Patterns

Article

Machine learning for predicting readmission risk among the frail: Explainable AI for healthcare

Highlights

- We report a 0.79 AUROC for 30-day readmission prediction
- We use frailty, comorbidity, high-risk medications, and demographics for improved accuracy
- We identify clusters of high-risk patients based on sets of patient features
- Explainability is a prime focus for model predictions at different levels of granularity

Authors

Somya D. Mohanty, Deborah Lekan, Thomas P. McCoy, Marjorie Jenkins, Prashanti Manda

Correspondence

p_manda@uncg.edu

In brief

Automated prediction of readmission risk has the potential to save millions of dollars in healthcare costs and can improve patient care. Our work presents machine learning models that take into account various facets of patients, such as demographics, comorbidities, and frailty parameters, to accurately estimate their risk of being readmitted within 30 days. We place high importance on explainability, thereby enhancing confidence in the automated models.

Mohanty et al., 2022, Patterns *3*, 100395
January 14, 2022 © 2021 The Authors.
https://doi.org/10.1016/j.patter.2021.100395

CellPress

Article

Machine learning for predicting readmission risk among the frail: Explainable AI for healthcare

Somya D. Mohanty,[1,5] Deborah Lekan,[2] Thomas P. McCoy,[2] Marjorie Jenkins,[3] and Prashanti Manda[4,*]
[1]Department of Computer Science, University of North Carolina at Greensboro, Petty Building, Greensboro 27403, NC, USA
[2]School of Nursing, University of North Carolina at Greensboro, Petty Building, Greensboro 27403, NC, USA
[3]Cone Health, North Church St, Greensboro 27401, NC, USA
[4]Informatics and Analytics, University of North Carolina at Greensboro, 500 Forest Building, Greensboro 27403, NC, USA
[5]Lead contact
*Correspondence: p_manda@uncg.edu

https://doi.org/10.1016/j.patter.2021.100395

THE BIGGER PICTURE Unplanned readmission currently costs the United States millions of dollars. Predicting whether an incoming patient is at a high risk of readmission can help target healthcare efforts better to reduce this risk. In this age of big data, we can use machine learning to analyze a cohort of variables to pinpoint the risk of readmission. Our work does exactly that. One of the hindrances for adoption of artificial intelligence in healthcare is the lack of explainability. To combat that, we provide several mechanisms for understanding reasons for the model's predictions, starting at a global level across the entire dataset and down to individual patient observations. These explanations enhance confidence in the model's decision making.

1 2 3 4 5 **Development/Pre-production:** Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Healthcare costs due to unplanned readmissions are high and negatively affect health and wellness of patients. Hospital readmission is an undesirable outcome for elderly patients. Here, we present readmission risk prediction using five machine learning approaches for predicting 30-day unplanned readmission for elderly patients (age $\geq$ 50 years). We use a comprehensive and curated set of variables that include frailty, comorbidities, high-risk medications, demographics, hospital, and insurance utilization to build these models. We conduct a large-scale study with electronic health record (her) data with over 145,000 observations from 76,000 patients. Findings indicate that the category boost (CatBoost) model outperforms other models with a mean area under the curve (AUC) of 0.79. We find that prior readmissions, discharge to a rehabilitation facility, length of stay, comorbidities, and frailty indicators were all strong predictors of 30-day readmission. We present in-depth insights using Shapley additive explanations (SHAP), the state of the art in machine learning explainability.

INTRODUCTION

Hospital readmission can be defined as the unplanned re-hospitalization of a patient after a specific period of being discharged from a medical unit. Unplanned patient readmissions lead healthcare systems to incur substantial financial burdens and result in a diminished level of patient care. Therefore, readmissions have gained scrutiny as an important patient care quality metric. Federal mechanisms such as the Affordable Care Act (ACA) also place financial penalties on healthcare organizations with higher rates of readmissions. These factors, in conjunction with each other, have led to numerous scientific studies exploring factors for reducing unplanned hospital readmissions (see section "related work").

The Healthcare Cost and Utilization Project (HCUP) estimates that unplanned 30-day readmission costs the United States $41.3 billion.[1] Approximately 18% of patients on Medicare were readmitted within 30 days of discharge, a number that remained relatively unchanged between 2007 and 2010.[2] These unplanned hospital readmissions are both a burden on the US healthcare system as well as a strong indicator of sub-par quality of care.[2]

In a bid to reduce healthcare costs, the ACA established the Hospital Readmission Reduction Program (HRRP) in 2012. This program aimed to financially penalize healthcare organizations for higher-than-expected readmission rates for certain health conditions. Specifically, the 30-day mark was identified as the threshold for unplanned readmissions. The focus on 30-day readmission and its reduction was not an arbitrary choice. It stemmed from a policy based on the fact that the 30-day time period was observed most often and contributed to the largest share of costs.[3] Eventually, risk-adjusted 30-day readmission measures were used to measure hospital performance and quality of patient care. The decision to focus on 30 days for measuring readmissions has been criticized by some in the medical community. It is understood that readmissions occurring a few days after discharge might reflect poor care and a misjudgment of the patient's post-discharge needs.[3] Readmissions at 4 weeks or later might be likely due to the underlying intensity of the patient's condition requiring further care, factors that the hospital might not be able to control. However, readmissions closer to the discharge date have the greatest likelihood of reflecting patient care.[3] Taking these factors into consideration, the 30-day mark was chosen as the defining window for measuring readmission as an indication of patient care and quality. This decision by the HRRP to identify the 30-day window for readmissions has propagated through the medical and scientific community. The majority of scientific literature investigating readmissions does indeed focus on the 30-day mark (see section "related work"), lending strong precedent to the decisions made in this study.

One of the first steps in reducing readmissions is understanding and determining the ever-evolving key causes that lead to instances of readmission and developing predictive tools that assess risk of readmission. Consistent factors that lead to unplanned readmissions include premature discharge, length of stay in the hospital, and lack of post-discharge treatments, and might include other factors.[2] These other factors include advanced age; use of high-risk medications; specific disease diagnoses; presence of comorbidities; demographics, including socioeconomic status and race; and insurance/healthcare utilization.

While readmission within 30 days of discharge is an undesirable outcome for all patients, the outcomes can be particularly critical for the medically frail. The American Medical Association Council on Scientific Affairs wrote that "one of the most important tasks that the medical community faces today is to prepare for the problems in caring for the elderly in the 1990s and the early 21st century."[4] The report places particular emphasis on the growing population of frail older adults and notes that this group presents unforeseen challenges for healthcare systems. Older adults can be identified by comorbidities, frailty, and disability, and studies in geriatric medicine have concluded that these entities are often causally related.[5]

Frailty, a syndrome that is marked by decreased physiological reserve, poor resilience, and increased vulnerability to stressors, is gaining recognition as an important risk factor and predictor of poor patient outcomes. It is conceivable that the factors driving readmission for non-frail patients are different compared with those that are frail. The majority of prior work studying readmission risk has squarely focused on cohorts that include non-frail patients or focused on specific diseases such as heart failure or chronic obstructive pulmonary disease (COPD) but not on frail and elderly populations. For this study, frailty was conceptually defined as a clinical syndrome resulting from age-associated declines, physiologic impairments, and failed integrative responses across multiple organ systems with diminished capacity to cope with stressors.[6,7] The Frailty Risk Score (FRS) is based on a biopsychosocial conceptual model and operationalized using evidence-based risk factors and blood biomarker laboratory tests extracted from the Electronic Health Records (EHRs) of hospitalized older adults.[8,9]

Our contributions in this study are 2-fold: (1) we present machine learning models for predicting readmission risk for the medically frail by creating an integrated portfolio of important variables, including frailty, comorbidities, high-risk medications, demographics, disease diagnoses, and healthcare utilization; and (2) we delve deep into interpreting the predictions using model explainability tools. Interpretability and explainability are crucial for machine learning models applied to healthcare, and we address this need to provide confidence in the model's findings.

Background

Toward the development of machine learning models capable of predicting readmission, we extract a wide range of feature variables from patient EHRs. In the following, we discuss and provide relevant background for the different categories of EHR features utilized in our modeling:

Frailty

Frailty, a syndrome that is marked by decreased physiological reserve, poor resilience, and increased vulnerability to stressors, provides a new way to capture the combined impacts of acute illness and other factors on health status and recovery. There is growing consensus that frailty is a state of high vulnerability leading to adverse health outcomes, including disability, readmissions, need for long-term care, and mortality. The American Medical Association estimates that 40% of adults aged 80 years and older are frail.[4] The majority of the 1.6 million nursing home residents in the United States are considered to be frail.[5] Estimates of the prevalence of frailty in the acute care setting ranges from 50% to 94%.[10–12] Based on these numbers, it is safe to estimate that frailty is prevalent in a substantial proportion of older adults. Research indicates that frailty is associated with a range of adverse outcomes such as falls, functional and cognitive decline, disability, increased healthcare utilization, and premature mortality; thus, frailty is increasingly viewed as a salient aspect of patient health status, and inclusion of frailty in risk prediction models is increasing. Note that frailty is an aggregate estimation of risk as a product of advanced age or disease-associated complications resulting in the weakening of multiple physiological systems. Recent work has indicated that frailty as a syndrome can be detected by examining various clinical, functional, behavioral, and biological markers.[5] It is important to stress that the clinical definition of frailty emphasizes that multiple physiological systems and conditions must be present, further motivating the need for building a comprehensive repertoire of

Table 1. List of variables utilized in predicting 30-day readmission risk and their corresponding category

Frailty
FRS-26-ICD; malnutrition; abnormal weight; dysphagia; delirium; dementia; depression; vision; weakness; fatigue; dyspnea; difficulty walking; falls; chronic pain; urine incontinence; fecal incontinence; decubitus ulcer; material resources; social support problems; smoking; WBC; albumin; C-reactive protein; hemoglobin; glucose; creatinine; sodium item 21
Comorbidity
ECI
High-risk medications
Anticholinergics, antispasmodics; benzodiazepines nonbenzodiazepine hypnotics; cardiovascular; central nervous system; endocrine; pain meds; H2 receptor blockers, proton pump inhibitors; antipsychotics; anti-infective; genitourinary
Disease diagnosis (ICD-10)
Infectious and parasitic diseases; neoplasms; disease of the blood and blood-forming organs and certain disorders involving the endocrine, nutritional, and metabolic diseases; mental, behavioral, and neurodevelopmental disorders; diseases of the nervous system; diseases of the eye and adnexa; diseases of the ear and mastoid process; diseases of the circulatory system; diseases of the respiratory system; diseases of the digestive system; diseases of the skin and subcutaneous tissue; diseases of the musculoskeletal system and connective tissue; diseases of genitourinary system; pregnancy, childbirth, and the puerperium; certain conditions originating in the perinatal period; congenital malformations, deformations, and chromosomal abnormalities; symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified; injury, poisoning, and certain other consequences of external causes; external causes of morbidity; factors influencing health status and contact with health services; num diagnosis; num unique diagnosis
Demographic
Age; sex; race/ethnicity; marital status
Healthcare and insurance utilization
Num prior readmits; quad length of stay; hospital service; admit source; discharge disposition; primary payor; secondary payor; hospital (de-identified, 1–5); hospital unit; hospital admit time

patient features for understanding the risk of readmission in frail patients.

Frailty can be measured in many ways and we discuss a few of them below. The Fried phenotype framework classifies frailty on the basis of the patient having at least three of five criteria (slow walking speed, weak grip strength, low physical activity, unintended weight loss, and exhaustion).[13] This framework does not include measures of cognition and mood, and there is controversy that frailty consists of more than physical components. The Rockwood deficit accumulation model identifies frailty based on the number of deficits identified from the history and physical examination to calculate a frailty index (FI) score.[14] The FI consists of between 30 and 90 deficits (signs, symptoms, diseases, activities of daily living and disabilities, and physical and cognitive impairments). The FI is based on the premise that the greater the number of deficits, the more likely that person is to be frail. The third method is based on geriatric assessment and subjective determination of frailty by the healthcare provider and includes clinical tools such as the Clinical Frailty Scale, Identification of Seniors at Risk, and the Tilburg Frailty Indicator (TFI). Widespread adoption of these tools has been limited due to implementation issues such as patient and provider burden due to the need for training, time to administer, and special equipment.

Recently, a Hospital FRS (HFRS) based on 109 International Classification of Diseases, Tenth Revision (ICD-10) diagnosis codes from EHR data were validated in the UK National Health Service.[15] Increasing HFRS was associated with significantly increased risk for longer length of stay, 30-day urgent readmission, and 30-day mortality (C-statistics for these outcomes ranging from 0.56 to 0.68). Further examination of the HFRS in a retrospective cohort Canadian study compared its performance with another administrative data-based algorithm, the Hospital-patient One-year Mortality Risk (HOMR) Score, which was developed to predict the 1-year risk of death after admission to hospital and was also considered a proxy for frailty.[16] The HFRS was calculated by assigning point values to any of 109 ICD-10 codes listed in each patient's index admission or any admissions in the prior 2 years. Using existing methods to detect frailty in the acute care setting is challenging since frailty-related diagnosis codes are subject to under-coding, whereas frail patients with other comorbidities, such as cardiac conditions or cancer, might be grouped with non-frail patients.[17]

Our study utilized a proxy measure for frailty (FRS-26-ICD)[18] drawn from ICD-10, Clinical Modification (ICD-10-CM) disease diagnosis codes that encompass, common geriatric syndromes, psycho-social factors, and blood biomarkers. The FRS-26-ICD defines frailty as a clinical syndrome resulting from multi-system physiologic impairments and failed integrative responses with diminished capacity to resist and recover from stressors.[8,19] In addition to the FRS-26-ICD composite score, 32 additional variables were incorporated to capture the measure of frailty for every patient. Variables used to compute the FRS-26-ICD include but are not limited to malnutrition, abnormal weight, fatigue, and difficulty walking (Table 1 shows a full list). Using ICD-10 codes to detect frailty is challenging since frailty-related diagnosis codes are subject to under-coding, whereas frail patients with other comorbidities, such as cardiac conditions or cancer, might be grouped with non-frail patients; nursing flowsheet data reflect other aspects of patient health status, such as the impact of medical diagnoses on symptom burden and function.[20] Shortcomings of the FRS include using clinical data for research in which documentation may vary across providers and medical diagnoses may be recorded preferentially over psycho-social problems. ICD-10 codes and coding practices do not reflect all of the patient's needs during hospitalization. The binary classification of biomarkers versus quantification as quartiles or a continuous measure may under-estimate risk.

Comorbidity

In the early 1970s, comorbidity was beginning to be seen as a critical factor for understanding the prognosis and outcome of patient health.[21] Comorbidity, defined as the coexistence of two or more medical conditions in a patient, is commonly included in risk prediction models to account for patient heterogeneity and chronic disease burden that is associated with increased complexity in care and poor clinical outcomes.[22] Comorbidity has been found to be associated with mortality, quality

Table 2. Summary of eight comorbidity measures

Comorbidity index	Description
Cornoni-Huntley[24]	used to investigate hypertension and associated conditions
Duke Severity of Illness[25]	evaluate ambulatory primary care patients
Hallstrom[26]	predicting outcomes of cardiac arrest
Hurwitz[27]	estimate the influence of comorbidity on different types of patient care for back problems
Incalzi[28]	uses 52 conditions weighted for strength of association to mortality
Kaplan	uses comorbidity and pathophysiologic of the comorbid conditions
Liu[18]	combines 38 conditions and used for stroke outcomes
Shwarz[29]	combines 21 conditions based on association with mortality

of life, and healthcare.[5] While the number of comorbid conditions has been found to be a factor for poor health outcomes, the impact of specific combinations has not been an object of widespread inquiry. Since comorbidities can act as confounders, and, given that the large number of comorbidities would not be practical to control for individually, it is useful to control for the overall burden of comorbidity using an index.[23] Several measures have been developed to estimate comorbidity. Here we describe two widely used metrics, the Charlson index and Elixhauser index, and provide a brief summary of other metrics (Table 2).

The Charlson index is the most extensively studied index for reporting and estimating comorbidities.[21,30] The index includes 19 different diseases that were selected and weighted based on how strongly they were associated with patient mortality. The Charlson index includes, but is not limited to, diabetes and congestive heart failure.

A more recent model of capturing comorbidities, the Elixhauser measure,[31] contains 31 conditions, some of which are not accounted for in the Charlson index.[31] The Elixhauser Comorbidity Index (ECI)[31] aggregates selected medical diagnoses yielding a sum score, in which higher scores confer a higher risk status that cannot be captured by individual medical diagnosis codes alone.[23]

The disadvantages of the indices mentioned in Table 2 are that they typically have been applied for specific disease states for mortality outcomes, and less often for readmission. The majority of those indices rely on clinical judgment and subjective assessment, in contrast to our goal of using existing data in the EHR. In addition, those metrics seem to be used substantially less frequently than the Charlson or Elixhauser, which have almost become *de facto* ways of measuring comorbidities. The ECI and Charlson indices are advantageous since they can also be applied in administrative data that include only demographics, insurance/billing data, and ICD codes while not requiring additional data collection or clinical judgment.

More recently, studies have shown the efficacy of using aggregate measures such as the Charlson and Elixhauser for measuring comorbidity. Austin et al.[32] investigate why summary measures such as ECI and Charlson have been so widely used in health services research and provide a mathematical proof confirming the utility of the Charlson and ECI scores. Interestingly, they point out that the variables used to construct the score are of utmost importance and, sometimes, strong predictors might be omitted, leading to poor performance.[32] This underlines the motivation of this study to build a comprehensive portfolio of patient features and variables to obtain the most accurate predictive power.

In a comparison of ECI and Charlson, prior work has shown that the ECI provided a relative improvement of 60% on prediction of mortality in the hospital.[33] Based on this precedent and taking into consideration that the ECI is a more recent model that accounts for more conditions compared with Charlson, we chose ECI as the comorbidity index in this study.

High-risk medications

Age-related changes in drug pharmacokinetics/pharmacodynamics and the greater prevalence of multiple comorbidities contribute to increased susceptibility to adverse drug events in older adults. High-risk medications are drugs that have an increased risk of causing substantial harm to patients. These medications include drugs with low therapeutic indices and present heightened risk when used in error. Medications classified as high risk exert numerous adverse effects on patients' health status and are associated with new morbidity, mortality, and readmission. The use of high-risk medications is common in older adults. A retrospective cross-sectional study of 456 patients 65 years of age and older found that slightly more than half of the patients (53.5%, n = 244) had at least one potentially inappropriate medication identified by the Beers 2015 Criteria.[34]

The Beers Criteria are a list of potentially inappropriate medications that are typically best avoided in older adults in most circumstances or under specific situations, such as in certain diseases or conditions.[35] Fifty-three medications feature in the Beers Criteria, divided into three categories: (1) potentially inappropriate medications to avoid, (2) potentially inappropriate medications and classes to avoid in older adults with certain diseases, and (3) medications to be used with caution. The recommendations in the Beers Criteria are based on expert consensus based on extensive literature review and surveys by experts in geriatric care, clinical pharmacology, and psychopharmacology.[36] These criteria have found extensive use to guide clinical medication use to decrease medication problems in older adults. The Healthcare Effectiveness Data and Information Set (HEDIS) also provides information about high-risk medications for elderly patients by grouping medications into high-level categories. A trademark of the National Quality Committee for Quality Assurance (NCQA), HEDIS provides standardized performance measures to compare the performance of healthcare plans.

Research on high-risk medications and hospital readmission is equivocal. In a study examining high-risk medications in hospitalized older adults, exposure to certain high-risk medication classes, such as benzodiazepines and opioids, were associated with increased odds of readmission.[37] Wang et al.[24] also found a high prevalence of high-risk medications (66.7%) in Chinese older adults, whereby proton pump inhibitors (42.6%) and benzodiazepine (34.4%) were most common, and having at least one prescribed high-risk medication per the Beers Criteria was a significant risk factor for all cause readmission.

In the older adult population, age-related changes in drug absorption, metabolism, distribution, and excretion, as well as drug-drug and drug-disease interactions, increase the risk for adverse drug events and potential for harm. The high-risk medications, such as those identified in the Beers Criteria (AGS, 2019), should be avoided or used with caution because they are associated with numerous negative consequences, such as falls and fractures, delirium, depression, mobility impairments and functional decline, urinary and bowel abnormalities, and nutritional issues (anorexia, nausea, dehydration).

In another study, Blachman et al. report that, other than old age, high-risk medications are the most important risk factors for falls among frail patients.[38] The strength of the association between falls and high-risk medications increases both with the number of high-risk medications prescribed and the dosage of those medications.[38] Another study found that high-risk medication categories such as steroids and narcotics, anticholinergics, and medications were associated with readmission.[39]

In this study, the medical experts on our team manually integrated information from the Beers Criteria with high-level categories from HEDIS to determine medications considered high risk for elderly patients.

Disease diagnoses

ICD-10 codes[40] are recorded in the EHR data to describe the principal problem attributed to the hospital admission as well as any additional diagnoses for each patient not directly related to the principal problem. According to the Centers for Disease Control and Prevention (CDC), these codes are important for classifying diseases, recording inpatient procedures, and estimating healthcare utilization. The use of a consistent and controlled set of codes for the description of patient conditions enables tracking of health conditions, severity of illness, and comorbidities, and measuring patient care/outcomes, to make clinical decisions.

The ICD-10 code list contains approximately 69,000 diagnosis codes and 71,000 codes for procedures. Each ICD-10 code contains three to seven characters. The first character is a letter followed by a numeric character. Characters three to seven can be letters or numbers. The leading letter in an ICD-10 code indicates the overall disease group. The thousands of ICD-10 codes are arranged in the form of hierarchical classification or an ontology that groups codes into higher-level codes that are further grouped.

Several studies have used and found ICD-10 codes to be important factors for predicting readmission for all causes and for specific conditions.[25,41,42] Lee et al.[41] report that the category of principal diagnosis was the most important predictor for patients with a hospital stay longer than 2 days, indicating that readmission can be more accurately predicted by analyzing the type of disease. Similar results were seen in Chirapongsathorn et al.,[25] where the first listed ICD-10 code for hospitalization was used to identify reasons for readmission. Another study reported prediction of readmission solely by considering ICD codes and a small set of background variables.[42]

We incorporate ICD-10 codes in a number of ways: we map each code into a higher-level category (see section "experimental procedures"), and use individual categories, the total number of absolute diagnoses, as well as the total number of unique diagnoses for prediction.

Demography and healthcare utilization

Patients admitted to hospital units have differences based not just on their clinical characteristics but also on non-clinical aspects, such as age, race, marital status, socioeconomic status, and insurance status, which might affect their risk of readmission. These factors can be broadly grouped into the umbrella of demographics and healthcare utilization.

The vast majority of studies discussed in the section "related work" incorporate self-reported demographic and healthcare utilization data. The importance of including these factors has also been well studied.[26,27,43] While some studies[43] found that socioeconomic, health status, and psychosocial variables are not dominant factors for predicting readmission, these features have found utility in numerous other studies. In a study of 30-day readmission for patients aged 65 years and older, Silverstein et al.[44] found that age, race, Medicare status, and other demographic factors predicted 30-day readmission.

In this study, we created two categories of variables: demographic, and healthcare and insurance utilization. Variables such as age, sex, race/ethnicity, and marital status are considered in the demographic category. The healthcare and insurance utilization category contains variables such as primary payor, hospital unit, and hospital service (see full list in Table 1).

Related work

The background of related work for this study is vast. There are investigators that studied readmission risks with and without employing machine learning, those that explored readmission risks for specific disease conditions, and those that investigate readmission risk for the frail but not including other factors such as comorbidities and high-risk medications. Common to the majority of these studies, though, is the use of the 30-day threshold for investigating readmission. As discussed in section "introduction," the 30-day threshold has stemmed from the federal policies in the HRRP program and has subsequently been adopted by the medical and research communities.

The use of machine learning for the prediction of readmission risk is ubiquitous and well studied.[1,2,28,45,46] A number of studies (described below) employ different machine learning models for understanding readmission risks in different patient cohorts and disease conditions. Some of these studies generated complex models that use thousands of features,[2] while others limit themselves to patient information that can be easily collected within the hours of initial admission.[45]

A comparison of commonly used models for predicting readmission risk studied a set of four models (LACE, Stepwise logistic, least absolute shrinkage and selection operator (LASSO) logistic, and AdaBoost).[1] The study finds that LACE has moderate predictive power, with area under the curve (AUC) scores around 0.65. Variables include number of emergency room visits in the last year, Braden pressure ulcer risk score, polypharmacy, employment status, and discharge disposition (patient's anticipated location or status following a hospital visit). LASSO was found to be the best model for both small and large data sizes (0.73 AUC). In a comparison of different models, a 2020 study in New Zealand showed that XGBoost, random forests, and AdaBoost achieve better predictive performance compared with LACE and PARR, with F-1 score improvements of 12.7% and 23.2% respectively.[47] Another study conducted on patients within the Maine Health

Information Exchange created a risk model for identifying patients at risk for readmission within 30 days post discharge.[2] The model achieved a C-statistic (AUC) of 0.72 in predicting readmission. While the above study uses a wide range of features and complex models to predict readmission, other studies have focused on identifying early hospital readmission factors in diverse patient populations using simple models with limited features.[45]

The above studies predict readmission across all patient samples without discrimination for disease diagnoses leading to readmission. On the other hand, a number of studies have been conducted to identify risk factors for readmission after specific disease conditions. Goto et al.[48] used machine learning to predict 30-day readmission after hospitalization for COPD. Patient characteristics and inpatient care data were used with logistic regression, LASSO regression, and deep neural network models to predict readmission after COPD. Tube feeding duration, blood transfusion, use of thoracentesis, and sex were found to be important predictors for the machine learning models. Machine learning has been used to study readmission among patients hospitalized with ischemic heart disease.[49] Results showed that length of stay and the ECI were the top predictors, with an AUC of 88%. In another case of predicting readmission for specialized disease conditions, Mahajan et al.[46] explored logistic regression, random forest, gradient boost, and neural networks for predicting 30-day readmission for heart failure.

More recently, a study now considered the state of the art from scientists at Google, Stanford, and other institutions, conducted an analysis of EHR to predict mortality, 30 day readmission, and prolonged length of stay.[29] The study reports that deep learning models were found to outperform traditional machine learning and clinical models at predicting the above events of interest. The models were validated using EHR data from two US academic medical centers and reported a 0.75 AUC for 30-day readmission.

Machine learning and other statistical methods have been applied to predict the risk of 30-day readmission in older adults.[44,50–54] In 2020, Grana et al.[50] conducted experiments in a cohort of 645 frail patients for the study of readmission showing positive results for the application of machine learning and making the case for more studies in larger cohorts such as ours. Another study reports that frail geriatric trauma and emergency general surgery patients tend to have longer lengths of stay and more readmissions. In a cohort of 239 patients, they found that screening for frailty and establishing a frailty pathway resulted in decreased length of stay.[53] A retrospective cohort study of 230 frail older adults assessed whether self-reported symptoms predicted unplanned hospital readmission or emergency department care within 30 days of discharge.[54] Here, four indicators similar to the FRS were predictive (drowsiness, depression, shortness of breath, and anxiety).

In another 2020 study of 720 older patients (majority >75 years old), higher Charlson comorbidity and excessive polypharmacy were among the features that were associated with increased odds of readmission. Silverstein et al.[44] developed predictors of 30-day readmission using administrative data for 29,000 adults aged 65 years or older. Results indicated that age, male sex, African American race, Medicare, and major comorbid conditions were important predictors (C-statistic 0.65). Another study[52] with one of the largest cohorts (479,854 patients aged 65 years and older) from Danish public hospitals reported that acute admission, number of days since previous hospital discharge, comorbidity, increased drug use, and greater utilization of hospital services were strongly associated with readmission (C-statistic 0.70).

Several prior studies have defined frailty using geriatric syndromes comparable with the FRS used in this study. Geriatric risk factors were defined as health status characteristics common in older adults who have potential to be intervened upon and ameliorated if identified in a timely manner. These risk factors are analogous to 10 of the 26 indicators in the FRS (malnutrition, weight loss, falls, urinary issues, bowel issues, difficulty walking, dementia, vision, decubitus ulcer, and social support) and are represented in the Fried frailty phenotype[13] and Johns Hopkins University ACG frailty indicator.

Many readmission risk prediction models exist,[55,56] but these models are limited in being able to predict readmission within a high-risk population, such as acutely ill frail older adults with multiple comorbidities and medications. Another factor that is critically important to consider when using machine learning models to explore readmission risk is the information used for the prediction. Most studies incorporate the standard patient data such as demographics, drugs, and medical diagnosis information. The importance of building a comprehensive cohort of variables for the specific cohort of patients cannot be overstated. Indeed, a new study from 2020 that developed predictive models for 30-day readmission and mortality on 3,000 patients underscores and explicitly makes the case for compiling a comprehensive set of variables to achieve the best predictive performance.[57]

RESULTS

The EHR dataset prior to pre-processing and manipulation (Figure 1) contained 145,148 observations corresponding to 76,294 patients. A series of seven exclusion criteria (Figure 2) were sequentially applied to result in 128,581 observations for 68,152 patients. These observations consisted of 18,840 readmissions and 109,741 non-readmissions. The data contained 458 variables that were used for prediction of readmission. Table 3 provides descriptive statistics of a set of salient variables. Pairwise collinearity tests were conducted for all pairs of features in the data. Based on the correlations, the FRS-19 feature was eliminated from the data.

The data in the EHR were severely imbalanced, with a disproportionate number of non-readmissions compared with readmissions. We tested three sampling techniques to address this imbalance: (1) under-sampling, (2) over-sampling, and (3) no sampling. Table 4 shows the mean F-1 and AUC scores for the three sampling strategies applied to five models. The last column of the table (mean AUROC) shows that under-sampling performs the best or comparably with the other two strategies. However, the real difference in the sampling strategies can be observed in the differences between mean F-1 score for the two class labels (0, no readmission; 1, readmission). Most models show a stark difference in F-1 score for predictions between the two classes when no sampling is applied, showing that the imbalanced data are affecting model performance. For example, the F-1 for label 0 for a random forest model is 0.93, while that of label 1 is 0.37. The same trend is observed when over-sampling is

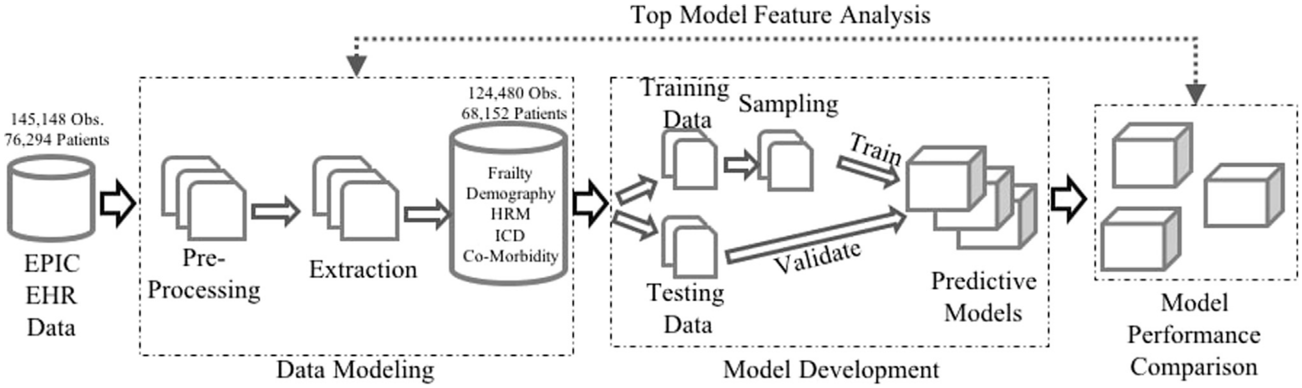


Figure 1. A breakdown of the components toward data pre-processing, model tuning/development, and explainability of the readmission risk prediction model

used across all the models tested. In contrast, under-sampling the data results in consistent F-1 scores for both classes (except for logistic regression). Based on these results, under-sampling was selected to create a balanced dataset. All subsequent results reported are on the dataset created using under-sampling.

Model selection

The four machine learning models along with the stacked classifier were modeled with the under-sampled balanced dataset. Of the four models tested (Figure 3), CatBoost outperformed the other models with an AUROC of 0.79. Random forest and XGBoost trailed behind with AUCs of 0.77 and 0.77 respectively. Logistic regression, which is widely used for prediction of readmission, performed substantially worse than the other three models, with an AUC of 0.68. The XGBoost-based meta-learner was selected for the stacking classifier because it showed the highest AUROC score. While the stacking classifier performed slightly better on the AUROC score (0.7964 versus 0.7948), the gain is negligible. Additionally, as the stacking classifier is a combination of multiple models (and dual stages), it is much harder to explain or interpret its findings using the SHAP explainability mechanisms described above. Considering that explainability is of the utmost importance for AI models in healthcare, we selected the CatBoost model for further analysis.

Feature importance

First, we further explored the best model from our tests (CatBoost) to determine the most important features for predicting 30-day readmission. Figure 4 shows the top 10% features for this model. The strongest feature associated with readmission was discharge to a rehabilitation facility. Ninety-two percent (844 of 925) of the patients who were discharged to a rehabilitation hospital were readmitted within 30 days. This was followed by the number of diagnoses and number of prior readmits. Morbidity and comorbidity, represented by the ECI, was also found to be an important factor. Next, we see that the length of stay affects the risk of readmission. Specific disease categories such as poison injury, parasitic infections, and connective and musculoskeletal are seen to be important predictors. The presence of several individual FRS indicators in the top features (e.g., white blood count, abnormal weight, sodium, malnutrition, and decubitus ulcer, corresponding to the well-known frailty phenotype defined by Fried et al.[13]), indicates the importance of including indicators of frailty for prediction of readmission.

Model explainability

Explainability of AI models is important, especially in the context of healthcare decision making. In this section, we delve deeper into the CatBoost model to explain the model's predictions, the features that influenced the model, along with examining how predictions were made at an individual observation level.

We provide model explanations along the following aspects:

1. Global interpretability: global interpretability allows the reader to identify the most important features used to make predictions for the model, glance at the distribution of the data values for these features, and learn how these data values affected predictions.
2. Impact of the top variables toward readmission: we show how the top six features of the model affect the risk of readmission. We show the model's tendency to predict readmission or non-readmission for each observation in a feature and compare the model's prediction with the actual ground truth. This section helps verify if the model's predictions are in congruence with the ground truth, a verification mechanism to boost confidence in the model's findings.
3. Local interpretability: we show how the model's predictions were made at an individual observation level. These visuals show the synergistic effect of a subset of features that led to a prediction for this particular observation. These visuals can be used by healthcare experts to understand the model's findings at a patient level.
4. Identification of at-risk patient groups: we employed clustering methods to group the dataset into clusters of patients with different levels of readmission risk. These clusters make it easy to identify which group a group would be most similar to, thereby estimating their risk of readmission. We also provide a comparison of the clusters for the model's predictions versus the ground truth to verify if the predictions are in line with the actual data.

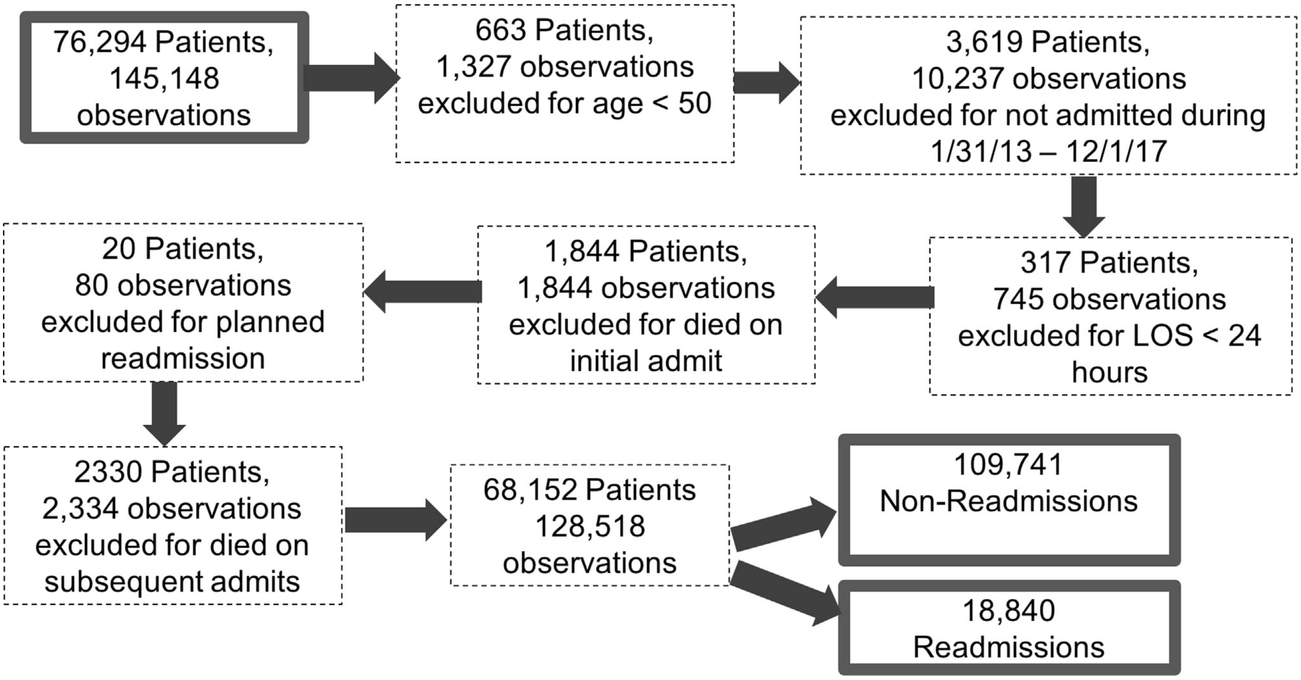

76,294 Patients, 145,148 observations
663 Patients, 1,327 observations excluded for age < 50
3,619 Patients, 10,237 observations excluded for not admitted during 1/31/13 – 12/1/17
317 Patients, 745 observations excluded for LOS < 24 hours
1,844 Patients, 1,844 observations excluded for died on initial admit
20 Patients, 80 observations excluded for planned readmission
2330 Patients, 2,334 observations excluded for died on subsequent admits
68,152 Patients 128,518 observations
109,741 Non-Readmissions
18,840 Readmissions

Figure 2. Exclusion criteria employed to compile the appropriate cohort of frail patients for prediction of readmission

Global interpretability of the model

We provide global interpretability of the CatBoost model using SHAP explanations. Figure 5 displays several key pieces of information that each provide insight into the model's predictions and what factors affected them. First, we explain the key facets of the SHAP feature importance plot and then describe the findings.

The SHAP feature importance plot displays the following information:

1. Feature importance (left y axis): similar to Figure 4, this plot shows the top 10% important features in descending order of importance.
2. SHAP value: impact on model's output (x axis). The x axis shows SHAP values, a quantitative measure of an observation's impact on model prediction. The SHAP value quantifies how much impact a particular observation makes toward the prediction of the target class. Higher SHAP values indicate an increased risk of readmission, while lower SHAP values indicate low readmission risk. The plot shows a vertical line running through the middle that acts as a baseline SHAP value. The baseline indicates observations that have no impact on the model's prediction.
3. Original values of variables (right y axis): the scale on the right (blue to red) indicates whether the values of a variable were low (blue) or high (red) in the data. Continuous variables are shown on a color gradient, while binary variables are indicated in blue or red. For example, the discharge disposition rehab facility variable is a binary variable where observations with a 0 value are shown in blue and those with a 1 value are shown in red. In comparison, the number of unique diagnoses is a discrete variable where we observe colors according to the grading scale of the color map; i.e., high observations have red, low observations have blue, and mean observations have a light color (mid-point on the color bar).
4. Variable impact: the horizontal location of a dot (blue or red) shows whether the effect of that variable's observations is associated with a higher or lower prediction of readmission risk based on the SHAP value. The plot shows a vertical line running through the middle that acts as a baseline. The farther away from the baseline an observation (blue or red) is, the more impact it has on the model's prediction of low or high readmission risk.

Synthesizing information about the original values (indicated by color) and the horizontal location, we can interpret red dots on the right of the baseline as the variable having high values in the data and the effect of those high values resulting in a high readmission risk. On the other hand, red dots on the left of the baseline indicate that the variable had high values in the observations but those high values resulted in low readmission risk. When blue dots appear on the left of the baseline, it indicates that the variable's low values were associated with lower readmission. When blue dots appear on the right of the baseline, it indicates that the lower values for that variable were associated with higher readmission risk. For example, in the variable Elixhauser, we observe a large number of red dots to the right of the baseline and large number of blue dots to the left of the baseline. This indicates that the higher the Elixhauser score, the greater the chances of readmission, while lower Elixhauser scores lead to lower readmission risk.

We see that each of the features contributes differently to the model's prediction. As expected, the top feature (discharge to a

Table 3. Descriptive statistics of salient variables in the EHR data used for predicting readmission risk

Variable	Statistic (standard error)
Mean readmissions	1.56 ± 0.011
Mean FRS-26-ICD	3.03 ± 0.08
Mean ECI	4.15 ± .09
Mean number of disease diagnoses	30.65 ± 0.07
Mean number of high-risk medication categories	2.32 ± 0.05
Mean number of prior readmits	2.1 ± 0.02
Mean length of stay (days)	4.85 ± 0.01

rehab hospital) has several observations with high values (red dots) that contribute to high readmission risk (indicated by their position to the far right of the baseline). Since, this variable is binary (1 = high, 0 = low), the red dots indicate observations for which the value is 1 and blue dots indicate that the value is 0. Note the variance in the red dots; this indicates that there is variability among these observations with high values in how impactful these observations are toward the prediction. The blue dots for this variable, on the other hand, do not exhibit similar variance, indicating that their contribution toward the prediction is relatively uniform. Moving to the next important feature, we see that the number of diagnoses had a large proportion of observations with low values compared with observations with higher values (blue dots outnumber the red dots). The concentration of the blue dots to the left of the baseline indicates that these lower numbers of diagnoses led to lower readmission risk. The small proportion of observations with higher diagnoses (shown in red) appear to the right of the baseline, indicating that these observations led to greater readmission risk. We observe a stark separation in observations (blue versus red) for the observations with prior readmissions. Observations with prior readmissions (1 = high) lead to higher risk of readmission (red dots to the right of the baseline), while observations with no prior readmissions (0 = low) lead to reduced readmission risk (blue dots to the left of the baseline). Elixhauser is the fourth most important variable, and the plot clearly shows that observations with high Elixhauser values (red dots on right) are associated with higher readmission risk and lower Elixhauser values (blue dots on the right) are associated with lower readmission risk. A very similar trend is observed with quadratic length of stay (LOS).

There are some variables where high values were associated with low readmission risk. The discharge disposition to hospice or medical facility variable has a large subset of observations with high numbers. This indicates that large numbers of patients were discharged to hospice or medical facility and that event was associated with low readmission (red dots to the left of the baseline). The feature representing discharge to home or self-care has an almost equal proportion of observations with high and low values. High rates of discharge to self-care indicate low readmission risk (red dots on left of baseline), while low rates of discharge to self-care lead to high readmission risk (blue dots on right of baseline).

The plot also shows that a large number of patients in the data were discharged to hospice or a medical facility and this event resulted in low readmission risk (red dots to the left of the baseline). On the contrary, observations with high values of being transferred to a short-term hospital saw an increased risk of readmission.

Impact of select variables toward readmission

We take an in-depth look at how a few select variables from Figure 5 affect the risk of readmission. In these figures, we compare the model's predictions with the ground truth to verify if the predictions align with the ground truth. In Figure 6, we show the eight variables of the model and how observations for each of those variables lead to readmission or no readmission. In each of the figures, we plot the original values of observations for a specific variable against their SHAP values. Each data point (corresponding to an observation) is marked red or blue to indicate readmission $(Class\_Label = 1)$ or no readmission $(Class\_Label = 0)$ respectively in the actual data. The dotted line shows the baseline for the SHAP value for the particular variable. Higher SHAP values (right of baseline) indicate that the model predicts readmission and lower SHAP value (left of baseline) predicts no readmission. A point of note here, while we are considering a single variable for analysis in this scenario, readmission can result from a combination of variables.

In Figure 6A, we see that observations with higher SHAP values (right top corner) are mostly colored red, indicating that these observations indeed had readmissions. Observations with lower SHAP values (bottom left) have a mix of observations (red and blue), indicating that some of these observations were readmissions, whereas some were not (new encounters or readmission after 30 days).

Looking at the number of diagnoses variable (Figure 6B), we see two trends: (1) as the number of diagnoses increases, the SHAP values increase, indicating that the model predicts higher readmissions for observations with a higher number of diagnoses; and (2) observations with higher SHAP values correspond to more readmissions compared with those with lower SHAP values, showing that the model aligns with ground truth in these cases. The behavior of the number of prior readmits variable (Figure 6D) is similar to trends seen for the discharge to rehab variable (Figure 6A).

In comparison, the behavior of the Elixhauser score is more complex (Figure 6C). Taking a look at the distribution of observations to the left and right of the baseline, we see that the majority of observations to the left of the baseline (with low SHAP values) were not readmissions and the majority of observations to the right of the baseline (with high SHAP values) did have readmissions. However, there are observations that the model assigns low SHAP values that were actually readmissions and vice versa. We attribute these instances of confusion to conflicting impacts from other variables.

In Figure 6D we observe records with prior readmissions, which is a strong predictor for readmissions. The observations marked in red indicate prior readmissions. The lower left of the plot shows observations with no prior readmission and are mostly marked by blue dots, indicating that the model predicts low chances of readmissions for these observations.

Figure 6E shows the observations for the quadratic length of stay variable. The figure also verifies that the majority of observations that were readmissions had high SHAP values that would lead to a readmission prediction. We do see a higher

Table 4. Performance metrics for the five models (random forest, XGBoost, CatBoost, logistic regression, and a stacking classifier)

	Mean F-1 score		Mean AUROC
	Random forest		
	Label 0	Label 1	
No sampling	0.94	0.38	0.77
Over-sampling	0.92	0.26	0.74
Under-sampling	0.68	0.69	0.77
	XGBoost		
	Label 0	Label 1	
No sampling	0.93	0.39	0.77
Over-sampling	0.92	0.38	0.76
Under-sampling	0.69	0.68	0.77
	CatBoost		
	Label 0	Label 1	
No sampling	0.93	0.38	0.79
Over-sampling	0.93	0.37	0.78
Under-sampling	0.71	0.70	0.79
	Logistic regression		
	Label 0	Label 1	
No sampling	0.91	0.07	0.56
Over-sampling	0.67	0.29	0.67
Under-sampling	0.01	0.66	0.67
	Stacking classifier		
	Label 0	Label 1	
No sampling	0.93	0.38	0.79
Over-sampling	0.93	0.36	0.79
Under-sampling	0.70	0.71	0.79

concentration of confusion instances where observations with lower SHAP values had readmissions and vice versa.

Figure 6F shows the relationship between SHAP values and the observations that had the ICD code category of poison. Here we do observe the majority occurrence of red (readmitted) with higher SHAP value and to the right of the SHAP baseline. However, there are a few observations where the SHAP values were higher and associated with no readmission and vice versa.

One of the most complex instances is the age variable outlined in Figure 6G. We do observe a large number of red instances to be at the right side of the plot, where the model is able to make correct decisions at different age values, but it is not as clear as other variables. All of the values of age >90 were set to 90 (for privacy) which is why we have a large distribution of points at 90. We also do not observe any relationship between the SHAP values and age; i.e. no linear trend of SHAP with increase in age.

Figure 6H shows the relationship between FRS26 and the SHAP values. We do observe there is a relationship between FRS26 and SHAP, where we observe a majority of readmitted patients to have higher FRS26 score, but there quite a few confused instances where the lower SHAP and FRS26 scores lead to readmission.

Higher-order interactions between top features

We briefly explored interactions between the top predictors identified in Figure 5 to gain a deeper understanding of the data and trends within. While a number of higher-order relationships can be modeled, we highlight two such interactions due to restrictions on space.

First, we show the interactions between number of diagnoses and FRS-26-ICD, both of which are among the top predictors for readmission risk (Figure 7). It can be observed that observations with low number of diagnoses (y axis) tend to have low FRS scores (colormap). We also see higher SHAP values (x axis) corresponding to higher number of diagnoses, which grows exponentially after the SHAP baseline of 0.0. This goes to show that the SHAP values of observations with low numbers of diagnoses and low FRS scores are also low, indicating that these observations are not at risk of readmission. Observations with high FRS and number of diagnoses carry high SHAP values, indicating that they are at risk of readmission.

Next, we highlight the interaction between ECI and the number of unique diagnoses (Figure 8). There is a stark contrast between observations with high versus low Elixhauser scores (y axis) and the number of unique diagnoses (colormap). Low Elixhauser scores are associated with lower SHAP values and also lower number of unique diagnoses. As we move higher in the Elixhauser score, we have a higher number of unique diagnoses as well as SHAP values above 0.0, indicating higher changes of readmission. Few observations on the lower left area of the figure show a higher number of diagnoses but low Elixhauser scores. Interestingly, these observations still have low SHAP values, indicating no readmission. It is the combined interaction of high Elixhauser and high number of unique diagnoses that truly affects the risk of readmission (observations on the right top).

Local interpretability

SHAP can be used to explain the model's decisions at an individual observation level. These insights can allow healthcare experts to identify and analyze the decision-making process for specific observations if need be. This level of detail allows to understand the particular features that played a role in the model's prediction and analyze whether those features align with human experts. Figures 9A and 9B show the observations with the highest SHAP scores (the observations also had high predicted readmission probability at 0.99). Figures 9C and 9D show the observations with the lowest SHAP scores and had high predicted no-readmission probability at 0.98.

The plots in Figure 9 show several insights about the prediction and the factors that affected the prediction for this specific observation. The model output value (in bold) is the cumulative SHAP score for an observation. The base value (0.1678, indicated above the number labels) is the baseline prediction that the model would make in the absence of any features. Features that push the output value higher (toward readmission) are shown in red, and those that lower the output value (toward non-readmission) are shown in blue. The data values for each of these features can also be seen in the plot. In Figure 9A, we see that the model predicts readmission with an output SHAP score of 7.02. We observe that the patient in this encounter has prior readmits, was discharged to rehab facility, had high-risk endocrine medication, 43 diagnoses, and was admitted to neurosurgery, leading to high confidence that the patient will be readmitted. Features shown in red contributed to the readmission prediction.

In comparison, Figure 9C, the observation had low number of diagnoses (1), was not discharged to rehab facility, and was admitted to orthopedics, leading to the low chance of

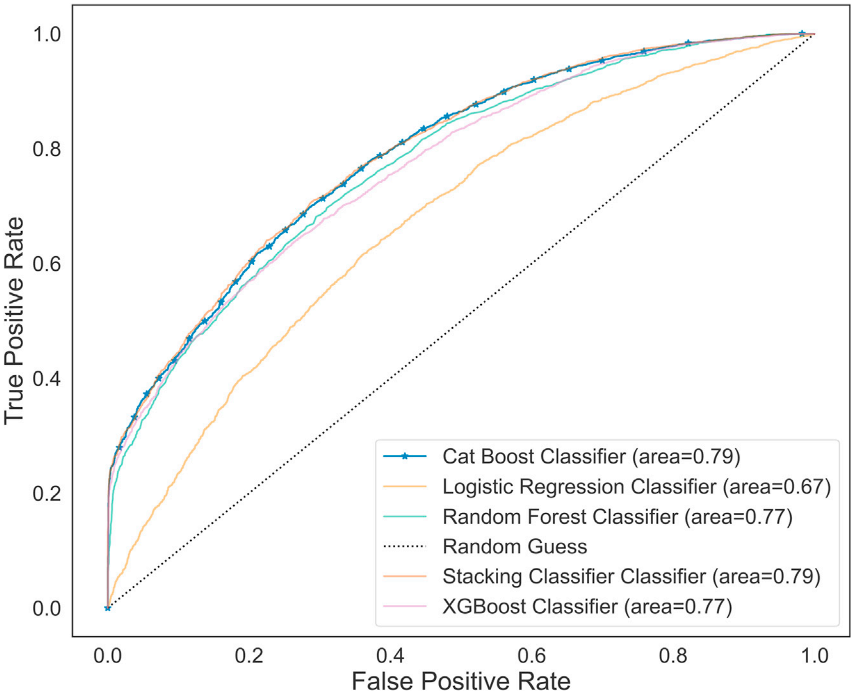


Figure 3. AUC comparison for the different machine learning models and the stacking classifier

readmission. Elixhauser score was unavailable for this observation. Similarly, for Figure 9D we observe lower number of diagnoses, lower number of unique diagnoses, were not discharged to a rehab facility, and were in low-risk hospital units such as orthopedics, each of which pushes the observations toward not getting readmitted. The blue indicates features that mitigated the effect of the red features. It is important to note that these observations did not have high values for any of the high-risk factors. It is the synergistic effects of these variables that result in the final prediction. This level of explainability is valuable to investigate the model's decisions at a patient-to-patient level for complete transparency.

Identification of at-risk patient groups

TreeSHAP associates a SHAP value to each observation for every variable in the dataset. These SHAP values are inferred from the model's decision process of making predictions about the target variable $(Class\_Label = \langle 0/1 \rangle)$. We utilized these SHAP values to cluster observations for identifying latent groupings of patients with varying risk of readmission.

K-means clustering was used to generate clusters of observations. In order to identify the ideal cluster size for the underlying data, we evaluated the inertia score for K-means clusters ranging from $K = 2, 3, \cdots 14$. As shown in Figure 10, the reduction in inertia score is no longer substantial after eight clusters. Based on this plot, we selected to cluster the data into eight clusters.

Observations in these eight clusters have different rates of readmission, ranging from 4.71% to 100% (Table 5). While all of the clusters have different variables that are influential for the observations in that cluster, overall, we found that the eight variables were important for all clusters: (1) discharge disposition rehab facility; (2) prior readmits; (3) number of diagnoses; (4) Elixhauser; (5) quadratic length of stay; and (6) ICD injury/poison. To understand what factors led to high readmissions in some of the clusters, we provide a detailed look at the distribution of values for these important variables for each cluster (Figure 11).

In Figure 11A, we see that cluster 2 and cluster 5 have high values for the discharge to rehab facility variable. Based on other results discussed above, we know that high values of this variable lead to readmission. Reaffirming this result, we see that the percentage of readmission in clusters 2 and 5 is 98.67% and 100% respectively. A similar trend can be seen in Figure 11B, where cluster 3 and cluster 5 have high numbers of prior readmissions, whereas observations in the other clusters have no readmissions. These two clusters also have high percentage of readmissions.

Clusters with high numbers of diagnosis (2, 3, 4, and 6) show higher percentages of readmission (Figure 11C). The same can be observed for the Elixhauser variable (Figure 11D), where clusters 2, 3, 6, and 5 have observations with a high comorbidity score, leading to high readmissions. Figure 11F shows the clusters (3, 5, 6) that have high prevalence of observations annotated with ICD category of injury/poison. This cluster also has a high degree of readmission percentage, with clusters 3 and 5 having 100% readmission along with cluster 6 at 60.55% readmission.

In Figure 11E, we see that observations in some clusters (2, 5, 7) have higher values compared with others (clusters 0, 1, 3, 4, 6). However, among clusters that have high lengths of stay, only clusters 2 and 5 have high rates of readmission. Cluster 7 is an anomaly, with only 17% readmission rate. Taking a look at the other features influencing cluster 7, we see that observations in this cluster have no discharges to a rehab facility, no prior readmissions, and relatively low number of diagnoses, leading to the low overall readmission rate.

It is important to note that it is the synergistic effect of multiple variables that causes observations to belong to clusters with high or low readmission scores. Looking at cluster 0, the cluster with the lowest readmission percentage (4.71%), we can see that observations in this cluster have no discharges to a rehab facility, have no prior readmissions, very low diagnoses, very low comorbidity scores, and similarly low values for length of stay and number of unique diagnoses.

These clusters clearly show how the observations can be segregated into meaningful groups that predict a patient's risk of readmission while identifying key variables for healthcare workers to track.

Note that these clusters are based on the model's prediction of readmission risk. Next, we compare the clustering of predicted risk with the actual ground truth as a validation exercise of the clustering method's efficacy.

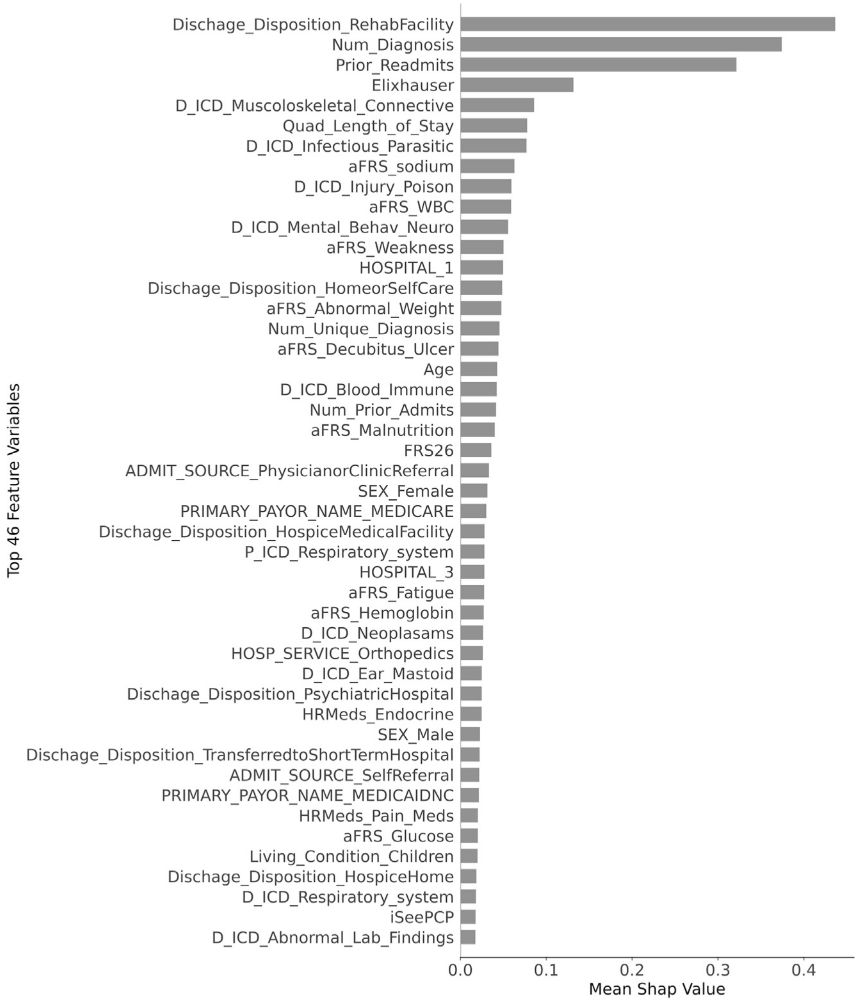


Figure 4. Feature importance ranking of the top 10% features

First, we visually represent the validation data to show the ground truth. We use UMAP to visually represent observations in our dataset across all dimensions reduced to a 2D space and denote whether that observation led to a readmission or no readmission (Figure 12A). We use the same technique to visually represent observations based on the model's predictions and denote which cluster each observation belonged to (Figure 12B). A comparison of these two clusters shows that the clusters based on the model's predictions aligns closely to the clusters based on the actual data. We can see clearly that observations associated with cluster 0 (blue color) are associated with non-readmitted observations (topmost group and right corner of the largest grouping). Cluster number 2, which has a high readmission percentage (98.67%), is highlighted in green (Figure 12B) and also has high ground truth readmissions (Figure 12A). Cluster 3, the rightmost group, has 100% readmission, as shown by the red dots in Figure 12A. Clusters 5 and 7 (Figure 12B) also have a high degree of readmitted observations. Within the largest group, we observe the presence of clusters 1, 6, 4, and 0, each segregating the higher and lower risk groups. However, we do observe this group to have a considerable overlap between the groups and the clusters. These two plots show striking similarities between the predictions made by our model in comparison with the ground truth in the data, lending credibility to the findings reported here.

DISCUSSION

This study examined a diverse array of health-related variables in a large EHR dataset to identify predictors for unplanned 30-day hospital readmission. The data used here capture various aspects of patient risk that are related to medical complexity and heterogeneity using novel indicators, including frailty (26-item FRS representing syndromes and psycho-social risk factors), high-risk medications (10 risk classification groups according to Beers Criteria, 2019), and comorbidity (ECI) in machine learning models. Using an ensemble of machine learning models, we found that the strongest features for risk of readmission were related to health-care utilization (prior readmissions, number of readmissions), length of stay, comorbidity (number of diagnoses, number of unique diagnoses, individual ICD-10 disease codes, ECI). Our results correspond with systematic reviews on risk prediction models for 30-day unplanned readmission that identify prior admissions, LOS, and comorbidities as the most frequently cited predictors related to their ability to classify high-risk patients for adverse outcomes.[55]

The most unexpected finding in this study was that almost all patients who were discharged to a rehabilitation hospital were readmitted within 30 days (844 of 925 patients, 92%); this was the strongest feature associated with readmission. These findings contrast with a study that tracked hospital readmissions for post-acute-care rehabilitation settings by the Medicare Payment Advisory Commission (MedPAC), where the 30-day readmission rate was substantially lower at 12%.[58] Our findings also contrast with a study of 1,365 inpatient rehabilitation facilities providing services to Medicare beneficiaries receiving post-acute hospitalization in which the 30-day readmission rate was 11.8%; however, patients with certain comorbidities had readmission rates as high as 26.3%.[58] This clinical setting, termed inpatient rehabilitation facilities (IRFs), may be freestanding facilities or specialized units within the acute care hospital designed to accommodate patients requiring rehabilitation services for problems such as lower extremity fracture and joint replacement, burns, neurological disorders, and stroke recovery (MedPAC, 2020). Additionally, higher readmission rates may pertain to continued eligibility requirements for inpatient rehabilitation therapy services reimbursed by the Centers for Medicare

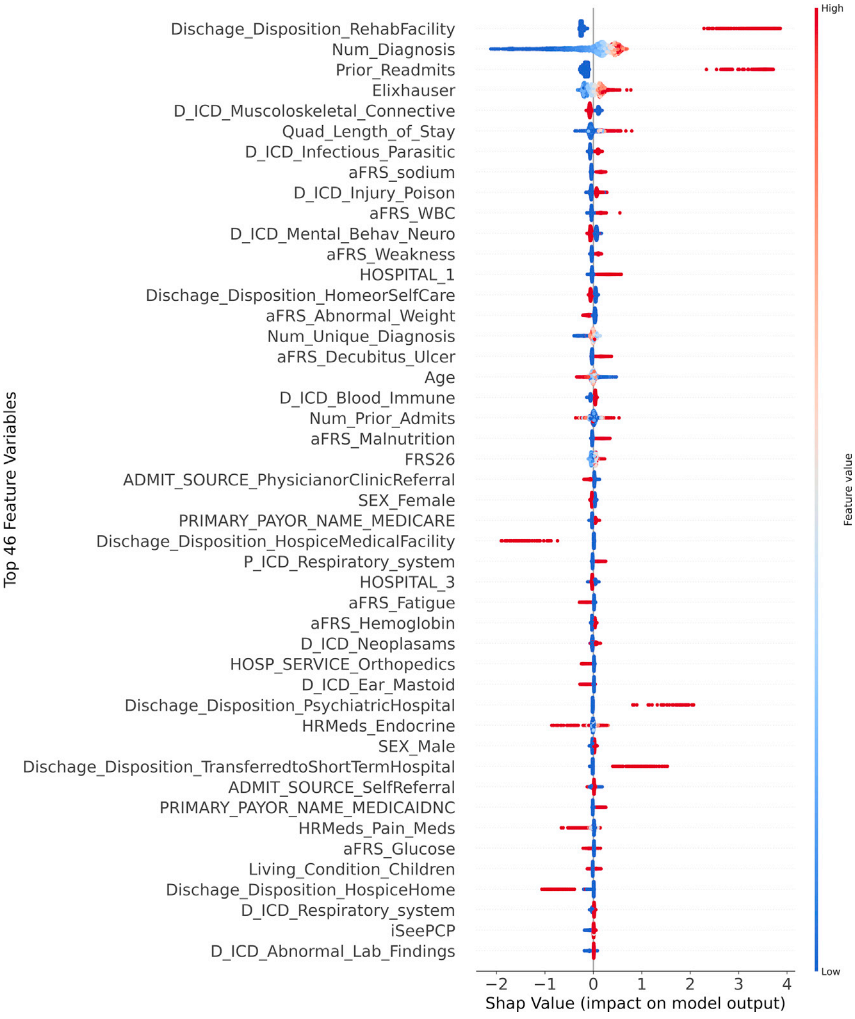


Figure 5. A depiction of model importance (top 10% features), along with a summary of individual impacts of observations for each variable on the target

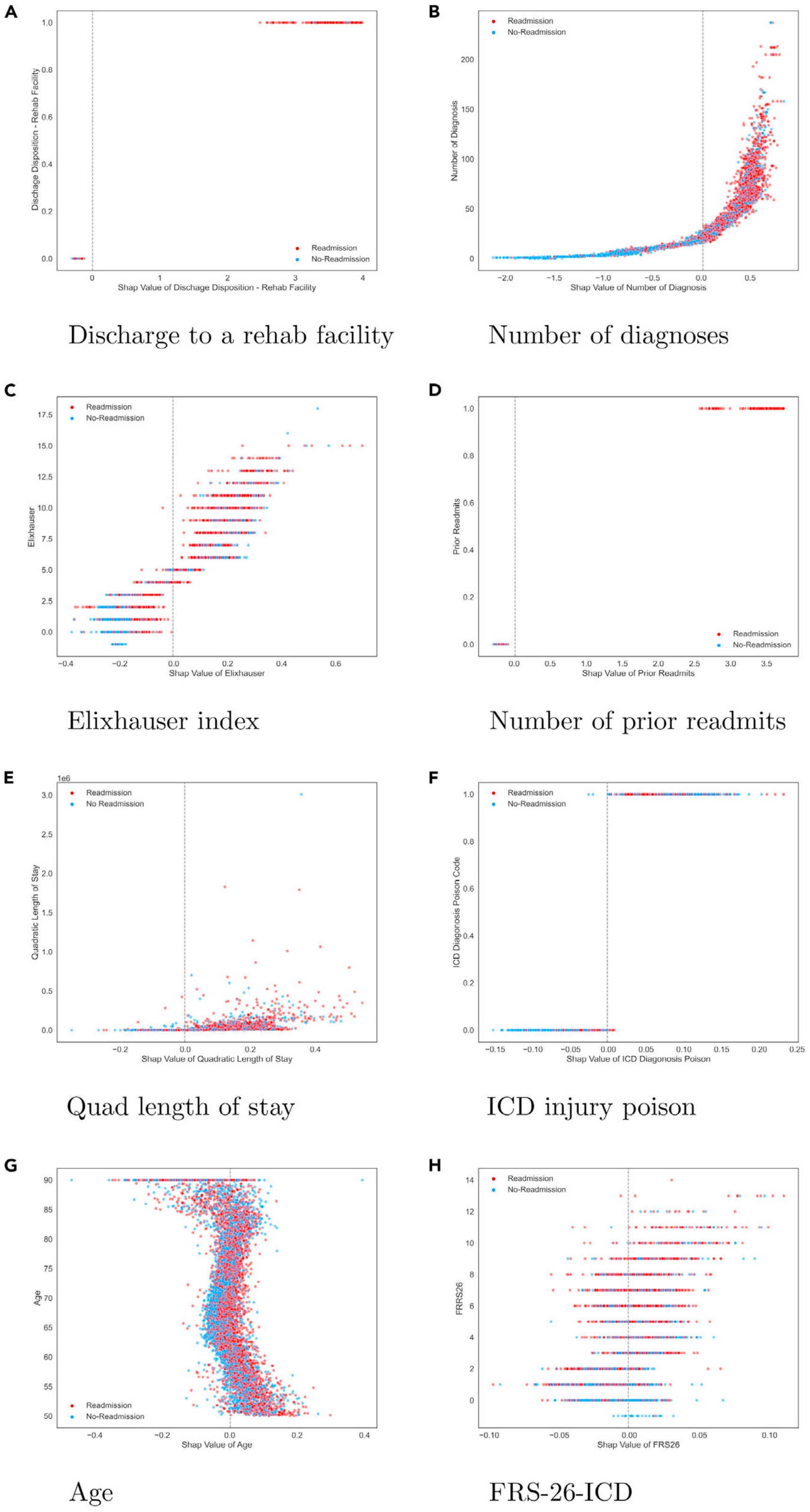
and Medicaid Services (CMS) (CMS, 2017), which stipulate that patients must continue to participate in intensive rehabilitation and demonstrate measurable improvement. The high readmission rate in our study signals the need for further investigation of organizational and population-related factors contributing to readmission risk. Complex care transitions and greater risk for clinical instability and illness exacerbation after hospitalization may contribute to early hospital readmission. Effective transitional care to reduce hospital readmission contains elements of care coordination, communication between providers across clinical settings, and intensive follow-up after hospital discharge[59]; hence, attention to these elements to bolster transitional care processes beginning in the hospital setting is warranted.

Frailty

A novel aspect of this study is inclusion of a proxy measure for frailty. The FRS-26-ICD includes geriatric syndromes and biomarkers that are associated with frailty and manifest across chronic disease conditions to reflect their combined impact on overall health status. We found that 17 of the 26 FRS-26-ICD indicators were strong features, ranking in the top 20%; notably, indicators that correspond to four of the five validated criteria for the physical frailty phenotype are top features: weakness, fatigue, malnutrition, and abnormal weight.[13] These findings underscore the importance of incorporating frailty when trying to predict readmission risk for older inpatients. Interestingly, the FRS-26-ICD composite score is not observed among the strongest features, although it is ranked 50th (top 11%), which suggests a nontrivial influence. This is a promising finding since research notes the uneven quality and low predictive accuracy of frailty instruments applied in prediction models in the acute care setting. We posit that there are differences in the impact of the 26 individual indicator features used to create the FRS-26-ICD score and it is of greater value to incorporate the individual features compared with the composite score. The absence of the composite FRS score as a strong predictor has been echoed by several systematic reviews that highlight the uneven quality and low predictive accuracy of frailty instruments applied in prediction models in the acute care setting.[60]

Several blood biomarkers from the FRS-26-ICD (WBCs, sodium, hemoglobin, glucose) and the ICD-10 codes for abnormal laboratory tests were also strong features, ranking in the top 20%. These results are in agreement with a study of 1,600 internal medicine inpatients in which an FI that assessed up to 27 laboratory tests was independently associated with readmission in comparison with a Clinical Frailty Scale based on the patient's chronic health status.[61] The use of laboratory data in risk models has limitations that hinder widespread application. Trends in healthcare reimbursement and provider practices have generally led to judicious ordering of laboratory tests in contrast to standing orders for blood panels; thus, missing data is a limitation and risks under-estimating frailty or biasing frailty to certain phenotypes. On the other hand, laboratory tests may become routine when these measures are demonstrated to have predictive or clinical value or when technology advances, clinical practice guidelines evolve, or policy changes facilitate or mandate their implementation.[62]

A plethora of tools are available to measure frailty; however, there is only modest overlap across the tools and little agreement on the best tools to use in acute care hospitals.[62] Many tools are complicated to use and require some form of direct clinical assessment, which can be time consuming, require special equipment and training, and be subject to poor inter-rater



Discharge to a rehab facility

Number of diagnoses

Elixhauser index

Number of prior readmits

Quad length of stay

ICD injury poison

Age

FRS-26-ICD

Figure 6. Relationship between select top features and the model's prediction

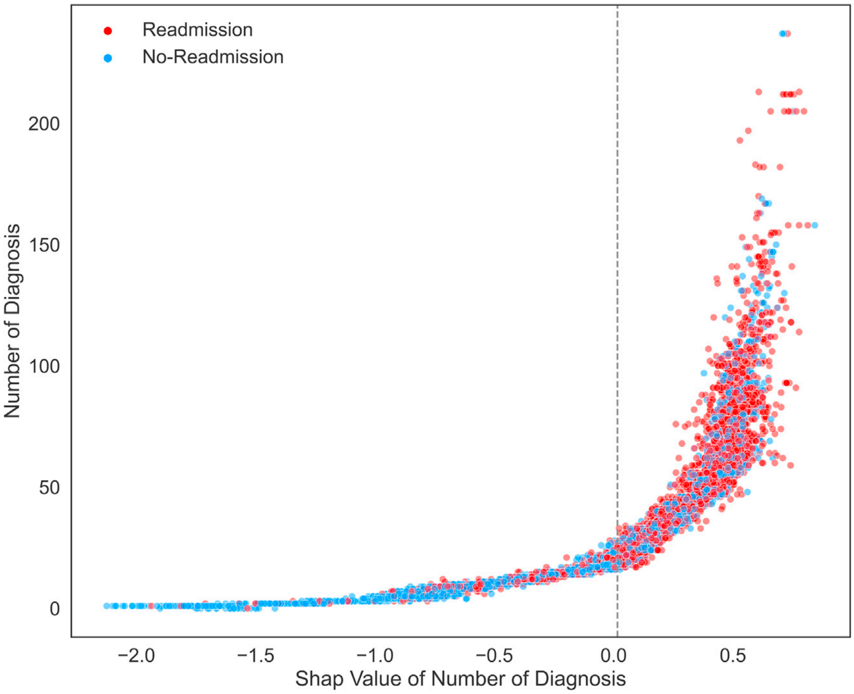


Figure 7. Higher-order interactions between number of diagnoses and FRS-26-ICD

reliability; others have limited clinical relevance to guide care management. These limitations as well as computational advances have spawned initiatives to use existing and readily available EHR data to identify proxy variables for frailty, as demonstrated here with the FRS-26-ICD. Clinicians and researchers can use the FRS-26-ICD in several ways to support care management and identify vulnerable populations for research. As a simple flag, patients who are frail or not frail can be identified. Detailed breakdown of the frailty indicators at the individual, population, and organizational level can identify frail patients that can be directed to care management programs. For example, targeting the FRS-26-ICD indicators weakness, walking difficulty, and fatigue can identify individual or groups of patients who may benefit from physical rehabilitation and/or chronic disease-specific programs (e.g., diabetes, arthritis, disease), or post-hospitalization referrals for transitional care. Effective decision making about clinical interventions can benefit from frailty assessment and consideration of patient and family priorities. At the population level, the FRS can be used to assess the overall health service needs and projected health expenditures.

A large number of studies consider individuals aged at least 65 years for assessment of frailty, in contrast with our study where the target population is aged 50 years and above. International and US-based studies indicate that frailty is not limited or unique to geriatric populations; although frailty prevalence increases with age, population-based studies indicate that frailty can also be detected in younger adults. Since many frailty parameters are already altered in middle age and predictive of adverse events, restricting frailty assessment to older age groups overlooks the needs of this vulnerable population and opportunities for prevention and risk mitigation.[63–65] Prior research using the FRS showed a weak effect size for age and FRS scores.[8,9]

Comorbidity

Our findings confirm the importance of comorbidity in readmission risk models. This finding also highlights the importance of including both frailty and comorbidity for prediction models. Indeed, recent studies indicate that while frailty and multimorbidity are related and overlapping, they are distinctly different constructs; most frail persons are also multimorbid, but fewer persons with multimorbidity also present with frailty.[66] Clinically, this is an important distinction since the aging process is associated with higher prevalence of multiple chronic conditions, geriatric syndromes, and functional impairments. While comorbidity measures of some type are commonly applied in risk prediction models with wide-ranging accuracy (C-statistic reported by 14 studies range from 0.55 to 0.80),[34] our results for the optimal model (CatBoost classifier [AUC = 0.79]) are promising and highlight the importance of comorbidity. A limitation of comorbidity counts and indices is they do not typically take into account individual disease severity or fully adjust for the adverse influence of co-occurring comorbidities on health status; thus, their impact may be attenuated in risk models and may partly explain their modest accuracy in risk prediction models and lack of consistency across studies.[23]

Multimorbidity, defined as the co-occurrence of at least two chronic conditions, increases with age and affects one-fourth of adults in the US.[67] Many chronic conditions also cluster together; however, there can be considerable variability with individuals not neatly fitting into groups that can be targeted for tailored interventions. The presence or count of chronic conditions also does not inform the level of care that may be needed since disease severity and symptom burden are not effectively represented in disease counts or in most comorbidity indices. Symptoms and syndromes included in the FRS provide a mechanism to capture disease impact, although a limitation is the inability to capture symptom severity. Numerous studies have incorporated comorbidity measures such as ECI in risk-adjusted methodologies and predictive models versus individual comorbidities and provides a way to condense comorbidity information into an easy-to-use metric.[68] It is difficult to model interaction between specific comorbidities at a general patient level, which is a motivating factor to develop composite measures such as ECI or CCI.[69] While the ECI and CCI metrics do not capture interactions between categories of comorbidities, they have been used extensively and shown to be strong predictors for readmission and mortality.

High-risk medications and polypharmacy

Prior research indicates that 3%–64% of hospital readmissions are drug related[70]; thus, attention to medication-related issues

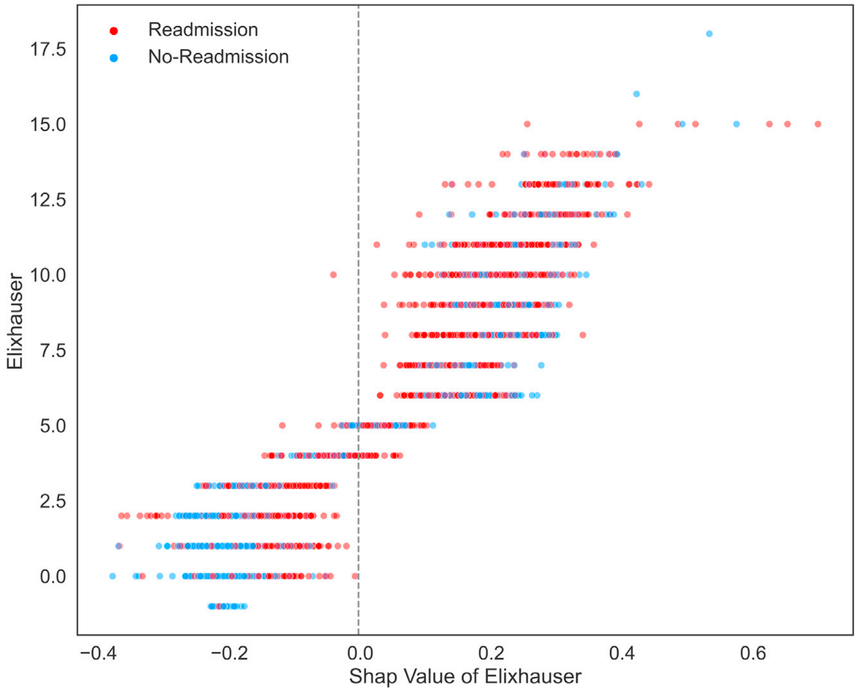


Figure 8. Higher-order interactions between ECI and number of unique diagnoses

such as polypharmacy and high-risk medications could help to improve readmission risk classification. In the present study, only one category of high-risk medication was represented in the top 40 features (endocrine medications; e.g., short-acting insulin); however, six other high-risk medication categories were identified in the top 100, suggesting their important role in readmission. Research on high-risk medications and hospital readmission is divided. In a study examining high-risk medications in hospitalized older adults, exposure to certain high-risk medication classes such as benzodiazepines and opioids was associated with increased odds of readmission.[37] However, in a similar study of hospitalized older adults, although an increased number of medications was significantly associated with unplanned 30-day hospital readmission, there was no significant association between the number of high-risk medications and 30-day re-hospitalization after controlling for covariates.[71] In one retrospective case control study of patients aged >75 years, five medication-related risk factors were associated with hospital readmissions, which included high-risk medications; however, the effects of these risk factors became insignificant in adjusted multivariable models with comorbidity.[72]

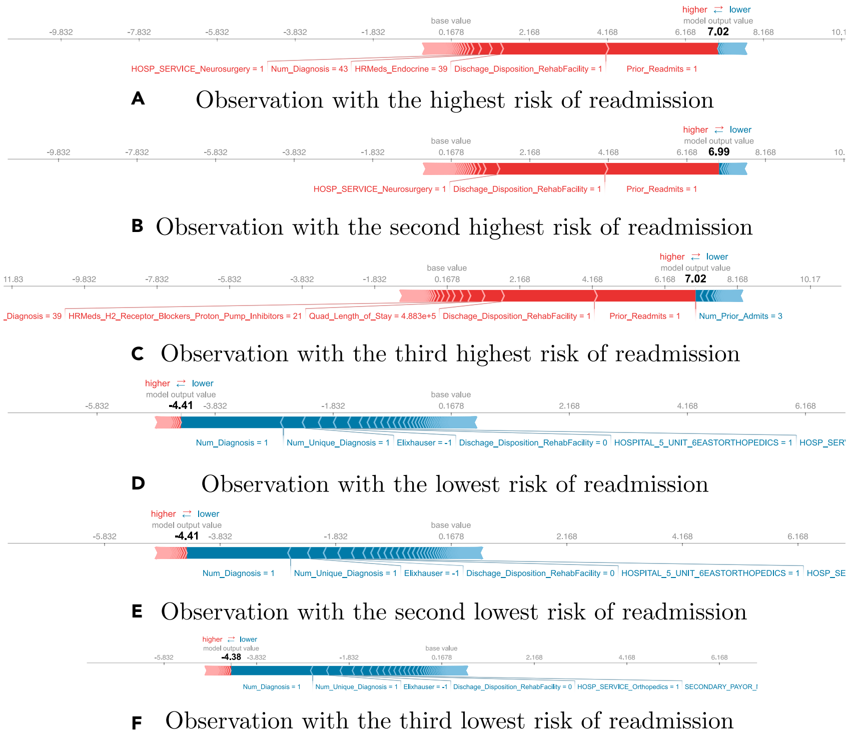
In the present study, polypharmacy was a strong feature ranking 51st (top 11%). Aging is associated with development of a number of chronic health conditions, which often require pharmacologic treatment with multiple medications, leading to polypharmacy.[73] The prevalence of polypharmacy is increasing globally and affects more than half of the older adult population, exposing them to adverse outcomes such as hospital readmission.[74]

Polypharmacy is relevant to readmission because (1) increasing medication use is likely related to disease severity, which is a marker for readmission risk; and (2) increased number of medications is associated with greater risk for medication non-adherence due to cost issues, side effects, and inability to keep track of medication use.[75]

Polypharmacy, especially regimens that include high-risk medications, can potentially cause more harm than benefit to older adults due to factors such as adverse drug reactions and drug-drug and drug-disease interactions.[35,74] Although polypharmacy was an important feature in our study, it is possible that how polypharmacy was defined (patient taking seven or more prescribed medications recorded by the nurse at admission) may have under-estimated its actual prevalence and influence. This definition is based on admission medications and may not have included over-the-counter medications, herbals, and supplements, which can interfere with drug metabolism and contribute to adverse drug interactions and reactions. Picker et al.[75] found that having more than six discharge medications was significantly associated with 30-day hospital readmission in models adjusted for a risk score among 5,507 internal medicine inpatients. In contrast, in a study of adult admissions to a university-affiliated hospital, for patients readmitted within 30 days of discharge the number of discharge medications was not a significant predictor for 30-day readmission.[76]

Research also suggests that the impact of the patient's medication regimen on readmission risk may not be fully captured by polypharmacy and high-risk medication exposure. A feature that should be considered for future research is medication regimen complexity, which can be quantified using a metric that considers the number of medications and at least one other parameter (e.g., dose form; dosing frequency; and special directions for medication use, such as take with food).[77] Refinement of prediction models to include medication regimen complexity in addition to high-risk medications and polypharmacy may improve precision in the identification of high-risk patients for discharge interventions to prevent hospital readmission.

It is plausible that certain disease conditions might have higher risk of readmission, in which case disease-specific models might be more appropriate. For example, surgery, heart failure, and oncology patients have historically had higher readmission rates.[78] However, it is difficult to build disease-specific models for different patient conditions, and even more cost-prohibitive to account for different combinations of comorbid disease conditions. It is important to understand that models developed from disease-specific cohorts will account for characteristics of the underlying cohort but complicate the model development process due to smaller cohort sizes, among other issues. On the other hand, general risk prediction models might



A Observation with the highest risk of readmission

B Observation with the second highest risk of readmission

C Observation with the third highest risk of readmission

D Observation with the lowest risk of readmission

E Observation with the second lowest risk of readmission

F Observation with the third lowest risk of readmission

Figure 9. A spotlight on individual observations

Here, we show the top two observations with the lowest number of readmissions and the highest number of readmissions

be built in a cost-effective manner while taking into account different disease conditions that are encountered in the patient population. A recent study focusing specifically on this question found that accurate prediction of readmissions can be possible through general disease-independent models such as the one we employ in this study.[78] These general models substantially decrease the cost of development, deployment, and maintenance of risk prediction models that can be used in daily clinical routine.

Limitations

Like all clinical prediction efforts, this study has several limitations. One important limitation is the composition of the dataset, which included a diverse patient population of adult hospital admissions 50 years of age and older to a health system that included a level 1 trauma center, community hospitals, and a women's health and behavioral health hospital. Precision of our risk models may vary when applied in more homogeneous datasets; i.e., based on medical or surgical service (e.g., cardiology, orthopedics) or patient characteristics (e.g., medical diagnosis such as heart failure, diabetes). Data used in this study come from a single hospital system (although sourced across five hospitals) in a relatively small geographic region. This factor should be considered when interpreting the results presented here with respect to the ability to generalize to other regions.

Using existing methods to detect frailty is challenging since frailty-related diagnosis codes are subject to under-coding, coding may over-represent frailty due to comorbidities such as dementia, whereas frail patients with other comorbidities such as cardiac conditions or cancer might be grouped with non-frail patients.[17] The FRS-26-ICD used in the present study is at risk for similar issues since coding for ambiguous syndromes such as weakness, fatigue, and dysphagia, which are indicators in the FRS-26-ICD, may not be a priority for the healthcare provider when considering the primary reason for admission when the number of codes possible for data entry is limited. However, in future research, capture of these important patient conditions that are salient in frailty and can be used in modeling frailty will make it possible to represent the patient's health status with higher accuracy. The present study included a diverse set of biopsychosocial risk factors and blood biomarkers associated with frailty that were obtained proximal to admission. Prediction models or early warning systems often include repeated vital signs and laboratory tests to signal clinical deterioration, but it is also important to predict the targeted risk group before a serious change in condition to allow for initiation of treatment. Considering the multifactorial nature of frailty and complex interrelationships among biological processes underlying it that may be complicated by acute illness or surgery in the acute care context, the FRS may provide a broader framework from which to operationalize patient vulnerability in the acute care context.[6] Additionally, the computation of the FRS-26-ICD metric is dependent upon availability of electronically recorded patient data. While the adoption of EHR is gaining momentum, it is not universal, and this lack of data availability is a hindrance to the identification of frailty using these composite metrics.

Finally, reproducing the FRS in other EHR systems and multicenter studies is necessary to evaluate prediction performance and increase the ability to generalize from the study findings.[79,80] Analyzing doctor notes using approaches such as natural language processing to identify additional frailty risk factors may improve the FRS. Future investigations of the FRS in acute care hospitals should also focus on its clinical utility at the individual level for care planning and decision making and at the population health level for program development and transitional care interventions for vulnerable patient subgroups. The challenges associated with the reuse of EHR data are well known and complicated by the massive volume and complexity of the data, unique measures for semantically equivalent concepts with different names, and local customization.[80] Problems due to poor interoperability across proprietary EHR systems have stymied replication of EHR-based research.[81]

The stacking classifier used here to combine predictions from individual models was successful in increasing prediction

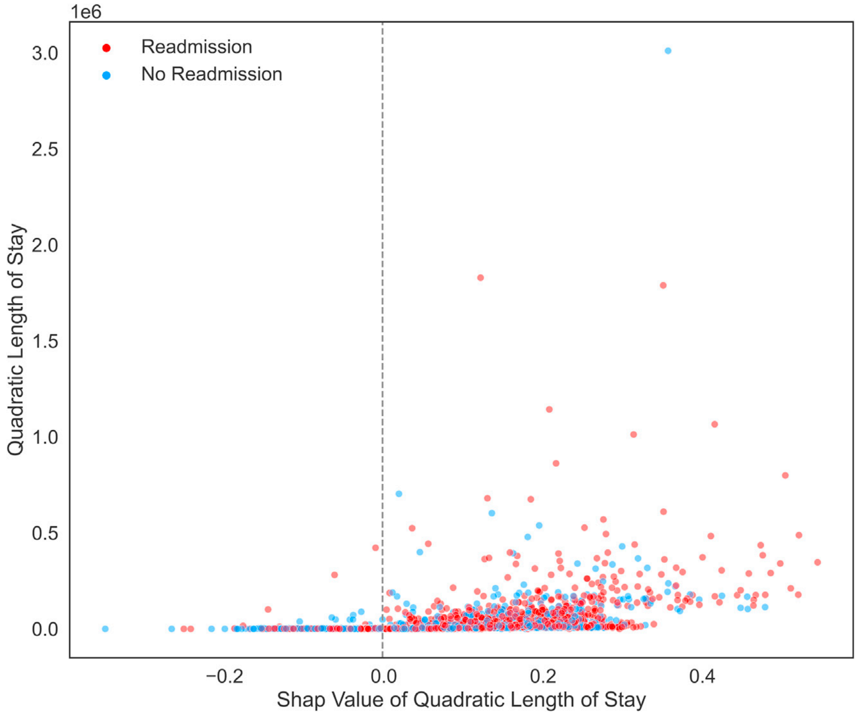


Figure 10. Determining the optimal number of clusters using changes in inertia score

accuracy; however, this increase comes at the cost of explainability. Since the stacking classifier takes predictions from individual models as input, it is agnostic to the features used as inputs for those models. Therefore, within the current study, it was not possible to investigate the model's predictions and trace back the predictions to the impact of individual features. Explainability as demonstrated in this study is infeasible when using a stacking classifier, and we opted to choose explainability over increased accuracy in this work. In future efforts, we plan to explore a dual-stage SHAP scoring mechanism; i.e., the SHAP scores of individual models' features weighted by the meta-learner SHAP score for each model. This can be done by recording the SHAP scores for each base model individually using the test data and then evaluating the stacked model with the same. As such, this will aid in both global and local explainability by enabling understanding of model prediction importance for each observation and the features that lead to the predicted outcome.

Another consideration is that although this paper focuses on patient-related risk factors, organizational and healthcare provider-related factors can also be influential in readmission within 30 days of hospital discharge.[34,82]

EXPERIMENTAL PROCEDURES

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Somya D. Mohanty (sdmohant@uncg.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The EHR data used in this study are bound by data-use agreements between the University of North Carolina at Greensboro (UNCG) and the hospital system it was gathered from, which preclude the authors from sharing the data even when de-identified. The code used in this analysis contains sensitive attribute names from the dataset, which prevents sharing.

Figure 1 shows the overall workflow of our study. The workflow consists of distinct stages where we extract, model, and evaluate the EHR data and develop machine learning models to predict readmission risk.

In the first stage, data mining/modeling, we extract relevant variables from the raw EHR data (from Epic dataset). We pre-process the data, where we analyze missing values, construct encodings, and calculate high-level scores for feature categories (such as Elixhauser and FRS). As the data are recorded within different data tables, we merge the variables based on their encounter identification (unique identifier for each observation) and patient identifier. Following this (in model development), we separate the data into model development (training) and validation datasets, where we also employ different sampling strategies. In the model performance comparison stage, we evaluate multiple models on their performance metrics and also perform hyper-parameter tuning with the available training dataset. A stacking model is also developed, which utilizes predictions of different models to create a super classifier. Following this, we conduct evaluations on model explanation where we observe characteristics of features and their role in model prediction. We conduct several studies to observe patient risk groups, feature to model relationship, feature importance, and observation risk assessment.

In the following we describe each of the components of these stages in further details:

Dataset and exclusion criteria

Data used in this analysis are from a collaborating health system in the southeastern United States. Data are sourced across five hospitals each with a capacity of 85–535 beds and extracted from the hospitals' Epic EHR patient data systems. Data were collected for the time period of 2013–2017, and were filtered to only retain all admissions for adults 50 years of age and older with an inpatient stay lasting longer than 24 h. The data were de-identified in accordance with the data-use agreement between UNCG and the health system.

We applied a set of exclusion criteria (Figure 2) to the data to arrive at the relevant set of patients for this analysis. Starting with the raw dataset of 145,148 observations (76,294 patients), the first step is to remove any observations where the patient's age was less than 50 years. In order to create a dataset for 30-day readmission, in the second step we created a buffer of 30 days at the start and end of earliest and latest admission times in the data. Third, fourth, and fifth steps removed observations that had less than 24 h for length of stay in the hospital, died on initial admission, and had planned readmissions respectively. The sixth step removed any subsequent observations where the patient expired. The final dataset had 128,581 observations recorded from 68,152 patients. Within the dataset, 18,840 observations comprised new encounters, and 109,741 did not (not readmitted or readmitted after 30 days).

Data cleaning and pre-processing

Demography

Demographic data contained information including sex, ethnic group, age, race, marital status, and date of birth. These patient-level variables were recorded at the time of admission and were extracted from the EHR data. For multi-category variables (such as sex, ethnic group, race, and marital status), one-hot encoding

Table 5. Readmission percentage of all observations in a cluster

Cluster number	Readmission percentage	# Of observations
0	4.71	382
1	43.71	2,786
2	98.67	453
3	100	307
4	25.02	1,059
5	100.00	31
6	60.55	2,434
7	17.85	84

Note the clustering is only run on validation data (20% of the original dataset).

was used to create individual variables for each category. Age was calculated from the date of birth and the time of admission, and any values above 90 were set to be 90 in order to preserve privacy of the EHR records.

Apart from the standard demographic variables, the dataset also contained patient living conditions at the time of admission. The variable was codified based on the type of living condition, where the original dataset consisted of 42 categories. Some of these categories were synonymous with each other (e.g. "children" and "child"). Subsequently, we manually identified 13 unique living conditions ("Not Available" [NA]), "Friends," "Other," "Family," "Nursing Facility," "Homeless," "Parent," "Children," "Relatives," "Spouse," "Alone," "Group Home," "Assisted Living") from the 42 categories and encoded into one-hot categories.

Healthcare and insurance utilization

Data about hospital, hospital use, admit source, admission time, discharge disposition, along with insurance payor information were extracted from the administrative records. Admission time was discretized into four groups 6 h apart: morning, afternoon, evening, and late night. Information about whether a patient has a primary care provider and if they see their primary care provider was also included.

Disease ICD

The data contained 12,592 unique ICD-10 codes. These codes were further aggregated to reduce dimensionality based on the first character of the code. For example, the code N40.0 is used to indicate benign prostatic hyperplasia without lower urinary tract symptoms. The N in the code indicates that the disease can be categorized into the class of genitourinary diseases. Each ICD-10 code was grouped into one of 19 high-level classes based on the first character of the code (listed in Table 6) as described by the CDC guidelines.[40] For each observation, the principal and additional diagnoses were mapped to the high-level classification. Observations that did not contain disease diagnoses were marked . The mapping of ICD-10 codes to the corresponding higher-level category was performed to reduce the dimensionality for machine learning and variable analysis. The total number of diagnoses and total number of unique diagnoses for each readmission instance were calculated based on the unique ICD-10 codes before the above steps were applied.

Frailty

Our study utilized a proxy measure for frailty (FRS-26-ICD) drawn from ICD-10-CM disease diagnosis codes that encompass common geriatric syndromes, psycho-social factors, and blood biomarkers. The FRS-26-ICD defines frailty as a clinical syndrome resulting from multi-system physiologic impairments and failed integrative responses with diminished capacity to resist and recover from stressors.[8,19] In addition to the FRS-26-ICD composite score, the 26 individual variables (see row 1, Table 1) that are used to calculate the composite score were incorporated. These variables include, but are not limited to, malnutrition, abnormal weight, fatigue, and difficulty walking. Laboratory values were discretized based on a reference range that indicates risk. Original indicators of frailty compiled from blood biomarkers were represented using laboratory reference ranges for abnormal (high or low) for factors such as albumin, hemoglobin, sodium, and white blood cells (WBCs).

Comorbidity

ECI scores were extracted and calculated for each observation using the ICD diagnosis codes present within the data. ECI consists of 30 comorbidities representing secondary diagnoses that were present on admission and not related to the principal diagnosis.[31] The ECI was computed for 30 unweighted comorbidities using the ICD-10-CM codes according to Quan et al.[83]

High-risk medications

High-risk medication mapping was conducted using two sources: (1) HEDIS, and (2) AGS Beers Criteria. The HEDIS dataset groups high-risk medications into a set of high-level drug classes such as anticholinergics and antipsychotics. The Beers Criteria also offer a similar classification but also include recommendations of use in older adults. First, the data from the HEDIS data were used to create a grouping of high-risk medication into 10 categories. Next, the Beers Criteria were consulted to include any medications/categories that were missing from the HEDIS data. After reconciling both data sources, a categorization that grouped all high-risk medications into 10 high-level categories (see row 3, Table 1) was created. This process was conducted manually by a gerontological nursing expert (D.L.) on our team. Each patient observation was marked containing the high-risk medication category if the prescribed drugs matched any of the criteria.

Specific drug names in the EHR were parsed to remove numerical dosage values, delivery method, type of medication (e.g., capsules, substances), and any other extra information. Next, the drug name was mapped to a generic name before matching to the drug categorizations described above. A string matching algorithm called Sequence Matcher (https://docs.python.org/2.4/lib/sequence-matcher.html) was used to link the drugs to the high-risk medications.

Label calculation

For each observation, a time delta $(\mathcal{T}_d)$ was calculated by taking the difference between the discharge time and the next admit time (or time to 12/1/17 if no readmission) for the same patient. If the $\mathcal{T}_d$ was within 30 days, the observation was marked as *Class_Label* = 1 (readmission), or *Class_Label* = 0 (non-readmission).

Finally, pairwise collinearity tests were conducted for all pairs of features in the data to eliminate highly correlated features. All variables were checked for missing values and missing values were filled with not available (NA). Subsequently, one-hot encoding was performed on all the categorical features.

Model data

The final data used for this analysis contain 458 variables (Table 1) that can be divided into six categories: (1) frailty, (2) comorbidity, (3) high-risk medications, (4) disease diagnosis, (5) demographic, and (6) healthcare and insurance utilization. These features across the above categories were used for the machine learning models.

Machine learning models

After data cleaning and variable extraction, the resulting dataset consists of 128,581 observations with 458 variables. Of these, 18,840 were readmissions (*Class_Label* = 1) and 109,741 non-readmissions (*Class_Label* = 0). These data were split into an 80:20 ratio (train-test), with 80% of the data being used for development of models, while the other 20% were used for validation of the models.

Three different strategies were used to address the imbalance between the two target classes (*Class_Label* = <1, 0>); (1) under-sampling, (2) over-sampling, and (3) no sampling. Under-sampling selects a same-sized random sub-sample from the majority class (*Class_Label* = 0) with respect to the size of the minority class (*Class_Label* = 1). Here, we obtained 18,840 observations each for *Class_Label* = 1 and *Class_Label* = 0 respectively. Over-sampling was done with the synthetic minority over-sampling technique (SMOTE),[84] which oversamples on minority instances by synthesizing new data points between real data instances. The strategy has been used successfully in augmenting low-instance classes for machine learning. We first split the original data to the aforementioned train-test split and then oversampled using only the training data, while keeping the test/validation data for model metrics. Specifically, we split the data into 109,741 training observations and 18,840 testing observations. We then resampled the training observations using SMOTE to have a balanced *Class_Label* = ⟨0 /1⟩ class, which increases the number of training samples to 175,796 to develop the models. In the no-sampling strategy, we used the original sample size of the data for training and testing.

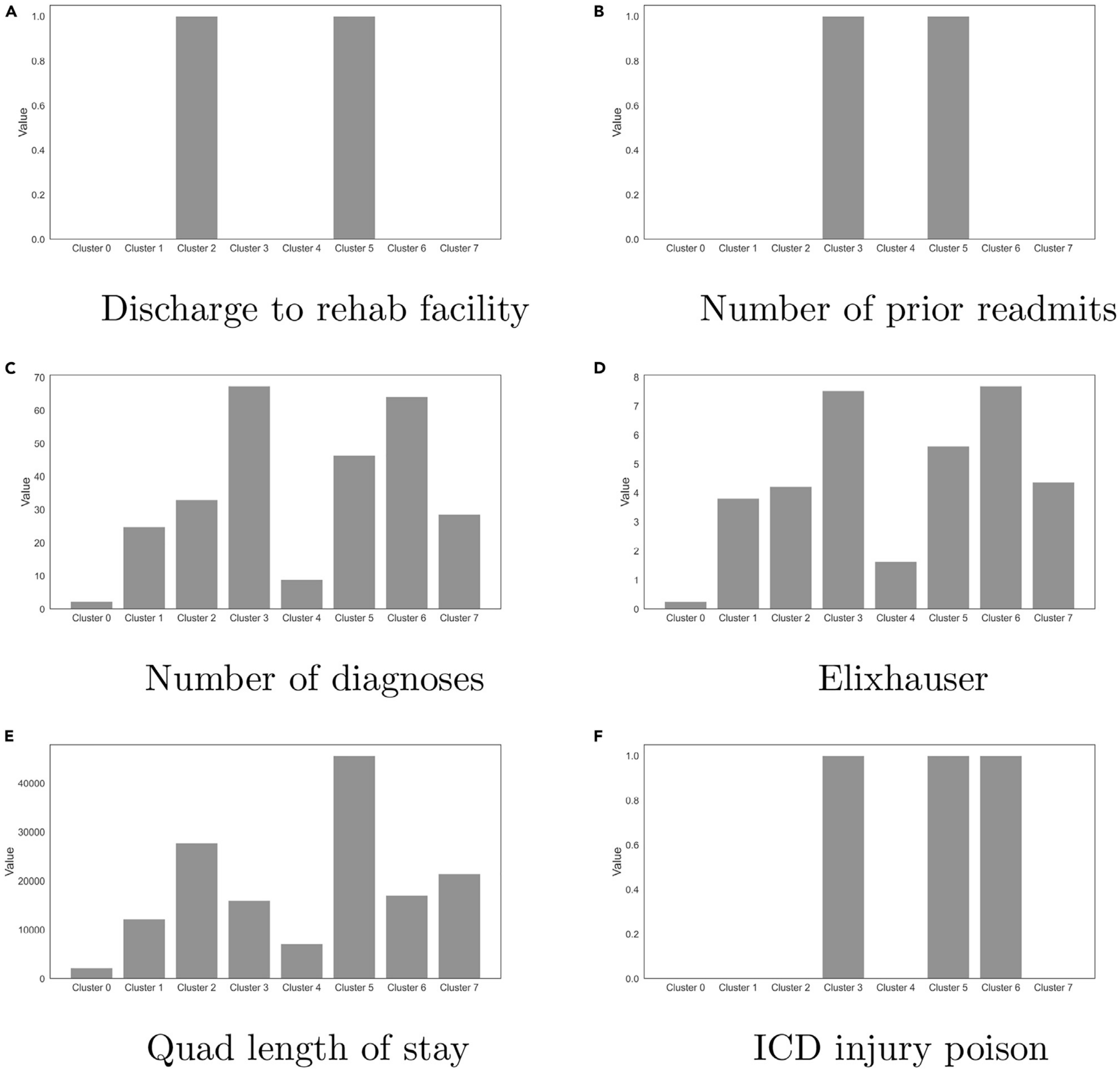


Figure 11. characteristics of different clusters based on data values of salient features

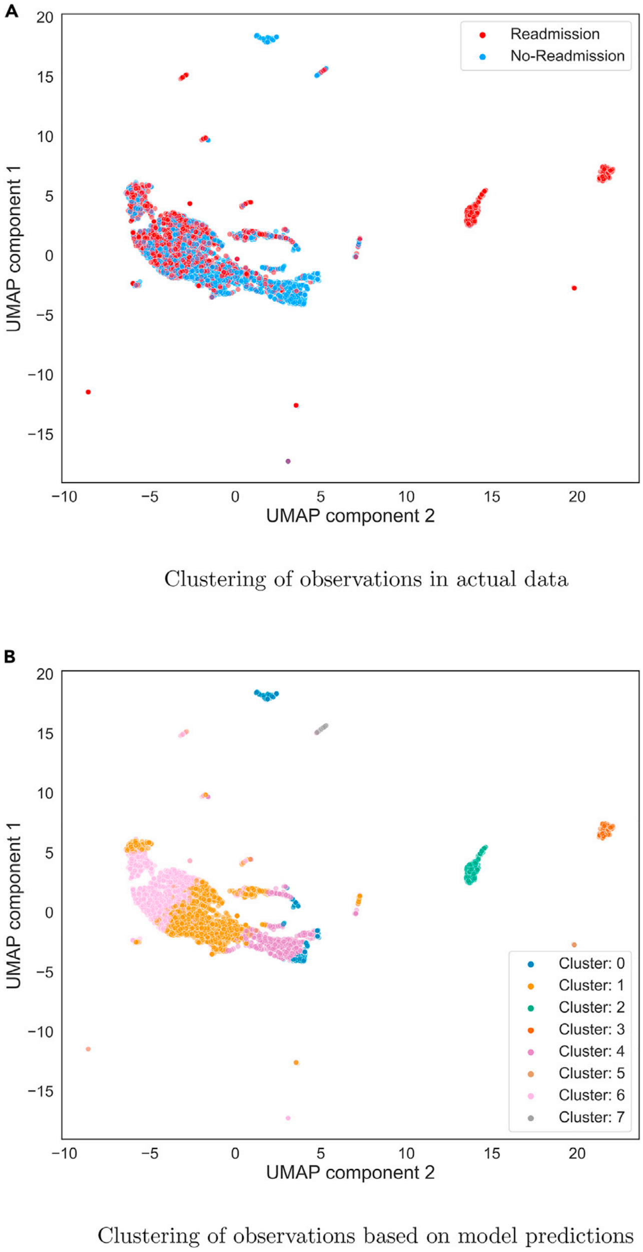
The datasets were fit to five machine learning models and compared for their metrics: (1) logistic regression, (2) random forest, (3) extreme gradient boosting (XGBoost), (4) category boost (CatBoost), and (5) stacking classifier. The choice of the machine learning models is based on baseline comparison with logistic regression, which has been evaluated by a large number of research studies and used in practice in hospital systems.[55] Random forest, XGBoost, and CatBoost are ensemble-based models that reduce variance in the models' learning while being able to provide explainability in their predictions. A stacking classifier was also developed as a super learner that combines the four models (logistic regression, random forest, XGBoost, and CatBoost) into a single model. Deep learning models were avoided in the study due to their inability to succinctly explain model predictions and the overall complexity versus performance benefits being lower than ensemble models. Below we describe the models used in our study in further detail:

Logistic regression

Logistic regression is one of the most widely used models for analyzing EHRs. An extension of linear regression, a basic logistic regression determines the probability of classification problems with a binary outcome.[85] The model uses a logistic function (sigmoid) to fit a linear equation between 0 and 1 using maximum likelihood estimator (MLE). A key reason for utilization of logistic regression in practice is its explainability, where the log-odds of the model can be used to explain the impact of individual features on the model's prediction. However, the model accuracy suffers in high-dimensional non-linear data, where ensemble-based methods outperform it.[86]

Random forest

Random forest is an ensemble prediction method that consists of a set of individual decision trees.[87] The decision trees are designed to have low correlation to each other to encourage diversity among the trees. The prediction of individual trees is aggregated to determine the prediction of the random forest.



Clustering of observations in actual data

Clustering of observations based on model predictions

Figure 12. A comparison of clustering prediction and the ground truth

Random forests use the principles of bootstrapping and aggregating to build trees based on different subsets of the training data using different subsets of features. Since random forest is an ensemble of decision trees, individual errors of the trees are reduced. Additionally, random forests result in good performance on imbalanced datasets while handling missing values well. These models are not substantially affected by outliers in data.

XGBoost

XGBoost is an implementation of gradient boosted decision trees whose main advantages are execution speed and model performance.[88] These models use boosting, an ensemble technique where each tree or model corrects errors made by previous trees. The technique adds more and more trees/models until the overall accuracy cannot be improved anymore. The predictions of the models are added together to make the final prediction. This reduces bias and variance compared with bagging methods such as random forest, as we train subsequent learners on the residuals. XGBoost requires low feature engineering, allowing steps such as normalizations and scaling to be omitted, and outliers have little impact. XGBoost models result in better speed compared with random forests while being more robust to overfitting.

CatBoost

CatBoost[89] is a variation of the boosting techniques typically used for machine learning. CatBoost improves on existing boosting techniques by using ordered boosting and a novel algorithm for effectively processing categorical features. Empirical studies have shown that CatBoost outperforms other publicly available boosting algorithms. CatBoost does not require any special pre-processing for categorical features such as encodings. Instead, the algorithm converts categorical values into numbers using statistics on combinations of categorical features. The algorithm has been shown to be more robust and therefore reduces the need for parameter tuning and optimization. It also reduces the chance of overfitting.

Stacking classifier

We combined predictions from the four models to create a stacked classifier. Model stacking was introduced in 1992 as a way of introducing generalizable prediction models that incorporate other learning models.[90] Model stacking is quite simply the process of combining multiple machine learning models in a sequence so the predictions from each model are formed into a new feature. Each model's predictions get transformed into a new feature, thereby ensuring that each model in the stack predicts a portion of the training data for this new feature. The final dataset obtained from the stack is fed into a final model, called a meta-learner, the purpose of which is to generalize all the features to generate a final prediction. Model stacking has been used to achieve better generalization compared with a single model. Wolpert[90] posits that model stacking deduces the bias in models so that the bias can be corrected in the meta-learner. Here, we used model stacking with the first stage consisting of logistic regression, XGBoost, CatBoost, and random forest classifiers and XGBoost as the meta-learner for the stacking classifier. The choice of the meta-learner was based on cross-evaluation of the stacked model performance between logistic regression, random forest, and XGBoost. The base (first-stage classifiers) and the meta-learner models are developed using the training data only.

Model metrics and hyper-parameter tuning

We evaluated the performance of the four machine learning models as well as the stacked classifier using the following criteria: (1) precision, (2) recall, (3) F-1 score, (4) area under receiver operating characteristics (AUROC) score. Defined as the ratio of true-positives (TPs) to the sum of TPs and the false-positives (FPs) for readmission; i.e., $Precision = TP/TP + FP$, and precision measures the positive predictive power. Similarly, recall or sensitivity is defined as the ratio of TPs to the total of TPs and the false-negatives (FNs); i.e., $Recall = TP/TP + FN$. F1-score combines the precision and recall into a harmonic mean defined as $F1 = 2(Precision * Recall\ /Precision\ + Recall)$. A receiver operating characteristics (ROC) curve describes the tradeoff between the TP rate (TPR)/recall and FP rate (FPR, where $FPR = FP/FP + TN$, and TN are true-negatives), across the different decision thresholds of a model. AUROC measures the AUROC curve to provide a score for the models. For each of the models, we record the aforementioned metrics across a k-fold (k = 5) validation and utilize mean metrics to compare between the models. Specifically, we perform five iterations of train-test split, sampling, and model development to record and present the metrics.

Hyper-parameter tuning was conducted on each of the ensemble models using a grid search approach. We evaluated the models on different number of estimators (500, 1,000, 5,000), maximum tree depth (3 … 10), and learning rate (0.02 … 0.05, 0.1, 0.5). The evaluation was done with the validation dataset and the best-performing model results are presented.

Model explanation

Explainability of artificial intelligence (AI) models for decision making and predictions is one of the most heavily debated topics, especially when it is applied to healthcare.[91,92] Amann et al.[91] argue that explainability of AI models invokes legal, ethical, and societal questions and deserves thorough investigation. The vast majority of prior studies in the area of health

Table 6. Categorization of ICD-10 codes into higher-level categories based on the first character of the code

First character of ICD code	Higher-level category
A, B	infectious parasitic
C	neoplasms (C–D50)
D	blood immune (above D50)
E	endocrine nutritional metabolic
F	mental behavior neuro
G	nervous system
H1	eye adnexa
H2	ear mastoid
I	circulatory system
J	respiratory system
K	digestive system
L	skin subcutaneous
M	musculoskeletal connective
N	genitourinary
N	penitourinary
O	pregnancy child puerperium
P	Perinatal
Q	congenital malformation deformations
R	abnormal laboratory findings
S, T	injury poison
V, W, X, Y	external morbidity
Z	health status

All codes containing the character in the left column are replaced with the category in the right column.

informatics present findings/predictions from machine learning but make no attempt to provide explainability to the model. Treating these models as black boxes greatly reduces confidence in their predictions no matter how high the accuracy measures are.

In this study, we provide extensive explanations and interpretability for the model's behavior and predictions using Shapley additive explanation (SHAP).[93] SHAP aims to provide explanations for models by evaluating the contribution of each feature to the predictions. It is able to do so by calculating the Shapley values[94] based on a game theoretic approach of feature coalitions. In the case of machine learning models, each feature (or a group of features) acts as a player in a cooperative game, where we calculate the marginal contributions that affect the overall prediction. In other words, we evaluate the contribution of each feature by observing how different the prediction of the model is from the expected prediction by observing the model predictions with coalitions of feature variables that include the target feature variable and ones that do not.

SHAP provides in-depth model explainability and can allow readers/healthcare experts to explore the impact of individual predictors within an ensemble-based machine learning model. We use a variant of SHAP called TreeSHAP[93] that was developed for specifically for tree-based machine learning models. TreeSHAP presents numerical scores called SHAP values that explain the prediction for an observation by quantifying the contribution of each feature toward readmission prediction. While standard feature importance graphs answer the "what" part of the prediction, SHAP answers the "why" by providing explanations that give the user an insight into why the model makes certain predictions, thereby increasing model transparency.

Below, we summarize the three benefits of SHAP:

1.Global interpretability: collectively, SHAP values show how much each predictor contributes to the outcome variable. In addition to this, the values indicate whether a predictor affects the outcome positively or negatively.
2.Local interpretability: each observation in the dataset can be explored using its own set of SHAP values. Therefore, it is easy to see the predictors that affected the prediction for that specific observation. These SHAP values can be easily visualized using individual plots for every observation.
3.Tree-based explainability: SHAP values can be computed for any tree-based model, including ensemble models, unlike other explainability methods that use regression models as a surrogate.

In this study, we utilize SHAP values for a wide range of analyses of the machine learning models employed here. First, we provide a global interpretation of the model's prediction by exploring important features that contributed to the model's prediction ranked by their SHAP values. We also explain the absolute values of important features in the data and show whether those values were associated with a higher or lower prediction value of the target variable. All of these insights are efficiently displayed using a single visualization called a SHAP plot.

Second, we explore global interpretability of the model, where we highlight relationships between a set of manually selected features using a SHAP partial dependence plot. These visualizations show the marginal effect of one or two features have on the prediction while also showing the relationship between features and a feature and the target variable. Third, we study local interpretability of SHAP values, where we highlight how the prediction was made for a few selected observations using an individual SHAP value plot. These plots pinpoint the features that played a critical role in the final prediction for that specific observation. The plots also display the effect (positive or negative) of each of those features on the prediction. These visualizations allow healthcare experts to delve deep into the data and analyze the model's decision-making process for specific observations.

Clustering and risk group analysis

An important criterion for hospital systems to understand the factors associated with readmission in their patient populations is to observe any latent patient groups that emerge from the data. To address this, we conduct risk group analysis on the SHAP data generated from the developed machine learning models.

After training the models and evaluating the best-performing one, we calculate the SHAP values for the validation observations. We then utilized an unsupervised machine learning approach, K-means clustering,[95] to group the observations into different patient groups. K-means is a distance-based clustering approach that has a goal to partition the given observations into K clusters. Each observation is associated with a cluster based on its Euclidean distance from the cluster centroid. The number of groups/clusters (K) was evaluated using the elbow approach,[96] where we observed the model perplexity across the patient groups of $K = 2 \cdots 10$, and choose $K = 6$ to be the number of groups to evaluate.

For each observation we then annotate the cluster number and study the statistical properties of the cluster observations. First, we calculate the distinguishing features for each cluster. This is done by measuring the relative score of each feature for its cluster by calculating relative feature score as $C_x = (C_x^i - \sum_{i=1}^{n} C_x^i - C_x^i / n - 1)^2$, where C_x is the cluster centroid value of the feature variable i and n = 458 for the total number of available variables. We then sort the C_x value for each cluster to observe the features that uniquely identify a particular cluster. These features are used to conduct comparisons between different patient risk groups to study which feature variables contribute toward an observation being placed in a particular group.

In order to visually validate the findings of the clustering, we also utilize the Uniform Manifold Approximation and Projection (UMAP)[97] dimensionality reduction technique to map the 458 dimensional vectors to two-dimensional (2D) space. UMAP is a non-linear manifold learning approach that is great at preserving both the local and global structure of high-dimensional data when transformed to lower-dimensional space. We utilize UMAP to transform the training SHAP data to a 2D space and plot cluster results for a visual comparison of the ground truth of readmitted and non-readmitted observations versus clusters of the model's predictions.

ACKNOWLEDGMENTS

Financial support was received from the UNCG Internal Faculty Research Grant and the UNCG Faculty First Summer Scholarship Grant, Office of the

Vice Chancellor for Research and Engagement, and Sigma Theta Tau International, Gamma Zeta Chapter (D.L.).

AUTHOR CONTRIBUTIONS

M.J. and D.L. participated in data compilation and retrieval. S.D., P.M., and T.P.M. conceived the experiments. S.D. and T.P.M. performed the analyses. P.M. wrote the manuscript and participated in data analysis. All authors have read and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 7, 2021
Revised: September 29, 2021
Accepted: November 2, 2021
Published: December 3, 2021

REFERENCES

1. Tong, L., Erdmann, C., Daldalian, M., Li, J., and Esposito, T. (2016). Comparison of predictive modeling approaches for 30-day all-cause non-elective readmission risk. BMC Med. Res. Methodol. *16*, 26.
2. Hao, S., Wang, Y., Jin, B., Shin, A.Y., Zhu, C., Huang, M., Zheng, L., Luo, J., Hu, Z., Fu, C., et al. (2015). Development, validation and deployment of a real time 30 day hospital readmission risk assessment tool in the Maine Healthcare Information Exchange. PLoS One *10*. https://doi.org/10.1371/journal.pone.0140271.
3. McIlvennan, C.K., Eapen, Z.J., and Allen, L.A. (2015). Hospital readmissions reduction program. Circulation *131*, 1796–1803.
4. Association, A.M., et al. (1990). Council on scientific affairs: American Medical Association white paper on elderly health. Arch. Int. Med. *150*, 2459–2472.
5. Gijsen, R., Hoeymans, N., Schellevis, F.G., Ruwaard, D., Satariano, W.A., and van den Bos, G.A. (2001). Causes and consequences of comorbidity: a review. J. Clin. Epidemiol. *54*, 661–674.
6. Rodriguez-Mañas, L., de Carvalho, I.A., Bhasin, S., Bischoff-Ferrari, H., Cesari, M., Evans, W., Hare, J., Pahor, M., Parini, A., Rolland, Y., et al. (2020). ICFSR task force perspective on biomarkers for sarcopenia and frailty. J. Frailty Aging *9*, 4–8.
7. Zaslavsky, O., Cochrane, B.B., Thompson, H.J., Woods, N.F., Herting, J.R., and LaCroix, A. (2013). Frailty: a review of the first decade of research. Biol. Res. Nurs. *15*, 422–432.
8. Lekan, D.A., Wallace, D.C., McCoy, T.P., Hu, J., Silva, S.G., and Whitson, H.E. (2017). Frailty assessment in hospitalized older adults using the electronic health record. Biol. Res. Nurs. *19*, 213–228.
9. Lekan, D.A., McCoy, T.P., Jenkins, M., Mohanty, S., Manda, P., and Yasin, R. (2021). Comparison of a frailty risk score and comorbidity indices for hospital readmission using electronic health record data. Res. Gerontological Nurs. *14*, 91–103.
10. Basic, D., and Shanley, C. (2015). Frailty in an older inpatient population: using the Clinical Frailty Scale to predict patient outcomes. J. Aging Health *27*, 670–685.
11. Forti, P., Maioli, F., Zagni, E., Lucassenn, T., Montanari, L., Maltoni, B., Pirazzoli, G.L., Bianchi, G., and Zoli, M. (2014). The physical phenotype of frailty for risk stratification of older medical inpatients. J. Nutr. Health Aging *18*, 912–918.
12. Wou, F., Gladman, J.R., Bradshaw, L., Franklin, M., Edmans, J., and Conroy, S.P. (2013). The predictive properties of frailty-rating scales in the acute medical unit. Age Ageing *42*, 776–781.
13. Fried, L.P., Tangen, C.M., Walston, J., Newman, A.B., Hirsch, C., Gottdiener, J., Seeman, T., Tracy, R., Kop, W.J., Burke, G., et al. (2001). Frailty in older adults: evidence for a phenotype. J. Gerontol. Ser. A Biol. Sci. Med. Sci. *56*, M146–M157.
14. Searle, S.D., Mitnitski, A., Gahbauer, E.A., Gill, T.M., and Rockwood, K. (2008). A standard procedure for creating a frailty index. BMC Geriatr. *8*, 24.
15. Gilbert, T., Neuburger, J., Kraindler, J., Keeble, E., Smith, P., Ariti, C., Arora, S., Street, A., Parker, S., Roberts, H.C., et al. (2018). Development and validation of a hospital frailty risk score focusing on older people in acute care settings using electronic hospital records: an observational study. Lancet *391*, 1775–1782.
16. van Walraven, C., McAlister, F.A., Bakal, J.A., Hawken, S., and Donzé, J. (2015). External validation of the hospital-patient one-year mortality risk (HOMR) model for predicting death within 1 year after hospital admission. CMAJ *187*, 725–733.
17. Shi, S.M., and Kim, D.H. (2018). The challenges of using the hospital frailty risk score. Lancet *392*, 2692.
18. Mohanty, S.D., McCoy, T.P., Manda, P., Lekan, D., and Jenkins, M. (2020). A multi-modal machine learning approach towards predicting patient readmission. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE), pp. 2027–2035.
19. Rodríguez-Mañas, L., and Sinclair, A. (2014). Frailty: the quest for new domains, clinical definitions and subtypes. is this justified on new evidence emerging? J. Nutr. Health Aging *18*, 92.
20. Kim, D.H. (2020). Measuring frailty in health care databases for clinical care and research. Ann. Geriatr. Med. Res. *24*, 62.
21. De Groot, V., Beckerman, H., Lankhorst, G.J., and Bouter, L.M. (2003). How to measure comorbidity: a critical review of available methods. J. Clin. Epidemiol. *56*, 221–229.
22. Zhao, P., and Yoo, I. (2017). A systematic review of highly generalizable risk factors for unplanned 30-day all-cause hospital readmissions. J. Heal. Med. Inform. *8*. https://doi.org/10.4172/2157-7420.1000283.
23. Yurkovich, M., Avina-Zubieta, J.A., Thomas, J., Gorenchtein, M., and Lacaille, D. (2015). A systematic review identifies valid comorbidity indices derived from administrative health data. J. Clin. Epidemiol. *68*, 3–14.
24. Wang, P., Wang, Q., Li, F., Bian, M., and Yang, K. (2019). Relationship between potentially inappropriate medications and the risk of hospital readmission and death in hospitalized older patients. Clin. Interventions Aging *14*, 1871.
25. Chirapongsathorn, S., Poovorawan, K., Soonthornworasiri, N., Pan-ngum, W., Phaosawasdi, K., and Treeprasertsuk, S. (2020). Thirty-day readmission and cost analysis in patients with cirrhosis: a nationwide population-based data. Hepatol. Commun. *4*, 453–460.
26. Lorei, T.W., and Gurel, L. (1973). Demographic characteristics as predictors of posthospital employment and readmission. J. Consulting Clin. Psychol. *40*, 426.
27. Munley, P.H., Devone, N., Einhorn, C.M., Gash, I.A., Hyer, L., and Kuhn, K.C. (1977). Demographic and clinical characteristics as predictors of length of hospitalization and readmission. J. Clin. Psychol. *33*, 1093–1099.
28. Tabak, Y.P., Sun, X., Nunez, C.M., Gupta, V., and Johannes, R.S. (2017). Predicting readmission at early hospitalization using electronic clinical data: an early readmission risk score. Med. Care *55*, 267.
29. Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., et al. (2018). Scalable and accurate deep learning with electronic health records. NPJ Digital Med. *1*, 18.
30. Charlson, M.E., Pompei, P., Ales, K.L., and MacKenzie, C.R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J. Chronic Dis. *40*, 373–383.
31. Elixhauser, A., Steiner, C., Harris, D.R., and Coffey, R.M. (1998). Comorbidity measures for use with administrative data. Med. Care, 8–27.
32. Austin, S.R., Wong, Y.-N., Uzzo, R.G., Beck, J.R., and Egleston, B.L. (2015). Why summary comorbidity measures such as the Charlson Comorbidity Index and Elixhauser Score work. Med. Care *53*, e65.
33. Menendez, M.E., Neuhaus, V., van Dijk, C.N., and Ring, D. (2014). The Elixhauser comorbidity method outperforms the Charlson index in predicting inpatient death after orthopaedic surgery. Clin. Orthopaedics Relat. Res. *472*, 2878–2886.

34. Zhang, X., Zhou, S., Pan, K., Li, X., Zhao, X., Zhou, Y., Cui, Y., and Liu, X. (2017). Potentially inappropriate medications in hospitalized older patients: a cross-sectional study using the Beers 2015 criteria versus the 2012 criteria. Clin. Interventions Aging *12*, 1697.

35. American Geriatrics Society Beers Criteria® Update Expert Panel, Fick, D.M., Semla, T.P., Steinman, M., Beizer, J., Brandt, N., Dombrowski, R., DuBeau, C.E., Pezzullo, L., Epplin, J.J., et al. (2019). American Geriatrics Society 2019 updated AGS Beers Criteria® for potentially inappropriate medication use in older adults. J. Am. Geriatr. Soc. *67*, 674–694.

36. Fick, D.M., Cooper, J.W., Wade, W.E., Waller, J.L., Maclean, J.R., and Beers, M.H. (2003). Updating the Beers Criteria for potentially inappropriate medication use in older adults: results of a US consensus panel of experts. Arch. Intern. Med. *163*, 2716–2724.

37. Pavon, J.M., Zhao, Y., McConnell, E., and Hastings, S.N. (2014). Identifying risk of readmission in hospitalized elderly adults through inpatient medication exposure. J. Am. Geriatr. Soc. *62*, 1116–1121.

38. Blachman, N.L., Leipzig, R.M., Mazumdar, M., and Poeran, J. (2017). High-risk medications in hospitalized elderly adults: are we making it easy to do the wrong thing? J. Am. Geriatr. Soc. *65*, 603–607.

39. Allaudeen, N., Vidyarthi, A., Maselli, J., and Auerbach, A. (2011). Redefining readmission risk factors for general medicine patients. J. Hosp. Med. *6*, 54–60.

40. The Centers for Medicare and Medicaid Services, et al., ICD-10-CM Official Guidelines for Coding and Reporting (2012).

41. Lee, E.W. (2012). Selecting the best prediction model for readmission. J. Prev. Med. Public Health *45*, 259.

42. Futoma, J., Morris, J., and Lucas, J. (2015). A comparison of models for predicting early hospital readmissions. J. Biomed. Inform. *56*, 229–238.

43. Krumholz, H.M., Chaudhry, S.I., Spertus, J.A., Mattera, J.A., Hodshon, B., and Herrin, J. (2016). Do non-clinical factors improve prediction of readmission risk? Results from the tele-HF study. JACC Heart Fail. *4*, 12–20.

44. Silverstein, M.D., Qin, H., Mercer, S.Q., Fong, J., and Haydar, Z. (2008). Risk factors for 30-day hospital readmission in patients ≥ 65 years of age. Baylor University Medical Center Proceedings, *Vol. 21* (Taylor & Francis), pp. 363–372.

45. Hasan, O., Meltzer, D.O., Shaykevich, S.A., Bell, C.M., Kaboli, P.J., Auerbach, A.D., Wetterneck, T.B., Arora, V.M., Zhang, J., and Schnipper, J.L. (2010). Hospital readmission in general medicine patients: a prediction model. J. Gen. Intern. Med. *25*, 211–219.

46. Mahajan, S., and Ghani, R. (2019). Using ensemble machine learning methods for predicting risk of readmission for heart failure. Stud. Health Technol. Inform. *264*, 243–247.

47. Baig, M.M., Hua, N., Zhang, E., Robinson, R., Spyker, A., Armstrong, D., Whittaker, R., Robinson, T., and Ullah, E. (2020). A machine learning model for predicting risk of hospital readmission within 30 days of discharge: validated with LACE index and patient at risk of hospital readmission (PARR) model. Med. Biol. Eng. Comput. 1–8.

48. Goto, T., Jo, T., Matsui, H., Fushimi, K., Hayashi, H., and Yasunaga, H. (2019). Machine learning-based prediction models for 30-day readmission after hospitalization for chronic obstructive pulmonary disease. COPD J. Chronic Obstructive Pulm. Dis. *16*, 338–343.

49. Okere, A.N., Sanogo, V., Alqhtani, H., and Diaby, V. (2020). Identification of risk factors of 30-day readmission and 180-day in-hospital mortality, and its corresponding relative importance in patients with ischemic heart disease: a machine learning approach. Expert Rev. Pharmacoecon. Outcomes Res. 1–6.

50. Grana, M., Lopez-Guede, J.M., Irazusta, J., Labayen, I., and Besga, A. (2020). Modelling hospital readmissions under frailty conditions for healthy aging. Expert Syst. *37*, e12437.

51. Glans, M., Ekstam, A.K., Jakobsson, U., Bondesson, Å., and Midlöv, P. (2020). Risk factors for hospital readmission in older adults within 30 days of discharge–a comparative retrospective study. BMC Geriatr. *20*, 1–12.

52. Pedersen, M.K., Nielsen, G.L., Uhrenfeldt, L., and Lundbye-Christensen, S. (2019). Risk assessment of acute, all-cause 30-day readmission in patients aged 65+: a nationwide, register-based cohort study. J. Gen. Intern. Med. *34*, 226–234.

53. Engelhardt, K.E., Reuter, Q., Liu, J., Bean, J.F., Barnum, J., Shapiro, M.B., Ambre, A., Dunbar, A., Markzon, M., Reddy, T.N., et al. (2018). Frailty screening and a frailty pathway decrease length of stay, loss of independence, and 30-day readmission rates in frail geriatric trauma and emergency general surgery patients. J. Trauma Acute Care Surg. *85*, 167–173.

54. Borkenhagen, L.S., McCoy, R.G., Havyer, R.D., Peterson, S.M., Naessens, J.M., and Takahashi, P.Y. (2018). Symptoms reported by frail elderly adults independently predict 30-day hospital readmission or emergency department care. J. Am. Geriatr. Soc. *66*, 321–326.

55. Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., and Kripalani, S. (2011). Risk prediction models for hospital readmission: a systematic review. JAMA *306*, 1688–1698.

56. Zhou, H., Della, P.R., Roberts, P., Goh, L., and Dhaliwal, S.S. (2016). Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. BMJ Open *6*, e011060.

57. Beecy, A.N., Gummalla, M., Sholle, E., Xu, Z., Zhang, Y., Michalak, K., Dolan, K., Hussain, Y., Lee, B.C., Zhang, Y., et al. (2020). Utilizing electronic health data and machine learning for the prediction of 30-day unplanned readmission or all-cause mortality in heart failure. Cardiovasc. Digital Health J. *1*, 71–79.

58. Ottenbacher, K.J., Karmarkar, A., Graham, J.E., Kuo, Y.-F., Deutsch, A., Reistetter, T.A., Al Snih, S., and Granger, C.V. (2014). Thirty-day hospital readmission following discharge from postacute rehabilitation in fee-for-service Medicare patients. JAMA *311*, 604–614.

59. Verweij, L., van de Korput, E., Daams, J.G., Ter Riet, G., Peters, R.J., Engelbert, R.H., Scholte op Reimer, W.J., and Buurman, B.M. (2019). Effects of Postacute Multidisciplinary Rehabilitation Including Exercise in Out-Of-Hospital Settings in the Aged, PhD thesis (Hogeschool van Amsterdam).

60. Theou, O., Squires, E., Mallery, K., Lee, J.S., Fay, S., Goldstein, J., Armstrong, J.J., and Rockwood, K. (2018). What do we know about frailty in the acute care setting? A scoping review. BMC Geriatr. *18*, 139.

61. Ellis, H.L., Wan, B., Yeung, M., Rather, A., Mannan, I., Bond, C., Harvey, C., Raja, N., Dutey-Magni, P., Rockwood, K., et al. (2020). Complementing chronic frailty assessment at hospital admission with an electronic frailty index (FI-laboratory) comprising routine blood test results. CMAJ *192*, E3–E8.

62. Buta, B.J., Walston, J.D., Godino, J.G., Park, M., Kalyani, R.R., Xue, Q.-L., Bandeen-Roche, K., and Varadhan, R. (2016). Frailty assessment instruments: systematic characterization of the uses and contexts of highly-cited instruments. Ageing Res. Rev. *26*, 53–61.

63. Santos-Eggimann, B., Cuénoud, P., Spagnoli, J., and Junod, J. (2009). Prevalence of frailty in middle-aged and older community-dwelling Europeans living in 10 countries. J. Gerontol. Ser. A *64*, 675–681.

64. Bandeen-Roche, K., Seplaki, C.L., Huang, J., Buta, B., Kalyani, R.R., Varadhan, R., Xue, Q.-L., Walston, J.D., and Kasper, J.D. (2015). Frailty in older adults: a nationally representative profile in the United States. J. Gerontol. Ser. A *70*, 1427–1434.

65. Segaux, L., Broussier, A., Oubaya, N., Leissing-Desprez, C., Laurent, M., Naga, H., Fromentin, I., David, J.-P., and Bastuji-Garin, S. (2021). Several frailty parameters highly prevalent in middle age (50–65) are independent predictors of adverse events. Sci. Rep. *11*, 1–10.

66. Vetrano, D.L., Palmer, K., Marengoni, A., Marzetti, E., Lattanzio, F., Roller-Wirnsberger, R., Lopez Samaniego, L., Rodríguez-Mañas, L., Bernabei, R., Onder, G., et al. (2019). Frailty and multimorbidity: a systematic review and meta-analysis. J. Gerontol. Ser. A *74*, 659–666.

67. Whitson, H.E., Johnson, K.S., Sloane, R., Cigolle, C.T., Pieper, C.F., Landerman, L., and Hastings, S.N. (2016). Identifying patterns of multimorbidity in older Americans: application of latent class analysis. J. Am. Geriatr. Soc. *64*, 1668–1673.

68. Moore, B.J., White, S., Washington, R., Coenen, N., and Elixhauser, A. (2017). Identifying increased risk of readmission and in-hospital mortality using hospital administrative data. Med. Care *55*, 698–705.

69. Thompson, N.R., Fan, Y., Dalton, J.E., Jehi, L., Rosenbaum, B.P., Vadera, S., and Griffith, S.D. (2015). A new Elixhauser-based comorbidity summary measure to predict in-hospital mortality. Med. Care *53*, 374.

70. El Morabet, N., Uitvlugt, E.B., van den Bemt, B.J., van den Bemt, P.M., Janssen, M.J., and Karapinar-Çarkit, F. (2018). Prevalence and preventability of drug-related hospital readmissions: a systematic review. J. Am. Geriatr. Soc. *66*, 602–608.

71. Basnet, S., Zhang, M., Lesser, M., Wolf-Klein, G., Qiu, G., Williams, M., Pekmezaris, R., and DiMarzio, P. (2018). Thirty-day hospital readmission rate amongst older adults correlates with an increased number of medications, but not with Beers medications. Geriatr. Gerontol. Int. *18*, 1513–1518.

72. Cheong, V.-L., Sowter, J., Scally, A., Hamilton, N., Ali, A., and Silcock, J. (2020). Medication-related risk factors and its association with repeated hospital admissions in frail elderly: a case control study. Res. Soc. Admin. Pharm. *16*, 1318–1322.

73. Saum, K.-U., Schöttker, B., Meid, A.D., Holleczek, B., Haefeli, W.E., Hauer, K., and Brenner, H. (2017). Is polypharmacy associated with frailty in older people? Results from the ESTHER cohort study. J. Am. Geriatr. Soc. *65*, e27–e32.

74. Wastesson, J.W., Morin, L., Tan, E.C., and Johnell, K. (2018). An update on the clinical consequences of polypharmacy in older adults: a narrative review. Expert Opin. Drug Saf. *17*, 1185–1196.

75. Picker, D., Heard, K., Bailey, T.C., Martin, N.R., LaRossa, G.N., and Kollef, M.H. (2015). The number of discharge medications predicts thirty-day hospital readmission: a cohort study. BMC Health Serv. Res. *15*, 1–8.

76. Robinson, R., Bhattarai, M., Hudali, T., and Vogler, C. (2019). Predictors of 30-day hospital readmission: the direct comparison of number of discharge medications to the hospital score and lace index. Future Healthc. J. *6*, 209.

77. Wimmer, B.C., Bell, J.S., Fastbom, J., Wiese, M.D., and Johnell, K. (2016). Medication regimen complexity and number of medications as factors associated with unplanned hospitalizations in older people: a population-based cohort study. J. Gerontol. Ser. A Biomed. Sci. Med. Sci. *71*, 831–837.

78. Sutter, T., Roth, J.A., Chin-Cheong, K., Hug, B.L., and Vogt, J.E. (2021). A comparison of general and disease-specific machine learning models for the prediction of unplanned hospital readmissions. J. Am. Med. Inform. Assoc. *28*, 868–873.

79. Goldstein, B.A., Navar, A.M., Pencina, M.J., and Ioannidis, J. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J. Am. Med. Inform. Assoc. *24*, 198–208.

80. Westra, B.L., Christie, B., Johnson, S.G., Pruinelli, L., LaFlamme, A., Sherman, S.G., Park, J.I., Delaney, C.W., Gao, G., and Speedie, S. (2017). Modeling flowsheet data to support secondary use. Comput. Inform. Nurs. CIN *35*, 452.

81. Hersh, W.R., Weiner, M.G., Embi, P.J., Logan, J.R., Payne, P.R., Bernstam, E.V., Lehmann, H.P., Hripcsak, G., Hartzog, T.H., Cimino, J.J., et al. (2013). Caveats for the use of operational electronic health record data in comparative effectiveness research. Med. Care *51*, S30.

82. Bailey, M.K., Weiss, A.J., Barrett, M.L., and Jiang, H.J. (2019). Characteristics of 30-day all-cause hospital readmissions, 2010–2016 [statistical brief# 248]. Rockv MD Agency Healthc. Res. Qual. https://www.hcup-us.ahrq.gov/reports/statbriefs/sb_readmission.jsp.

83. Quan, H., Li, B., Couris, C.M., Fushimi, K., Graham, P., Hider, P., Januel, J.-M., and Sundararajan, V. (2011). Updating and validating the Charlson Comorbidity Index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. Am. J. Epidemiol. *173*, 676–682.

84. Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. *16*, 321–357.

85. Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M., and Klein, M. (2002). Logistic Regression (Springer).

86. Couronné, R., Probst, P., and Boulesteix, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinform. *19*, 1–14.

87. Shi, T., and Horvath, S. (2006). Unsupervised learning with random forest predictors. J. Comput. Graphical Stat. *15*, 118–138.

88. Chen, T., He, T., Benesty, M., Khotilovich, V., and Tang, Y. (2015). XGBoost: Extreme Gradient Boosting, R Package Version 0.4-2, pp. 1–4.

89. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., and Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. In Advances in Neural Information Processing Systems, pp. 6638–6648.

90. Wolpert, D.H. (1992). Stacked generalization. Neural Netw. *5*, 241–259.

91. Amann, J., Blasimme, A., Vayena, E., Frey, D., and Madai, V.I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med. Inform. Decis. Making *20*, 1–9.

92. Ignatiev, A. (2020). Towards trustable explainable ai. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI), pp. 5154–5158.

93. Lundberg, S.M., Erion, G.G., and Lee, S.-I. (2018). Consistent Individualized Feature Attribution for Tree Ensembles, arXiv:1802.03888.

94. Shapley, L.S. (1953). A value for n-person games. Contrib. Theor. Games *2*, 307–317.

95. Likas, A., Vlassis, N., and Verbeek, J.J. (2003). The global k-means clustering algorithm. Pattern Recognit. *36*, 451–461.

96. Hamerly, G., and Elkan, C. (2004). Learning the k in k-means. Adv. Neural Inf. Process. Syst. *16*, 281–288.

97. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv:1802.03426.