Development and Prospective Validation of a Machine Learning-Based Risk of Readmission Model in a Large Military Hospital

Carly Eckert[1] Neris Nieves-Robbins[2] Elena Spieker[3] Tom Louwers[1] David Hazel[1]
James Marquardt[1] Keith Solveson[3] Anam Zahid[1] Muhammad Ahmad[1] Richard Barnhill[3]
T. Greg McKelvey[1] Robert Marshall[3] Eric Shry[3] Ankur Teredesai[1]

[1] KenSci Inc., Seattle, Washington, United States
[2] Office of the U.S. Army Surgeon General, Defense Health Headquarters (Health Information Technology/CMIO Office), Falls Church, Virginia, United States
[3] Clinical Informatics Division, Madigan Army Medical Center, Joint Base Lewis-McChord, Tacoma, Washington, United States

Address for correspondence Carly Eckert, MD, MPH, KenSci, 615 2nd Avenue Suite 700, Seattle, WA 98104, United States (e-mail: carly@kensci.com).

Appl Clin Inform 2019;10:316–325.

Abstract

Background Thirty-day hospital readmissions are a quality metric for health care systems. Predictive models aim to identify patients likely to readmit to more effectively target preventive strategies. Many risk of readmission models have been developed on retrospective data, but prospective validation of readmission models is rare. To the best of our knowledge, none of these developed models have been evaluated or prospectively validated in a military hospital.

Objectives The objectives of this study are to demonstrate the development and prospective validation of machine learning (ML) risk of readmission models to be utilized by clinical staff at a military medical facility and demonstrate the collaboration between the U.S. Department of Defense's integrated health care system and a private company.

Methods We evaluated multiple ML algorithms to develop a predictive model for 30-day readmissions using data from a retrospective cohort of all-cause inpatient readmissions at Madigan Army Medical Center (MAMC). This predictive model was then validated on prospective MAMC patient data. Precision, recall, accuracy, and the area under the receiver operating characteristic curve (AUC) were used to evaluate model performance. The model was revised, retrained, and rescored on additional retrospective MAMC data after the prospective model's initial performance was evaluated.

Results Within the initial retrospective cohort, which included 32,659 patient encounters, the model achieved an AUC of 0.68. During prospective scoring, 1,574 patients were scored, of whom 152 were readmitted within 30 days of discharge, with an all-cause readmission rate of 9.7%. The AUC of the prospective predictive model was 0.64. The model achieved an AUC of 0.76 after revision and addition of further retrospective data.

Conclusion This work reflects significant collaborative efforts required to operationalize ML models in a complex clinical environment such as that seen in an integrated health care system and the importance of prospective model validation.

Keywords
- machine learning
- operationalization
- patient readmission
- ROC curve

received
April 24, 2018
accepted after revision
March 22, 2019

Stuttgart · New York

DOI https://doi.org/10.1055/s-0039-1688553.
ISSN 1869-0327.

Background and Significance

Decreasing 30-day hospital readmissions is motivated by efforts to improve patient care and reduce penalties for avoidable readmissions, such as those levied by health care payers including the Centers for Medicare and Medicaid (CMS).[1] Accurately predicting patients who are likely to be readmitted should enable a targeted approach for more effective interventions.[2] Efforts to accurately predict 30-day readmissions have been met with moderate success,[3–6] as summarized in two systematic reviews.[7,8] The authors of the 2011 review[7] note the overall modest performance of readmission prediction models. Indeed, only one study, focused on a cohort of heart failure patients,[9] exhibited even moderate discriminative capability with an area under the curve (AUC) of 0.72. The 2016 review[8] similarly concluded that while several of the tools predicting readmissions did exhibit moderate discriminative ability (AUC $>$ 0.70), none of the studies with either a prospective or an external validation cohort surpassed this threshold. The authors of this review caution that more rigorous validation of models is needed.[8]

The widespread use of electronic health records (EHRs) and their underlying data have enabled new opportunities for applied research. In a review of these tools by Goldstein et al, the authors note the overall shortcomings of much of the current work, including the limited use of available features and longitudinal data.[10] The application of artificial intelligence and predictive analytics to health care is fitting given the vast amount of potential data for analysis.[11] Analytic methods, particularly machine learning (ML)-based ensemble methods, are well suited for the complexity of health care data which includes a large number of features and missing data.[12] Futoma et al describe the utility of ML-based models in predicting readmissions[13] and several other studies now populate the academic literature offering ML-based solutions to address this issue.[14–17] Recent work by Hao et al demonstrates prospective validation of a 30-day readmission risk tool using health information exchange data from Maine.[18]

Thirty-day hospital readmissions are a priority for all health care payers, including the federal government. The 2014 Military Health System (MHS) Review, a special report at the behest of the Secretary of Defense to address MHS access, quality, and safety, highlighted reducing readmissions as a priority for military treatment facilities (MTFs) and cited an all-cause MTF readmission rate of 8.8%.[19] While rates are lower than those seen in most other patient populations,[20] penalties still apply. These penalties, which are aligned with CMS Hospital Readmissions Reduction Program penalties and potentially apply to the 25% of Medicare eligible Department of Defense (DoD) beneficiaries,[21] are levied against DoD facilities when readmissions occur. These penalties, along with the continual desire to improve patient care, spurred interest in developing and prospectively validating ML models aimed at reducing readmissions.

Objectives

Beginning in 2016, MAMC Clinical Informatics (CI) and KenSci, a ML company, collaborated to develop, test, and validate ML-based predictive readmission models. Operating within the DoD health care environment presents unique opportunities and challenges for ML applications. Specifically, considerations such as health system size, complexity, and necessary safeguards introduced several variables that distinctly impact health information technology (HIT) projects. The aim of this work is to develop and externally validate ML-based models for all-cause risk of readmission within a DoD health care facility: a large, complex, and integrated health care delivery network.

Methods

Study Design and Setting

MAMC is a tertiary care DoD medical facility located in Joint Base Lewis-McChord in Tacoma, Washington, United States. The MTF serves over 110,000 active duty service members, their families, and military retirees. MAMC, through an agreement with community hospitals, also provides emergency and inpatient care services for civilians whose proximity and condition require activation of the local trauma system. The hospital has 243 inpatient beds and approximately 15,000 inpatient admissions each year. MAMC, and other MTFs, functions as integrated delivery networks and offer inpatient, outpatient, and pharmaceutical services to beneficiaries. Several DoD-specific approvals were required to begin the project and this project was considered Quality Improvement by the MAMC Institutional Review Board.

The initial retrospective cohort included all MAMC inpatient encounters from January 2014 to January 2016. A readmission occurred if a patient was hospitalized at MAMC within 30 days of discharge from the index hospital stay. Given the available data, we were unable to exclude planned readmissions for the analysis. For all included encounters, the patient was alive at the time of discharge. A discharge status flag was present in the data and distinguished discharged patients from transfers. Thirty-day readmissions were only captured if they occurred at MAMC, as opposed to other DoD or civilian hospitals. Patient encounters were excluded in cases where encounter discharge date was later than 30 days prior to the date of data extraction (to ensure 30 days of follow-up).

Data Sources and Data Preprocessing

The data sources that comprise the DoD EHR include several noninteroperable health information systems with disparate naming conventions and ontologies. The data includes patient comorbidities, health care utilization elements, and pharmaceutical details. Data captured at the point of care, such as vital signs and certain laboratory results, is available in the EHR. Data from the outpatient documentation system was not available for this project. Free-text clinical notes were not codified when data sources were identified and were omitted.

Raw data elements from the EHR and other sources required significant preprocessing prior to model building. Twenty-one laboratory test parameters were selected for model inclusion (►Table 1). The parameters were selected based on information available in the literature and input from the project's subject matter experts. There was considerable

Table 1 Features included in the AdaBoost v1 and v2 models

Model	Initial model	Revised model (additional features)
Admission details	Admission diagnosis Date/Time Admitting physician Admitting ward Admitting division HCDP code Primary care provider (PCP) PCP location	
Sociodemographics	Year of birth Ethnicity/Race Sex/Gender Marital status Military grade Insurance coverage Home Zip code	Military rank
Comorbidities	Acute coronary syndrome Alcohol abuse Anemia Arrhythmia Asthma Atherosclerosis Cancer Cardiorespiratory failure CHF CKD Cerebrovascular disease Complicated diabetes Connective tissue disorder COPD Dementia Depression Fluid disorders Gastrointestinal disorder Other liver disease HIV Lung disorder Malnutrition Myocardial Infarction Nephritis Other heart disease Paralysis Peptic ulcer PVD Other psychiatric disorder Renal failure Rheumatic Sepsis Solid tumor Ulcer Uncomplicated diabetes Urinary tract disorder Pneumonia	Charlson Comorbidity Index
Vital signs	Heart rate Respiratory rate Temperature Systolic blood pressure Diastolic blood pressure Pain score	
Pharmacy		Number of prescriptions Number of new prescriptions Rx filled (total) Rx dispensed (index admission) Rx dispensed (prior to admission) Medication type (NDC class)
Laboratory results	Sodium, glucose, calcium, potassium, creatinine, blood urea nitrogen, total protein, albumin, total cholesterol, white blood cell count, red blood cell count, hemoglobin, hematocrit, platelets, international normalized ratio (INR), alanine transaminase (ALT), aspartate transaminase (AST), bilirubin, brain natriuretic peptide (BNP), pro-BNP	

Abbreviations: CHF, congestive heart failure; CKD, chronic kidney disease; COPD, chronic obstructive pulmonary disease; HCDP, health care delivery program; HIV, human immunodeficiency virus; PCP, primary care provider; NDC, National Drug Class; PVD, peripheral vascular disease; Rx, prescriptions.

variability in the naming conventions for each laboratory test parameter. For example, "Glucose" is labeled: "GLUCOSE POCT," "GLUCOSE, COBAS," and "GLUCOSE (WRB-COBAS)," among other variations. This variation is due to lack of standardization at the data entry level. We relied on our subject matter experts to map each laboratory result name to its parent parameter, as with "GLUCOSE" variations above. Laboratory values were maintained as continuous variables. We did not utilize normal or abnormal result flags in model building.

Comorbidities were encoded according to International Statistical Classification of Diseases and Related Health Problems (ICD) 9 and 10 codes.[22] These diagnoses were then grouped according to Clinical Classification Software (CCS) groups. The 38 most common CCS groups, verified by clinical domain experts, were used in the model (►Table 2). CCS group selection was guided by our subject matter experts' experience and evidence from the peer-reviewed literature. Patient records containing current and historical encounter information (vital signs, laboratory values, and diagnostic codes) were combined with derived features, such as length of inpatient stay, to provide additional inputs. For some patients, this could be many years of included data and features. Prior admission count for each unique encounter was defined as the number of MAMC inpatient encounters the patient had prior to the current encounter, that is, the encounter being scored. Prior emergency visit count was defined as the number of MAMC emergency visits for the patient in the 6 months prior to the current encounter. Length of stay from prior inpatient encounters was calculated as the number of days from date of admission to date of discharge. This variable was encoded as an integer and fractions of days were not considered. Thus, the ceiling function was used for considering fractions of days. Observational stays were not encoded differently by the model. Encounters with invalid ICD code fields were excluded as were laboratory results with nonnumeric data in the results field. Multiple inpatient encounters for the same patient during the study period were considered independent observations by the model. The ensemble methods utilized do not consider the longitudinal nature of subsequent encounters and the correlation between intrapersonal encounters was managed by the learning algorithm.[23]

Feature selection was informed by literature review, discussions with subject matter experts, and prior work by KenSci.[16,24–27] Transformed features were automated and features from multiple sources were flattened into a single record for modeling. Multiple methods of imputation were explored to manage null or missing values, including k-nearest neighbor (KNN),[28] carry forward, and mean value imputation. In KNN imputation, the k closest neighbors (based on other data values) to the encounter with the missing field are identified. The mean of the feature under evaluation is then abstracted from the nearest neighbors and imputed. In carry forward imputation, the last complete value for that field is used to complete the missing field. For example, if the laboratory result for albumin is missing, the patient's most recent albumin prior to the missing one is

Table 2 CCS codes used in feature construction

CCS group	CCS code
CHF	108
COPD	127
Asthma	128
Renal failure	157
Uncomplicated diabetes	49
Complicated diabetes	50
Septicemia	2
Pneumonia	122, 129, 130
Chronic kidney disease	158
Anemia	59, 60, 61
Cerebrovascular disease	109, 111, 113
Cancer	11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 41
Arrhythmia	106, 107
Connective tissue	210, 211
Dementia	653
Depression	311
Drug / Alcohol disorder	660
Fluid disorders	55
HIV	5
Nutrition deficiencies	52
Nephritis	156
Paralysis	82
Gastroduodenal ulcer	139
Atherosclerosis	114
Acute coronary syndrome	101
Cardiorespiratory failure/ Shock	131, 249
Rheumatic	96
Gastrointestinal disorder	153, 154, 155
Hematological	62, 63, 64
Lung disorder	133
Myocardial infarction	100
Other heart disease	104
Psych disorder	650, 651, 652, 655, 656, 657, 658, 659, 660, 661, 662, 663, 670
Peripheral vascular disease	115
Solid tumor	39
Ulcer	199
Urinary tract disorder	159, 160
Other liver disease	151

Abbreviations: CCS, Clinical Classification Software; CHF, congestive heart failure; COPD, chronic obstructive pulmonary disease; HIV, human immunodeficiency virus.

carried forward into the missing field. Mean value imputation involves calculating the mean value for that patient for the missing parameter. In this case, the patient's calculated mean albumin for all nonmissing results is used as the value for the missing field. Twenty-one laboratory test parameters were included as defined above. The feature space was reduced to reduce the noise of the predictive model. We applied principal component analysis to identify the top binary features to capture $> 80\%$ of the variance in the data.[29] After feature selection, 54 features were used in the model (summarized in ►Table 1).

We modeled 30-day risk of readmission as a binary classification task, with the negative class *nonreadmitted*, and the positive class *readmitted*. The model generated a risk score between 0 and 1 for each patient based on the available data, with a score closer to 1 indicating a higher likelihood of readmission. Specificity, sensitivity/recall, precision/positive predicted value, accuracy, and C-statistic of the receiver operating characteristic (AUC) curve were used to evaluate model performance.[30] Four ML algorithms were evaluated for this task: decision tree,[31] AdaBoost,[12] random forest,[32] and logistic regression.[33] Decision trees are commonly utilized in health care ML, as they provide easily interpretable algorithms and results for clinical use. Logistic regression is a classic statistical and ML algorithm to utilize when modeling a binary outcome (*readmitted* or *nonreadmitted*). The output of logistic regression algorithms is similar to that of decision trees in that it is typically readily interpretable without specialized knowledge. Boosting methods, such as AdaBoost and random forest, are commonly used with health care data.[12] These methods refer to an ensemble of decision trees constructed through boosting. We used a stochastic gradient boosting with 50 iterations where at each iteration of the algorithm, a base learner fits on a subsample of the training set drawn at random without replacement.[34]

Outcomes

The primary outcome of this work is to predict the risk of 30-day readmission for each inpatient encounter. The input feature vector is $X_i = x_{i1}, x_{i2}, \ldots x_{in}$ for an encounter i that contains the n features mentioned in table. Each row is designed to represent an encounter at the hospital for every patient. The goal is to predict Y (1/0) if the patient will be readmitted to MAMC within 30 days of discharge.

Analysis

After building and testing the ML model on the retrospective cohort, we attempted to validate our model with prospectively scored inpatient encounters. Within the initial MAMC retrospective inpatient cohort, a patient was predicted to be in the *readmitted* class if the resulting risk score was greater than 0.25, the Youden threshold determined by the model.[35] We optimized the precision, recall, and AUC with 10-fold cross-validation. Cross-validation is considered to be a superior method for assessing model performance as compared with the holdout method since cross-validation gives a better estimate of the generalization error.[36] Cross-validation was also used to test model generalizability and overfitting.[37] Risk scores were only generated once for the retrospective cohorts (both original and revised) at the time of patient discharge. For the prospective cohorts, risk scores were generated at the time of patient admission, once daily throughout the hospital stay, and at the time of discharge. The risk score at the time of discharge was then utilized to evaluate model performance. Only data available at the time of scoring was used for all scoring models. For the prospective cohort, as more data from laboratory tests and other results become available, the models are updated to reflect patient status (►Fig. 1).

Model performance assessment on prospective inpatient data at MAMC began in June 2017. Data validation was performed by the MAMC CI database analyst with daily data inspection for inconsistencies. The prospective performance of the model was evaluated after 3 months by comparing model predictions, based on discharge readmission risk score, to actual readmissions. The prediction of readmission was returned as a continuous value between 0 and 1. The risk of readmission (RoR) tool actively scored admitted patients every 24 hours as their clinical records updated.

Results

Initial Retrospective Results

We utilized data from 32,659 inpatient admissions involving 24,499 individual patients admitted to MAMC from January 2014 to January 2016 for the initial retrospective analysis. For each of our predictor variables all missing data was imputed using mean imputation.

Multiple methods of imputation were initially evaluated; however, the required computing power of KNN imputation was prohibitive for the near-real time scoring scenario required by the clinical users. The prevalence of missing data, based on the number of inpatient encounters, for select features is show in ►Table 3. There were 3,085 all-cause 30-day readmissions (9.4%) during this time. Four different ML approaches were applied to this classification problem; of these, AdaBoost exhibited the best performance across metrics including accuracy, recall, and AUC (►Table 4). Model performance was compared with the LACE model, a frequently used rules-based risk of readmission score.[38] The LACE score incorporates four variables when predicting risk of readmission: length of stay, acuity of admission, comorbidities, and number of emergency department visits in the prior 6 months. The AdaBoost model surpassed the LACE model in all performance metrics, as shown in ►Table 4.

Initial Prospective Results

Prospective model evaluation began in June 2017. Patients with recent inpatient stays were labeled into *readmitted* or *nonreadmitted* classes according to their RoR scores. This model utilized the same identified variables as the retrospective evaluation. Mean imputation was similarly used. The performance of the model was evaluated after 11 weeks of clinical use (June 12–August 24). Admissions from June 12 to July 20 were included in the analysis, to ensure at least 30 days of follow-up. During this period, 1,574 patients were admitted to MAMC and 152 were readmitted within 30 days

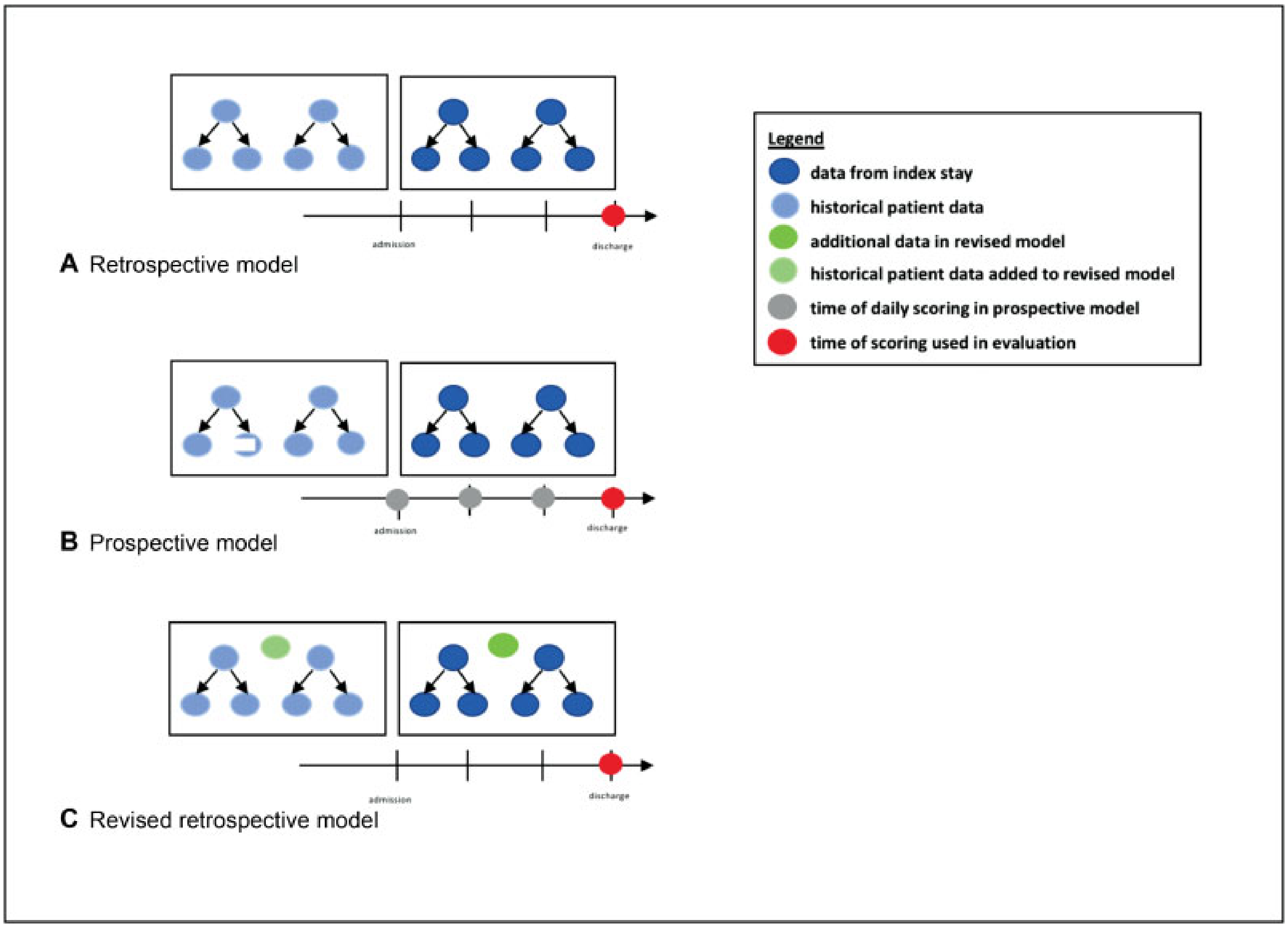
A Retrospective model
admission
discharge
B Prospective model
admission
discharge
C Revised retrospective model
admission
discharge
Legend
data from index stay
historical patient data
additional data in revised model
historical patient data added to revised model
time of daily scoring in prospective model
time of scoring used in evaluation

Fig. 1 Data incorporated and time of scoring for retrospective (**A**), prospective (**B**) and revised (**C**) models.

Table 3 Prevalence of missing data for patient vital signs (initial retrospective cohort)

Data element	% missing
Heart rate	5.5
Systolic blood pressure	17.2
Diastolic blood pressure	17.1
Pulse oxygen	20.7
Temperature	6.6
Respiratory rate	5.7

Table 4 Retrospective performance metrics of ML algorithms and LACE

Model	Accuracy	Precision	Recall	AUC
AdaBoost	0.64	0.18	0.73	0.68
LACE	0.61	0.17	0.70	0.65
Decision tree	0.59	0.21	0.65	0.60
Random forest	0.62	0.20	0.71	0.65
Logistic regression	0.55	0.15	0.51	0.54

Abbreviations: AUC, area under the curve; ML, machine learning.

of discharge (9.7%). The AdaBoost model discriminated between readmissions and nonreadmissions with an accuracy of 0.60, precision (positive predictive value) of 0.15, recall (sensitivity) of 0.66, and an AUC of 0.64 (►Table 5).

Revised Model Results

The performance of the model on prospective data prompted involved clinicians, data scientists, and engineers from MAMC and KenSci to reevaluate the risk of readmission model. Several working sessions followed to discuss, revise, and augment the model including advanced feature development over the following weeks. A revised version with additional features was then tested on retrospective data reflecting inpatient admissions that occurred from January 2014 to June 2017. This data set included 42,392 admissions of 32,219 patients. There were 3,894 30-day readmissions within this group, 514 of which occurred in patients less than 18 years of age. The revised features were based on the availability of data sources and discussions with subject matter experts. These additional features included Charlson Comorbidity Indices for modeling patient comorbidities,[39] military rank as a surrogate for socioeconomic status, and pharmaceutical data incorporated using National Drug Code drug class flags. Pharmaceutical data, which is available from MTF-provided inpatient and outpatient pharmaceutical services, included data related to present and

Table 5 Comparison of AdaBoost model performance on retrospective and prospective data

Model	Data type	Accuracy	Precision	Recall	AUC
AdaBoost v1	Retrospective	0.64	0.18	0.73	0.68
AdaBoost v1	Prospective	0.60	0.15	0.66	0.64
AdaBoost v2	Retrospective	0.73	0.23	0.76	0.76

Abbreviation: AUC, area under the curve.

historical prescriptions ordered, filled, and dispensed. Pharmaceutical data was incorporated into the model as drug class (benzodiazepines, antidepressants, etc.) and included flags for new prescriptions over the past 1 and 6 months. Military rank of the service member was imputed to dependents. Two specialized models were included in the ensemble: one for pediatric patients (age $<$ 18 years) and a second model for adult patients (age $\geq$ 18 years). The revised model was retested on the retrospective patient cohort and exhibited improved performance metrics (►Table 5). The revised model achieved an AUC of 0.76 with these modifications.

Discussion

The purpose of this work was to develop and prospectively validate the performance of an ML-based solution for prospectively predicting 30-day readmissions. This work represents significant collaboration between a government entity and a private company, each providing key expertise in an attempt to solve a significant problem in health care. Similar to the work of Escobar et al, buy-in from the HIT team was critical to the success of this project and the MAMC CI team was instrumental.[40] This work builds on that of Choudhry et al which highlights the need for health care organizations to partner with information technology companies to achieve lofty goals.[41] The Health Information Technology for Economic and Clinical Health Act of 2009 aims to improve coordination of patient care, of which reducing readmissions is a key component.[42]

The derived model reported throughout this article is based on AdaBoost. AdaBoost is an ensemble learning method that utilizes multiple models to build the strongest possible model.[12] AdaBoost's performance is robust in the setting of missing values, making it particularly useful with clinical data. As an ensemble method, AdaBoost can accept a variety of base algorithms and decision trees were used in this situation. One advantage of using an ensemble of decision trees is the ability to incorporate multiple collinear features into models, such as comorbidity flags and Charlson scores, without concern for model convergence.[43] The ML methods used here also enable the incorporation of a large feature space. While many EHR-based models include a relatively low number of features, our models exploit more of the available data and included 54 features.[10] Though we allowed a high number of features in our model, we did consider the computing power required for operationalizing this work into a clinical workflow. Decisions such as the use of mean imputation emphasize to the need to streamline the computational load of these models as a key component of applied ML in health care.

Many published works recount predictive RoR models but few report prospective validation metrics.[7,8] The work of Amarasingham et al is a rare example of a prospectively validated predictive model.[44] The authors of this study prospectively stratified heart failure patients by readmission risk and intervened accordingly, resulting in reduced readmissions. The work of Hao et al is notable as an operationalized, externally validated model of readmission risk using data from a health information exchange in Maine.[18] Our plan for data and clinical validation presented a similar opportunity to prospectively evaluate model performance and to revise the model as needed. The need for model revision is not unexpected, but rather points to the critical nature of model validation.[45,46] An iterative process should be expected as a key component of operationalizing ML. Although prospective validation is rare in health care research, it is more common in other industries such as tax accounting and gaming.[47,48] Altman and Royston discussed the differences between statistical and clinical validation, and the importance of each.[49] The authors write that it is insufficient to describe a predictive model's performance based solely on retrospective data, yet it is too commonly found in the published literature.[49] Model performance is frequently described as being "excellent" or "very good," based solely on its performance on retrospective data. Overfitting and other problems can occur, often giving a false depiction of model performance.[45] Furthermore, the dynamic nature of real-world data introduces computer science problems such as "concept drift."[50] This is particularly true in health care where patient populations shift, and the distributions of disease, patients, and patient characteristics constantly evolve. Predictive models must be continuously tested on new data to address concept drift and related issues, which includes data collected asynchronously.[51] Risk model validation, especially models that may influence clinical care, is an emerging field of research of vast importance as the interface between artificial intelligence and health care grows.[45]

In this work, model performance improved after feature enhancement. The primary set of features used for model building was based on availability during model construction. The inclusion of military rank, pharmaceutical data, and Charlson Comorbidity Indices in our revised model substantially improved predictive performance. The role of socioeconomic features, or their surrogates, has been discussed in the literature with mixed results.[52,53] Military rank is an interesting surrogate for socioeconomic status, as it is nearly uniformly collected among military beneficiaries and can be extended to a spouse or other family members.[54] Other DoD research has reported that military rank does not significantly correlate with outcomes in military health care, citing the single payer system as reducing socioeconomic disparities.[55,56] Prior work by Forster et al demonstrated that drug classes can be predictive of adverse events, particularly in elderly patient cohorts.[57] Other additional features included incorporating a patient's Charlson Comorbidity Index to model comorbidities, a scale commonly used in other such models.[39] Additionally, the initial model attempted to score all inpatients (from pediatrics to adults) using one model. This was revised to two models in the ensemble, one for pediatric patients ($<$ 18 years old) and one for adults.

While the revised model's performance metrics show improvement, as discussed above, these results are based solely on retrospective patient data. The reported metrics for our revised model are based on a sample of inpatient encounters that overlap with the populations used in both the original retrospective evaluation and the prospective study. This model requires exposure to new data to truly evaluate its performance and clinical value. We plan to validate this model on prospective data using a similar approach as described in this article. Additionally, by nature of our access to the data under evaluation, we are unable to determine if the improvement in the AUCs across our models is statistically significant.

Limitations of this work include the single-center nature of the encounter data. Right censoring is a common problem in longitudinal data collection and is often not addressed in health care studies.[10] MTFs provide comprehensive care to their population, have an integrated payment and billing systems, and have a particularly strong relationship to the DoD beneficiaries they serve. As such, we assumed data capture to be high. This assumption could be more fully evaluated by examining third-party billing data and tracking patient encounters, including admissions and readmissions, at other MTFs. Additionally, the inclusion of planned readmissions is a limitation of this work. It is reasonable to expect that around 10% of readmissions in a large hospital system are planned, although rates vary.[58] The data sources incorporated into the present work did not allow for the exclusion of these readmissions. Future work will involve the exclusion of this type.

As we iterate on this work, we will make available to the clinical user the patient-specific explanations associated with each derived risk score. Explainable ML models are a key component to engender user trust and to ensure the validity and accuracy of an individual patient score.[59] These associated factors may provide additional insight to the end users in terms of actions to take to reduce a patient's future readmission risk.

Conclusion

Hospital readmissions and successful methods to prevent them, is a central problem for health systems and a problem many attempt to solve with predictive analytics. Introducing a predictive analytics solution within an MTF adds increasing complexity to this problem due to infrastructure and HIT requirements. Our project highlights the challenges of developing and externally validating ML models in health care and the potential for significant collaboration between civilian and military partners to advance military health care. Our work also demonstrates the importance of iteratively improving ML algorithms using prospectively collected data.

Clinical Relevance Statement

This study demonstrates a successful collaboration and implementation of advanced predictive analytics in a large military hospital. Further development of this predictive tool has the potential to augment clinical care in health care settings to achieve the quadruple aim of better care, improved patient and provider satisfaction, and reduced health care costs.

Multiple Choice Questions

1. Evaluating a prognostic model for clinical use involves all of the following EXCEPT:
   a. Model validation using prospective data.
   b. Awareness of data nonstationarity.
   c. Removing all records with missing values.
   d. Consideration of model calibration.

   Correct Answer: The correct answer is option c. Two very important aspects of a clinically directed model are model and data health. To ensure model and data health, the model must be evaluated in terms of calibration and discrimination using prospective data, or data that was not used in model training. A plan for evaluating model drift, or nonstationarity, must also be considered, as the distribution of the population and associated events is likely to change over time. The handling of missing data must also be addressed; however, the blanket removal of all records with missing values is not usually done.

2. An advantage of machine learning models over more traditional statistical models used in health care and medicine includes which of the following?
   a. The ability to model relationship linearly.
   b. The ability to automatically incorporate a variety of complex relationships between available features.
   c. The ability to generate human understandable risk scores.
   d. The ability to discern the weight of different features in the model output.

   Correct Answer: The correct answer is option b. Advantages of machine learning strategies to address health care problems include the ability to efficiently analyze "big data," to employ multiple models, to model nonlinearity and complex interactions between variables, and to develop local models for specific cohorts or populations. Both machine learning models and traditional statistical methods like those that are regression-based enable linear modeling of the model and its variables. Both machine learning models and traditional statistical methods can generate risk scores that are understandable to a human. Also, both approaches have the ability to discern the weight of different features in the model output.

Protection of Human and Animal Subjects

This study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects and was reviewed by MAMC Institutional Review Board.

Funding

This work was funded by a research grant from the Army Medical Department (AMEDD) Advanced Medical Technology Initiative (AAMTI) provided by the Telemedicine and Advanced Technical Research Center (TATRC).

Conflict of Interest

The authors C.E., T.L., D.H., A.Z., M.A., J.M., T.G.M., and A.T. are employed by and/or have equity ownership in KenSci, the private firm that collaborated on this work. The authors N.N., E.S., K.S., R.B., R.M., and E.S. do not report any conflict of interest. KenSci, the partner in this work, is a private machine learning-based company in Seattle, Washington, United States.

Acknowledgments

We are grateful for the numerous collaborators from both MAMC and KenSci that contributed to this project.

References

1 Boccuti C, Casillas G. Aiming for fewer hospital U-turns: the Medicare hospital readmission reduction program. Policy Brief 2015

2 Kripalani S, Theobald CN, Anctil B, Vasilevskis EE. Reducing hospital readmission rates: current strategies and future directions. Annu Rev Med 2014;65(01):471–485

3 Frizzell JD, Liang L, Schulte PJ, et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. JAMA Cardiol 2017;2(02):204–209

4 Allaudeen N, Schnipper JL, Orav EJ, Wachter RM, Vidyarthi AR. Inability of providers to predict unplanned readmissions. J Gen Intern Med 2011;26(07):771–776

5 Billings J, Blunt I, Steventon A, Georghiou T, Lewis G, Bardsley M. Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30). BMJ Open 2012;2(04):e001667

6 Donzé JD, Williams MV, Robinson EJ, et al. International validity of the hospital score to predict 30-day potentially avoidable hospital readmissions. JAMA Intern Med 2016;176(04):496–502

7 Kansagara D, Englander H, Salanitro A, et al. Risk prediction models for hospital readmission: a systematic review. JAMA 2011;306(15):1688–1698

8 Zhou H, Della PR, Roberts P, Goh L, Dhaliwal SS. Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. BMJ Open 2016;6(06):e011060

9 Amarasingham R, Moore BJ, Tabak YP, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Med Care 2010;48(11):981–988

10 Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017;24(01):198–208

11 Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff (Millwood) 2014;33(07):1163–1170

12 Freund Y, Schapire RE. A decision-theoretic generalization of online learning and an application to boosting. In: European Conference on Computational Learning Theory. 1995:23–37

13 Futoma J, Morris J, Lucas J. A comparison of models for predicting early hospital readmissions. J Biomed Inform 2015;56:229–238

14 Shameer K, Johnson KW, Yahi A, et al. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using Mount Sinai Heart Failure Cohort. In: Pacific Symposium on Biocomputing. 2017:276–287

15 Yang C, Delcher C, Shenkman E, Ranka S. Predicting 30-day all-cause readmissions from hospital inpatient discharge data. In: e-Health Networking, Applications and Services (Healthcom), 2016 IEEE 18th International Conference on IEEE. 2016:1–6

16 Basu Roy S, Teredesai A, Zolfaghar K, et al. Dynamic hierarchical classification for patient risk-of-readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015:1691–1700

17 Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015:1721–1730

18 Hao S, Wang Y, Jin B, et al. Development, validation and deployment of a real time 30-day hospital readmission risk assessment tool in the Maine Healthcare Information Exchange. PLoS One 2015;10(10):e0140271

19 Final Report to the Secretary of Defense: Military Health System Review; 2014

20 Barrett M, Wier L, Jiang J, Steiner C. All-Case Readmissions by Payer and Age, 2009–2013. Healthcare Cost and Utilization Project. Statistical Brief #199. Agency for Healthcare Research and Quality; 2015

21 Evaluation of the TRICARE Program. Fiscal Year. 2017 Report to Congress. Available at: https://health.mil/Military-Health-Topics/Access-Cost-Quality-and-Safety/Health-Care-Program-Evaluation/Annual-Evaluation-of-the-TRICARE-Program. Accessed March 30, 2018

22 World Health Organization. International Statistical Classification of Diseases and Related Health Problems. Geneva, Switzerland: World Health Organization; 2004

23 Wang W. Some fundamental issues in ensemble methods. In: Proceedings of IEEE World Congress on Computational Intelligence. 2008:2244–2251

24 Zolfaghar K, Meadem N, Teredesai A, Roy SB, Chin SC, Muckian B. Big data solutions for predicting risk-of-readmission for congestive heart failure patients. In: Big Data, 2013 IEEE International Conference. 2013:64–71

25 Hon CP, Pereira M, Sushmita S, Teredesai A, De Cock M. Risk stratification for hospital readmission of heart failure patients: A machine learning approach. In: Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 2016:491–492

26 Liu R, Zolfaghar K, Chin SC, Roy SB, Teredesai A. A framework to recommend interventions for 30-day heart failure readmission risk. In: Data Mining (ICDM), 2014 IEEE International Conference. 2014:911–916

27 Rao VR, Zolfaghar K, Hazel DK, Mandava V, Roy SB, Teredesai A. Readmissions Score as a Service (RaaS)

28 García-Laencina PJ, Sancho-Gómez JL, Figueiras-Vidal AR, Verleysen M. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing 2009;72(7–9):1483–1493

29 Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev Comput Stat 2010;2(04):433–459

30 Verbiest N, Vermeulen K, Teredesai A. Evaluation of Classification Methods. CRC Press; 2014

31 Hastie TJ, Tibshirani RJ, Friedman JH. The Elements of Statistical Learning: Data Mining Inference and Prediction. 2nd ed. Springer; 2009

32 Breiman L. Random forests. Mach Learn 2001;45(01):5–32

33 Cox DR. The regression analysis of binary sequences. J R Stat Soc [Ser A] 1958:215–242

34 Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal 2002;38(04):367–378

35 Fluss R, Faraggi D, Reiser B. Estimation of the Youden index and its associated cutoff point. Biom J 2005;47(04):458–472

36 Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI (U S) 1995;14(02):1137–1145

37 Hawkins DM, Basak SC, Mills D. Assessing model fit by cross-validation. J Chem Inf Comput Sci 2003;43(02):579–586

38 van Walraven C, Dhalla IA, Bell C, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. CMAJ 2010;182(06): 551–557

39 Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40(05):373–383

40 Escobar GJ, Turk BJ, Ragins A, et al. Piloting electronic medical record-based early detection of inpatient deterioration in community hospitals. J Hosp Med 2016;11(Suppl 1):S18–S24

41 Choudhry SA, Li J, Davis D, Erdmann C, Sikka R, Sutariya B. A public-private partnership develops and externally validates a 30-day hospital readmission risk prediction model. Online J Public Health Inform 2013;5(02):219

42 Blumenthal D. Launching HITECH. N Engl J Med 2010;362(05): 382–385

43 Pan F, Converse T, Ahn D, Salvetti F, Donato G. Feature selection for ranking using boosted trees. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. 2009: 2025–2028

44 Amarasingham R, Patel PC, Toto K, et al. Allocating scarce resources in real-time to reduce heart failure readmissions: a prospective, controlled study. BMJ Qual Saf 2013;22(12):998–1005

45 Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162(01):W1–73

46 Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol 2015;68(01):25–34

47 Hsu KW, Pathak N, Srivastava J, Tschida G, Bjorklund E. Data mining-based tax audit selection: a case study of a pilot project at the Minnesota Department of Revenue. In: Real World Data Mining Applications; 2015:221–245

48 Borbora Z, Srivastava J, Hsu KW, Williams D. Churn prediction in MMORPGs using player motivation theories and an ensemble approach. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing. 2011:157–164

49 Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19(04):453–473

50 Widmer G, Kubat M. Learning in the presence of concept drift and hidden contexts. Mach Learn 1996;23(01):69–101

51 Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc 2017;24(06):1052–1061

52 Krumholz HM, Chaudhry SI, Spertus JA, Mattera JA, Hodshon B, Herrin J. Do non-clinical factors improve prediction of readmission risk? JACC Heart Fail 2016;4(01):12–20

53 Calvillo-King L, Arnold D, Eubank KJ, et al. Impact of social factors on risk of readmission or mortality in pneumonia and heart failure: systematic review. J Gen Intern Med 2013;28(02):269–282

54 Taylor AJ, Meyer GS, Morse RW, Pearson CE. Can characteristics of a health care system mitigate ethnic bias in access to cardiovascular procedures? Experience from the Military Health Services System. J Am Coll Cardiol 1997;30(04):901–907

55 Clark JY, Thompson IM. Military rank as a measure of socioeconomic status and survival from prostate cancer. South Med J 1994;87(11):1141–1144

56 Chaudhary MA, Sharma M, Scully RE, et al. Universal insurance and an equal access healthcare system eliminate disparities for Black patients after traumatic injury. Surgery 2018;163(04): 651–656

57 Forster AJ, Murff HJ, Peterson JF, Gandhi TK, Bates DW. Adverse drug events occurring following hospital discharge. J Gen Intern Med 2005;20(04):317–323

58 Jencks SF, Williams MV, Coleman EA. Rehospitalizations among patients in the Medicare fee-for-service program. N Engl J Med 2009;360(14):1418–1428

59 Lipton ZC. The Mythos of Model Interpretability. arXiv; 2016