

Predicting In-Hospital Death from Derived EHR Trajectory Features

Rajeev BOPCHE^a, Lise Tuset GUSTAD^b, Jan Egil AFSET^a, Jan Kristian DAMÅS^a
and Øystein NYTRØ^{a1}

^aNorwegian University of Science and Technology, Trondheim, Norway

^bNord University, Levanger, Norway

ORCID ID: Rajeev Bopche <https://orcid.org/0000-0003-4835-9488>, Lise Tuset Gustad <https://orcid.org/0000-0003-2709-3991>, Jan Egil Afset <https://orcid.org/0000-0003-2489-4179>, Jan Kristian Damås <https://orcid.org/0000-0003-4268-671X>, Øystein Nytrø <https://orcid.org/0000-0002-8163-2362>

Abstract. Medical histories of patients can predict a patient's immediate future. While most studies propose to predict survival from vital signs and hospital tests within one episode of care, we carried out selective feature engineering from longitudinal medical records in this study to develop a dataset with derived features. We thereafter trained multiple machine learning models for the binary prediction of whether an episode of care will culminate in death among patients suspected of bloodstream infections. The machine learning classifier performance is evaluated and compared and the feature importance impacting the model output is explored. The extreme gradient boosting model achieved the best performance for predicting death in the next hospital episode with an accuracy of 92%. Age at the time of the first visit, length of history, and information related to recent episodes were the most critical features.

Keywords. Electronic Health Records, health trajectory analysis, machine learning, bloodstream infection

1. Introduction

Electronic Health Records (EHR) data is used for secondary applications like disease progression modeling [1], patient trajectory modeling [2], disease inference [3], risk stratification, and survival prediction [4]. These data-driven analyses are increasingly needed in all kinds of health services and research. However, the problem with EHRs is mostly the sparseness and context-dependent interpretation of EHRs can cause incompleteness and to a lesser extent inconsistency and inaccuracy [5]. Patient discharge summaries provide a longitudinal perspective of patients' interactions with hospital service. In Norway, with predominantly public specialist healthcare, patients often have long and continuous histories within one hospital's records. Thus, we can conduct a retrospective analysis of medical histories with poor outcomes.

In this study, we selectively derived the most relevant count and temporal features and used them to train classifiers for predicting in-hospital mortality for the next episode. We then studied the features that led to improvement in model performance as predictors

¹Corresponding Author: Øystein NYTRØ, Oystein.Nytrø@ntnu.no

of in-hospital deaths. Similar work on prediction models using longitudinal medical data by Chicco et al. (2020) showed that traditional machine-learning (ML) models predicted the survival of the patients diagnosed with sepsis using minimal clinical records of patients [4]. Other work mainly focused on visualizing the medical history or building patient disease trajectories [6]. Some studies have worked on tackling the problems with the representation of medical data and codes. For example, Tran et al. (2015) worked on building a low-dimensional representation of medical events using a modified Restricted Boltzmann Machine (RBM). Thereafter they trained a logistic regression classifier for suicide risk stratification [7]. Similarly, Choi et al. (2015) disease-specific applications have treated the medical history as a sequence of events and then trained machine learning (ML) models to predict an outcome [8]. While Jia et al. (2020) used patient similarity-based frameworks to group similar patient histories together [9]. The aim of this study was to investigate patterns of events leading to death in the hospital and determine if these patterns in prior episodes can help predict an impending medical episode with a risk of death.

2. Methods

2.1. Data and Feature Engineering

The data for this study includes individual episodes of care from St. Olavs university hospital between 2015-2020 for 35,594 patients that had at least one episode of suspected bloodstream infection (BSI). The episodes range from the introduction of EHR in 1999 until 2020 but do not include primary care or visits to other specialist care. The mean age of the complete cohort is 63.6 years, and the gender distribution is 52.5% males to 47.4% females. The data contains information on a total of 1.2 million medical episodes. Diagnosis is coded using ICD-10 (International Classification of Diseases 10th Revision). Python's Pandas library was used to preprocess and clean the dataset. Patient histories with only one episode or with a death date more than 30 days from the last episode were also excluded. New features were calculated by aggregating time under each ICD-10 code group and counting the number of respective episodes. To capture the repetitiveness of episodes, the time between episodes was calculated and normalized by the total number of episodes. Features such as age at the time of the first and last visit, the total length of medical history, and the total number of hospital visits were also included in the analysis. Table 1 gives a description of these derived features.

Table 1. Description of the derived features.

Features	Unit
Age at first visit and last visit	Years
Time to recent episodes	Hours
Time between episodes	Hours
Time Hospitalized for implicit sepsis	Hours
Time Hospitalized for explicit sepsis	Hours
Time Hospitalized for cancer disease	Hours
Time Hospitalized for cardiovascular disease	Hours
Total time under acute care	Hours
Total length of stay	Hours
Length of history	Days
Implicit sepsis episodes	Count
Number of cancer episodes	Count
Number of explicit sepsis episodes	Count

Number of cardiovascular episodes	Count
Number of infection episodes	Count
Number of hospital visits	Count

2.2. Prediction Modelling

In this study, we derive temporal and count features like exposure and repetitiveness not otherwise available to non-temporal or process-blind ML methods. All derived features are taken as continuous values and empty cell values were imputed to zero. The values were then standardized by removing the mean and scaling to unit variance. The target feature was labeled as 1, if death occurred within 30 days of final episodes, and 0, if the patient was alive. The dataset consisted of 22,637 patients labeled 0 and 7,294 patients labeled 1. The dataset was split into a training set (80%) and a testing set (20%). SMOTE (Synthetic Minority Over-sampling Technique) was used to address the class imbalance in the training set. For our binary classification problem, we used Logistic Regression (LR) as the linear model, Gaussian Naïve Bayes (GNB) as the probabilistic model, K-Nearest Neighbors (KNN) as the non-parametric model, Random Forest (RF), Bagging and Boosting decision tree classifiers (BG and ADB), Voting Classifier as the ensemble model, multi-layer perceptron (MLP) as neural network based model, and eXtreme Gradient Boosting (XGBoost). We interpreted the machine learning models by using SHapley Additive exPlanations (SHAP) values.

3. Results

3.1. Model performance

This study compared ten different machine learning models on our dataset. Performance metrics for each of these models are given in Table 2. The results indicate that all models achieved high accuracy scores with the XGBoost classifier showing the overall best performance. The receiver operating characteristic (ROC) plot for all models and the confusion matrices on testing data for the Random Forest classifier (classifier used to study feature importance) and the XGBoost classifier (the best classifier) are given in Figure 1.

Table 2. Prediction performance metrics.

Model	Accuracy	Sensitivity	Specificity	F1 Score	AUROC
Logistic Regression	0.883	0.868	0.887	0.851	0.940
Naïve Bayes	0.823	0.631	0.885	0.760	0.849
K-Nearest Neighbors	0.830	0.826	0.831	0.792	0.887
Support Vector Machine	0.885	0.859	0.893	0.853	0.950
Bagging Decision Trees	0.900	0.774	0.941	0.863	0.949
Boosting Decision Trees	0.902	0.776	0.942	0.863	0.949
Random Forest	0.882	0.854	0.891	0.865	0.947
Ensemble	0.895	0.702	0.958	0.852	0.948
Neural Network	0.899	0.768	0.942	0.861	0.946
XGBoost	0.918	0.827	0.947	0.888	0.964

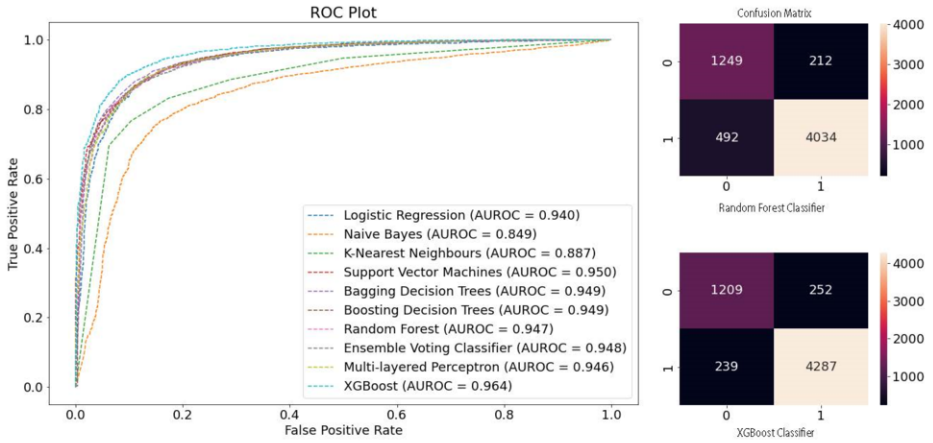


Figure 1. ROC curve for all models and Confusion matrices Random Forest model (top right) and XGBoost model (bottom right).

3.2. Feature Importance

To study the feature importance, SHAP values were estimated for each feature. SHAP values indicate the impact on model output. Figure 2(a) gives the feature importance as the impact on model prediction using the Random Forest classifier. Age at the first visit, length of history, and information relating to recent visits were the top features for the model followed by time hospitalized for various disease episodes. Figure 2(b) summarizes the SHAP value by combining feature importance with feature effects. The plot depicts the features’ overall influence on the model prediction. Each data point represents an individual case and data points shifted to the right indicate features that contribute to a high risk of death, while those shifted to the left indicate features that contribute to a lower risk of death.

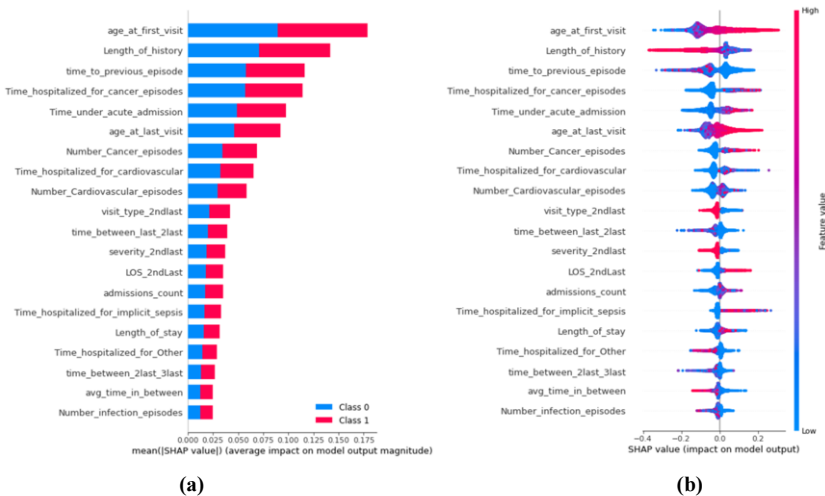


Figure 2. SHAP values for feature importance bar plot and dot plot. In Figure 2 (a) Class 0 stands for alive patients and class 1 for patients who died.

4. Discussion

It was discovered during a thorough analysis of the patient histories that the last part of the medical history of the patients who died usually consisted of not one single episode but a series of episodes. Moreover, these patients also had longer length of stays for major diseases. This made our prediction model for in-hospital mortality achieve high-performance metrics as the derived features successfully capture these critical indicators of individual trajectory and rapid succession of episodes leading to a fatal episode. The age at the first visit is the most important feature indicating that patients arriving at the hospital for the first time very late in life are at very high risk. The length of history is the second most important feature with a shorter length of history contributing towards a higher risk of in-hospital death. One major limitation of this study is that live patient histories were taken as complete instead of trimming histories up to some critical episodes. The major contribution of this study is that these early warning models can be easily implemented in the current and developing digital health platforms to predict adverse outcomes enabling proactive and precautionary care.

5. Conclusions

Accurate prediction of mortality can assist in better hospital resource allocation, enabling healthcare providers to prioritize patients who are at higher risk of adverse events. In this study, we were able to predict if an impending disease episode entailed a risk of death. The derived features can be further well formulated with coordination and discussions with the hospital staff and stakeholders.

References

- [1] Mould DR. Models for disease progression: new approaches and uses. *Clin Pharmacol Ther.* 2012 Jul;92(1):125-31, doi: 10.1038/clpt.2012.53.
- [2] Ebadollahi S, Sun J, Gotz D, Hu J, Sow D, Neti C. Predicting patient's trajectory of physiological data using temporal trends in similar patients: a system for near-term prognostics. *AMIA Annual Symposium Proceedings*; 2010 Nov 13; Hawthorne, NY. p.192-6.
- [3] Zhao D, Weng C. Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. *J Biomed Inform.* 2011 Oct;44(5):859-68, doi: 10.1016/j.jbi.2011.05.004.
- [4] Chicco, D., Jurman, G. Survival prediction of patients with sepsis from age, sex, and septic episode number alone. *Sci Rep.* 2020 Oct;10(1):17156, doi: 10.1038/s41598-020-73558-3.
- [5] Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. *Summit Transl Bioinform.* 2010 Mar;2010:1-5.
- [6] Beck MK, Jensen AB, Nielsen AB, Perner A, Moseley PL, Brunak S. Diagnosis trajectories of prior multi-morbidity predict sepsis mortality. *Sci Rep.* 2016 Nov;6(1):1-9, doi: 10.1038/srep36624.
- [7] Tran T, Nguyen TD, Phung D, Venkatesh S. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). *J Biomed Inform.* 2015 Apr;54:96-105, doi: 10.1016/j.jbi.2015.01.012.
- [8] Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: predicting clinical events via recurrent neural networks. *Proceedings of the 1st Machine Learning for Healthcare Conference, JMLR Workshop Conference Proceedings*; 2016 Dec 10; Los Angeles, CA. p. 301-18, doi: 10.48550/arXiv.1511.05942.
- [9] Jia Z, Zeng X, Duan H, Lu X, Li H. A patient-similarity-based model for diagnostic prediction. *Int J Med Inform.* 2020 Mar;135:104073, doi: 10.1016/j.ijmedinf.2019.104073.