informatics

MDPI

Article

Development and Internal Validation of an Interpretable Machine Learning Model to Predict Readmissions in a United States Healthcare System

Amanda L. Luo [1], Akshay Ravi [2], Simone Arvisais-Anhalt [3], Anoop N. Muniyappa [2], Xinran Liu [2,*,†] and Shan Wang [1,4,†]

[1] Master of Science in Data Science Program, University of San Francisco, San Francisco, CA 94117, USA
[2] Division of Hospital Medicine, Department of Medicine, University of California, San Francisco, CA 94143, USA
[3] Department of Laboratory Medicine, University of California, San Francisco, CA 94143, USA
[4] Department of Mathematics and Statistics, University of San Francisco, San Francisco, CA 94117, USA
* Correspondence: xinran.liu@ucsf.edu
† These authors contributed equally to this work as Co-senior authors.

Abstract: (1) One in four hospital readmissions is potentially preventable. Machine learning (ML) models have been developed to predict hospital readmissions and risk-stratify patients, but thus far they have been limited in clinical applicability, timeliness, and generalizability. (2) Methods: Using deidentified clinical data from the University of California, San Francisco (UCSF) between January 2016 and November 2021, we developed and compared four supervised ML models (logistic regression, random forest, gradient boosting, and XGBoost) to predict 30-day readmissions for adults admitted to a UCSF hospital. (3) Results: Of 147,358 inpatient encounters, 20,747 (13.9%) patients were readmitted within 30 days of discharge. The final model selected was XGBoost, which had an area under the receiver operating characteristic curve of 0.783 and an area under the precision-recall curve of 0.434. The most important features by Shapley Additive Explanations were days since last admission, discharge department, and inpatient length of stay. (4) Conclusions: We developed and internally validated a supervised ML model to predict 30-day readmissions in a US-based healthcare system. This model has several advantages including state-of-the-art performance metrics, the use of clinical data, the use of features available within 24 h of discharge, and generalizability to multiple disease states.

Keywords: machine learning; hospital readmission; patient readmission; risk assessment

Citation: Luo, A.L.; Ravi, A.; Arvisais-Anhalt, S.; Muniyappa, A.N.; Liu, X.; Wang, S. Development and Internal Validation of an Interpretable Machine Learning Model to Predict Readmissions in a United States Healthcare System. *Informatics* **2023**, *10*, 33. https://doi.org/10.3390/informatics10020033

Academic Editor: Jiang Bian

Received: 30 January 2023
Revised: 10 March 2023
Accepted: 17 March 2023
Published: 27 March 2023

1. Introduction

In the United States, the Centers for Medicare and Medicaid Services (CMS) has effectively mandated a focus on hospital readmission by publicly reporting hospital performance and reducing payments for unplanned hospital readmissions. To date, CMS includes six conditions and procedures for 30-day risk-standardized unplanned readmission measures, including acute myocardial infarction, chronic obstructive pulmonary disease, heart failure, pneumonia, coronary artery bypass graft surgery, and elective primary total hip arthroplasty and/or total knee arthroplasty [1]. For these six conditions and procedures, CMS calculates payment reductions for hospitals based on their readmission performance [1]. Beyond these conditions and reimbursement implications, it is commonly understood that unplanned hospital readmissions may indicate poor quality of care, and it has been shown that one in four readmissions is potentially preventable [2,3].

In order to address readmission risk factors and improve patient outcomes, hospitals perform a variety of interventions to help patients make successful transitions out of the hospital. Some pre-discharge interventions include patient education, discharge

planning, medication reconciliation, and appointment scheduling before discharge. Some post-discharge interventions include timely follow-up, timely primary care provider communication, follow-up telephone calls, access to patient hotlines, and home visits [4]. Some bridging interventions include transition coaches, patient-centered discharge instructions, and provider continuity [4]. These interventions can be provided singly or in combination [4]; however, given that there are often limited institutional resources in deploying such interventions to patients, there is great interest in predicting patients at highest risk for readmission to best understand where to devote limited resources and coordination efforts.

In an effort to best allocate resources, much effort has been placed on developing machine learning models to predict hospital readmission and to risk-stratify patients [5–8]. Prior work has used either administrative data alone or combined with clinical data to make these predictions. Li et al. built a model using administrative data in Taiwan with a high area under the receiver operator curve (AUC), but using similar datasets in the United States for a more specific disease state did not yield promising results [6,7]. This contrast highlights the variability of administrative data, especially between countries. Furthermore, these datasets are often compiled weeks or months after a patient's discharge, limiting their immediate post-discharge utility. Mišić et al. have applied these models to predict 30-day readmission for postoperative patients using administrative and clinical data [8]. Others have demonstrated similarly effective models for predicting readmission when restricted to a single-use case or disease state, but these models are not generalizable to all discharged patients. Lo et al. addressed this with a model that predicts 14-day unplanned admissions using administrative and clinical data, although these results are built using data from Taiwan. However, the strength of a generalizable readmission model using administrative and clinical data in the United States is unknown.

We extend these results using state-of-the-art machine learning modeling techniques including XGBoost to predict 30-day all-cause readmissions in a US-based patient population. Our model was built using clinical and administrative data available within 24 h of hospital discharge to allow for better operationalization and clinical implementation of the model. We see the development of this model as the first step in a long journey. Ultimately, we hope to deploy our model within a US healthcare system so that it can be used to risk-stratify patients after hospital discharge. This risk stratification can then be used to drive enrollment in targeted post-discharge support interventions in order to decrease readmission rates.

2. Materials and Methods

2.1. Patient Selection

The patient health records were extracted from the University of California, San Francisco (UCSF) De-Identified Clinical Data Warehouse (DEID CDW) database. This database collects deidentified demographic and clinical data from patients at UCSF, a tertiary care academic medical center with 861 staffed beds and 34,105 annual admissions [9]. As part of the deidentification process, dates associated with individual patients were randomly date-shifted 1–365 days into the past. For this study, patients who were 18 years old or older and had an inpatient or observation encounter status between January 2016 and November 2021 were included. Each patient encounter was treated as a row for modeling purposes, with columns for each row representing features from the encounter.

2.2. Outcome Variable

The primary study outcome was 30-day all-cause readmission. Readmission was defined as admission (inpatient or observation status) to a UCSF-affiliated hospital within 30 days of the index discharge. Each index patient encounter was assigned an outcome value of "1" if it led to a 30-day readmission, or "0" if it did not.

2.3. Feature Engineering, Selection, and Imputation

After a review of the literature, a multidisciplinary team of clinicians, informaticians, data scientists, and operational leaders at UCSF created the initial set of features to input in the model through several steps. The first step included determining what raw features to include in the dataset (e.g., age, demographics, labs, vitals). The second step was creating additional engineered features from the raw features. Some of these engineered features were created to capture information that we suspected was important for the use case, but not directly captured from the raw features. Other features were created to reduce the dimensionality of certain data types such as labs and vitals, which can have numerous values per encounter. Specifically, for each encounter, all lab and vital sign data were aggregated into mean, minimum, maximum, first, and last values. For raw features including text, such as diagnosis codes, chief complaint, reason for admission, etc., we tried two methods. One was using BioSentVec [10,11] to transform the raw text into sentence embeddings that were then used in modeling. The other was treating the most common diagnoses as their own categories, while bucketing less common diagnoses into an "other" category. Features broadly included demographic data, admission metadata, and clinical data that were extractable within 24 h of discharge from the hospital.

The data were then split into train/validation and test sets (detailed further in next section). We examined missing values in the processed data. Features with an abundant number of missing values (>99%) or with only one unique value in the column were excluded. For features that were not excluded but that still had missing values, we applied imputation as we believed these data still contained potentially useful information (i.e., the absence of a specific lab might still be clinically important). Missing values in categorical features and nonderived numerical features (e.g., labs, vitals) were assigned a "missing" label of "−1". Missing values in the derived numeric features (e.g., number of admissions in the last year) were imputed with the median of the column. Imputation was done separately within train/validation and test sets.

After feature engineering and missing value imputation, we performed feature selection from the train/validation set with the drop column feature importance method [12,13]. We used the area under the precision-recall curve (AUC-PR) metric of the model with all columns as the baseline, and then dropped a column entirely, retrained the model, and recomputed the AUC-PR. The importance value of a feature is the difference between the baseline and the score from the model missing that feature. For our study, features that, when dropped, led to an increase from the baseline AUC-PR were excluded. We chose AUC-PR because the prevalence of the outcome of interest (readmissions) was relatively low, and in this situation AUC-PR might be a more practical representation of the usefulness of a model compared to the AUC [14].

2.4. Modeling Process

To predict the probability of a patient being readmitted within 30 days, we compared four supervised machine learning models for binary classification, including logistic regression, random forest, gradient boosting, and XGBoost. The pre-processed data were split into train/validation and test datasets, respectively. The test set included data from the most recent year and was used to judge the final performance of our model. The rest of the data were used for training and validation using the expanding-window-based 3-fold cross-validation method [15,16].

Expanding-window cross-validation applies a cross-validation logic that accounts for the sequenced nature of the dataset. In this study, we created three iterations, each with a split of the training and validation sets. Each validation set consists of records from the most recent one-year period. The corresponding training set consists only of records that occurred before the time of the validation set (Figure 1). All models were trained using the training set of each cross-validation iteration, and performance metrics were obtained from the respective validation set. The performance of each model was evaluated by averaging AUC-PR scores over the three validation sets.

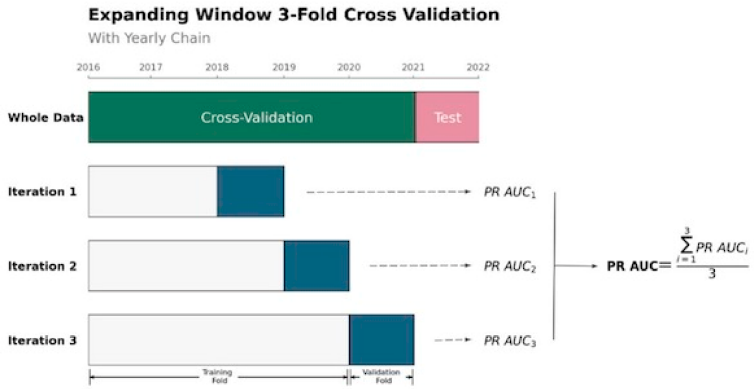


Figure 1. Expanding window 3-fold cross-validation. Data were split into training/validation and test datasets. Data between 2021 and 2022 were used as the final test set. The remaining data were split into 3 iterations of training and validation sets using the expanding-window 3-fold cross-validation method, in which the most recent 1-year period was treated as the validation set and the remaining data in that iteration was used for training. Model performance was evaluated using average AUC-PR of all 3 iterations.

The best-performing model from the above process would be selected as our final model, and it would be applied to the test set to determine its performance. Important outcome metrics to be measured from the test set include AUC-PR, AUC, accuracy, precision, recall, and F1-score [14]. We also used the SHAP Shapley Additive exPlanations (SHAP) Feature Importance method [17] to examine the significance of features included in the final model. SHAP feature importance connects local interpretable model-agnostic (LIME) [18] and Shapley value [19] and calculates a kernel-based estimation of the Shapley value on each instance of a feature. The Shapley value of a feature gives the average marginal contribution of a feature value across all the possible combinations of features. This main property of the Shapley value, the efficiency property, distinguishes the Shapley value from other feature-importance methods. It provides a fair contribution of features with mathematically proven theory. Each feature's final SHAP value was calculated by averaging the SHAP values obtained from each training set from the cross-validation iterations. Features with larger SHAP values are considered more important [17,20,21].

3. Results

The development cohort consisted of 147,358 patients, of which 20,747 (13.9%) were readmitted within 30 days of discharge. Their baseline characteristics are summarized in Table 1.

Table 1. Baseline patient characteristics.

Characteristics	Total Cohort	Patients Readmitted	Patients Not Readmitted
		Age	
Mean (SD)	54.02 (18.56)	53.41 (18.73)	54.12 (18.53)
		Gender	
Male	66,482 (45.12%)	10,405 (50.15%)	56,077 (44.29%)
Female	80,827 (54.85%)	10,332 (49.80%)	70,495 (55.68%)
Nonbinary	28 (0.02%)	7 (0.03%)	21 (0.02%)
Unknown	21 (0.01%)	3 (0.01%)	18 (0.01%)
		Ethnicity	
Hispanic or Latino	23,487 (15.94%)	3980 (19.18%)	19,507 (15.41%)
Not Hispanic or Latino	120,426 (81.72%)	16,505 (79.55%)	103,921 (82.08%)
Unknown	3445 (2.34%)	262 (1.26%)	3183 (2.51%)

Table 1. *Cont.*

Characteristics	Total Cohort	Patients Readmitted	Patients Not Readmitted
		Race	
American Indian or Alaska Native	1390 (0.94%)	230 (1.11%)	1160 (0.92%)
Asian	23,871 (16.20%)	3528 (17.00%)	20,343 (16.07%)
Black or African American	13,128 (8.91%)	2278 (10.98%)	10,850 (8.57%)
Native Hawaiian	65 (0.04%)	13 (0.06%)	52 (0.04%)
White or Caucasian	79,816 (54.17%)	10,330 (49.79%)	69,486 (54.89%)
Other Pacific Islander	1605 (1.09%)	212 (1.02%)	1393 (1.10%)
Other	27,469 (18.64%)	4156 (20.03%)	23,313 (18.42%)
		Admission Type	
Emergency/Urgent	90,564 (61.88%)	14,278 (68.80%)	76,286 (60.73%)
Routine/Elective	54,241 (37.06%)	6270 (30.22%)	47,971 (38.19%)
Other	1553 (1.06%)	203 (0.98%)	1350 (1.08%)
		Insurance	
Commercial	51,388 (34.82%)	6031 (29.05%)	45,357 (35.77%)
Medi-Cal	38,464 (26.06%)	6646 (32.01%)	31,818 (25.09%)
Medicare	56,235 (38.11%)	7922 (38.16%)	48,313 (38.1%)
Other	1488 (1.01%)	163 (0.79%)	1325 (1.04%)
		Length of Stay	
Mean (SD)	6.13 (9.45)	7.53 (9.94)	5.89 (9.35)

Most characteristics such as age, gender, race, ethnicity, etc., were similar between readmitted and not readmitted patients. Readmitted patients, however, seemed to have a higher average hospital length of stay and were more likely to be admitted as "emergency/urgent" admission type. Thirty-seven raw features were initially extracted from the De-ID CDW that included information on patient demographic information, medical history, ancillary orders, procedures, flowsheet values, lab tests, and patient utilization. We created 3796 engineered features based on the original data, including 3500 sentence-embedded features from five textual columns including patient admission and discharge diagnoses. A total of 230 aggregation type features were derived from lab and flowsheet values, and 66 features were created based on physician insight. Table 2 summarizes the features at a macro level.

Table 2. Features used for modeling.

Type of Feature	Examples of Features Created for Model(s)
Patient utilization	Binary target variable: readmission status within 30 days (0 = No, 1 = Yes);Number of days since last admission; Number of admissions in the past 90 days, 180 days, 1 year, and 2 years; Number of emergency visits in the past 90 days, 180 days, 1 year, and 2 years;
Demographic information	Age; sex; race; ethnicity; marital status; preferred language; financial class; postal code; smoking status; BMI
Procedure information	Number of procedures performed during encounter; Binary indicator for if any procedure was performed (0 = No, 1 = Yes)

Table 2. *Cont.*

Type of Feature	Examples of Features Created for Model(s)
Lab tests	Mean, minimum, maximum, the first and last value of each unique type of lab test (e.g., creatinine, hemoglobin) that was resulted during the encounter; Binary indicators for if amphetamine, barbiturates, benzo, cocaine, opiates, THC, and Utox was ordered (0 = No, 1 = Yes); Binary indicators for if amphetamine, barbiturates, benzo, cocaine, opiates, THC, and Utox were positive (0 = No, 1 = Yes);
Flowsheet values	Mean, minimum, maximum, the first and last value of each unique type of flowsheet value (e.g., heart rate, respiratory rate, nursing mobility scores) that was recorded during the encounter
Ancillary orders	Binary indicators for if each given ancillary order (e.g., palliative care consult, DNR/DNI order, social work) was placed (0 = No, 1 = Yes)
Textual information such as diagnosis and primary chief complaints	Sentence-embedded vectors generated from textual columns, categorical features that treat unique diagnoses as their own categories
Other features that have remained the same as in EHR	Admission source; admission type; inpatient length of stay in days; discharge disposition; department; hospital service; admitting provider type; admitting provider primary specialty; arrival method; acuity level

We assessed the 3500 sentence-embedded columns with drop-column feature importance and found that the improvement in AUC-PR added by the sentence-embedded columns was not significant (0.01 increase) compared to the extra computational burden and complexity in model interpretation they added. Thus, we excluded diagnosis-related features that were created using word embeddings. However, these diagnosis-related features were still included in the final model as categorical features described in Section 2.3. Of the remaining 296 features, 50 features were determined to be unimportant by drop-column importance, and thus were also excluded. After this process, 246 features were included in the final model. This full process can be seen in Figure 2. A list of all 246 of our features can be found in Appendix A.

Only 1 feature had more than 99% missing values and was dropped; 25 features had no missing values. The rest of the features had between 0.007 and 95.63% missing values, with only 42 features having more than 80% missing values.

Data from January 2016 to December 2020 was used to perform expanding-window 3-fold cross-validation for model selection (Figure 1). We evaluated four classification models (parameters found in Appendix B), and Table 3 shows each model's average AUC-PR score and running time. Gradient boosting reached the highest average AUC-PR score of 0.444, but we chose XGBoost as the final model considering its comparative performance (AUPRC = 0.434) and lower computational complexity (running time = 305.9 s).

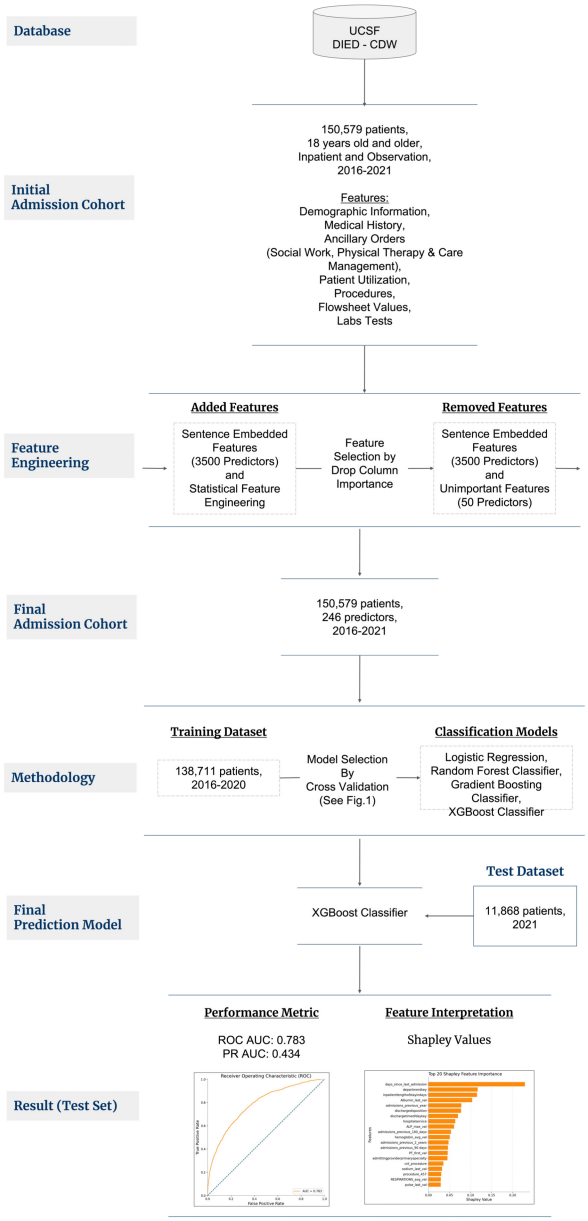


Figure 2. Cohort selection and model selection. UCSF: University of California, San Francisco. DEID CDW: Deidentified clinical data warehouse.

Table 3. Model performance from cross-validation.

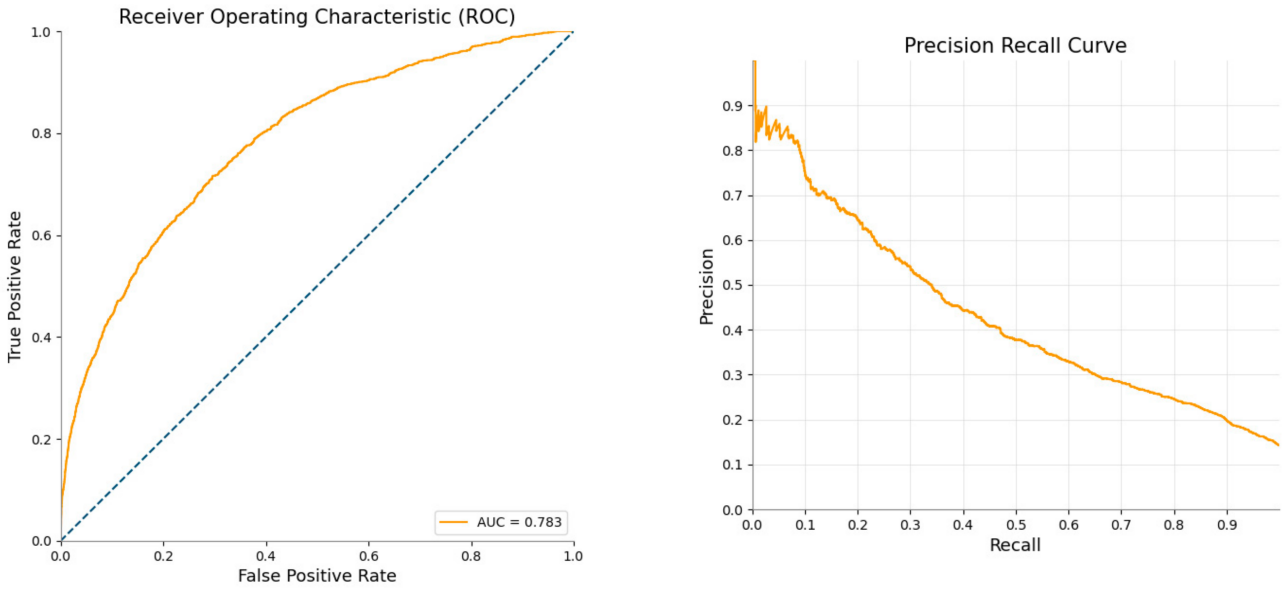
Machine Learning Model	Average Area under the Precision-Recall Curve	Average Training Time (Seconds)
Logistic Regression	0.2403	81.522
Random Forest	0.4116	106.875
Gradient Boosting	0.4435	489.752
XGBoost	0.434	305.884

We tested the XGBoost classifier on the test dataset between January 2021 and November 2021, and Table 4 summarizes the performance. The XGBoost classifier had an AUC-PR score of 0.434 on the test set, which is the same as in the validation set. This gives us confidence that our model is capturing the important relationships in the data rather than overfitting to random noise. The AUC was 0.783. The exact precision (positive predictive value) and recall (sensitivity) of the model can be tuned based on the needs of

the user. We have highlighted that at a set recall of 0.701, our model had a precision of 0.283. The overall ROC and PR curves can be seen in Figure 3. These results are highly favorable when compared to results from other papers [22], especially when compared to US-based datasets.

Table 4. Performance of XGBoost classifier on test set.

Test Characteristic	Value
AUC	0.783
AUC-PR	0.434
Accuracy	0.713
Precision	0.283
Recall	0.701
F1	0.403
Threshold	0.486
True positives	903
True negatives	5738
False positives	2286
False negatives	384



(**a**) (**b**)

Figure 3. Receiver operator characteristic and precision-recall curves: (**a**) receiver operator characteristic curve for XGBoost classifier; (**b**) precision-recall curve for XGBoost classifier.

We also applied SHAP feature importance to highlight the top 20 importance features in our training data. The plot can be seen in Figure 4. These include a mix of utilization (e.g., length of stay, number of admissions in the past year), disposition (e.g., discharge disposition, admitting department or hospital service), laboratory (e.g., last albumin or sodium value, first prothrombin time), and vitals-based features (e.g., average respiratory rate, last heart rate).

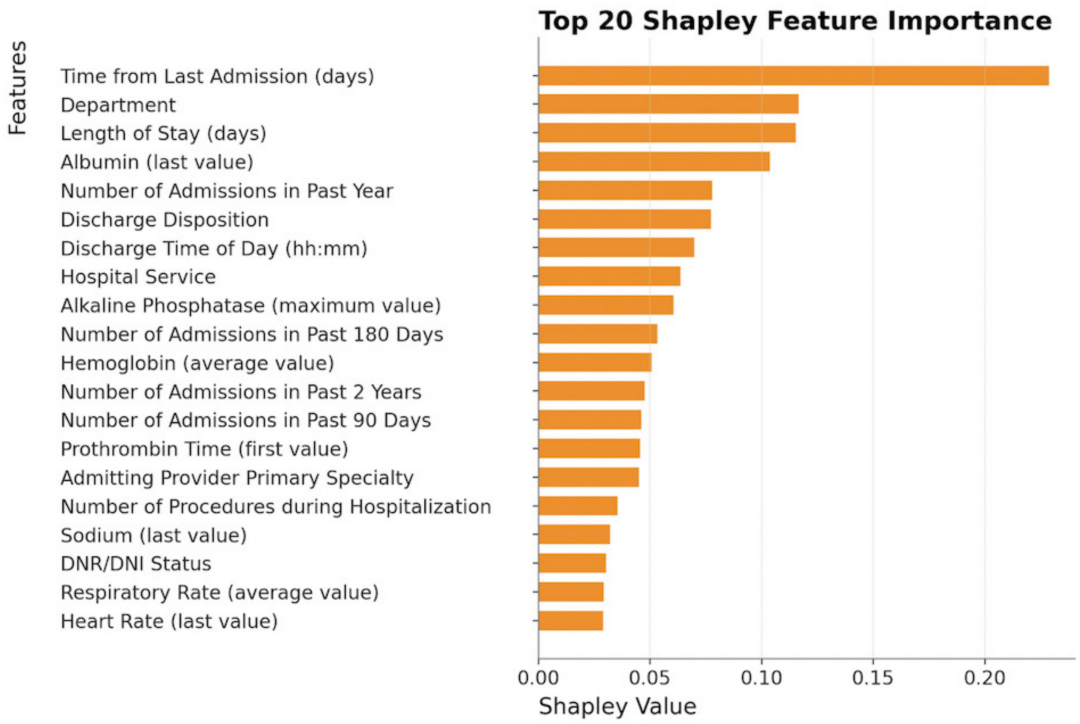


Figure 4. Shapley features importance. Top 20 most important features by Shapley Additive Explanations.

4. Discussion

We developed a ML model using clinical and administrative data from a US-based healthcare system to predict 30-day all-cause readmissions with an AUC of 0.783 and AUC-PR of 0.434. These AUCs and AUC-PRs were consistent between our validation and test sets, which gives us confidence that our model did not overfit to either dataset. Our work is novel in several ways. First, we incorporate a range of clinical features from nursing-based risk scores, to vital signs and lab data, to relevant admission metadata rather than using administrative or billing data alone. To maximize operational utility, we purposefully selected features for the model that are available within 24 h of discharge, so that the model could be used to generate predictions the day after patient discharge. Claims data, while often used to develop ML models, are not ideal for time-sensitive predictions such as readmissions, as there is a delay between when the patient is discharged and when the data might become available [23]. Predictions that are made after a patient has already been readmitted, even if accurate, offer little operational utility.

Second, our model achieves a higher AUC than other US-based readmission prediction models, which have had AUCs between 0.62 and 0.76 [22,24–26]. Of these, the model described by Ko et al. has the best performance for general readmissions, although they incorporated administrative score data that may not be available at the time of discharge for operationalization [25]. We do acknowledge that there are non-US-based readmission prediction models that cite similar or better results in terms of AUC [5,6,26]. However, there may be significant differences in the patient populations, healthcare delivery systems, societal priorities, cultures, resources, etc., between the US and other countries [27]. Thus ML models built on non-US populations might not generalize well to the US setting.

An important question to ask when evaluating the utility of ML models is not just the AUC, but whether or not that AUC can translate to recall (sensitivity) and precision (positive prediction values) levels that are clinically useful [13]. At a set recall of 0.70, our model achieves a precision of 0.283. We choose to highlight this threshold, as we believe it does pass the "eye test" in terms of meeting meaningful levels for recall and precision to be considered for operationalization. The exact model threshold, and hence the resulting precision and recall, that should be used varies based on the intended intervention,

resources available, and institution. For example, if the model is used to determine which patients will receive automated phone calls after discharge, higher recall might be preferred. If the model is used to determine which patients will receive personal case management outreach, a higher precision might be preferred.

A third area of innovation in our work is the focus on all types of patients, rather than only specific patient populations. There are numerous ML models to predict hospital readmission for specific patient populations, such as postoperative patients [8,28–30], stroke patients [31,32], hypertensive disorders of pregnancy [33], heart failure [33–35], and more [36,37]. However, by focusing on a single disease state of interest, these models are able to use highly disease-specific features that may not be applicable for a broader population, limiting the generalizability of these models. Furthermore, from a health system perspective, it might be preferred to implement and maintain one general model as opposed to numerous models specific to different populations.

Finally, our model uses SHAP feature importance to provide some insight into how our model makes its predictions at the global level. One of the biggest concerns for using ML in medicine is the lack of interpretability of some models [38], especially compared to traditional statistical models such as linear or logistic regression. However, tree-based models such as random forests, gradient boosting, and XGboosting do have methods to explain how they work [39–44], such as SHAP feature importance. Using this method, we were able to highlight the top 20 most influential features in our model. These include a mix of utilization, disposition, laboratory, and vitals-based features. The relevance of utilization and disposition-related features are well described in the literature [45,46], and it is reassuring that our model highlighted the importance of these features as well. Our model also picked up on less well-described risk factors for readmission, such as nutrition [47], DNR/DNI code status, last heart rate value during hospitalization, average respiratory rate during hospitalization, last serum sodium value, number of procedures performed during the hospitalization, and average hospital hemoglobin. Methods such as SHAP do not prove causation and suffer from collinearity, confounding and other biases. Despite these limitations, it is reassuring to see that the features that our model identified as important seem to pass the clinical "sniff test." The ability to understand how a model made its predictions may go a long way toward improving clinicians' trust in ML models and, ultimately, improve buy-in for using these models for patients. We have demonstrated that this can be accomplished at the global ML model level using methods such as SHAP. Future work will focus on explaining how ML models make their predictions at the individual prediction level.

It is worth noting that we attempted to incorporate diagnostic information by using word embeddings in addition to treating diagnoses as categorical features. We chose to try word embeddings because we wanted to include as much information as possible from the diagnostic text [48]. We found that doing so did not significantly increase our model performance compared to treating diagnostic information as categorical features only. As a result, we decided to omit word embedding features from our final model, as the negligible increase in performance was not worth the increased complexity and loss of interpretability. In future work, we can try incorporating diagnostic groupers such as Clinical Classification Software [23] or try other word embedding frameworks such as BERT.

Our study has several limitations. First, our data were pulled from a single center, which limits generalizability of our model to other organizations. However, given the relative ubiquity of our most significant features, this may be a blueprint for training similar models at other centers. Second, our data come from a tertiary care academic medical center, which may not generalize to private or county hospital systems. Third, although we attempted to use broad categories of features in our dataset, it does not include unstructured clinical note data, which may contain key information. Fourth, our current model was trained and tested on retrospective data, which may not be applicable to current practice, although we mitigate this limitation by using the most recent admission data as the test set. Fifth, the random date-shift method used to deidentify the dataset makes

it impossible to determine exactly when events related to COVID-19 started. Sixth, we did not exclude encounters with an AMA discharge disposition (<1% of the encounters). Seventh, we did not use LASSO or ElasticNet when comparing logistical regression to tree-based ML models.

Future work will focus on prospectively validating our model at our local institution. Our original work was conducted on deidentified data, which limited the availability of some data types. We are already in the process of retraining our model on live EHR data, which will give us the ability to differentiate between planned and unplanned readmissions as well as incorporate more data on social determinants of health and discharge metadata (e.g., discharge on weekend or holiday, month of year). Once implemented, we plan to use the model to risk-stratify patients based on their readmission risk after hospital discharge and enroll high-risk patients into targeted post-discharge support programs. Our hope is that this will lead to significant decreases in 30-day hospital readmissions and act as a template for other health systems in the US.

5. Conclusions

Accurately predicting the risk of readmissions for hospitalized patients can enable targeting of post-discharge interventions to reduce readmissions and improve quality of care, patient experience, and hospital reimbursement. We developed and internally validated a supervised ML model using XGBoost to predict 30-day readmissions in a US healthcare system. Our model achieves a higher AUC than other US-based readmission prediction models. Major advantages of the model include the use of clinical and administrative data rather than administrative data alone, selection of features available within 24 h of discharge, generalizability to multiple disease states, and high level of interpretability based on SHAP feature importance. These unique strengths make our model clinically relevant and feasible to operationalize within a health system to reduce readmissions.

Author Contributions: Conceptualization, X.L.; methodology, A.L.L. and S.W.; software, A.L.L.; validation, S.W., A.L.L. and X.L.; formal analysis, A.L.L.; investigation, X.L.; resources, X.L.; data curation, A.L.L., A.R., S.A.-A. and X.L.; writing—original draft preparation, A.R., A.L.L. and S.A.-A.; writing—review and editing, X.L., A.N.M., S.W. and A.R.; visualization, A.N.M.; supervision, S.W. and X.L.; project administration, A.N.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study because only deidentified data were used and human subjects were not involved.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to institution policy around deidentified data.

Acknowledgments: The authors acknowledge the use of the UCSF Information Commons computational research platform, developed and supported by UCSF Bakar Computational Health Sciences Institute.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Features Included in the Final Model.

Features		
patientage	PT_first_val	venous_last_val
financialclass	PT_last_val	Amphetamine

Table A1. *Cont.*

Features		
postalcode	PaCO2_min_val	Benzo
sex	PaCO2_max_val	SBP_min_val
firstrace	PaCO2_avg_val	SBP_max_val
ethnicity	PaCO2_first_val	SBP_avg_val
maritalstatus	PaCO2_last_val	SBP_first_val
preferredlanguage	PaO2_min_val	SBP_last_val
smokingstatus	PaO2_max_val	DBP_min_val
Readmission	PaO2_avg_val	DBP_max_val
admissionsource	PaO2_first_val	DBP_avg_val
dischargetimeofdaykey	PaO2_last_val	DBP_first_val
admissiontype	PvCO2_min_val	DBP_last_val
inpatientlengthofstayindays	PvCO2_max_val	oxygen_amt_min_val
dischargedisposition	PvCO2_avg_val	oxygen_amt_max_val
departmentkey	PvCO2_first_val	oxygen_amt_avg_val
hospitalservice	PvCO2_last_val	oxygen_amt_first_val
admittingprovidertype	Urea_min_val	oxygen_amt_last_val
admittingproviderprimaryspecialty	Urea_max_val	5_class_oxygen_device_min_val
principalproblemdiagnosisname	Urea_avg_val	5_class_oxygen_device_avg_val
cnt_procedure	Urea_first_val	oxygen_device_min_val
days_since_last_admission	Urea_last_val	oxygen_device_max_val
admissions_previous_year	WBC_min_val	oxygen_device_avg_val
admissions_previous_2_years	WBC_max_val	oxygen_device_first_val
admissions_previous_90 days	WBC_avg_val	oxygen_device_last_val
admissions_previous_180_days	WBC_first_val	SP_O2_min_val
arrivalmethod	WBC_last_val	SP_O2_max_val
acuitylevel	arterial_min_val	SP_O2_avg_val
primarychiefcomplaintname	arterial_max_val	SP_O2_first_val
primaryeddiagnosisname	arterial_avg_val	SP_O2_last_val
edvisits_last_year	arterial_first_val	pulse_min_val
edvisits_last_2_years	arterial_last_val	pulse_max_val
edvisits_last_90_days	creatinie_min_val	pulse_avg_val
SLP consult	creatinie_max_val	pulse_first_val
Nutrition consult	creatinie_avg_val	pulse_last_val
SLP plan order	creatinie_first_val	r_number_ppl_assist_min_val
Observation status	creatinie_last_val	r_number_ppl_assist_max_val
Palliative care consult	eGFRhigh_min_val	r_number_ppl_assist_avg_val
5150 order	eGFRhigh_avg_val	r_number_ppl_assist_first_val
Psych consult	eGFRhigh_first_val	r_number_ppl_assist_last_val
Social work consult	eGFRhigh_last_val	R ED RISK OF FALL ADULT SCORE_min_val
DNR/DNI order	eGFRlow_min_val	R ED RISK OF FALL ADULT SCORE_first_val
Home health order	eGFRlow_max_val	R IP STRATIFY MOBILITY SCORE_avg_val
Cardiology consult	eGFRlow_avg_val	R IP STRATIFY MOBILITY SCORE_first_val
SNF discharge order	eGFRlow_first_val	R IP STRATIFY TOTAL SCORE_max_val
Inpatient psychiatry order	eGFRlow_last_val	R IP STRATIFY TOTAL SCORE_avg_val
SNF discharge attending contact	glucose_min_val	R IP STRATIFY TOTAL SCORE_first_val
ALP_min_val	glucose_max_val	R IP STRATIFY TRANSFER AND MOBILITY SUM_min_val
ALP_max_val	glucose_avg_val	R IP STRATIFY TRANSFER AND MOBILITY SUM_avg_val

Table A1. *Cont.*

Features		
ALP_avg_val	glucose_first_val	R IP STRATIFY TRANSFER AND MOBILITY SUM_first_val
ALP_first_val	glucose_last_val	R IP STRATIFY TRANSFER SCORE_min_val
ALP_last_val	hemoglobin_min_val	R IP STRATIFY TRANSFER SCORE_max_val
ALT_min_val	hemoglobin_max_val	R IP STRATIFY TRANSFER SCORE_avg_val
ALT_max_val	hemoglobin_avg_val	R IP STRATIFY TRANSFER SCORE_first_val
ALT_avg_val	hemoglobin_first_val	R NU-DESC DISORIENTATION_max_val
ALT_first_val	hemoglobin_last_val	R NU-DESC DISORIENTATION_avg_val
ALT_last_val	pH_min_val	R NU-DESC DISORIENTATION_first_val
AST_min_val	pH_max_val	R NU-DESC DISORIENTATION_last_val
AST_max_val	pH_avg_val	R NU-DESC INAPPROPRIATE BEHAVIOR_avg_val
AST_avg_val	pH_first_val	R NU-DESC INAPPROPRIATE BEHAVIOR_last_val
AST_first_val	pH_last_val	R NU-DESC INAPPROPRIATE COMMUNICATION_max_val
AST_last_val	platelets_min_val	R NU-DESC INAPPROPRIATE COMMUNICATION_avg_val
Albumin_min_val	platelets_max_val	R NU-DESC PSYCHOMOTOR RETARDATION_avg_val
Albumin_max_val	platelets_avg_val	R NU-DESC PSYCHOMOTOR RETARDATION_first_val
Albumin_avg_val	platelets_first_val	R NU-DESC SCORE V2_max_val
Albumin_first_val	platelets_last_val	R NU-DESC SCORE V2_avg_val
Albumin_last_val	potassium_min_val	R NU-DESC SCORE V2_first_val
BNP_min_val	potassium_max_val	R NU-DESC SCORE V2_last_val
BNP_max_val	potassium_avg_val	RESPIRATIONS_min_val
BNP_avg_val	potassium_first_val	RESPIRATIONS_max_val
BNP_first_val	potassium_last_val	RESPIRATIONS_avg_val
Bicarb_min_val	sodium_min_val	RESPIRATIONS_first_val
Bicarb_max_val	sodium_max_val	RESPIRATIONS_last_val
Bicarb_avg_val	sodium_avg_val	TEMPERATURE_min_val
Bicarb_first_val	sodium_first_val	TEMPERATURE_max_val
Bicarb_last_val	sodium_last_val	TEMPERATURE_avg_val
Bilirubin_min_val	troponin_min_val	TEMPERATURE_first_val
Bilirubin_max_val	troponin_max_val	TEMPERATURE_last_val
Bilirubin_avg_val	troponin_avg_val	year_discharge_date
Bilirubin_first_val	troponin_first_val	
Bilirubin_last_val	venous_min_val	

Table A1. *Cont.*

Features	
PT_min_val	venous_max_val
PT_max_val	venous_avg_val
PT_avg_val	venous_first_val

Appendix B

XGBoost model parameters:

- Learning objective: 'binary:logistic'
- Learning rate: 0.1
- Maximum depth: 5
- Number of trees: 100
- Scale_pos_weight: 6.08
- Evaluation Metric: AUC-PR

Gradient boosting model parameters:

- Minimum sample leafs: 98
- Maximum features: 0.152
- Maximum depth: 8
- Number of trees: 100
- Learning rate: 0.1

Random forest parameters:

- n_estimators: 250
- min_samples_leaf: 98
- max_features: 0.152
- max_depth: 8

Logistic regression parameters:

- default parameters from sklearn library, LogisticRegression module.

References

1. Hospital Readmissions Reduction Program (HRRP) | CMS. Available online: https://www.cms.gov/Medicare/Medicare- (accessed on 20 July 2022).
2. Auerbach, A.D.; Kripalani, S.; Vasilevskis, E.E.; Sehgal, N.; Lindenauer, P.K.; Metlay, J.P.; Fletcher, G.; Ruhnke, G.W.; Flanders, S.A.; Kim, C.; et al. Preventability and causes of readmissions in a national cohort of general medicine patients. *JAMA Intern. Med.* **2016**, *176*, 484–493. [CrossRef] [PubMed]
3. Becker, C.; Zumbrunn, S.; Beck, K.; Vincent, A.; Loretz, N.; Müller, J.; Amacher, S.A.; Schaefert, R.; Hunziker, S. Interventions to Improve Communication at Hospital Discharge and Rates of Readmission: A Systematic Review and Meta-analysis. *JAMA Netw. Open* **2021**, *4*, e2119346. [CrossRef] [PubMed]
4. Kripalani, S.; Theobald, C.N.; Anctil, B.; Vasilevskis, E.E. Reducing hospital readmission rates: Current strategies and future directions. *Annu. Rev. Med.* **2014**, *65*, 471–485. [CrossRef] [PubMed]
5. Lo, Y.-T.; Liao, J.C.; Chen, M.-H.; Chang, C.-M.; Li, C.-T. Predictive modeling for 14-day unplanned hospital readmission risk by using machine learning algorithms. *BMC Med. Inf. Decis. Mak.* **2021**, *21*, 288. [CrossRef]
6. Li, Q.; Yao, X.; Échevin, D. How Good Is Machine Learning in Predicting All-Cause 30-Day Hospital Readmission? Evidence From Administrative Data. *Value Health* **2020**, *23*, 1307–1315. [CrossRef]
7. Allam, A.; Nagy, M.; Thoma, G.; Krauthammer, M. Neural networks versus Logistic regression for 30 days all-cause readmission prediction. *Sci. Rep.* **2019**, *9*, 9277. [CrossRef]
8. Mišić, V.V.; Gabel, E.; Hofer, I.; Rajaram, K.; Mahajan, A. Machine learning prediction of postoperative emergency department hospital readmission. *Anesthesiology* **2020**, *132*, 968–980. [CrossRef]
9. AHA Guide. Available online: https://guide.prod.iam.aha.org/guide/hospitalProfile/6930043 (accessed on 14 November 2022).
10. Chen, Q.; Peng, Y.; Lu, Z. BioSentVec: Creating sentence embeddings for biomedical texts. In Proceedings of the 2019 IEEE International Conference on Healthcare Informatics (ICHI), Xi'an, China, 10–13 June 2019; pp. 1–5. [CrossRef]
11. Zhang, Y.; Chen, Q.; Yang, Z.; Lin, H.; Lu, Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci. Data* **2019**, *6*, 52. [CrossRef]

12. Parr, T.; Turgutlu, K.; Csiszar, C.; Howard, J. Beware Default Random Forest Importances. Available online: https://explained.ai/rf-importance/ (accessed on 27 October 2022).
13. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
14. Liu, X.; Anstey, J.; Li, R.; Sarabu, C.; Sono, R.; Butte, A.J. Rethinking PICO in the Machine Learning Era: ML-PICO. *Appl. Clin. Inform.* **2021**, *12*, 407–416. [CrossRef]
15. Hyndman, R.J.; Athanasopoulos, G. Forecasting: Principles and Practice, 2nd ed. Available online: https://otexts.com/fpp2/ (accessed on 27 October 2022).
16. Omphalos. Uber's Parallel and Language-Extensible Time Series Backtesting Tool | Uber Blog. Available online: https://www.uber.com/blog/omphalos/ (accessed on 27 October 2022).
17. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**. Available online: https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (accessed on 20 July 2022).
18. Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; Lakkaraju, H. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. *arXiv* **2019**, arXiv:1911.02508. [CrossRef]
19. Shapley, L.S. A value for n-person games. In *The Shapley Value: Essays in Honor of Lloyd S. Shapley*; Roth, A.E., Ed.; Cambridge University Press: Cambridge, UK, 1988; pp. 31–40. [CrossRef]
20. Sundararajan, M.; Najmi, A. The many Shapley values for model explanation. *arXiv* **2019**, arXiv:1908.08474. [CrossRef]
21. Štrumbelj, E.; Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **2014**, *41*, 647–665. [CrossRef]
22. Huang, Y.; Talwar, A.; Chatterjee, S.; Aparasu, R.R. Application of machine learning in predicting hospital readmissions: A scoping review of the literature. *BMC Med. Res. Methodol.* **2021**, *21*, 96. [CrossRef] [PubMed]
23. Cohen, M.F.; Irie, S.M.; Russo, C.A.; Pav, V.; O'Connor, S.L.; Wensky, S.G. Lessons Learned in Providing Claims-Based Data to Participants in Health Care Innovation Models. *Am. J. Med. Qual.* **2019**, *34*, 234–242. [CrossRef] [PubMed]
24. Eckert, C.; Nieves-Robbins, N.; Spieker, E.; Louwers, T.; Hazel, D.; Marquardt, J.; Solveson, K.; Zahid, A.; Ahmad, M.; Barnhill, R.; et al. Development and Prospective Validation of a Machine Learning-Based Risk of Readmission Model in a Large Military Hospital. *Appl. Clin. Inform.* **2019**, *10*, 316–325. [CrossRef]
25. Ko, M.; Chen, E.; Agrawal, A.; Rajpurkar, P.; Avati, A.; Ng, A.; Basu, S.; Shah, N.H. Improving hospital readmission prediction using individualized utility analysis. *J. Biomed. Inform.* **2021**, *119*, 103826. [CrossRef]
26. Schiltz, N.K.; Dolansky, M.A.; Warner, D.F.; Stange, K.C.; Gravenstein, S.; Koroukian, S.M. Impact of instrumental activities of daily living limitations on hospital readmission: An observational study using machine learning. *J. Gen. Intern. Med.* **2020**, *35*, 2865–2872. [CrossRef]
27. Papanicolas, I.; Riley, K.; Abiona, O.; Arvin, M.; Atsma, F.; Bernal-Delgado, E.; Bowden, N.; Blankart, C.R.; Deeny, S.; Estupiñán-Romero, F.; et al. Differences in health outcomes for high-need high-cost patients across high-income countries. *Health Serv. Res.* **2021**, *56*, 1347–1357. [CrossRef]
28. Shah, A.A.; Devana, S.K.; Lee, C.; Bugarin, A.; Lord, E.L.; Shamie, A.N.; Park, D.Y.; van der Schaar, M.; SooHoo, N.F. Prediction of Major Complications and Readmission After Lumbar Spinal Fusion: A Machine Learning-Driven Approach. *World Neurosurg.* **2021**, *152*, e227–e234. [CrossRef]
29. Hassan, A.M.; Lu, S.-C.; Asaad, M.; Liu, J.; Offodile, A.C.; Sidey-Gibbons, C.; Butler, C.E. Novel Machine Learning Approach for the Prediction of Hernia Recurrence, Surgical Complication, and 30-Day Readmission after Abdominal Wall Reconstruction. *J. Am. Coll. Surg.* **2022**, *234*, 918–927. [CrossRef] [PubMed]
30. Li, L.; Wang, L.; Lu, L.; Zhu, T. Machine learning prediction of postoperative unplanned 30-day hospital readmission in older adult. *Front. Mol. Biosci.* **2022**, *9*, 910688. [CrossRef]
31. Darabi, N.; Hosseinichimeh, N.; Noto, A.; Zand, R.; Abedi, V. Machine Learning-Enabled 30-Day Readmission Model for Stroke Patients. *Front. Neurol.* **2021**, *12*, 638267. [CrossRef]
32. Lineback, C.M.; Garg, R.; Oh, E.; Naidech, A.M.; Holl, J.L.; Prabhakaran, S. Prediction of 30-Day Readmission After Stroke Using Machine Learning and Natural Language Processing. *Front. Neurol.* **2021**, *12*, 649521. [CrossRef]
33. Hoffman, M.K.; Ma, N.; Roberts, A. A machine learning algorithm for predicting maternal readmission for hypertensive disorders of pregnancy. *Am. J. Obstet. Gynecol. MFM* **2021**, *3*, 100250. [CrossRef] [PubMed]
34. Frizzell, J.D.; Liang, L.; Schulte, P.J.; Yancy, C.W.; Heidenreich, P.A.; Hernandez, A.F.; Bhatt, D.L.; Fonarow, G.C.; Laskey, W.K. 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*, 204–209. [CrossRef]
35. Mortazavi, B.J.; Downing, N.S.; Bucholz, E.M.; Dharmarajan, K.; Manhapra, A.; Li, S.X.; Negahban, S.N.; Krumholz, H.M. Analysis of machine learning techniques for heart failure readmissions. *Circ. Cardiovasc. Qual. Outcomes* **2016**, *9*, 629–640. [CrossRef] [PubMed]
36. Wu, Y.-K.; Lan, C.-C.; Tzeng, I.-S.; Wu, C.-W. The COPD-readmission (CORE) score: A novel prediction model for one-year chronic obstructive pulmonary disease readmissions. *J. Formos. Med. Assoc.* **2021**, *120*, 1005–1013. [CrossRef]
37. Goto, T.; Jo, T.; Matsui, H.; Fushimi, K.; Hayashi, H.; Yasunaga, H. Machine Learning-Based Prediction Models for 30-Day Readmission after Hospitalization for Chronic Obstructive Pulmonary Disease. *COPD J. Chronic Obstr. Pulm. Dis.* **2019**, *16*, 338–343. [CrossRef] [PubMed]
38. Rajkomar, A.; Dean, J.; Kohane, I. Machine learning in medicine. *N. Engl. J. Med.* **2019**, *380*, 1347–1358. [CrossRef]

39. Burkov, A. *The Hundred-Page Machine Learning Book*; Burkov, A., Ed.; Anton Burkov: Québec, QC, Canada, 2019; p. 160.
40. Breiman, L. Arcing the Edge. *Ann. Prob.* **1998**, *26*, 1683–1702.
41. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **2001**, *29*, 1189–1232. [CrossRef]
42. Mason, L.; Baxter, J.; Bartlett, P.; Frean, M. *Boosting Algorithms as Gradient Descent*; MIT Press: Cambridge, MA, USA, 1990.
43. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, NY, USA, 2009; pp. 106–119. [CrossRef]
44. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
45. Gryczynski, J.; Nordeck, C.D.; Welsh, C.; Mitchell, S.G.; O'Grady, K.E.; Schwartz, R.P. Preventing hospital readmission for patients with comorbid substance use disorder: A randomized trial. *Ann. Intern. Med.* **2021**, *174*, 899–909. [CrossRef] [PubMed]
46. Kaya, S.; Sain Guven, G.; Aydan, S.; Toka, O. Predictors of hospital readmissions in internal medicine patients: Application of Andersen's Model. *Int. J. Health Plann. Manag.* **2019**, *34*, 370–383. [CrossRef] [PubMed]
47. Cruz, P.L.M.; Soares, B.L.d.M.; da Silva, J.E.; Lima, E.; Silva, R.R.d. Clinical and nutritional predictors of hospital readmission within 30 days. *Eur. J. Clin. Nutr.* **2022**, *76*, 244–250. [CrossRef]
48. Arnaud, É.; Elbattah, M.; Gignon, M.; Dequen, G. Deep Learning to Predict Hospitalization at Triage: Integration of Structured Data and Unstructured Text. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 4836–4841. [CrossRef]