A Comparative Analysis of Data-Driven Modelling Techniques for 30-Day Heart Failure Readmission Prediction

Ryan Missel[1], Jesdin Raphael[1], Christopher Haggerty[2],
Dustin Hartzel[2], Jeffery Ruhl[2], Brandon Fornwalt[2], Linwei Wang[1]

[1] Rochester Institute of Technology, Rochester, NY, USA
[2] Geisinger Health System, Danville, PA, USA

Abstract

Unplanned readmissions due to heart failure are a major contributor to the overall annual healthcare costs associated with heart failure. It is anticipated that these costs will increase over 100% by the year 2030. Identifying patients who are at risk of readmission or mortality presents an opportunity to alleviate the burden on hospitals and improve patient outcomes. While there have been numerous studies exploring the use of machine-learning techniques on Electronic Health Records (EHR) data, only a subset of them have considered the temporal aspects of the data, and even fewer have conducted dedicated investigations. In this study, we utilize a dataset of EHR information obtained from a large-scale multi-center healthcare system to assess the effectiveness of extracting trends from time-series medical data using a range of machine-learning methods, spanning various model families. Our investigation incorporates statistical features, such as regression line coefficients and volatility metrics, and reveals notable enhancements in the predictive performance of all tested machine-learning models.

1. Introduction

Unplanned heart failure readmissions contribute to 80% of the yearly medical cost related to heart failure and is expected to rise from 30.7$ to 69.7$ billion dollars annually by 2030 [1]. By identifying at-risk patients for unplanned readmission or mortality, the burden on hospital resources may be reduced and avoidable suffering prevented [2]. Systemic advancements in the digitization of patient health recordings, called Electronic Health Records (EHR), have allowed for opportunities in large-scale modelling with machine-learning techniques.

Although there exists an extensive body of literature dedicated to exploring statistical methods for risk prediction using electronic medical records, many of these studies have failed to meet the standards of clinical acceptability in terms of metrics and interpretability. Furthermore, a significant issue is the lack of methodological consistency and variability in the candidate variables considered [3]. In many instances, researchers have employed sets of features that do not overlap, originating from disparate data sources and possessing distinct meanings. To illustrate, one study may focus on demographic and socioeconomic factors, while another may concentrate on qualitative mental health records and drug usage, despite both studies ultimately reporting equivalent predictive outcomes. This inconsistency in feature selection and utilization has hindered understanding the overall contribution of specific data types and modalities across the literature in the context of predictive tasks.

Among the various types of data in electronic health records (EHR), temporal EHR features, such as vital chart records and laboratory test results, have not been extensively studied. Recognizing the importance of investigating these features, our study empirically examines two sets of temporal EHR features in the context of predicting unplanned 30-day heart failure readmissions within a large-scale patient cohort. We derived predictive features from electronic health records, including both static variables and time-series measurements. Subsequently, we constructed a suite of 18 machine learning algorithms, notably including AdaBoost and Bagging. These algorithms were trained and evaluated using a 5-fold cross-validation approach across all possible combinations of static and temporal data sets. Our results are presented quantitatively for each feature combination, categorized by model families (e.g., non-parametric methods, ensemble methods, etc.), and highlight the top-performing methods within each category. Notably, our experimental findings underscore the significance of leveraging temporal candidate variables, as they led to a notable average improvement of 0.0325 in the Area Under the Curve (AUC), a commonly used metric, across all tested models.

Despite the growing popularity of deep learning in various domains, classical machine learning methods still hold relevance in readmission risk prediction. This emphasizes the need for further research into methods that can effectively incorporate and extract complex temporal information from EHR data.

2. Related Work

In a survey examining the characteristics of EHR data in risk prediction models [3], when specifically focusing on heart failure models, it was found that only a limited number of studies incorporated vital signs and laboratory results as potential candidate variables. Furthermore, among those studies, only a small subset considered the temporal aspects of these variables in any way. Most of them utilized features that represented average values [4] or the frequency of occurrences during the patient encounter [5].

An insightful analysis conducted by Kennedy et al. [6] delved into the extraction of temporal features from EHR data collected in pediatric intensive care units, particularly to model and predict cardiac arrest in short timeframes. This analysis explored various data formats, including binning techniques at different time resolutions (e.g., seconds before an event, per-minute, or hourly intervals). Additionally, the study discussed the extraction of latent features, such as identifying trends through regression line fitting.

Building on this research, Lin et al. [7] investigated statistical features derived from vital chart events and laboratory test results. These features were used as inputs for classical machine learning methods to predict heart failure readmission. Specifically, the study considered parameters obtained from regression lines fit to each time-series variable and sample, along with measures of how well these fits matched the data (i.e., R^2) and volatility metrics. Their findings demonstrated improvements over baseline methods when incorporating these extracted latent features into the prediction model.

3. Design, Setting, and Participants

A cohort of 8,802 patients in a set of 16,216 unplanned emergency department admissions between January 1, 2000 to January 1, 2020 with encounter diagnoses related to a set of heart failure-based ICD-9 codes were collected from the Geisinger Health System, a healthcare system composed of regional hospitals in the Greater Pennsylvania region. Analysis was performed between June 1, 2020 and March 16, 2023.

3.1. Inclusion/Exclusion Criteria

We excluded certain types of patient encounters from our dataset for the purpose of our analysis. Specifically, encounters involving patients under 21 years of age or pregnant, those who left against medical advice, patients transferred from another hospital, those with no recorded temporal features, or individuals who passed away during the visit were excluded. Additionally, we excluded patients who were solely treated in the ICU due to their considerably shorter average stays and representing a distinct patient distribution.

For an encounter to be included in our dataset, it had to be associated with an ICD-9 code related to heart failure, as defined by the eMERGE Heart Failure Phenotype [8], even if heart failure was not the primary diagnosis. We treated each encounter from a patient as an independent sample for our readmission prediction analysis.

3.2. End Point

In the predictive task, we defined a positive case as an encounter that either led to an unplanned emergency admission within 30 days after the recorded discharge date to an outpatient setting or resulted in mortality within 30 days. Within this specific patient cohort, we observed a positive readmission rate of 8.5%, which corresponds to 1,393 encounters, and a mortality rate of 6.0%, which accounts for 981 encounters. These rates are consistent with the typical ranges reported in existing literature [3].

3.3. Predictive Candidate Variables

In this section, we present the predictive features that have been extracted for training our models. To develop a practical readmission prediction model, it is essential to establish a foundational set of features beyond temporal attributes. While there are various data types that could potentially enhance the accuracy of our predictions, such as medication and procedure records, we deliberately focused on a more limited set of features to streamline our scope and concentrate on the most relevant variables. Consequently, we have categorized our candidate variables into three groups:

1. Static demographic features and encounter history derived from patients' general records, forming the core set of basis features.
2. Temporal vital chart events recorded by healthcare providers, either by digitized notes or direct electronic recordings.
3. Temporal laboratory test results processed by out-of-unit laboratories.

Of the larger set of available data archetypes (e.g. medication or procedures), vital chart recordings and lab results emerge as robust temporal candidates due to their frequent collection during patient encounters [7, 9].

For static features, we selected a set of 24 variables, including demographics (e.g. race, sex, age), socioeconomic factors (e.g. billed insurance type), and encounter history statistics (e.g. number of emergency visits in the last year).

For temporal features, we extracted 14 vital signals (e.g. diastolic/systolic blood pressure and SpO2) and 8 laboratory results (e.g. glucose, potassium, and hemoglobin). With an average admission length of stay at 3.5 ± 4.9 days of admission, we considered the last 72 hours of EHR data for temporal features, as the last 48 hours of admission prior to discharge are found to be the most informative [10].

3.4. Data Processing and Extraction

An important aspect of including temporal features into machine-learning methods is that most are not equipped to deal with matrix-level inputs but rather just a vector per sample. As such, care must be taken when leveraging them. Hourly binning and flattening the temporal features into a 1D vector leads to the computational and representational issues as the feature dimensions can quickly surpass the number of available samples in an EHR setting. Instead, we followed [7] and [6], in which statistical trends are extracted from the time-series. These included the parameters (slope a and intercept b of $y = ax + b$) of a linear regression fit on each sequence as well as the mean, maximum, and minimum values to represent a feature's volatility.

Data missingness in static features was set as the mean of the cohort for nominal features and to 0 for ordinal features. The cohort here refers only to the training cohort as that is the available information from which to impute at test-time. Given the multi-center nature of this cohort, for vital and laboratory results that were taken using more than one method, care was taken to ensure that the measurements were in the same unit scale and of compatible reporting. All features were z-score normalized across each feature. Individual samples were a feature measurement exceeded a threshold of 3 in a z-score test were removed.

Following feature extraction, we end up with three sets of features, representing the static, temporal vital, and temporal lab value results, which we respectively denote as S, V, and L in the following experiments. The total number of static features remained 24 while feature extraction increased the dimensionality of the vitals and lab results to 70 and 40 features respectively.

4. Methodology

For a thorough evaluation, we assessed a diverse set of 18 machine-learning methods drawn from different categories. These methods encompassed a wide spectrum, ranging from non-parametric models to ensemble tree methods. We include 4 families of interest: *(i)* non-parametric models (e.g. k-nearest neighbors classification), *(ii)* methods under the generalized linear framework (e.g. Ridge Classifier), *(iii)* vector machine methods (e.g. Nu-Support Vector Machine), and *(iv)* ensemble-based methods (e.g. Random Forest). We implemented and ran these methods using the Scikit-Learn library [11].

For each method, we conducted a grid search to fine-tune its hyperparameters. Each configuration underwent training using a 5-fold cross-validation procedure. The selection of the best configuration within each grid search was determined by the highest mean test score across the cross-validation folds. This scoring function was based on the area under the receiver operating curve (AUC).

5. Results

To provide clear and concise results, we calculated the mean and standard deviation of the AUC (Area Under the Curve) across all test-folds, specifically focusing on the highest-performing method on average within each model family. To assess the influence of temporal features on model performance, we trained each method on various combinations of feature sets $\{S, V, L\}$, including individual sets as well. As previously mentioned, we established a baseline model, denoted as the model suite trained solely on feature set S, which exhibited performance comparable to methods found in the literature [3]. In Table 1, we present the results achieved by the top-performing method within each model family.

The inclusion of both V and L trends demonstrated significant improvement over just using S for risk prediction. It is interesting to note that purely using these trends is not sufficient for meaningful prediction and that it is the combination of these features that is crucial, as evidenced by the disparity between individual S, V, L performances when compared to $S + V + L$. Including the V component on S, i.e. $S + V$, showed an average AUC improvement of 0.02 while including the L component, i.e. $S + L$, showed an average AUC improvement of 0.015. Including both temporal sets, $S + V + L$, improved the AUC on average by 0.0325.

Among the different model families examined, the ensemble methods stood out as the most robust overall performers. They consistently achieved a significantly higher average AUC across all the tested methods. This finding aligns with existing literature, which generally demonstrates the superior performance of ensemble methodologies compared to classical regression techniques [3, 12], albeit with varying levels of significance.

6. Conclusion

We conducted an empirical analysis to assess how features derived from temporal Electronic Health Records (EHR) affect the predictive performance of various machine-learning methods in the context of 30-day heart failure readmission risk prediction. We examined 18 dif-

Table 1. Results of adding temporal features to the feature set.

	Model Family			
	Non-Parametric	**GLMs**	**Vector Machines**	**Ensemble Methods**
Best Method	Gaussian Naïve Bayes	Ridge Classifier	LinearSVM	Bagging
S	0.615(0.013)	0.610(0.013)	0.610(0.013)	0.741(0.019)
V	0.587(0.010)	0.606(0.009)	0.607(0.009)	0.602(0.014)
L	0.604(0.011)	0.585(0.008)	0.585(0.008)	0.610(0.009)
S + V	0.618(0.008)	0.641(0.010)	0.642(0.008)	0.756(0.019)
S + L	0.629(0.008)	0.626(0.004)	0.626(0.004)	0.757(0.018)
V + L	0.615(0.011)	0.631(0.012)	0.633(0.011)	0.634(0.009)
S + V + L	**0.632(0.008)**	**0.653(0.009)**	**0.654(0.009)**	**0.766(0.020)**

ferent techniques spanning a range of model families, both with and without the inclusion of temporally-focused candidate variables. These candidate variables encompassed data from vital chart events and laboratory test results. Additionally, we included a set of static features derived from a patient's general record as a reference model. Across all model families, we observed a noteworthy enhancement of 0.0325 in the Area Under the Curve (AUC) as the outcome measure when both feature sets were combined. This finding aligns with previous studies that incorporated temporally-focused variables in predictive models, albeit lacking an investigation on their impact. Future research endeavors will delve into more advanced feature extraction techniques applied to these datasets. Exploring other types of temporal features and their potential utility is another direction for this predictive task.

Acknowledgments

This work is supported by NIH/NINR award R01NR018301.

References

[1] Ponzoni CR, Corrêa TD, Filho RR, Serpa Neto A, Assunção MS, Pardini A, Schettino GP. Readmission to the Intensive Care Unit: Incidence, Risk Factors, Resource Use, and Outcomes. A Retrospective Cohort Study. Annals of the American Thoracic Society 2017;14(8):1312–1319.

[2] McIlvennan CK, Eapen ZJ, Allen LA. Hospital Readmissions Reduction Program. Circulation 2015;131(20):1796–1803.

[3] Mahmoudi E, Kamdar N, Kim N, Gonzales G, Singh K, Waljee AK. Use of Electronic Medical Records in Development and Validation of Risk Prediction Models of Hospital Readmission: Systematic Review. British Medical Journal 2020;369.

[4] Shameer K, Johnson KW, Yahi A, Miotto R, Li L, Ricks D, Jebakaran J, Kovatch P, Sengupta PP, Gelijns S, 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. World Scientific, 2017; 276–287.

[5] Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, Hisamitsu T, Kojima G, Felsted J, Kakarmath S, et al. A Machine Learning Model to Predict the Risk of 30-Day Readmissions in Patients with Heart Failure: a Retrospective Analysis of Electronic Medical Records Data. BMC Medical Informatics and Decision Making 2018;18(1):1–17.

[6] Kennedy CE, Turley JP. Time Series Analysis as Input for Clinical Predictive Modeling: Modeling Cardiac Arrest in a Pediatric ICU. Theoretical Biology and Medical Modelling 2011;8(1):1–25.

[7] 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):e0218942.

[8] Bielinski S. Heart Failure (HF) with Differentiation Between Preserved and Reduced Ejection Fraction— PheKB, 2019.

[9] Cheung BLP, Dahl D. Deep Learning from Electronic Medical Records using Attention-Based Cross-Modal Convolutional Neural Networks. In 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). IEEE, 2018; 222–225.

[10] Brown SE, Ratcliffe SJ, Halpern SD. An Empirical Derivation of the Optimal Time Interval for Defining ICU Readmissions. Medical Care 2013;51(8):706.

[11] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research 2011;12:2825–2830.

[12] Zolbanin HM, Delen D. Processing Electronic Medical Records to Improve Predictive Analytics Outcomes for Hospital Readmissions. Decision Support Systems 2018; 112:98–110.

Address for correspondence:

Ryan Missel
1 Lomb Memorial Drive
Rochester NY, 14623
Email: rxm7244@rit.edu