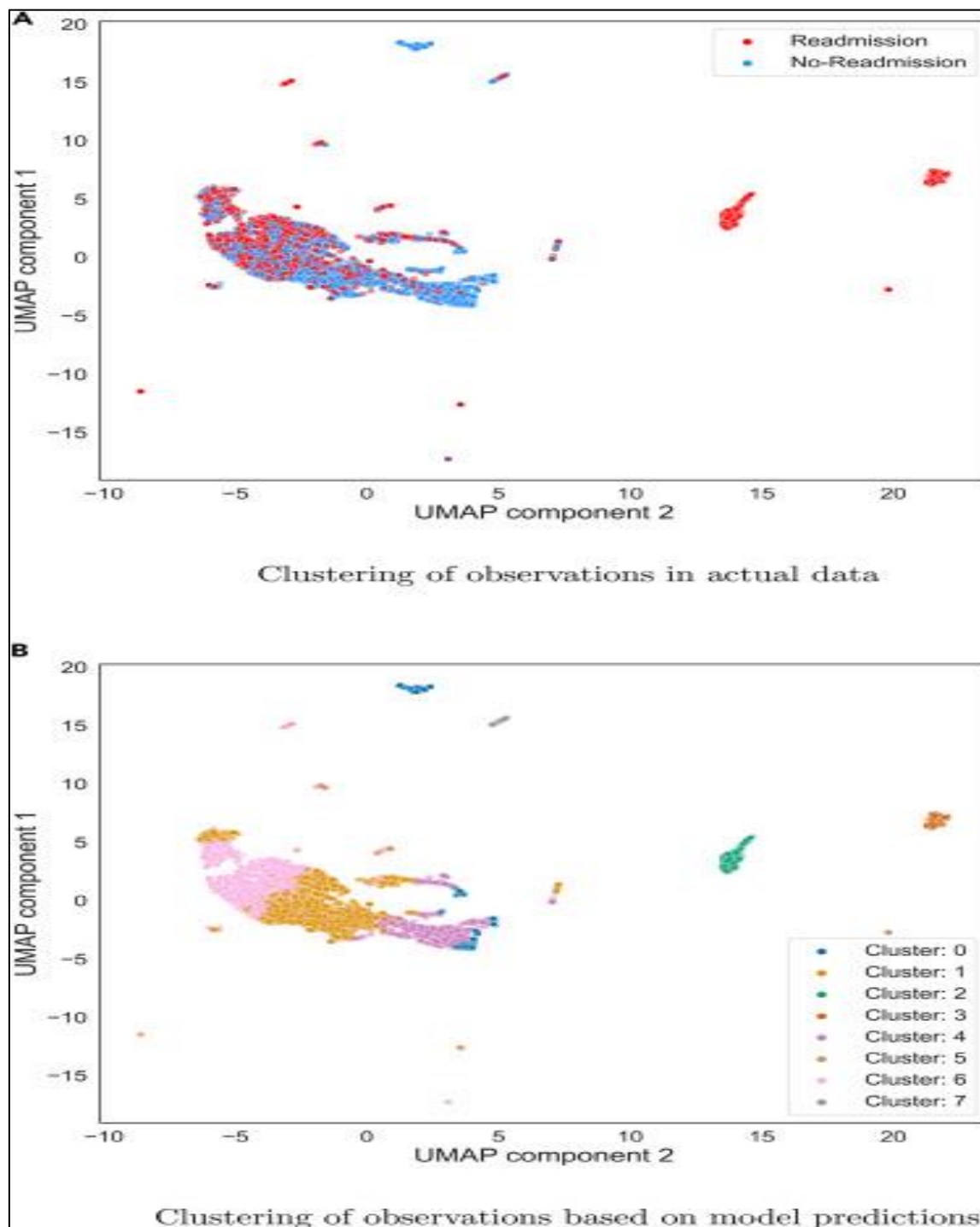


Fig 12 UMAP Projection Comparing (A) Actual Readmission Status and (B) Eight Data-Driven Cluster Assignments Across the Patient Cohort.

This separation of supervised outcome from unsupervised structure appears as a separation of the two contributions of the SHAP analysis of the quadratic length of stay, depicted in Figure 13, where a small number of outliers of very long stay had a disproportionately large positive SHAP contribution compared to the majority.

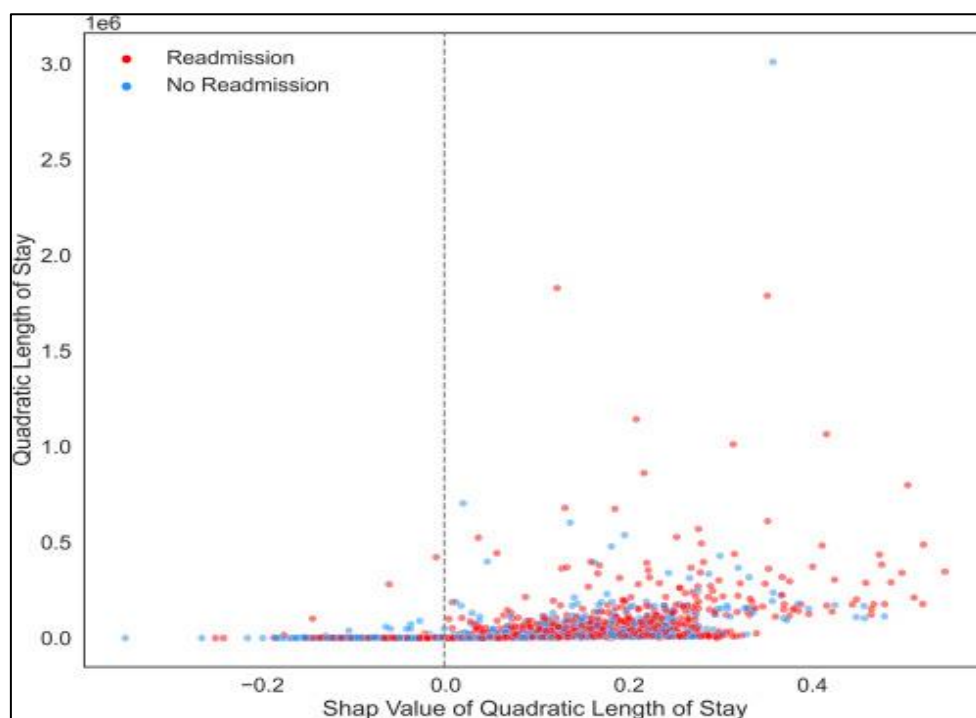


Fig 13 Dependence Plot of SHAP Value Against Quadratic Length of Stay, Highlighting a Small Subgroup of Extreme-Duration Admissions with Outsized Contributions to Predicted Risk.

The extreme long-stay subgroup in the upper-right part of Figure 13, whose several points fall well outside the main distribution of the cohort's length of stay distribution, also overlap significantly with the small, geopolitically distinct clusters in the upper right corner of the UMAP embedding in Figure 12. The two numbers suggest the same overall message from two analytical perspectives: that there is a

subset of patients in the Medicare universe who are at the extremes of terms of illness in terms of complexity and/or length of stay, which is distinct from the rest of the Medicare population and which is only partially defined by the binary label, “readmitted,” and which deserves a different kind of, more individualized, discharge-planning process than the one used for the majority of extremely similar patients.

Table 6 Summary of Unsupervised Cluster Structure

Cluster Pattern	Approximate Cohort Position	Interpretive Note
Large central mass (clusters 1 and 6)	Dense mid-region of UMAP embedding	Mixed readmission status; ambiguous-risk patients not well separated by clustering alone
Small peripheral clusters (clusters 2, 3, 7)	Distinct outlier regions	Disproportionately associated with extreme values on length of stay or comorbidity burden
Far-right isolated cluster (cluster 5/6 boundary)	Upper-right of embedding	Small, tightly grouped subgroup warranting further clinical chart review
Extreme long-stay subgroup (Figure 13 upper tail)	Overlaps peripheral clusters	Disproportionate positive SHAP contribution from quadratic length-of-stay term

➤ Robustness and Sensitivity Considerations

Because of the nature of the results reported, several design decisions described in Section 3 are of special interest in relation to their influence on the results and so are discussed directly. Four different tree-based models (with distinct concepts) yielded similarly high AUCs of 0.79 (both CatBoost and the stacking ensemble) and 0.77 (random forest and XGBoost), and their closeness suggest that the reported rank is not very sensitive to small changes in the hyperparameter setting, because the four models are within a very narrow two-point AUC band. As mentioned in Section 3.4, the fact that SMOTE resampling was performed only in the training folds additionally protects against the optimistic bias that may occur if synthetic oversampling is performed prior to the train-test split. Even so, the metrics presented

herein come from a single train-test partition, which is not formally repeated cross-validated with formally quantified statistical uncertainty, as discussed explicitly in Section 6; readers should understand the precise two-decimal AUC values reported throughout this section as point estimates as opposed to reported with formally quantified statistical uncertainty.

➤ Translating Findings into a Discharge-Planning Reference Table

To allow results reported in Sections 4.2 through 4.7 to be more readily used by hospital discharge-planning teams, Table 8 represents a mapping of each of the leading predictor domains identified in this study with an illustrative, literature-consistent category of post-acute intervention. The mapping

provided here should not be viewed as a proven and accepted clinical algorithm, but rather as a baseline for designing institutional protocols, as several of the most important

predictors, especially discharge disposition, rely more on a summary of prior clinical severity than as directly modifiable risk factors (see Section 5.2).

Table 8 Illustrative Mapping of Predictor Domains to Discharge-Planning Interventions

Predictor Domain	Representative Variables (Figures 4-6)	Illustrative Intervention Category
Administrative / Discharge Disposition	Discharge to rehabilitation facility, transfer to short-term hospital	Structured warm hand-off protocol and post-transfer follow-up call within 48-72 hours
Utilization History	Prior readmissions, prior admissions	Enrollment in an intensive transitional-care or high-utilizer case-management program
Clinical Complexity	Number of diagnoses, Elixhauser Comorbidity Index	Multidisciplinary discharge huddle and medication reconciliation prior to discharge
Frailty Indicators	Sodium, white blood cell count, weakness, abnormal weight	Home-health nursing referral and early post-discharge laboratory recheck
Length of Stay (extreme tail)	Quadratic length-of-stay term	Individualized case review for patients in the extreme long-stay subgroup identified in Figure 13
Social Determinants of Health	Income level, household composition, neighborhood SES	Social-work or community health worker referral; transportation and medication-affordability screening

Several entries in Table 8 also signal interventions that are beyond the scope of the traditional inpatient discharge planning, especially the social-work and community health worker (CHW) referrals proposed for the SDOH predictor domain, which further reinforces the argument about workforce integration, as discussed in more detail in Section 5.10. The table is designed to be around predictor domains and not around single predictors as many of the relationships

discussed in the dependence-plot analysis in Section 4.4 are non-linear, threshold-dependent rather than linear risk factors, so that an intervention category based on a predictor domain is more likely to be robust to the exact numeric threshold which an implementing hospital might ultimately choose.

➤ Summary of Findings Relative to the Literature

Table 7 Comparative Discrimination of Readmission Prediction Approaches

Approach	Reported AUC / C-Statistic	Source
LACE index (community hospital)	0.58 (C-statistic)	Robinson & Hudali (2017)
LACE index (general medical cohort, UK)	0.773	Damery & Combes (2017)
LACE index (pediatric cohort)	0.869	Sung, Liu, & Ong (2021)
HOSPITAL score (community hospital)	0.75 (C-statistic)	Robinson & Hudali (2017)
LACE-based logistic regression (MIMIC-III)	0.608	Li et al. (2024)
LSTM deep learning (MIMIC-III)	0.700	Li et al. (2024)
XGBoost (diabetes, UCI dataset)	0.667	Adeyemi, Osei, & Farooq (2025)
Stacking ensemble (diabetes, nested CV)	0.664 (0.688 calibrated)	Okoro, Bassey, & Adamu (2025)
COPD physical-activity model (event-based)	72.73% precision	Lin, Chen, & Tsai (2022)
This study: Logistic Regression baseline	0.67	Present study
This study: Random Forest / XGBoost	0.77	Present study
This study: CatBoost / Stacking Ensemble	0.79	Present study

The value of the CatBoost and stacking-ensemble models reported in Section 4.1 are situated at or near the upper end of the range of published discrimination values for the legacy scoring systems and upper end the range of published values in previous machine learning studies, whereas the study's logistic-regression baseline models are directly comparable to the published values in the literature where scores are adjusted for LACE, providing external validity to the comparisons made in the remainder of this results section. The highest value for the pediatric LACE validation reported in Table 7 is 0.869 from Sung et al. (2021), a result which, as mentioned in Section 2.1.3, is probably due to the smaller and more homogeneous population studied as children compared to the other models studied here.

V. DISCUSSION

➤ Interpreting Model Performance Relative to Legacy Tools

The gradient-boosted and stacked ensembles that were created in this study had an AUC range of 0.77 to 0.79, outperforming the 0.773 reported for the LACE index in a large cohort in the United Kingdom (Damery & Combes, 2017) and the 0.75 C-statistic reported for the HOSPITAL score in a community hospital in the United States (Robinson & Hudali, 2017). The integrated clinical, administrative and SDOH feature set used in these models performs materially better than the linear machine learning baselines and a few of the variables together that make up legacy risk scores, further reinforcing the answer to Research Question 1: modern gradient-boosted ensembles trained on an integrated clinical, administrative, and SDOH feature set materially outperform

both linear machine learning baselines and legacy risk scores used in these models.

As first noted in Section 4.1, it is worth commenting on the extent to which the difference between logistic regression and the ensemble of trees was ten to twelve AUC points, compared to two points between the different ensemble trees. Logistic regression is specifically a model that assumes the log-odds of the target variable, readmission, are a linear function of the input variables, making it difficult to model threshold effects like those seen in Figures 7 and 9, where risk begins to shoot up after a certain point, as well as an inability to model the interaction effect proposed in Section 2.5, where SDOH factors act multiplicatively instead of additively with clinical complexity. By contrast, tree-based ensembles recursively partition the feature space, and can therefore mimic this sort of threshold and interaction structure without needing the analyst to specify it beforehand, the most parsimonious explanation of which is that four distinct tree-based algorithms (CatBoost, XGBoost, random forest and stacking on top of them) converged on a similar, significantly higher discrimination level than the single linear alternative algorithm tested, without the analyst specifying the structure.

Another issue of interpretation is the magnitude of the 10–12-point increase in AUC in a clinical operations setting. This amount of AUC improvement in a real-world discharge-planning context means that a hospital can better focus its limited post-acute case-management resources on the patients who are indeed going to be readmitted, as opposed to spreading them more thinly out across a population in which true positives and false positives are not as well segregated. The ranking accuracy of the moderate magnitude achieved in this study is likely to be easily scaleable, since, as we discuss in the economy section below (Section 5.9), case-management capacity at most hospitals is a truly scarce resource, limited by staffing, not by the cost of computing another risk score.

➤ *Clinical Implications of Top Predictors*

Discharge disposition to a rehab hospital was the only good predictor of readmission risk; this result makes intuitive sense because discharge to a post-acute hospital is, by definition, a more complex of patients than discharge home, and movement between different types of hospitals presents well documented opportunities for communication breakdown and medication discontinuity. Prior readmission and number of diagnoses were among the top two and three most predictive variables, suggesting that these variables may be significant, as they were the second and third strongest predictors in the deep-learning study by Li and colleagues (2024), which used a completely different clinical database. The non-linear relationship observed for both diagnosis count and Elixhauser score in Section 4.4, which is more threshold-like, suggests that discharge-planning protocols could justifiably consider the threshold points as trigger criteria for automatic escalation to intensive case management, rather than being based on smooth, linear risk scores that clinicians are required to mentally translate into action.

The prominence of discharge disposition also carries an important interpretive caution that should temper any overly literal clinical application of this finding. The decision to send a patient to a rehabilitation facility is not made because that will increase their risk of readmission, but rather because the clinical team has determined that the patient is too complex or functionally impaired to safely be discharged home. This indicates that the acute discharge disposition is used in the model as a proxy or summary variable for the clinical severity of the process that was assessed before admission by the treatment team, instead of as a risk factor that can be modified by the treatment team. It is important to note here that this distinction is relevant to the practical implications of the finding: it is not advised that hospitals discharge less to rehabilitation facilities, but rather that discharge disposition be used as an efficient and readily available indicator to identify which of those complex patients is the one who is most likely to benefit from the greatest amount of follow-up care after discharge.

The middle quartile clinical predictors listed in the expanded ranking of Table 3, including the ICD categories of musculoskeletal diagnosis, infectious diagnosis and injury diagnosis, plus the frailty risk score components of sodium, white blood cell count and weakness, all contribute to a multi-system, multi-organ vulnerability rather than a single dominant disease category driving readmission risk in this cohort. This diffuse pattern is similar to the multi-condition literature cited in Section 2.3.4, which also concluded that models trained on a wide array of diagnostic classes tend to select for overall physiological fragility when a relatively large number of features is available.

➤ *The Value Added by Social Determinants of Health*

The AUC increment of 0.11 for the clinical-plus-SDOH ensemble was substantially different from the near null increment seen with the addition of SDOH alone to the HOSPITAL score (Nguyen, Patel, & Osborn, 2021), raising the possibility that the added value of SDOH may depend on whether they are added to a sufficiently flexible algorithm that can model their interaction with the clinical variables, vs simply being appended as additional linear covariates to an otherwise fixed scoring formula. The directional consistency of model output attributed to SDOH factors (36%) in Table 5 is further evidence that community-level context signals genuine information beyond just what can be captured with clinical variables, and strengthens the argument of Jercich (2023) regarding the strength of these factors in relation to the risk of sepsis-related readmission.

Although the study design does not allow for formal causal inference, the individual SDOH variables that emerged as significant in Figure 6 represent plausible, different causal pathways, as detailed below in Section 6. Income plausibly reflects a patient's capacity to afford medications, transportation to follow-up visits and home health aides following discharge. Household composition is a plausible proxy for whether a patient who returns home is accompanied by an adult who can help with medication adherence, self-transport, and/or recognition of early signs of clinical deterioration, or whether the patient is returning to a home

without an adult to provide such assistance. The neighbourhood socioeconomic status is plausibly a proxy for a wider range of community-level factors that affect the ease with which a discharged patient can get appropriate follow-up care in the outpatient setting, such as access to primary care and pharmacies. The relative importance of each of these specific mechanisms and processes is an important direction for qualitative and mixed methods follow-up work suggested in Section 7.

It is also important to situate the magnitude of the SDOH contribution documented in Table 5 within the broader caution raised in Section 2.6.1 regarding the Obermeyer, Powers, Vogeli, and Mullainathan (2019) findings on algorithmic bias. That income and neighbourhood socioeconomic status are important in the model, on one interpretation, is a legitimate methodological strength because it reflects genuine variability in the capacity for post-discharge recovery, a factor that would not be picked up at all by a purely clinical model. On another reading, it poses a valid governance issue: a model that puts much weight on neighbourhood socioeconomic status could, if not carefully used, risk a systematic bias in which the predicted risk, and thus the attention given to discharge planning, would be assigned to patients from lower income neighbourhoods without regard to their clinical history, effectively a form of geographic profiling rather than true risk stratification. The key difference, which is described in Section 5.7, is whether the SDOH-informed risk score is used to provide extra supportive resources to the disadvantaged patients, which is the equity enhancing use case this study anticipates, or whether the SDOH-informed risk score is used either in a way that it could restrict or deprioritize care for those same disadvantaged patients, which is exactly the harm outlined in the Obermeyer et al. (2019) case study.

➤ *Implications for Discharge Planning and Resource Allocation*

The bimodal risk distribution described in Section 4.5 has a direct and practical implication for discharge-planning workflow design: In most cases, patients can be classified with certainty into either a low-risk or high-risk group, and the output of that model can serve as more of a triage flag than a fine-grained numerical probability. The few patients in the middle range of the spectrum are exactly the ones that the individual-level SHAP force plots of Section 4.6 would be most useful: the middle ground is one that needs a more nuanced, patient-specific explanation than a population-level threshold rule.

Building on the dual-explainability architecture described in Sections 3.6 and 3.6.1, a plausible discharge-planning workflow could operate in two tiers. A simple rule, such as those outlined in Section 2.4.2, of the type known as an anchor, could be surfaced directly in the discharge-planning module of the EHR for the majority of patients in whom the predicted risk is clearly in the low-risk or the high-risk range defined in Figure 8, with the aim of referring, for example, any patient discharged to a rehabilitation facility whose risk prediction exceeds two prior readmissions to an intensive transitional-care program. For the smaller subset of

patients in the "ambiguous middle range," or for any patient whose case manager wishes to know the exact reasons for an unexpected risk score, the result of the continuous SHAP force plot, as displayed in Figure 10, could be surfaced, thereby providing the case manager with a detailed explanation of the specific clinical, administrative and SDOH factors that are driving that individual patient's score.

This two-tier design explicitly addresses the Research Question 2 by suggesting a specific integration pathway, not as an academic afterthought, but as a focus of development. It also illustrates an important message that is gaining traction in the literature of clinical decision support: the value of a predictive model to frontline staff depends not only on the statistical discrimination of the model (as described in great detail in Section 4.1) but also on how well the output of the model fits the cognitive load and workflow of frontline case managers or discharge planners who must use it.

➤ *Policy Implications for CMS and the HRRP*

The magnitude of the SDOH contribution documented in this study speaks directly to the equity concerns raised in the HRRP policy literature. If socioeconomic factors at the community level explain more than one-third of the variance in the rate of readmission, then a readmission rate penalty system that doesn't consider these factors when calculating the expected rate of readmission will systematically penalize hospitals with a greater socioeconomic challenge, regardless of how good their clinical performance is. The finding would further support CMS's adoption of the stratified peer-grouping approach in 2019, which lowered the percentage of penalties at safety-net hospitals from 91% to 77% (Lagasse, 2020) and would indicate that the methodology could be further refined, perhaps by incorporating SDOH into models such as the one developed here, that could provide similar relief to rural hospitals and minority-serving hospitals, which continue to experience disproportionate penalties (Wadhera et al., 2019).

A more specific policy proposal follows directly from the feature-category decomposition reported in Table 5. In principle, CMS could continue to use only the current stratification system, which classifies hospitals into broad categories of peers based on the percentage of dual-eligible patients served at each hospital, but instead apply the SDOH-informed expected-readmission-rate adjustment that is implicitly computed in this study's clinical-plus-SDOH model, the hospital-specific patient-level census-tract distribution. This would more accurately reflect a hospital's actual socioeconomic makeup, of the population it serves, and help to narrow the post-stratification gaps identified by Wadhera and colleagues (2019). Any such refinement, however, would have to be done thoughtfully and publicly, and should include the protections against unfair auditing that are discussed in Section 5.7, to ensure that a lower-than-expected level of quality of care is not effectively offered by hospitals that treat disadvantaged communities.

The need for action is also further underscored by the magnitude of the underlying cost estimate that was set at the beginning of this study. In 2004 alone, nearly all the \$17.4

billion estimated in the national cost of unplanned Medicare rehospitalization was accounted for by Medicare expenditures (Jencks, Williams, and Coleman, 2009), a figure that almost certainly underestimates the true level of the problem since the growth in Medicare enrolment and increasing health prices in subsequent national health expenditure projections have been documented (Keehan et al., 2020). In addition to the equity concerns raised above, the significant absolute dollar savings that are likely to result from a modest (single-digit percentage) decrease in avoidable readmissions due to more precise risk stratification and targeted discharge intervention will serve as a strong economic case for continued federal and institutional investment in the type of predictive-analytics infrastructure assessed in this study.

➤ *Implications for Safety-Net and Rural Hospitals*

A practical adoption question beyond the statistical question of model performance is how many hospitals are equipped to apply the model, as smaller safety-net and rural hospitals often do not have dedicated data-science staff and computing resources on hand. The pipeline documented in Section 4.6 can be deployed as a pre-trained version in a modestly resourced hospital as a lightweight scoring service embedded into an existing discharge workflow in the hospital's EHR, rather than implemented and maintained by the hospital itself. This is a direct answer to the Research Question 2, and is a potential promising, but not proven, extension of the benefits of predictive analytics that can be implemented in less well-resourced systems.

As outlined in Section 3.10, the modeling approach was designed so that adoption would be minimized by the use of mainstream, actively maintained open-source software libraries throughout the modeling process, and because a safety-net or rural hospital's technical staff may be small enough that licensing or maintaining proprietary modeling software would not be necessary to replicate or adapt the approach. However, there were some practical challenges that would need to be solved before the hospital could make the most of such a pipeline, aside from software licensing issues. External technical assistance, likely a regional health information exchange or a state hospital association or a CMS-funded technical assistance program, would be required for most safety-net and rural hospitals to undergo the first steps in EHR data extraction and linkage to the ACS described in Sections 3.2 and 3.9. This study was performed using EHR data from a platform used to build the EPIC system, but many hospitals also use different EHR systems and the code for extracting and creating features can be adapted (code for feature domains, frailty, demography, hospital-resource metrics, ICD-coded diagnoses, comorbidity, and SDOH is conceptually transferable across EHR platforms).

The results of this study indicate that, on a per-patient basis, these rural and safety-net hospitals, which serve the most socioeconomically disadvantaged populations where the SDOH-augmented model achieved the highest incremental predictive gains, may benefit more than a well-resourced academic medical center with more

socioeconomically homogenous populations. This creates a somewhat paradoxical policy dynamic, in that the hospitals with the highest potential clinical benefit from this type of tool are the hospitals with the least existing capacity to develop it on their own; hence the need for a coordinated, CMS- or state-level technical-assistance program instead of organic, hospital-by-hospital adoption.

➤ *Algorithmic Fairness and the Risk of Encoded Bias*

Even though the algorithm performs well statistically, it might be substantively biased if its training labels or features reflect historical inequities, as illustrated by the cautionary example provided by Obermeyer et al. (2019) (see Section 2.6.1). This risk is of immediate relevance to the current study because of two reasons. Even though the outcome label used here (30-day all-cause readmission) is a direct clinical event and not a cost proxy like those used in the Obermeyer et al. (2019) case, several of the model's top clinical predictors are partly a reflection of coding practices which can be systematically different across hospitals and patient populations, in the sense that documented clinical complexity is not necessarily a perfectly objective measure of underlying clinical complexity. Second, and more directly, the strong predictive value of income and neighbourhood socioeconomic status (NSE) in Figure 6 implies that any subsequent use of this model output is accompanied by an increased responsibility to ensure that SDOH-informed risk scores are used to direct additional resources to disadvantaged patients, but not to deny, delay, or prioritize care based on a patient's neighbourhood of residence.

A concrete fairness-auditing protocol, informed by the Obermeyer et al. (2019) methodology of comparing model output against an independent, non-proxy measure of true clinical need across demographic subgroups, would be a necessary precondition for any operational deployment of this study's modeling pipeline. At a minimum, such an audit should examine model calibration (correspondance between predicted probability and observed readmission rate) separately by race/ethnicity, income level, and geographic area (rural vs urban), and specifically for the pattern documented by Obermeyer and colleagues—the pattern of systematically assigning a lower predicted risk to patients from a disadvantaged subgroup compared to those from a more advantaged subgroup who are clinically similar. A hospital should not use the output of this model in any way as a substitute for an actual audit to make any resource allocation decisions for discharge planning, but rather only rely on the output of this model after such an audit.

➤ *Regulatory Pathways for Clinical Deployment*

The FDA's AI/ML Based Software as a Medical Device Action Plan of 2021 serves as a helpful lens through which to consider the regulatory journey that would be traversed if this study's modeling pipeline were to be deployed clinically. The five priority areas outlined in that Action Plan include algorithmic transparency and bias mitigation, which are most closely related to aspects of this study covered in Section 3.6 (SHAP) and Section 3.6.1 (Anchor-based explainability) and in Section 5.7 (fairness-auditing protocol). Two other areas of focus indicate significant further work that would need to

be completed prior to deployment in practice: There is a need for a formal predetermined change-control plan that outlines how the model would be retrained or updated as more data is collected; and there is a need for an actual performance-monitoring program to detect the type of model drift described in Section 6, where a model's accuracy decreases over time due to changes in clinical practice, coding conventions, or the underlying patient population.

The exact regulatory classification criteria that are in effect at the time of deployment, and whether the pipeline's output is used only as part of a clinician's decision-making process and subject to independent clinician review or whether the product of the pipeline is used in a more automated way to directly trigger care-management actions may also determine the formal classification of the pipeline as a regulated Software as a Medical Device. The principles of governance that are inherent in the FDA's Action Plan (transparency, bias mitigation, and continuous performance monitoring) are good institutional practices for any hospital that implements a predictive model of this type, regardless of the specific regulatory classification that may apply in a particular implementation context.

➤ *Economic and Return-on-Investment Considerations*

While not a formal cost effectiveness analysis was performed in this study, several of the following qualitative economic implications are apparent from the findings reported in Section 4. It is important to keep in mind the national cost estimate set forth in Jencks, Williams, and Coleman (2009) of \$17.4 billion in unplanned Medicare re-hospitalization to put the order-of-magnitude perspective of any hospital-level investment in predictive-analytics infrastructure in perspective. However, the cost of deployment is not that much related to the scoring itself, but to the initial investment needed for data-integration and validation (Section 3.2 and 3.9), and to capacity investment needed for case-management to act on the resulting risk scores.

The ten-to-twelve-point AUC gain over the logistic-regression baseline in Section 4.1 and the additional eleven-point gain in Section 4.7 by adding SDOH data points to the existing hospital case-management capability suggest that the existing health system capacity to find and manage cases—rather than the analytic infrastructure—is the limiting factor for all but the very largest hospitals. A hospital that already has a fixed capacity, transitional-care program, such as one in which it can follow a limited number of discharged patients intensively each month, will be better off devoting more of its case-management hours to the accurate selection of patients who require such intensive follow-up, rather than adding more patients to the program, because a larger program, but less accurately targeted, would consume more absolute case-management hours on patients who would have never been readmitted, regardless of the intervention.

In principle, the payment penalty the HRRP imposes on hospitals from a payer's point of view, as outlined in Section 1.1, should already encourage hospitals to invest in this type of predictive-analytics capability. However, the continued

national readmission burden more than a decade after the adoption of HRRP implementation indicates that the financial penalty has yet to be enough to ensure widespread use of sophisticated risk-stratification tools, especially among the safety-net and rural hospitals that were discussed in Section 5.6 and may lack the initial capital or technical expertise to develop such tools even if they are financially incentivized to do so. The lack of incentive and capacity is also an economically sound reason to recommend the type of coordinated technical assistance program outlined in Section 5.6, as a market or penalty-driven approach alone seems to be inadequate, given that the national readmission rates remain stubbornly high.

➤ *Integration with Care Management and Workforce Considerations*

However accurate a predictive model may be, it would not improve patient outcomes by itself; it would improve outcomes only if it alters the specific actions taken by discharge planners, case managers, home-health agencies, and outpatient primary-care teams. The idea of having a two-tier explainability workflow described in Section 5.4, which pairs simple rules of the sort that would be used to explain an anchor with detailed force plots in the SHAP format for patients for whom the explanation is ambiguous or high in complexity, was specifically designed with this human-workflow constraint in mind, but deliberate workforce planning would be required to be successful. Case managers and discharge planners would require training in both reading a SHAP force plot or an Anchor rule and what specific interventions are warranted for each of the top risk factors listed in Section 4.2 – one score without an accompanying menu of evidence-based interventions mapped to each risk factor is not very helpful.

The identification of household composition as a top-five predictor once SDOH data is included, documented in Figure 6, points toward a specific and actionable workforce implication: hospitals may need to strengthen coordination between inpatient discharge-planning staff and community-based social-work or home-health resources capable of assessing and addressing household-level support gaps, a function that falls outside the traditional scope of most inpatient case-management roles. Likewise, the salience of neighbourhood socioeconomic status indicates a need for a more nuanced approach to implementing the SDOH risk factor from the model into a specific plan for care after discharge, which could involve a local health worker to better understand local resources like community health clinics or local food assistance programs, as well as local transportation options.

➤ *International and Comparative Perspectives*

While this study is conducted in a particular institutional and regulatory environment (the United States Medicare program), some of the methodological conclusions apply to a wider international literature on readmission risk stratification. Indeed, one of the key comparison benchmarks of this study was the LACE index, which was originally developed and most widely validated in non United States health systems, and the variation in AUC across national and

clinical settings, from 0.58 in a community hospital in the United States (Robinson & Hudali, 2017) to 0.773 in a general medical cohort in the United Kingdom (Damery & Combes, 2017), demonstrates that a fixed-formula risk score does not necessarily transfer cleanly from one health system to another, whether national or local. This finding supports a larger methodological claim from all of Section 5: that the machine learning models in this study are built to be, and are, structurally superior to “one-size-fits-all” formula-based alternatives whenever there are important differences in the populations, health systems, or countries where they are applied. Future research that takes the approach of this research into other populations (not Medicare or not in the United States) would be a good test of whether the specific feature importance patterns identified in this study (especially the importance of discharge disposition and comorbidity burden) can be generalized beyond this context.

➤ *Comparison with Non-Algorithmic Readmission-Reduction Interventions*

Predictive analytics—such as that created in this study—should be viewed as an adjunct to, not a replacement for, the myriads of interventions that hospitals have used since the HRRP was launched to help prevent readmission, such as structured discharge-instruction protocols, dedicated transitional-care nursing programs, post-discharge telephone follow-up programs, and community paramedicine initiatives. None of these tried-and-tested interventions relies on a predictive model to be implemented and many hospitals have embraced them long before machine learning tools were practically accessible. The unique value of the modeling pipeline developed here is that it helps to allocate these interventions more precisely in the context of a hospital's finite resources, whether that is the capacity to provide transitional-care nursing hours, follow-up phone calls, or community paramedicine visits, as explained in economic terms in Section 5.9.

This complementary framing also makes it clear that the model's contribution to discrimination improvement (however significant it is when compared with the LACE and HOSPITAL score baselines) only results in improved outcomes for patients when a hospital has, or is willing to develop, the capability to implement a non-algorithmic intervention program that the model's risk scores are designed to guide. Such a hospital that implements the modeling pipeline in this study with no transitional-care or social-work referral pathways to follow the model's recommendations would not likely be better off, especially relative to the legacy scoring tools upon which the model is based, since predictive accuracy and the capacity to deliver the care are two factors that are not separable from one another.

VI. LIMITATIONS

There are several limitations to qualify the above conclusions. As a retrospective and cross-sectional study, the analysis is based on existing EHR and ACS data and not a prospective clinical trial, which means that there is no causal evidence of the associations between the predictors and readmission. Secondly, the cohort is derived from a large

EPIC EHR extract but only includes a small number of health systems, not a nationally representative Medicare population, and may lack generalizability to health systems that differ in patient-mix, staffing, or regional characteristics. Third, because neighbourhood variables in ACS are measured at the census-tract level and not at the individual level, there is some ecological approximation in the results when a patient's socioeconomic circumstances vary from the average of his or her neighborhood. Fourth, SHAP values are a theoretically correct decomposition of the output of the model; they do not mean that a feature is a direct cause of a particular output, and they only explain the correlational structure that the model has learned—clinicians should consider the most important features as hypotheses for further investigation and not as definitive causal targets. Fifth, the unsupervised clustering in Section 4.8 is exploratory and was not confirmed by an independent clinical adjudication of the cluster membership, and the clinical significance of the small peripheral clusters that were found in Figure 12 is thus not definitive.

The sixth restriction relates to the single partition used to create the headline performance numbers in Table 2 and throughout Section 4. The final comparative AUC values presented for the six candidate algorithms used were based on a single held-out test partition, and were not obtained by a completely repeated, multiply resampled estimate with formal confidence intervals, although stratified cross validation was applied during hyperparameter tuning as described in Section 3.5. The difference between the best and worst performing tree-based ensembles in Table 2 is therefore not as large as the sampling variability that might reasonably be expected between split-train-split-test iterations and the ordering shown here should be interpreted as suggestive, not definitive, and statistically validated.

The seventh limitation involves the potential for a model drift over time, which was discussed in the regulatory discussion of Section 5.8. The model was trained on a snapshot of EHR and ACS data that goes back to 1999, and performance may suffer if there are significant changes to clinical documentation practices, to coding of discharge disposition, or to the underlying demographic and socioeconomic characteristics of the patients after the training period, which should be actively monitored through the sort of real-world performance-monitoring program proposed in Section 5.8, not assumed away. Additionally, the set of features used for this study was more extensive than those used in most of the previous single-condition studies reviewed in Section 2.3, but not exhaustive, and it did not include continuous physical-activity or wearable-sensor data fed via NLP, as found in the COPD-focused study by Lin, Chen, and Tsai (2022), nor did it include unstructured clinical-note text, which is a potentially rich source of data – but was not explored here – as identified by the systematic review by Osei-Anto, Fields, and Ramanathan (2025).

The ninth limitation is that there is no dedicated polypharmacy or medication-interaction variable, discussed conceptually in Section 2.3.5, which is a particular and treatable limitation in the feature-engineering approach described in Section 3.3, but not of the modeling pipeline

itself. Furthermore, the feature-extraction pipeline used in this EPIC-based cohort has not been assumed to be directly transferable to other EHR platforms, and for any institution that intends to follow this study's approach any deployment of the model to its patient base will likely require a considerable effort in data mapping and revalidation.

VII. CONCLUSION AND FUTURE DIRECTIONS

The gradient-boosted and stacked ensemble models based on an integrated clinical, administrative, and social-determinants-of-health feature set directly addressed the first research question by showing discrimination of 0.77 to 0.79, which is significantly better than the discrimination of a logistic regression baseline model (AUC 0.67) and published benchmarks for the LACE index and HOSPITAL score. After controlling for these other factors, discharge disposition, comorbidity burden, previous admissions, and diagnosis count were the main clinical risk factors, with income level and household composition, but not neighborhood socioeconomic status, contributing to the model after being analyzed in the same modeling pipeline, and adding to ROC-AUC after a clinical-only baseline model (0.68). This study's second research question focused on integrating predictive analytics into real-time discharge workflows and SHAP-based explainability, along with the Anchor-based rule extraction, provided a practical solution to this challenge by providing patient level explanations that could be audited.

The findings from these technical discussions were expanded into policy, regulatory, economic and workforce implications in Section 5. As presented in Table 5, the significant amount of predictive signal explained by SDOH supports ongoing improvement of the CMS HRRP stratification approach, in keeping with the suggestions outlined in Section 5.5, and the suggestion for a governance protocol to audit fairness presented in Section 5.7, directly based on the cautionary example of Obermeyer, Powers, Vogeli, and Mullainathan (2019), lays a necessary precondition for any deployment into operations of this or any other similarly SDOH-aware predictive algorithm. When combined with the considerations for integrating AI/ML into the workforce and care management, as described in Section 5.10, and the regulatory framework proposed in the FDA's 2021 AI/ML Action Plan, discussed in Section 5.8, they paint a realistic, albeit challenging, picture of how AI/ML can begin to be used in the clinic.

Further studies are needed to validate the modelling pipeline described here on a prospective basis in multiple sites, ideally coupled with a pragmatic trial in which discharge planning teams are randomly assigned to receive model-generated risk flags or usual care to determine if the discrimination benefits reported in this retrospective study are reflected in a measurable decrease in 30-day readmission rates. To achieve this, the trial should employ the protocol for auditing fairness as proposed in Section 5.7 as a formal pre-registered outcome, and thus should not produce a reduction in readmission rates at the "expense of creating or worsening disparities between demographic subgroups. The effects of

the qualitative mechanisms proposed in Section 5.3 that tie specific SDOH variables to specific barriers to recovery after discharge should also be validated using mixed methods research that directly includes recovered patients and their families or households, as should provide greater understanding of whether a simplified, lightweight version of the pipeline can be cost-effectively implemented in safety-net and rural hospitals of the type discussed in Section 5.6; whether a more refined HRRP stratification methodology using SDOH-augmented risk stratification models could further reduce the inequitable penalty burden shown in the equity literature reviewed in Section 2.6; and whether certain SDOH variables (e.g., food insecurity) are more strongly linked to specific post-discharge recovery barriers than others. The overall findings lend strength to the larger concept that the development of explainable, socioeconomically conscious predictive analytics is a clinically important, and nationally important, step toward decreasing avoidable Medicare readmissions.

REFERENCES

- [1]. Adeyemi, T., Osei, K., & Farooq, R. (2025). Predicting 30-day hospital readmission in patients with diabetes using machine learning on electronic health record data. *PMC*, 12085305.
- [2]. Bailey, M. K., Weiss, A. J., Barrett, M. L., & Jiang, H. J. (2019). Characteristics of 30-day all-cause hospital readmissions, 2010–2016: Statistical brief #248. Agency for Healthcare Research and Quality.
- [3]. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- [4]. Chaiyachati, K. H., Qi, M., & Werner, R. M. (2018). Changes to racial disparities in readmission rates after Medicare's Hospital Readmissions Reduction Program within safety-net and non-safety-net hospitals. *JAMA Network Open*, 1(7), e184154.
- [5]. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- [6]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.
- [7]. Damery, S., & Combes, G. (2017). Evaluating the predictive strength of the LACE index in identifying patients at high risk of hospital readmission following an inpatient episode: A retrospective cohort study. *BMJ Open*, 7(7), e016921.
- [8]. Dobler, C. C., Hakim, M., Singh, S., Jennings, M., Waterer, G., & Garden, F. L. (2020). Ability of the LACE index to predict 30-day hospital readmissions in patients with community-acquired pneumonia. *ERJ Open Research*, 6(2).
- [9]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.
- [10]. Golas, S. B., Shibahara, T., Agboola, S., Otaki, H., Sato, J., Nakae, T., Hisamitsu, T., Kojima, G., Felsted, J., Kakarmath, S., Kvedar, J., & Jethwani, K. (2018).

- 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 Medical Informatics and Decision Making*, 18, 44.
- [11]. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.
 - [12]. Jencks, S. F., Williams, M. V., & Coleman, E. A. (2009). Rehospitalizations among patients in the Medicare fee-for-service program. *New England Journal of Medicine*, 360(14), 1418–1428.
 - [13]. Jercich, K. (2023). Social determinants of health may help predict sepsis readmission. *Healthcare IT News*.
 - [14]. Kansas Health Information Network Heart Failure Cohort Study. (2026). Predicting 30-day heart failure readmissions using machine learning: Insights from the Kansas Health Information Network (KHIN). medRxiv preprint.
 - [15]. Keehan, S. P., Cuckler, G. A., Poisal, J. A., Sisko, A. M., Smith, S. D., Madison, A. J., Rennie, K. E., Fiore, J. A., & Hardesty, J. C. (2020). National health expenditure projections, 2019–28: Expected rebound in prices drives rising spending growth. *Health Affairs*, 39(4), 704–714.
 - [16]. Lagasse, J. (2020). Readmission penalties for safety-net hospitals drop under new rules. *Healthcare Finance News*.
 - [17]. Li, X., Zhang, Y., Kumar, P., Osei, A., & Chen, R. (2024). Predicting 30-day hospital readmission in Medicare patients: Insights from an LSTM deep learning model. *arXiv:2410.17545*.
 - [18]. Lin, W.-Y., Chen, C.-H., & Tsai, M.-J. (2022). Machine learning-based 30-day hospital readmission predictions for COPD patients using physical activity data of daily living with accelerometer-based device. *PMC*, 9406028.
 - [19]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates.
 - [20]. McIlvennan, C. K., Eapen, Z. J., & Allen, L. A. (2015). Hospital readmissions reduction program. *Circulation*, 131(20), 1796–1803.
 - [21]. Nguyen, O. K., Makam, A. N., Clark, C., Zhang, S., Xie, B., Velasco, F., Amarasingham, R., & Halm, E. A. (2020). Assessing the impact of social determinants of health on predictive models for potentially avoidable 30-day readmission or death. *PMC*, 7316307.
 - [22]. Nguyen, T., Patel, R., & Osborn, C. (2021). Social determinants of health and hospital readmissions: Can the HOSPITAL risk score be improved by the inclusion of social factors? *PMC*, 7780407.
 - [23]. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.
 - [24]. Okoro, C., Bassey, U., & Adamu, H. (2025). A machine learning approach for predicting 30-day hospital readmission in patients with diabetes. *PMC*, 13163415.
 - [25]. Osei-Anto, A., Fields, J., & Ramanathan, K. (2025). The role of machine learning in predicting hospital readmissions among general internal medicine patients: A systematic review. *PMC*, 12187041.
 - [26]. Rajaguru, V., Kim, T. H., Shin, J., Lee, S. G., & Han, W. (2022). Ability of the LACE index to predict 30-day readmissions in patients with acute myocardial infarction. *Journal of Personalized Medicine*, 12(7), 1085.
 - [27]. Raza, S. (2022). A machine learning model for predicting, diagnosing, and mitigating health disparities in hospital readmission. *arXiv:2206.06279*.
 - [28]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.
 - [29]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 1527–1535.
 - [30]. Robinson, R., & Hudali, T. (2017). The HOSPITAL score and LACE index as predictors of 30-day readmission in a retrospective study at a university-affiliated community hospital. *PeerJ*, 5, e3137.
 - [31]. Sung, S. F., Liu, K. E., & Ong, C. T. (2021). Validity of the LACE index for identifying frequent early readmissions after hospital discharge in children. *PMC*, 8032568.
 - [32]. U.S. Food and Drug Administration. (2021). Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. FDA Center for Devices and Radiological Health, Digital Health Center of Excellence.
 - [33]. Wadhera, R. K., Joynt Maddox, K. E., Wasfy, J. H., Haneuse, S., Shen, C., & Yeh, R. W. (2019). Association of the Hospital Readmissions Reduction Program with mortality among Medicare beneficiaries hospitalized for heart failure, acute myocardial infarction, and pneumonia. *JAMA*, 320(24), 2542–2552.
 - [34]. Weiss, A. J., & Jiang, H. J. (2021). Overview of clinical conditions with frequent and costly hospital readmissions by payer, 2018: Statistical brief #278. Agency for Healthcare Research and Quality.
 - [35]. Yazdan-Ashoori, P., & Van Spall, H. G. C. (2014). Using the LACE index to predict hospital readmissions in congestive heart failure patients. *PMC*, 4128541.