Comparing Global with Disease specific Machine-learned Readmission Prediction Models

Aansh Shah, MSc,[1] Jerome Ramos, ScB,[1]
Michael Littman, PhD,[1] Laura Mercurio, MD,[2,3] Indra Neil Sarkar, PhD, MLIS[2], Carsten Eickhoff, PhD[1,2]
[1]Department of Computer Science, Brown University, Providence, RI, USA
[2]Center for Biomedical Informatics, Brown University, Providence, RI, USA
[3]Department of Emergency Medicine, Brown University, Providence, RI, USA

Keywords: readmissions, prediction, machine learning, patient outcome, clustering

ABSTRACT

Objective: To create disease specific machine learning models to improve 30-day readmissions and better understand readmission patterns of disease specific subgroups using Electronic Healthcare Records.

Materials and Methods: Three different groupings of data from the Medical Information Mart for Intensive Care (MIMIC-III) database were generated: global clusters including all patients, unsupervised clusters created with KModes using CCS codes, and clinical clustering based on hierarchical ICD9 codes. LACE, Logistic Regression, Random Forest, Neural Network, and XGBoost classifiers were trained on the global cluster as baselines. The best performing model across all patients was used to train on each of the individual clusters. The performance differences between the global and disease specific models were then evaluated in terms of precision, recall and F1-score.

Results: A global XGBoost model outperformed all the baseline classifiers across most of the metrics and was used to generate disease specific models for each cluster. Disease specific models do not increase the overall readmission classification performance compared to a global model trained on the entire dataset. When the patients were separated into disease specific clusters, a universally better disease specific model did not emerge. For the positive (i.e., readmission) class, XGBoost performed better in K-modes across F1-scores and recall (0.33 vs. 0.25; 0.64 vs. 0.34) but was slightly less performant in terms of precision (0.23 vs. 0.26). For the negative class, XGBoost performed better in the clinical, disease specific, cluster than in K-modes across F1-score (0.91 vs. 0.79) and precision (0.89 vs. 0.7) but was equally performant in recall (0.94 vs. 0.94). The global model had the highest recall (0.66) and F1-score (0.33) among readmitted patients but had a slightly lower precision (0.22). Similarly, among the negative class, the global model was tied for the highest precision (0.94) and reported the second highest F1-score (0.81).

Discussion: Models based on clusters of patients with similar diagnoses did not yield substantially better results than modeling the entire dataset. In particular, performance was not improved when patients were clustered by CCS codes. K-modes performed marginally better than the clinical clusterings, but the difference between global and disease specific models was negligible. The performance of disease specific clustering does not justify the additional computational and methodological costs involved with this type of approach. Our study found XGBoost to be the preferred algorithm for the global model, as compared to neural network or deep learning-based approaches, due to faster training times and higher precision-recall and ROC scores.

Conclusion: The global XGBoost scheme was found to be better at predicting readmissions than the LACE, Logistic Regression, Random Forest, and Neural Networks baseline classifiers. Disease specific models did not yield significantly better results than modeling the entire dataset in terms of precision, recall, and F1-score. Overall, the results of this study demonstrate that unsupervised clustering approaches, in combination with classification techniques, do not improve overall model performance and, consequently, do not identify better features of interest.

BACKGROUND AND SIGNIFICANCE

The intensive care unit (ICU) provides complex, life-saving interventions for critically ill patients; in the United States, ICU stays can cost over $81.7 billion per year.[1,2] As such, unnecessary admissions to ICUs come at significant cost not only to the medical system, but also to patients, including hospital-acquired infection and over-treatment.[3] Prior literature suggests that up to 76% of ICU readmissions are avoidable.[4] Yet, premature discharge from the ICU exposes patients to insufficient monitoring and an increased potential for delay in diagnosis, recognition of complication, and timely treatment.[3] Enhanced predictive models could save on treatment costs and improve the health outcomes for patients. To date, researchers have explored a wide range of global readmission classifiers that conflate all possible diagnoses into a single model. Clinical intuition would suggest that disease specific models should offer a more differentiated view and possibly result in increased predictive performance. To date, no such disease specific readmission prediction models have been evaluated broadly.

The industry standard readmission scoring model is LACE, which stands for length of stay, acuity of admission (i.e., if the patient was emergently admitted to the hospital), comorbidity scores (i.e., Charlson Comorbidity score) and the number of emergency department visits in the past six months.[5] Past work has extensively examined ICU readmission predictions using traditional approaches, including logistic regression and random forest classifiers, while recent advances suggest that neural networks outperform these traditional methods.[6] Studies to date also generally analyze entire datasets, with minimal evaluation of potential influences within sub-cohorts. Unsupervised learning techniques represent a novel way to cluster a study population into sub-cohorts, partitioning the full dataset into clusters of patients with similar diagnoses. This technique has been applied in other medical contexts such as identifying pregnancy comorbidities, but not towards predicting ICU readmission rates.[7] This approach enables researchers to more directly correlate feature importance to a particular patient diagnosis in contrast to the global models, which have historically failed to capture the complexity of relationships across the patient population.

The aim of this study is to create disease specific machine learning models to improve recognition accuracy of 30-day readmissions and better understand readmission patterns of disease specific subgroups. We used unsupervised learning methods to create disease specific models and compared them to both global model performance and clinically relevant clusters.

MATERIALS AND METHODS

In this study, three different groupings of data were generated: global clusters (a single cluster encompassing all patients), unsupervised data-driven clusters, and clinical clustering based on hierarchical ICD9 codes (Figure 1). The best performing model across all patients, the single global cluster, was trained on each of the individual clusters. The performance differences between the global and disease specific models were then evaluated.

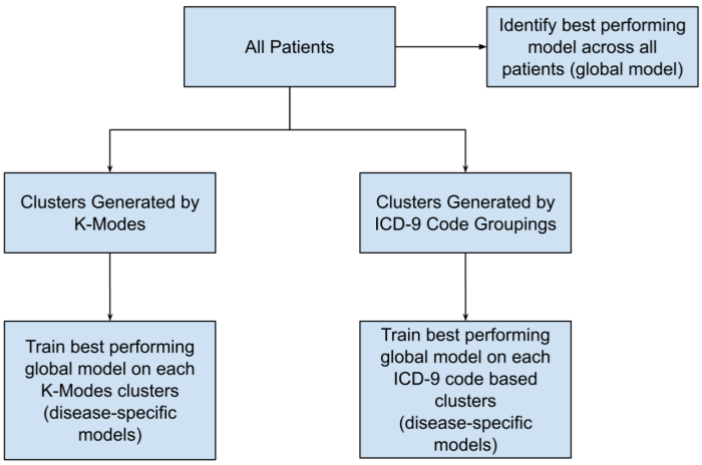

Figure 1. Dataset Partitioning and Model Training Methodology Summary

Dataset Construction

The readmission dataset was created from the MIMIC-III (Medical Information for Intensive Care III) database,[8] which consists of health-related data from ICU episodes at Beth Israel Deaconess Medical Center in Boston between 2001 and 2012. The original dataset contained more than 60,000 hospital visits with about 40,000 patients. Entries that had missing or invalid values, duplicate entries, were not adults (18 years or older) and cases where the patient died during the stay were removed. Patients were then

categorized based on whether they were readmitted to the hospital within 30 days of their discharge. The patient demographics after preprocessing are displayed in Figure 2.

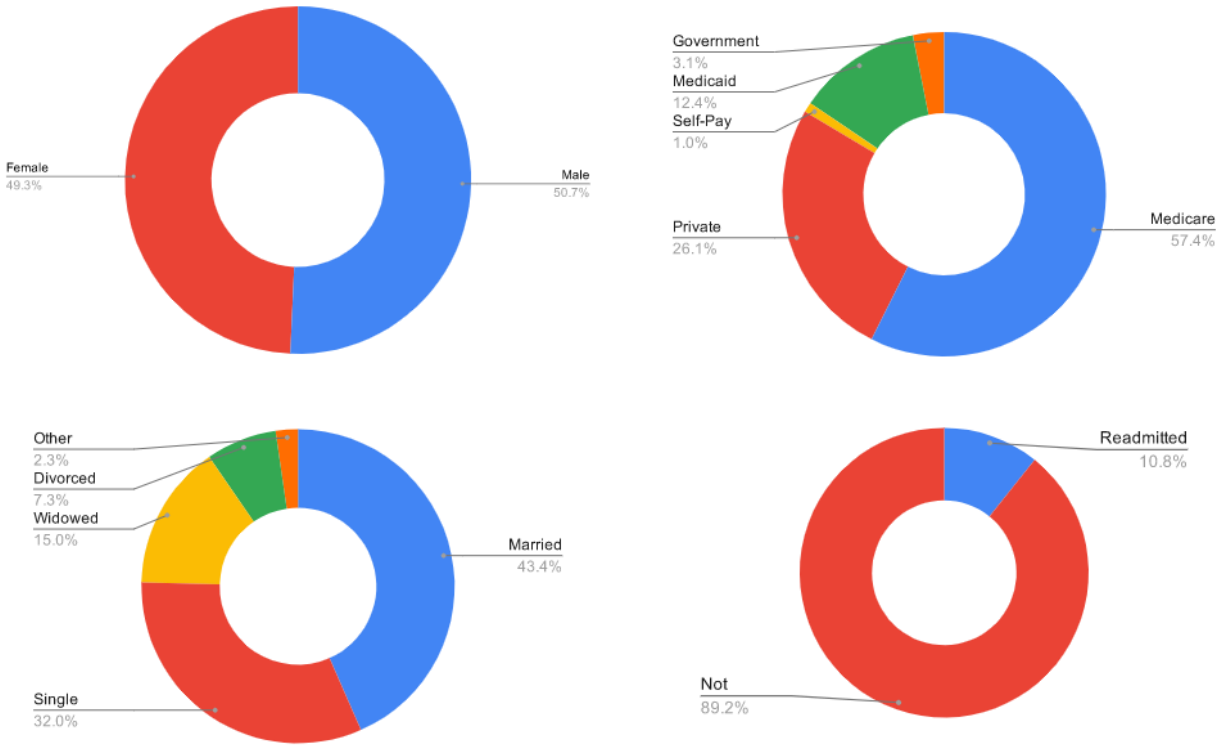

Figure 2. Patient Demographic Features (N_P=36,076) and Hospital Readmission Rate (N_R=58,702)

Feature Construction

Based on previously-published research,[5, 6, 9-13] a list of 61 features (Table 1) were considered for developing a readmission model. These features included information such as vital signs, laboratory studies, medication orders and dosages from the MIMIC III dataset. Since some clinical variables were measured multiple times throughout the admission (e.g., heart rates), we report minimum, maximum and mean values of those respective variables as features. Other features remained static over the course of the admission and only include a single feature (e.g., age).

Table 1. Mean and Standard Deviation for Continuous Features

Feature Name	Minimum Value	Mean Value	Max Value

Age	-	62.63 ± 16.45	-
Urea	14.95 ± 11.40	23.20 ± 16.47	33.66 ± 25.65
Platelets	172.27 ± 90.50	239.74 ± 110.07	335.61 ± 176.24
Magnesium	-	-	2.40 ± 0.87
Calcium	7.85 ± 0.79	-	-
Respiratory rate	10.38 ± 3.48	18.83 ± 3.38	30.66 ± 8.26
Glucose	91.80 ± 28.05	136.11 ± 109.40	272.66 ± 7744.38
Heart rate	65.69 ± 13.8	84.60 ± 12.98	111.36 ± 22.83
Systolic Blood Pressure	86.57 ± 18.07	121.51 ± 15.26	161.29 ± 26.97
Diastolic Blood Pressure	39.39 ± 11.70	61.29 ± 9.83	95.10 ± 25.11
Body Temperature	35.83 ± 0.79	36.89 ± 0.49	37.84 ± 0.80
Albumin	3.05 ± 0.73	-	-
Urine	33.55 ± 72.49	140.28 ± 694.30	677.30 ± 24904.36
Dobutamine	25.33 ± 61.32	100.02 ± 143.89	309.50 ± 590
Norepinephrine	0.1 ± 0.23	0.96 ± 1.44	5.24 ± 9.52
Phenylephrine	0.97 ± 2.88	8.61 ± 16.8	42.98 ± 103.26
Vasopressin	23 ± 35.77	43.06 ± 35.03	71.70 ± 57.52
Epinephrine	0.09 + 0.17	0.41 ± 0.73	2.35 ± 4.28
Dopamine	17.14 ± 78.27	95.38 ± 143.89	348.53 ± 590

Demographic information including age, race, assigned sex, and insurance information were also added for each patient. Aggregated metrics (SAPS-II, SOFA, LACE) were not included in the features, but instead were used as baseline comparators for model performance. SAPS-II (Simplified Acute Physiology) measured the severity of a patient's disease during their ICU stay.[15] SOFA (Sequential Organ Failure Assessment) determined the extent of a person's organ function or rate of failure.[16] As noted in the introduction, the LACE index is a metric used by hospitals to calculate a patient's risk of 30-day hospital readmission using the length of stay, acuity of admission, comorbidity scores and the number of emergency department visits in the past six months.[5]

In addition, chronic diseases are an important factor in determining later readmissions.[13] ICD-9-CM codes were used to capture diseases and procedures for each patient. Using Clinical Classification Software (CCS) codes from the Agency for Healthcare Research and Quality,[17] which clusters patient diagnoses and procedures into a manageable number of clinically meaningful categories, ICD-9-CM codes were grouped together. Mirroring the approach used in related work,[18] this "clinical grouper" made it easier to quickly understand categories of diagnoses and procedures for each ICU stay.

A 30-day timeframe was chosen for direct comparison with the widely used LACE score.

Model Generation

Four candidate models were generated. The first model was a logistic regression baseline model restricted to only LACE scores containing the four features that comprise LACE. For the second model, a multi-layer perceptron (MLP) neural network was used with the same architecture presented in Jamei et al.[9] The architecture consisted of a two-layer neural network, containing one dense hidden layer with half the size of the input layer, and dropout nodes between all layers to prevent overfitting. The neural network was trained in batches of 64 samples using the Adam optimizer, limiting training to 5 epochs. The third model consisted of a random forest classifier. For the fourth model, Extreme Gradient Boosting (XGBoost), a gradient-boosted decision tree, was used.[19] Gradient boosting is a method used to learn optimal tree hyperparameters. The method uses features in the data to determine the best leaves to split on in the tree, starting at a depth of 0, and uses regularization to improve model generalization.[19]

Stratified K-folds (k=5) were used for cross validation to generate our training and testing set. The models were fitted on the training sets and their performance evaluated on the testing set for each fold. The precision, recall, receiver operating curve (ROC), and F1-score were computed for the testing data at each fold for each of the models. A precision-recall curve was used for the primary analysis to compare the models and to account for the strong class imbalance. Area under the receiver operating curve (AUROC)

was recorded to compare to previous work. A grid search was then used to identify hyperparameters for each of the classifiers.

After model generation, independent feature importances were calculated using SHAP values, which use a 'game theory' approach to explain the output of arbitrary machine learning models.[6] In order to select samples to generate feature importances, K-means clustering was used to select 50 samples from the training set. These samples were used to compute the optimal local explanations (via TreeExplainer) for XGBoost with the entire test set, resulting in SHAP value feature importances.[20]

Clustering

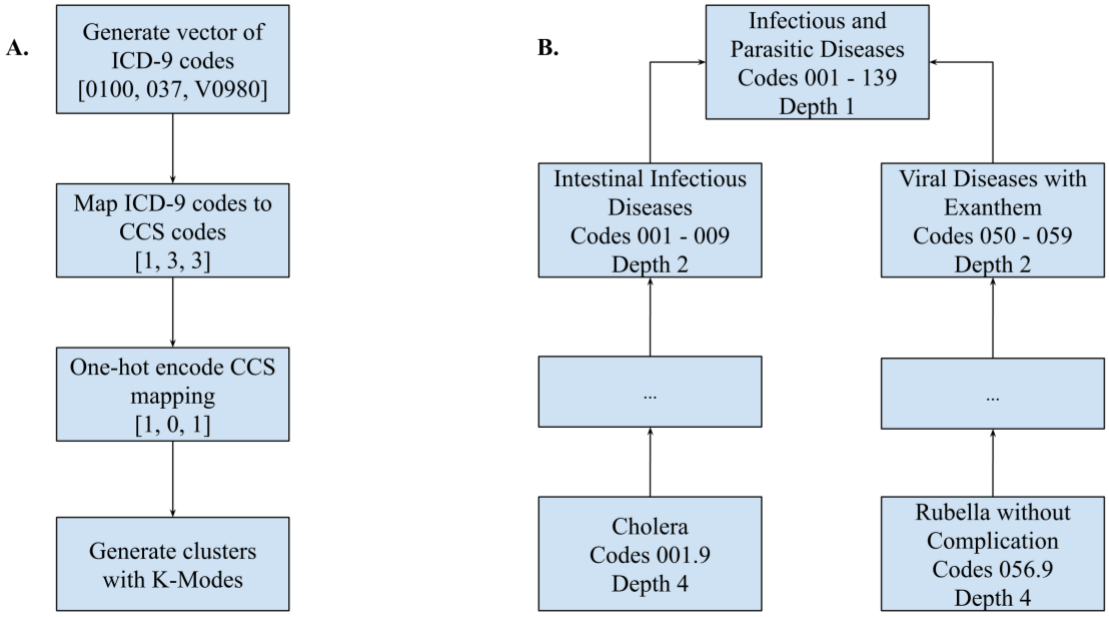

Figure 3. (A) K-Modes clustering and **(B)** Clinically informed ICD-9 code-based clustering

Unsupervised clustering algorithms were used to group patients by their ICD-9-CM codes (Figure 3A). 15,072 diagnostic ICD-9-CM codes contained within the patient population were mapped to 275 CCS codes and were one-hot encoded. Data clustering was performed using K-modes,[21] which tries to minimize the sum of "within-cluster Hamming distances" from the mode of the cluster, summed over all clusters. The best performing K value was selected (k=3) via the elbow method, a heuristic that

determines the appropriate amount of clusters by finding the point at which there is a marginal decrease in distortion if another cluster were to be added.[22] The Gini-coefficient, which measures the heterogeneity of the collection of clusters and explains how well the cluster collections reveal the true underlying cluster patterns, was used to determine cluster validity.[23]

In order to determine the clinical meaningfulness of each K-Modes cluster, patients were also clustered according to their most severe ICD-9-CM codes and were then grouped into clusters based on the disease hierarchy (Figure 3B).[24] For example, a patient with an ICD-9-CM code of 001.0 (Cholera due to vibrio cholerae) was grouped into 001-139 (Infectious and Parasitic Diseases). The top two depth levels of ICD-9 medical coding was considered.

RESULTS

The results of this study suggest that the use of disease specific models does not increase the overall readmission prediction performance compared to a global model trained on the entire dataset.

The K-Modes approach yielded three clusters with 10878, 19390, and 5808 patients, respectively. To understand the validity of the resulting clusters, we computed the Gini Coefficients over the patient's ICD-9 code distribution (Figure 4). The value of the Gini coefficient is between 0 and 1, with higher values indicating higher inequalities between the clusters. A Gini coefficient less than 0.4 indicates adequate equality, whereas a value above 0.4 shows a large gap.[23] The results suggest that the clusters contain a homogenous patient distribution because the values reported are below 0.4. The most common ICD-9 code groups per cluster are reported in Table 2. Clusters 0 and 1 contained the same ICD-9 code descriptions with varying frequencies, which means there was a different distribution of the same ICD-9 codes within the two clusters. Cluster 2, however, contained more Infectious Parasitic and Diseases Circulatory than the other two clusters. When patients were sorted into clusters based on disease hierarchy (Figure 3B), 16 and 92 clusters emerged at depths of one and two, respectively.

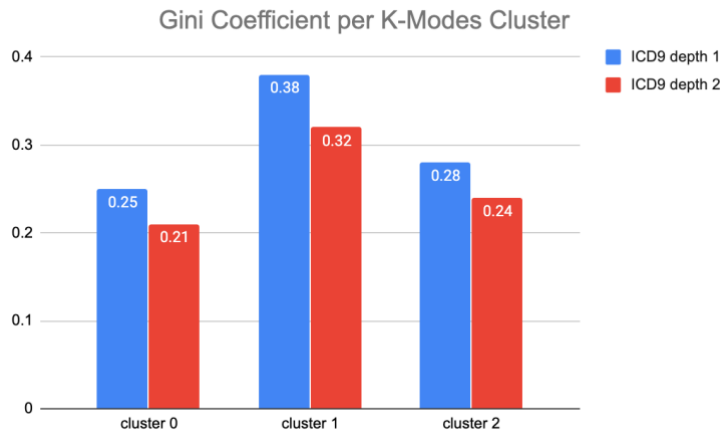

Figure 4. Cluster purity per k-modes cluster using Gini-coefficients and silhouette scores

Table 2. Most common ICD-9 descriptions ranked by frequency in each K-modes cluster

	Cluster 0	Cluster 1	Cluster 2
Rank	ICD-9 Description	ICD-9 Description	ICD-9 Description
1	Diseases Circulatory	Injury Poisoning	System Disease
2	Circulatory Systems	System Diseases	Diseases Circulatory
3	System Diseases	Diseases Circulatory	Circulatory System
4	Injury Poisoning	Circulatory System	Diseases Respiratory
5	Poisoning Diseases	Poisoning Diseases	Respiratory System
6	Diseases Digestive	Diseases Digestive	Injury Poisoning
7	Digestive System	Digestive System	Infectious Parasitic

Table 3. Global model classification broken down by class label

Name	Class	Precision	Recall	F1-Score
Logistic Regression	0	0.90	0.96	0.93
	1	0.27	0.11	0.16
	total	0.59	0.54	0.55
Random Forest	0	0.90	0.86	0.88
	1	0.13	0.17	0.14
	total	0.51	0.51	0.51
LACE	0	0.90	0.99	0.94
	1	0.34	0.03	0.05
	total	0.64	0.51	0.495
XGBoost	0	0.94	0.71	0.81
	1	0.22	0.66	0.33

	total	0.58	0.685	0.57
Neural Network	0	0.89	1.0	0.94
	1	0	0	0
	total	0.45	0.5	0.46

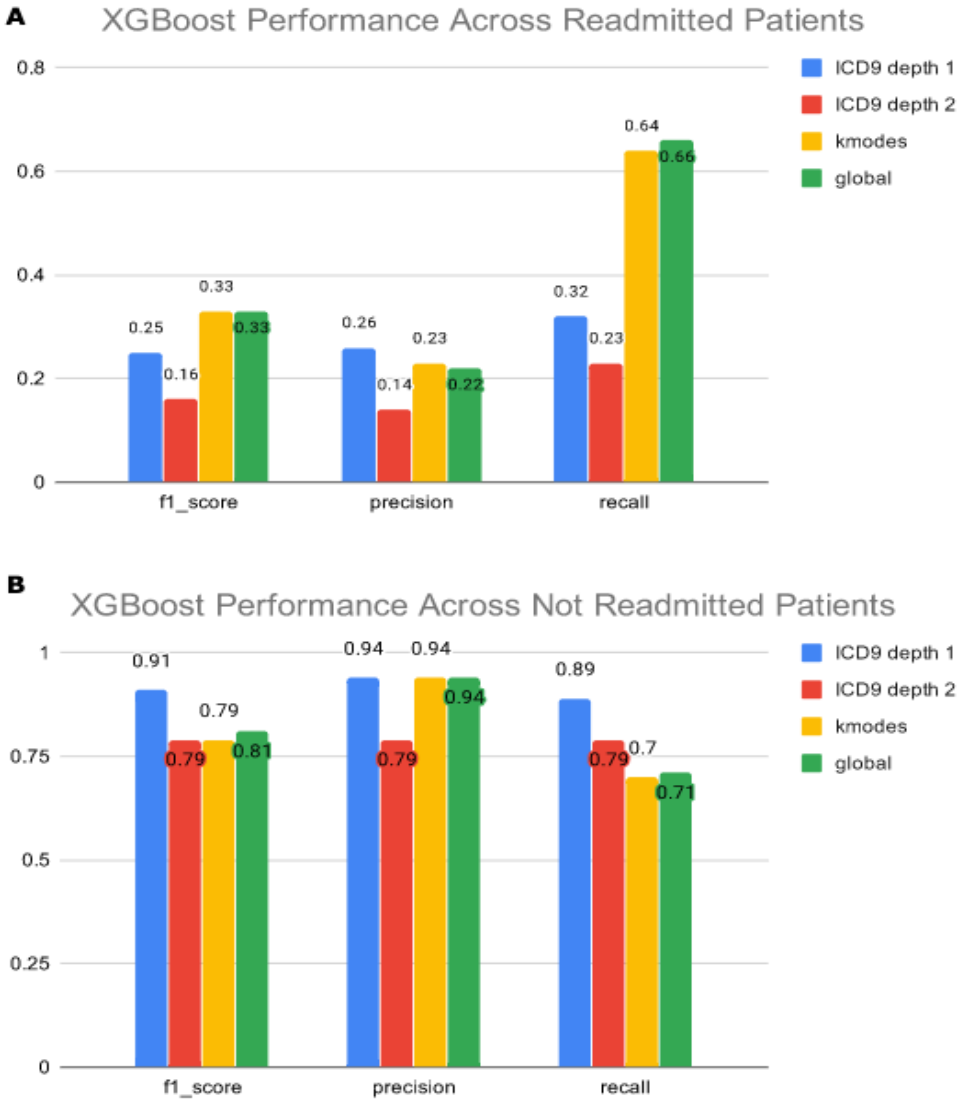

Figure 5. XGBoost model performance across different clusters.

A global XGBoost model outperformed other recently proposed approaches, as well as all the baseline classifiers across most of the metrics (Table 3). As such, XGBoost was chosen as the leading model and used to generate disease specific models for each cluster. After separating patients into disease specific

clusters, a universally better disease specific model did not emerge (Figure 5). For the positive (i.e., readmission) class, XGBoost performed better in K-modes across F1-scores and recall but was slightly less performant in terms of precision (Figure 5A). For the negative class, XGBoost performed better in the clinical, disease specific, cluster than in K-modes. The global model had the highest recall and F1-score among readmitted patients but had a slightly lower precision. Similarly, among the negative class, the global model was tied for the highest precision and reported the second highest F1-score (Figure 5B). The global and disease specific models placed varying importances on different features, though certain features emerged as important across all the clusters (Table 4). For example, Max Urea was ranked highly across all the clusters.

Table 4. Feature importances ranked by absolute SHAP values in each global and K-modes cluster

	Global	Cluster 0	Cluster 1	Cluster 2
Rank		Feature Name	Feature Name	Feature Name
1	Max Urea	Max Diastolic Blood Pressure	Min Albumin	Min Temperature
2	Max Respiratory Rate	Max Platelets	Max Diastolic Blood Pressure	Min Albumin
3	Max Heart Rate	Min Heart Rate	Max Urea	Mean Temperature
4	Max Diastolic Blood Pressure	Max Magnesium	Mean Temperature	Max Platelets
5	Min Diastolic Blood Pressure	Max Heart Rate	Min Temperature	Max Urea
6	Min Systolic Blood Pressure	Max Urea	Min Systolic Blood Pressure	Age
7	Max Platelets	Mean Temperature	Max Respiratory Rate	Min Glucose

DISCUSSION

Models based on clusters of patients with similar diagnoses did not yield significantly better results than modeling the entire dataset. In particular, performance was not improved when patients were clustered by CCS codes. K-modes performed marginally better than the clinically informed clusterings, but this difference between global and disease specific models was negligible. The observed performance of disease specific clustering does not justify the additional computational and methodological costs involved with this type of approach. Our study found XGBoost to be the preferred algorithm for the global model, as compared to neural network or deep learning-based approaches, due to faster training times and higher precision-recall and ROC scores.

Past research suggests that neural networks outperform traditional methods for predicting readmission rates in global models. Barbieri et. al benchmarked several deep learning architectures using attention based mechanisms, recurrent layers, neural ordinary differential equations and medical concept embeddings with time-aware attention using the MIMIC-III dataset.[26] A recurrent neural network, with time dynamics of code embeddings computed by neural ODEs, achieved the highest average precision of 0.331 (AUROC: 0.739, F1-score: 0.372). Rafi et al. used mimic learning, a Long Short Term Memory (LSTM) model for temporal features, and a feed-forward layer with many hidden layers for static features and reported an AUROC of 0.709.[10] The global XGBoost model had a higher precision-recall, 0.58, and ROC scores, 0.744, than the deep learning based architecture (Table 3). This suggests that XGBoost is better for modeling complex interactions of disease processes in critically ill patients than deep learning-based approaches. Deep learning approaches require a large number of samples; in these studies, often *N* is small (i.e., N < 100,000).

Past work has utilized weighted linear regressions and Gini importances to calculate the most important features.[19] However, these methods are known to suffer from inconsistency and are only approximations.[24] The TreeExplainer used in this study calculates the exact optimal local explanations

for features, which enables the measurement of interaction effects; this allows for more consistent, accurate, and faster feature importance calculations than linear and other tree-based models.[24]

Our study has several potential limitations. The preprocessed and filtered MIMIC dataset contained substantially more negative cases (i.e. 89.2% were not readmitted) than positive cases (i.e. 11.8% were readmitted). Models learned to predict 'no readmission' when they were optimized for accuracy, leading them to perform quite poorly for patients that should have been readmitted (i.e. false negatives); a model could learn to predict "no" for every patient and achieve an accuracy of 88%. Models that were optimized for sensitivity and specificity improved the classification for positive-case classification (i.e. readmissions). Class imbalances are more adequately captured in precision-recall curves than traditional ROC curves;[14] ROC metrics (false positive and true positive rate) measure the ability to distinguish between classes, which can remain falsely elevated despite class imbalance. In contrast, precision measures the probability of correct detection of positive values. As such, our precision-recall curves more accurately reflected the models' limitations.

Given that diagnosis codes are correlated with certain features, building a global model that can capture the complexity of this relationship is difficult. For each patient, one hot encoding was used to indicate whether the patient encounter contained a particular ICD-9-CM code. The resulting one hot encoded matrix was very sparse and high-dimensional because there were 15,072 total ICD-9-CM codes in the patient population, In order to reduce the dimensionality, similar diagnoses were *first* grouped together by mapping ICD-9-CM codes to the higher order CCS-code categories, followed by one-hot encoding. This reduced the number of codes in the population from 15,072 ICD-9 codes to a more manageable 275 CCS codes. By reducing the dimensionality of diagnosis, and then using K-modes, patients with similar diagnoses were clustered at a lower cost.

Comparing the study model with past research is difficult because risk models are highly dependent on their specific patient populations. Jamei et al.[9] reported a 24% average precision-recall on their top-performing model, but it was based upon a different dataset (SutterHealth), so the results are not directly

comparable. Applying the architecture created by Jamei et al. to the MIMIC-III dataset studied here yielded very poor results; the neural network was not able to predict any readmissions (Table 3). Finally, our study performed an extensive hyperparameter grid search, but was unable to find stable parameters. This finding suggests that neural networks are not always better than traditional machine learning classifiers.

Other papers that used the MIMIC-III dataset either predicted readmission rates within a different time period than this study, or insufficient information was provided to enable a proper comparison. Lin et al., Rojas et al., and Ricardo [5, 25, 11] used time-series information to make predictions within a more specific time period. Lin et al. examined predictions within a 48-hour window, and Ricardo used a 48-72-hour window. A 30-day timeframe was chosen in this study because LACE calculates readmissions within 30 days and this model was created to have a comparison with an industry standard scoring index. This limited comparability between MIMIC-III derived studies has been noted for other tasks such as mortality prediction.[27]

The global model outperformed LACE, in general, in recall (0.685 vs. 0.51) and F1-score (0.57 vs. 0.495), but had a lower precision (0.58 vs. 0.64) (Table 3). When separated by class label, the global model demonstrated significantly better readmission predictions across precision (0.22 vs. 0.27), recall (0.66 vs. 0.03), and F1-score (0.33 vs. 0.05). LACE was slightly more performant than the global model at predicting non-readmissions across the three metrics (0.94 vs. 0.90; 0.71 vs. 0.99; 0.81 vs 0.94). This suggests that the global model may be better than LACE due to its higher sensitivity towards patients that should be readmitted.

A large limitation in ICD-9-CM code-based clustering is that the codes may be incorrect due to a wide range of errors in diagnostic and procedure coding. The quantity and quality of communication between the patient and admitting clerk influence the accuracy of the admitting diagnosis. The patient may only describe a subset of symptoms or withhold information, resulting in compromised diagnostic accuracy. Likewise, clinicians may fail to ask the right questions, properly elicit the patient's history, or

misunderstand the patient's presentation. The clinician's diagnosis is further biased by diagnostic testing and procedures; interpretation of these test results may be compromised by test-specific limitations, errors in recording, or interpretation of test results. Finally, given the constantly evolving nature of medical knowledge and best practices, errors may occur in physicians using outdated diagnostic criteria.[21]

This paper focused on modeling readmissions within a 30-day time period; it would be interesting to explore other timeframes, especially models predicting readmission within a 24-72-hour time period. Data analyzed in this study come from a single site over an eleven-year period (2001-2012). Classifiers may not be generalizable to patients in another hospital or a time frame because of various differences in patient population, hospital resources, and availability of treatments or interventions. Further exploration of a multi-institutional, longitudinal data set would be required to generate a more comprehensive model.

Another future direction could involve reducing the feature space and reexamining classifiers on the reduced feature space; selecting a high number of variables may bias the dataset towards selecting patients having similar conditions that require their specific measurement/testing. Finally, given K-modes bias towards convex clusters, it would be worthwhile to explore other clustering methods such as density-based methods or self-organizing maps.

CONCLUSION

In this study, unsupervised learning methods were used to create disease specific models for predicting ICU readmissions within a 30-day period and were compared to global readmission classifiers. An XGBoost model was found to be better at predicting readmissions than LACE, Logistic Regression, Random Forest, and the Neural Networks described in state-of-the-art literature. Despite the additional effort invested into their creation, disease specific models did not yield significantly better results than modeling the entire dataset. Overall, the results of this study suggest that disease specific models derived

either via clinically meaningful or data-driven grouping of conditions does not positively affect readmission prediction performance and does therefore not justify the required effort.

References

1. Halpern NA, Pastores SM. Critical care medicine in the United States 2000–2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. Critical care medicine. 2010 Jan 1;38(1):65-71.
2. Johnson AE, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine learning and decision support in critical care. Proceedings of the IEEE. 2016 Feb;104(2):444-66.
3. Lin YW, Zhou Y, Faghri F, Shaw MJ, Campbell RH. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. PloS one. 2019;14(7).
4. Gerhardt G, Yemane A, Hickman P, Oelschlaeger A, Rollins E, Brennan N. Data shows reduction in Medicare hospital readmission rates during 2012. Medicare Medicaid Res Rev. 2013;3(2):E1-1.
5. Damery S, Combes G. Evaluating the predictive strength of the LACE index in identifying patients at high risk of hospital readmission following an inpatient episode: a retrospective cohort study. BMJ open. 2017 Jul 1;7(7):e016921.
6. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. InAdvances in neural information processing systems 2017 (pp. 4765-4774).
7. Chang J, Sarkar IN. Using Unsupervised Clustering to Identify Pregnancy Co-Morbidities. AMIA Summits on Translational Science Proceedings. 2019;2019:305.
8. Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Scientific data. 2016 May 24;3:160035.
9. Jamei M, Nisnevich A, Wetchler E, Sudat S, Liu E. Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. PloS one. 2017;12(7).
10. Rafi P, Pakbin A, Pentyala SK. Interpretable deep learning framework for predicting all-cause 30-day ICU readmissions. Texas A&M University. 2018.
11. Afonso RB. Feature Extraction and Selection for Prediction of ICU Patient's Readmission Using Artificial Neural Networks.
12. Xue Y, Klabjan D, Luo Y. Predicting ICU readmission using grouped physiological and medication trends. Artificial intelligence in medicine. 2019 Apr 1;95:27-37.
13. Brown SE, Ratcliffe SJ, Halpern SD. An empirical derivation of the optimal time interval for defining ICU readmissions. Medical care. 2013 Aug;51(8):706.
14. Fawcett T. ROC graphs: Notes and practical considerations for researchers. Machine learning. 2004 Mar 16;31(1):1-38.
15. Le Gall JR, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. Jama. 1993 Dec 22;270(24):2957-63.
16. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart CK, Suter P, Thijs LG. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure.
17. Wei WQ, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, Cox NJ, Roden DM, Denny JC. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PloS one. 2017;12(7).
18. Choi Y, Chiu CY, Sontag D. Learning low-dimensional representations of medical concepts. AMIA Summits on Translational Science Proceedings. 2016;2016:41.
19. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794).
20. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. Explainable ai for trees: From local explanations to global understanding. arXiv preprint arXiv:1905.04610. 2019 May 11.
21. Chaturvedi A, Green PE, Caroll JD. K-modes clustering. Journal of classification. 2001 Jan 1;18(1):35-55.

22. Aranganayagi S, Thangavel K. Improved k-modes for categorical clustering using weighted dissimilarity measure. International Journal of Computer, Electrical, Automation, Control and Information Engineering. 2009 Mar 25;3(3):729-35.
23. Han J, Zhu L, Kulldorff M, Hostovich S, Stinchcomb DG, Tatalovich Z, Lewis DR, Feuer EJ. Using Gini coefficient to determining optimal cluster reporting sizes for spatial scan statistics. International journal of health geographics. 2016 Dec 1;15(1):27.
24. O'malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health services research. 2005 Oct;40(5p2):1620-39.
25. Rojas JC, Carey KA, Edelson DP, Venable LR, Howell MD, Churpek MM. Predicting intensive care unit readmission with machine learning using electronic health record data. Annals of the American Thoracic Society. 2018 Jul;15(7):846-53.
26. Barbieri S, Kemp J, Perez-Concha O, Kotwal S, Gallagher M, Ritchie A, Jorm L. Benchmarking Deep Learning Architectures for predicting Readmission to the icU and Describing patients-at-Risk. Scientific Reports. 2020 Jan 24;10(1):1-0.
27. Johnson AE, Pollard TJ, Mark RG. Reproducibility in critical care: a mortality prediction case study. InMachine Learning for Healthcare Conference 2017 Nov 6 (pp. 361-376).