FEATURE SELECTION AND PERSONALIZED MODELING ON MEDICAL ADVERSE OUTCOME PREDICTION

By

QINGQING DAI

Bachelor of Economics in Statistics
University of Central Finance and Economics
Beijing, China
2011

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2020

FEATURE SELECTION AND PERSONALIZED MODELING ON MEDICAL ADVERSE OUTCOME PREDICTION

Dissertation Approved:

Dr. Lan Zhu

Dissertation Adviser

Dr. Joshua Habiger

Dr. Ye Liang

Dr. Bing Yao

Outside Committee Member

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my gratitude to those who have provided so much help and support on my conducting this research and writing the thesis. First of all, my special thanks go to my advisor Dr. Lan Zhu for her guidance, support and patience during my five years of PhD study and I won't be able to finish this thesis without her. I also want to thank the other committee members, Dr.Joshua Habiger, Dr.Ye Liang and Dr.Bing Yao for providing their precious time, comments, suggestions and insights on my thesis. Secondly, I would like to thank the Center for Health Systems Innovation (CHSI), especially Dr. Zhuqi Miao for providing me with both financial support, research resources and insightful guidance. Also, I would like to thank all the faculty, staff and fellow students of the Department of Statistics that have provided endless support during my time in Oklahoma State University. Last but not least, I would like to thank my dear family and friends for supporting me with constant love and friendship.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: QINGQING DAI

Date of Degree: MAY, 2020

Title of Study: FEATURE SELECTION AND PERSONALIZED MODELING ON MEDICAL ADVERSE OUTCOME PREDICTION

Major Field: STATISTICS

Abstract: This thesis is about the medical adverse outcome prediction and is composed of three parts, i.e. feature selection, time-to-event prediction and personalized modeling. For feature selection, we proposed a three-stage feature selection method which is an ensemble of filter, embedded and wrapper selection techniques. We combine them in a way to select a both stable and predictive set of features as well as reduce the computation burden. Datasets on two adverse outcome prediction problems, 30-day hip fracture readmission and diabetic retinopathy prognosis are derived from electronic health records and exemplified to prove the effectiveness of the proposed method. With the selected features, we investigated the application of some classical survival analysis models, namely the accelerated failure time models, Cox proportional hazard regression models and mixture cure models on adverse outcome prediction. Unlike binary classifiers, survival analysis methods consider both the status and time-to-event information and provide more flexibility when we are interested in the occurrence of adverse outcome in different time windows. Lastly, we introduced the use of personalized modeling(PM) to predict adverse outcome based on the most similar patients of each query patient. Different from the commonly used global modeling approach, PM builds prediction model on smaller but more similar patient cohort thus leading to a more individual-based prediction and customized risk factor profile. Both static and metric learning distance measures are used to identify similar patient cohort. We show that PM together with feature selection achieves better prediction performance by using only similar patients, compared with using data from all available patients in one-size-fits-all model.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Adverse medical outcome refers to the suboptimal outcome of patients, like mortality, readmission and cancer diagnosis, and it has brought huge economic burden to the healthcare system. Hospital readmission is one important example of adverse outcome, which is not only costly but also life-threatening.[1–3] For example, hip fracture (HF) is one serious injury that frequently occurs in the geriatric population and leads to 20-25% first-year mortality rate[4, 5], even worse, readmission within 30 days after hip fractures can nearly double the first-year mortality rate [4]. Hospitals with more frequent readmission are believed to be providing poorer healthcare and will be penalized by the Centers for Medicare & Medicaid Services (CMS).[6] The onset of disease is another often-discussed topic in adverse outcome research. One example is the onset of diabetic retinopathy (DR), a complication of diabetes which can cause damage to patients' retina. Once it occurs the lost vision can't be regained and it has been a leading cause of blindness in American adults.[7] Fortunately, many of the adverse outcome can be prevented through predictive analysis and targeted treatment based on the predicted results, especially in such an era of advanced information technology and a large amount of rich, quality longitudinal patient data that is conducive to making accurate predictions.

Predictive modeling has been one of the most popular and important techniques in the study of clinical adverse outcome. [8–12] Generally speaking, predictive modeling involves four key steps, as shown in Figure 1.1, including data acquisition, feature selection, model development and evaluation.[13–19] Although there are other factors that can influence the prediction performance of models, like the treatment of missing data, they are often the limitations of all models and complicated tasks which varies by problem. The focus of this investigation is on improving the prediction accuracy by optimizing the listed steps of predictive modeling.

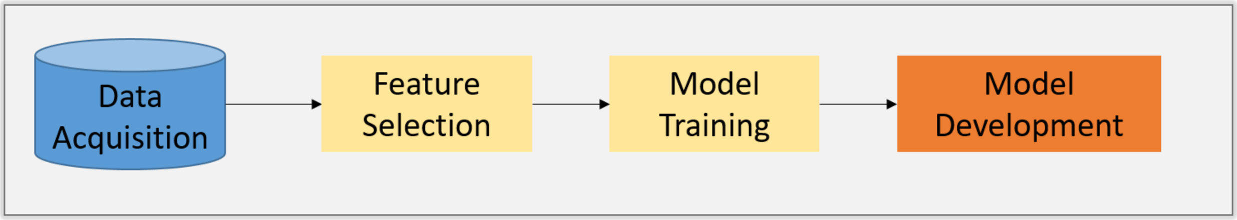

Figure 1.1. General Steps of Predictive Modeling

For the acquisition of data, with the passage of the Health Information Technology for Economic and Clinical Health (HITECH) Act in 2009 and the Affordable Care Act (ACA) in 2010, many hospitals have transitioned to the electronic health records (EHRs) system and the use of EHRs has grown dramatically. By 2014, 75.5% of hospitals were using EHRs.[20] The extensive adoption of EHRs has ensured that we can obtain a huge amount of data that contains various information about patients. EHRs record the patients' medical history over time, including demographics, medications, vital signs, laboratory data, physician's notes, etc. With such tremendous information and easier access, EHRs have become an important data source for medical research. Cerner Health Facts Database (Kansas City, MO) is such an EHRs database which we will conduct our analysis on. With this reliable data source, we will be able to obtain sufficient cohort size which is a key to accurate prediction.

Despite the rich information in EHRs data, we should also note that its exploration is challenging due to the variation in different data sources, data quality issues, inherent case heterogeneity and high dimensional nature.[21–24] Among all these problems, in this work, our focus will be on alleviating the high dimensional problem. Due to the high dimensionality of the data, it usually causes a huge computational burden, and also introduces a lot of redundant and irrelevant information which may affect the prediction performance. Feature selection is a technique to address this problem, by choosing features that are truly associated with the response variable.[25] It can not only improve accuracy but also speed up model fitting and enhance interpretability. In clinical research, it can even help to reduce the costs by identifying the unnecessary lab tests[26], and provide targeted guidance on post-discharge care planning. Besides fitting prediction models, the selected features can also be used to derive risk scores to quantify the risk of patients' developing adverse medical condition.[20, 27–29] A common requirement for medical risk score is its simplicity and good interpretability, which again emphasizes the importance of choosing the smallest set of significant features. Feature selection has been extensively studied in recent years and various algorithms have been derived.[30] Different methods, like machine learning models, information theory, designed experiments and fuzzy modeling, have been proposed to identify significant features.[26, 31–33] But there is a lack of general guideline for feature selection in adverse outcome prediction and the stability of selection process seems to be overlooked. A more detailed review on feature selection techniques will be present in Chapter II.

When predictive features are selected, the choice of an appropriate modeling framework according to the data types is critical to the final prediction performance. After reviewing the current models used in adverse outcome prediction [34–37], we identified two potential directions to improve the model development step, that is, the application of survival analysis and personalized modeling. Binary classifiers like logistic regression have often been used in the adverse outcome prediction since the outcome is usually dichotomous, like "readmitted" versus "non-readmitted". But actually, the time-to-event information are oftentimes available or can be converted from the data and the inclusion of this information may lead to more accurate prediction.[37, 38] For example, when we are trying to identify the 30-day readmission status of patients, we first need to calculate days from the index admission to the re-hospitalization and then label the patients as "readmitted" or "non-readmitted" based on whether the number of days exceed 30. If we analogize non-readmission as survival, the days to readmission can be taken as survival days and survival analysis is needed. In this work, we will explore the application of survival analysis on readmission prediction.

In adverse outcome prediction studies, we also notice that the most straightforward and popular approach is to build a single classifier using all of the available training observations and then predict the outcome of patients entering the system in the future, which has been referred to as one-size-fits-all approach or global model.[39] Global model can capture the patterns prevalent in the entire population but may miss the less popular information that is important for specific patients. Thus, global models usually perform well for the average patients but are sub-optimal for individual patients with unique characteristics[40]. Other researchers also point out that the global model is not robust to population shift.[41] Population shift happens when the distributions of the training and test sets are different, which is not unusual when we split the data into training and test sets chronologically. A remedy to this is personalized predictive modeling which selects only the similar cases to train the model on the individual basis. It has been a success in a variety of domains like the personalized product recommendation in e-commerce, which is based on the belief that users have similar tastes on some items may have the same taste on other items.[39] Its success indicates the potential of improving adverse outcome prediction by building individual-based prediction model with medical history of the most similar patients. Consequently, patients will be able to receive care and service based on individual needs and conditions, which will in turn improve disease prevention, management and drug prescription.

To summarize, this study intends to present a new automatic and stable ensemble feature selection process which can reduce the number of features as well as maintain or even improve prediction performance compared with the full model. It can also serve as a general feature selection guideline for similar prediction researches that aim to achieve the smallest stable feature set to build interpretable prediction model and derive simple risk scores. We also explore improving prediction accuracy by including the time-to-event information with survival analysis. To address the possible population shift problem, a personalized modeling process is proposed, which can also generates individualized risk factor profiles. The outline of this study is as follows. Chapter II briefly reviews the terminologies and concepts relevant to the models or techniques that will be used through this paper. Chapter III introduces the three-step ensemble feature selection method, the survival models and personalized modeling steps. Two adverse outcomes, 30-day hip fracture readmission and diabetic retinopathy prognosis, are exemplified to show the effectiveness of our proposed feature selection method in Chapter IV. The applications of survival models on readmission prediction and personalized modeling on adverse outcome prediction are also discussed. Chapter V provides the discussion and conclusions based on the results as well as future research directions.

CHAPTER II

LITERATURE REVIEW

The aim of this chapter is to investigate the recent trends in feature selection and prediction modeling.

2.1 Predictive Modeling

Predictive modeling refers to the process of applying statistical techniques to historical records to predict future outcomes. It is also defined as the process by which a model or a mathematical tool is created to predict the probability of an outcome.[16, 42] It has been in the intersection of statistical modeling, machine learning and database technology.[43] As the need for future event prediction continues to growing in various fields, the application of predictive modeling starts to be found everywhere, like the fraud detection systems of banks[44] and the recommender systems of commercial service providers[45]. The applications of predictive modeling has revolutionized many industries and at the same time, the development of various industries also fueled the advances of modeling techniques.[46] The predictive techniques have developed from regression models which emphasizes interpretability to more sophisticated machine learning methods that focus more on accuracy.[16] In general, there are two classes of predictive models, parametric and non-parametric models. The former type makes specific assumptions on the parameters to characterize some underlying distributions while the latter usually has no restriction on the distribution form.[47] The advantages of parametric models include better interpretability, better ability to quantify the feature effect as well as indicating the direction of the impact.

Predictive analytics identifies the not-readily-apparent trends, patterns or relationships among data and also produces insights to help practitioners understand how people behave as customers, buyers, patients, and so on.[48–50] It has been widely used to assist human decision-makers in various fields, including but not limited to, financial service[50, 51], insurance[52, 53] and healthcare[54, 55]. For example, banks can decide whether to issue an applicant the credit card based on fraud detection models.[56, 57] Predictive models are also used in applications like crime detection[58] and email spam filter[49], estimating the probability of a crime or email being spam. It is also one of the most important technologies used in healthcare and clinical research, which can be reflected by the growing publication rate of relevant papers.[59] This type of model in clinical practice is also referred to as prognostic model, which predicts future events or behaviors and provides data-driven decision support to healthcare providers[60–62]. For example, it helps doctors decide whether to provide intensive or mild treatment to patients based on the predicted probability of readmission risk.[63–65] Readmission is one type of medical adverse outcome, which refers to suboptimal or unwanted event for patients following medical care and other examples include morbidity, mortality, new disease or worsening symptoms, unscheduled physician visits and emergency department visits.[66–68] A single-institution adverse outcome analysis[69] showed that about 25% of patients in their study had an adverse outcome and 50% of the occurrences could have been prevented if predictive models are used appropriately and timely treatments are provided.

A systematic review paper[20] thoroughly reviewed 107 papers on medical outcome prediction with electronic health records from 15 different countries. 78.5% of them used the generalized linear models, including logistic regression and Cox regression, as the prediction models. The regularized regression, like ridge regression and the least absolute shrinkage and selection operator(LASSO) are also often incorporated for the purpose of variable selection. They also point out that machine learning models are also used but more likely to include all the variables. The c-statistic (i.e. the area under the receiver operating characteristic curve (AUC)), a most commonly used measurement to evaluate the model performance, ranges from 0 to 1 and the larger the value, the better the prediction performance.[70] They present the distributions of c-statistics by the modelled outcome type and it shows that the model performance varies greatly among the same outcome type as well as among different outcome types, and their c-statistic can be as low as less than 0.6 and as high as close to 1.

Although a large number of prediction models have been established, only a few are used in clinical practice.[62] Possible reasons behind this include the insufficient size of cohort which the model is built on, inappropriate processing of missing data, lack of interpretability, unsuitable model framework and lack of transparent reporting of the above steps.[71, 72] Missing data is universally encountered in clinical research since it's difficult to collect all data on all predictors for all patients. Handling missing values is a significant yet complicated task and the necessity of treatment on missing data varies by models, thus there is no consensus on how to deal with the missing problem.[73] The most popular and easiest approach in clinical research is the complete case analysis, where all patient records with a missing value on any of the considered features are excluded.[62, 74] In this study, we will take the complete case analysis approach on the data and focus on solving the other key problems, including enlarging cohort size, improving model interpretability and selecting suitable model framework, which corresponds to the three key aspects of developing a prediction model[74, 75], i.e. data acquisition, feature selection and model development.

2.2 Data Acquisition — Electronic Health Records (EHRs)

Advances in information technology contribute to the digitization of health and patient data. Electronic health records (EHRs), the electronic version of patient health records, was originally introduced to store and manage patient data effectively.[76] The authorities have speed up the adoption of EHRs with legislation like Health Information Technology for Economic and Clinical Health (HITECH) Act in 2009 and the Affordable Care Act (ACA) in 2010, which have pushed the industry to shift towards EHRs and there has been a dramatic increase in their usage. In 2014, 75.5% of hospitals were using EHRs.[20] The wide application of EHRs systems has guaranteed the access to rich, longitudinal admission records containing rich information of patients.[77, 78] EHRs record the patients' medical history over time, which include demographics, medications, vital signs, laboratory data, physician notes, etc. With such tremendous information and easy access, researchers have been actively exploring the secondary use of EHRs, such as patient risk stratification and disease detection.[20, 76, 79]

EHRs data is known to be rich in variety and large in volume. Variety means the various formats the patient information takes, including the structured value records, physicians' prescription, symptoms description as well as the photographic or digital images.[80] Volume refers to the fact that the databases are usually extremely large both in terms of the number of features and the number of samples considered.[81, 82] The high dimensionality results from improvements in data acquisition capacity, falling costs of data storage and computation, development of database and data warehousing technology.[83] The rich information and large volume of EHRs data present great potential since the access to more data instances and more data features is critical for predictive modeling.[17]

2.3 Feature Selection

Although the big volume of the EHRs database brings rich patient information, it also brings manifold challenges to predictive analytics. One challenge is that there are not only predictive features but also many irrelevant or redundant features, which often increase the size of search space and bring difficulty to the pattern learning in the data. There are two groups of dimensionality reduction techniques: One is feature extraction where a new and smaller set of features are usually constructed as the combination of existing features and has often been used in computer vision problems.[82, 84, 85] Another type is feature selection which entails choosing the smallest feature subset that can maximize the prediction accuracy. Feature selection can not only reduce dimensionality but also maintain the interpretability and it is believed to improve the robustness and usefulness of prediction algorithms by selecting a small set of relevant features to construct the final model.[14, 86, 87]

Feature selection is a research topic of long history and has been more extensively studied in the past decades because of the breakthrough in information technology and the need to deal with high dimensional data.[13, 88–91] By identifying the most relevant subset of features, feature selection techniques simplify the model and lead to better interpretability, shortening computation time, reducing overfitting, etc.[92] In clinical practice, risk index or prognostic index which measures the risk of developing the outcome, is often derived after fitting the final model with the selected most significant features.[61, 93] The index is defined as a weighted sum of the features with their coefficient values as weight.[94–97] Given these benefits, various feature selection techniques have been proposed, and can be classified in different ways.[98] Based on the automaticity, they can be divided into the expertise-based (knowledge-driven) and automatic (data-driven) approach[99]. The first type uses domain knowledge to help assess the significance of a predictor and whether to include it in the final model.[100, 101] On the contrary, data-driven approaches select features based on the analysis and interpretation on data and have been extensively studied, especially in the machine learning and artificial intelligence community.[89, 91] Although the expert knowledge is of great importance to the interpretation of results, it may not be available all the time and is subject to bias. In our study, we focus on the data-driven approaches. Depending on the relationship between selection process and model building, the data-driven approaches can be categorized into three major categories: filter, wrapper and embedded methods.[89, 93]

Filter methods select features independently from the classifier. In most cases, the relevance of features to the dependent variable is evaluated and the highly relevant features are selected as the input to the classification model. The common evaluation criteria include distance metrics, data variance, fisher score, correlation, mutual information, Kolmogorov-Smirnov test, T-test and Chi Square test.[102–104] Advantages of filter methods include their scalability to high-dimensional datasets and the fast computation. Due to the simplicity and fast-speed of univariate filter methods, they have been the most commonly used feature selection technique. Chi-square test has been popular in evaluating the relevance of categorical features and T-test for continuous ones.[29, 105–108]

Chi-square test was introduced by Pearson (1900), firstly used to test the goodness of fit and has been applied to test the dependence of two events. The test statistic is derived on the independence rule of two events $P(AB) = P(A)P(B)$. To test the dependence between the binary outcome Y and a feature X, we have the contingency table below, where $n_{ij}(i,j = 0,1)$ denotes the observed number of each event. Denote the expected number of each event as $e_{ij}(i,j = 0,1)$. If X and Y are independent, we

Event	Y=1	Y=0	Total
X occurs	n_{11}	n_{10}	$n_{11} + n_{10}$
X not occur	n_{01}	n_{00}	$n_{01} + n_{00}$
Total	$n_{11} + n_{01}$	$n_{10} + n_{00}$	n

Table 2.1. Contingency table of feature X and the outcome Y

would have $P(X = i|Y = j) = P(X = i)$. When sample sizes are large enough, we can compute $\frac{e_{ij}}{n_{1j}+n_{0j}} = \frac{n_{i1}+n_{i0}}{n}$, thus $e_{ij} = \frac{n_{i1}+n_{i0}}{n} * (n_{1j} + n_{0j})$. Chi-square test is defined as $c^2 = \sum_{i=0,1} \sum_{j=0,1} \frac{(n_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2(1)$.

An alternative to chi-square test is Fisher's exact test when the expected values in any cell of the contingency table are below 5.[108] For continuous variables, T-test is most commonly used and sometimes Wilcoxon rank sum test or Mann-Whitney U test is used instead.[107–110] Since filter methods consider each feature separately, they ignore the feature dependencies. Additionally, these methods also ignore the interaction between feature subspace and the classifier. As a result, filter methods often lack robustness against interaction among features and also that between features and the classifier.[111]

On the contrary, wrapper methods embed the feature selection within the model selection. Wrapper-type methods entail an exhaustive search of all possible subsets of features to guarantee finding the best subset of features.[86] They train a new model for each subset of features and test on a hold-out set to score feature subsets. The subset with the highest score will be chosen thus they are tailored for specific models. They have the advantage of considering both feature dependencies and the interaction between feature selection and model building.[89, 93] Wrapper methods can be grouped into sub-categories by their searching pattern, that is, greedy and randomized.[111] Examples of wrapper methods include the forward and backward selection, which sequentially include or exclude features by means of the loss function.[29, 112–114] The disadvantage of these methods is that they tend to overfit the classifiers and are more computationally intensive than filter techniques.[93] Since the wrapper methods are classifier-dependent, if we are interested in different classifiers, we need to perform feature selection for each of the classifiers separately, while we just need to do feature selection once for filter techniques. Wrappers have also been criticized for requiring massive amounts of computation, but efficient search strategies can be devised to address this problem.[115]

For the third class of feature selection methods, embedded methods, the feature selection process is integrated as part of the model construction, which means that they are specific to a given model, like the wrapper approaches.[98] Embedded methods discover the set of features from the model structure. Two popular examples are the decision tree algorithm[116, 117] and regression with the least absolute shrinkage and selection operator (LASSO)[118]. A decision tree can infer the variable importance by its node structure but a single tree can be unstable.[119] As an alternative, random forest ensembles a forest of trees on bagged samples and takes the average of variable importance to get an overall measure.[120] LASSO has obtained success in many applications of dimension reduction but is known to be unstable when features are correlated.[121] Actually, different feature selection techniques can be cooperated together to reach a smaller set of features and maintain the performance. Although most of the existing studies tend to use filter method or embedded method alone for selection, one study combined Support Vector Machines (SVM) with various feature selection methods, including a two-step one which used both the filter method F-score and the embedded method random forest.[99] F-score is a simple measure for the discriminability of a feature, defined as:

$$F(i) = \frac{(\bar{x_i}^{+} - \bar{x_i})^2 + (\bar{x_i}^{-} - \bar{x_i})^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{\bar{k},i}{}^{+} - x_i^{\bar{+}})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{\bar{k},i}{}^{-} - x_i^{\bar{-}})^2} \tag{2.1}$$

This type of feature selection method is also called the hybrid method, which combine different feature selection methods to integrate their advantages.[111] For example, the filter-wrapper algorithms are proposed to reduce the computation burden of using wrapper alone.[122] There are also empirical researches showing that we can benefit from feature ranking aggregation and the fusion of a set of feature selection techniques.[123] It's even claimed that the future opportunities for feature selection research lies on new combinations of existing techniques such as hybrid or ensemble methods and it's possible to distribute the data vertically by features to reduce computational burden when applying wrapper methods.[124, 125] This inspires us to propose the feature selection algorithm that integrates the three types of methods in a computation-efficient way.

To evaluate and compare the performance of different techniques, there are two important aspects, i.e. accuracy and stability. Accuracy measures the predictive power of the selected features while stability measures the degree of agreement of the feature subset selected by the method when used on different training sets drawn from the same distribution.[93]. Accuracy has been commonly used to evaluate the performance while stability is often ignored. Actually, the stability of feature selection is of great importance since it ensures the reliability in real-world practice.[98, 115] To quantify the stability, we can measure the similarity of two feature subsets and available methods include the Jaccard stability measure (JSM), Pearson's correlation coefficient (PCC) and Spearman's rank correlation coefficient (SRCC).[93] It has been shown that ensemble of different feature selection techniques and feature ranking aggregation can strengthen the stability of selection.[14, 126–129] One recent study explored the use of three embedded feature selection techniques together with six aggregation methods, including three simple aggregation methods and their weighted versions.[41] They concluded that the weighted mean rank performs the best in their scenario although the difference between those aggregation methods were not significant. Thus, we will choose two representative aggregation methods in our research, i.e. ensemble by simple mean and ensemble by weighted mean where the weight is the out-of-bag AUC.

2.4 Model Development

2.4.1 Time-to-event prediction

When defining the occurrence of an adverse outcome, it's usually necessary to first select an appropriate time frame. For example, different time frames are used to define readmission and there are readmissions within 15 days, 30 days, 60 days and 90 days of the initial admission to hospital.[130] Sometimes there is controversy on the choice of the time frame and the associated risk factors may vary by different postdischarge days.[131]

When researchers are interested in whether an adverse outcome will occur before a specific time, the process is often treated as a classification problem predicting whether an adverse outcome occurs before the specific time, without considering the time information, that is, the time length until the occurrence of adverse outcome.[132] Instead of treating this as a binary classification problem, if we take the time-to-event into account, the process can be viewed as a survival analysis problem.[133] Similar to classification problems, survival analysis problems are also about examining the associations between the adverse outcome and the clinical predictors, but they care about not just whether it occurs but also the time to the occurrence.[37, 134, 135] And we can predict the probability of an adverse outcome at any time rather than at a certain time point, with survival analysis. Since the outcome is not observed for all patients, censoring has to be considered and the standard regression methods are not appropriate here. Various survival models are proposed to address these problems and identify the associations[133], including the popular proportional hazards model[136] and the accelerated failure time model[137]. Other researchers explored extensions of these models by using specific parametric transformation[138] and time-varying covariates[37]. One extension is the mixture cure model, which assumes that the studied population is a mixture of susceptible individuals who may experience the event of interest, and cured individuals who may never experience the event.[139] In this study, we aim to investigate the use time-to-event prediction on adverse outcome problems with some classical survival analysis models.

2.4.2 Personalized modeling

In clinical studies, the most common practice of prediction modeling is to train a one-size-fits-all classification model on all the labelled training records and then use the single model to make prediction on any unlabeled new patient entering the system.[55, 100, 109] This kind of population-based model usually works well on the average patients but may not be able to capture the unique characteristics of specific patients.[40] Due to the specificity of each patient to the treatment and medication and the variability of available data per patient, the widely used one-size-fits-all models for risk evaluation are losing their power.[35, 48] Another big challenge for population-based prediction modeling in medical research is the study of commobidity, which is specific to patients and limits the generalization to common patients.[140] Its's also discussed that the global model aims to learn a decision function which can reach a low mistake rate on the whole data space thus may not work well on local space.[141] One remedy to this is the personalized modeling, which is also known by different names, like case-based training[142], instance-based training[143], transductive learning[48], customized training[36] etc. Its core idea is using both the labeled and unlabeled data and leveraging the similarity among individuals to build individualized prediction model with the most similar records, instead of one general model. The example that we are most familiar with is the neighbors-based model which doesn't construct a general model but simply detect the nearest neighbors of each point and make predictions from them.[144] It has been more often discussed in personalized product and service recommendation systems[145, 146] and personalized medicine[35], but not in adverse outcome prediction.

The relevant concept in healthcare delivery is the personalized medicine.[147] Researches on personalized medicine shows that the patient population is heterogeneous and each patient has unique characteristics thus requiring specific predictions, recommendations and treatment.[35] But currently this term focuses on genetic data and has often been equated to genomic medicine.[148] It's shown that personalized health care recognizes the dynamic relationship among genetic inheritance, environmental exposures and systems biology.[149] Personalized medicine has often been discussed as a possible application in genomics information practice. Besides the personalized modeling with genomic information alone, there is also study that integrated both clinical factors and gene expression data to predict disease outcome in a person-centered way.[150]

However, genomic information is not yet widely available in everyday clinical practice, while EHRs are more accessible in the healthcare systems. It's of great potential to personalize healthcare service by making use of EHRs' tremendous clinical information, which includes but not limited to demographics, diagnostic history, medications, laboratory test results and vital signs.[151] A recent work on predicting diabetic kidney disease (DKD) onset[41] concluded that the ignorance of population shift resulted in the performance drop on their temporal validation data. A comparative research on global and personalized models in bioinformatics problems pointed out that models on individualized patients were more adaptable to both new data and features.[152] Another research pointed out that for many machine learning techniques, the local structure of data space is more important than the global structure, thus considering patients' similarity and using most similar patients to build predictive model will help to address these problems.[104] Motivated by this, we aim to improve the adverse outcome prediction with person-based modeling techniques on EHRs data.

Patient similarity analytics have been used in various medical problems, like the target patient retrieval, medical prognosis, risk stratification and clinical pathway analysis.[151] More broadly, they can be applied to create personalized risk profile and disease management plan. Many empirical researches have shown the great potential of personalized data-driven prediction systems. Personalized modeling involves the similarity calculation and selection of similar instances, thus sensitive to the choice of features, number of nearest neighbor samples and distance metrics.[152] It's reported that the inclusion of irrelevant features would degrade the performance of personalized modeling[153], which emphasizes the importance of feature selection from another angle. To guarantee the benefits from personalization, we also need to choose appropriate similarity criteria and reach a good balance between training data size and the degree of similarity. The commonly used similarity or dissimilarity measures can be divided into two groups, that is, static metrics like Euclidean distance, cosine distance[154, 155] and the learned metrics[35]. The determination of sample size of the similar cohort is very critical and, it's been shown that predictive performance can degrade when the sample size of similar cohort is too small.[40]

Different distance metrics have been applied to the distance-based clustering methods, like the k means, to evaluate their effect on clustering.[156, 157] The commonly used distance metrics include Euclidean, Manhattan, cosine, Pearson correlation, city block distance, to name a few.[158–160] Previous studies reveal that the best distance or similarity measure is specific to clustering algorithm, the dimension of data and also the data types.[161–163] A dynamic distance framework was proposed for this, called metric learning, and it has been shown that the learned metrics on the local structure of data worked better than the static distance.[35, 164] A previous study proposed four new distance-based classification methods and compared their performance with that of the two well-known classifiers, k nearest neighbors (KNN) and the Parzen windows methods. L_1 and L_2 distance metrics were used to measure the dissimilarity between points.[165] They compared the methods with cross-validated accuracy on 10 benchmark datasets from the UCI machine learning repository.

CHAPTER III

METHODS

In this chapter, we propose a novel ensemble feature selection approach, named three-stage feature selection to improve the adverse outcome prediction performance by selecting informative features. After feature selection, we explore the use of survival analysis models to predict hip fracture readmission and introduce the personalized modeling on adverse outcome prediction.

3.1 Three-stage Feature Selection

In this section, a three-stage feature selection strategy is proposed, as shown in Figure 3.1. As stated earlier in this paper, each type of selection method included in this study has its own strengths and limitations. It's also noted that different selection methods may select different feature subsets on the same data set.[166] To reduce this high variability, to overcome the drawbacks of different methods, and to reduce the risk of overfitting are the main goals of this new feature selection method. It consists of three successive steps: *filter, embedded* and *wrapper*. Suppose that we have extracted the complete cases and they are split into the training and test sets, the feature selection process will be applied on the training set to select significant features and the selection result will be evaluated on the test set. The proposed method first excludes the features that are not associated with the response using filter methods. After this stage, hopefully the potential noise in the data will be removed, resulting in a smaller number of explanatory variables, thus lower computation burden for the next stage. Then it applies the embedded methods to select features that contribute most to the model accuracy by leveraging the structure of specific machine learning models. At the end of this step, most of the falsely significant features selected in the filter step are excluded. In the last wrapper step, we try to further reduce the feature space by the leave-one-covariate-out (LOCO) approach, which is a model-free method to evaluate variable importance.[167] This step also plays the role in testing the sufficiency of the first two steps. The ideal situation is that the feature size won't drop significantly from the second to the third step.

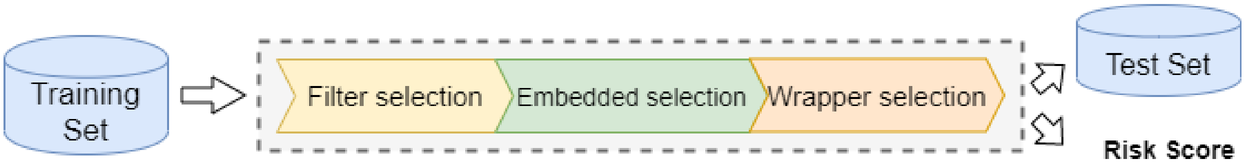

Figure 3.1. Flowchart of the three-stage feature selection

3.1.1 Filter Selection

Since the embedded and wrapper methods are more computationally intensive, it would be helpful if we can use filter methods to reduce the feature space first, and then use the feature selection techniques that are more computation-demanding but more helpful for improving prediction accuracy on the data. Here we use chi-square test to test the significance of categorical variables, or Fisher exact test when at least one cell has the expected cell count less than 5, and t test for continuous variables. The significant variables with $p - value < 0.1$ are selected for the next step. The filter step is a pre-screening of features to reduce the computation burden in the steps that follow, thus we don't want to miss any truly significant features but can tolerate some false positive features since they can be identified in the next two steps. That's why we select 0.1 as the significance level rather than the more popular 0.05. In this stage, suppose that the original feature dimension is decreased from p to p_1.

3.1.2 Embedded Selection

If the most non-informative features have been filtered out in the first step, the computation burden of the second step will be largely reduced. Here we focus on two embedded methods of different model structures: 1) Logistic regression with LASSO[118], which assumes additive structure and linear relationship, known for its good interpretability; 2) Gradient boosting machine (GBM)[168], which can deal with hierarchical structure and nonlinear relationship. To eliminate the characteristics of data sets [41, 166], we apply each ensemble method on 20 bootstrap samples. This step includes three sub-steps: firstly, we fit each of the embedded models on the training set and rank features by importance values; then we aggregate the feature rankings over the bootstrapped sub-samples and lastly we estimate the minimal feature size for close-to-optimal accuracy using golden-section search[51] together with the DeLong test[169]. The feature dimension is decreased from p_1 to p_2.

Feature ranking

The features are ranked by their importance value from each model, denoted by $IV(j)$ where $i = 1, 2, \dots n$.

- Logistic regression with LASSO

  Denote the probability of adverse outcome for the i^{th} patient by p_i, then we have

$$y_i \sim Bernoulli(p_i) \tag{3.1}$$

  We assume that there is a linear relationship between the predictors and the log-odds of the event $y_i = 1$. The linear relationship can be formulated as:

$$\log(\frac{p_i}{1-p_i}) = \mu(x_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \tag{3.2}$$

  where β_0 is the intercept and β_i's are the coefficients for each of the p predictors. The coefficients can be estimated by maximizing the log-likelihood:

$$\begin{aligned} l(\beta) &= \log[\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{(1-y_i)}] \\ &= \sum_{i=1}^{n}[y_i \log(p_i) + (1-y_i)\log(1-p_i)] \\ &= \sum_{i=1}^{n}[y_i\mu(x_i) - \log(1+\exp(\mu(x_i))] \end{aligned} \tag{3.3}$$

By adding an l_1 penalty term, we can derive the estimators of logistic regression with LASSO are

$$\hat{\beta} = arg \min_{\beta} [-l(\beta) + \lambda \beta_1] \tag{3.4}$$

For this model, we use the magnitude of the normalized coefficients to measure the feature importance, i.e. $IV_{LASSO}(j) = |\frac{\beta_j}{\sqrt{\sum_{j \in 1,2,..p_1} \beta_j^2}}|$ where β_j is the coefficient of feature j.

- Gradient Boosting Machine (GBM)

  GBM is an ensemble of weak learners and we use decision tree as the base learner here. Different from random forest, it iteratively builds additive models by sequentially fitting a base learner to current "pseudo" residuals which are the negative gradient of the loss function[168]. The process can be summarized as follows: Denote the loss function as $L(y, f(x))$,

  1. Initialize the model as $f_0(x) = c$ where c is a constant
  2. Set the number of iterations as M. For each iteration $m(m = 1, 2, \ldots M)$, the pseudo-residuals are computed as $r_{im} = -\frac{\partial L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)}$ for $i = 1, 2, \ldots, n$.
  3. Fit a base learner $h_m(x)$ to the pseudo-residuals
  4. Compute the multiplier γ_m such that $\gamma_m = argmin \sum_i^n L(y_i, f_{m-1}(x_i) + \gamma h_m(x_i))$.
  5. Update the model: $f_m(x) = f_{m-1}(x) + \gamma h_m(x)$
  6. After all iterations, output the final model

  In case of the overcapacity of base learners, at each iteration it randomly select a subsample of training data in place of the full sample to fit the learner.[170] For each decision tree, at each node i except the leaf node, the node importance NI_i is calculated as the reduction in node purity from the split at this node; the feature importance FI_j is calculated as the sum of node importance split on feature j divided by the sum of all node importance.

$$\begin{aligned}
&NI_i = w_i * \text{Impurity}_i - w_{\text{left}(i)} * \text{Impurity}_{\text{left}(i)} - w_{\text{right}(i)} * \text{Impurity}_{\text{right}(i)} \\
&FI_j = \frac{\sum_{i\text{: nodes split on feature } j} NI_i}{\sum_{m\text{: all nodes}} NI_m} \\
&\text{Impurity}_i = \sum_{c=1}^{2} p_{ic}(1 - p_{ic}) \\
&w_i = \text{weighted number of samples at node } i \\
&\text{left}(i) \text{ or right}(i) = \text{child node from split on node } i
\end{aligned}$$

p_{ic} is the frequency of label c at node i. The feature importance in GBM is the averaged importance over the individual trees, i.e. $IV_{GBM}(j) = \frac{\sum_{\text{all trees}} FI_j}{\text{No. of trees}}$

For both $IV_{LASSO}(j)$ and $IV_{GBM}(j)$, the higher the value, the more important the variable. Features are ranked by the importance value in descending order and the rank is denoted as r_j.

Ranking aggregation

Given a finite sample size, a small change in the data may result in big difference in the feature selection of machine learning models, so we bootstrap the training data for 20 times and repeat the feature ranking sub-step to obtain feature importance rank from different data samples. Then we aggregate the 20 feature rankings into a more stable ranking. An extensive study[171] on comparing different rank aggregation techniques concluded that different methods tend to be similar as the size of feature subset increases, so we use two representative rank aggregation methods in this step: One is the simple average of the rank from each bootstrap subsample and another is the weighted aggregation with the out-of-bag validation AUC as the weight.

The simple average is

$$F_j = \frac{1}{B} \sum_{b=1}^{B} r_j^b \tag{3.5}$$

The weighted average is denoted as

$$F_j = \frac{1}{\sum_{b=1}^{B} AUC_{OOB,b}} \sum_{b=1}^{B} AUC_{OOB,b} * r_j^b \tag{3.6}$$

Feature size estimation

Based on the aggregated feature ranking from the ranking aggregation sub-step, we need to find the feature size k_{opt} that achieves the highest AUC. Here we use the cross-validated AUC to evaluate model performance. Since feature size k has its corresponding AUC, the sequence of AUC for different k can be considered as a function of k, denoted by $f(k)$. To estimate the minimal size of features that can achieve close-to-optimal prediction accuracy, we use the golden section search procedure, a search technique that can find the maximum of a function over a specified interval. This algorithm searches for the maximum by updating a triplet of points with specified updating criteria.[51] In this sub-step, we update the triplet only when there will be significant change in the AUC and DeLong test is used to test the difference between $AUCs.(\alpha = 0.05)$

$$H_0 : AUC_1 = AUC_2 \quad v.s. \quad H_1 : AUC_1 \neq AUC_2 \tag{3.7}$$

In DeLong test, we first need to estimate AUC with Mann-Whitney statistic, which is a non-parametric unbiased estimator, denoted by $eAUC$:

$$eAUC = \frac{1}{n_0 n_1} \sum_{x_i \in D_0} \sum_{x_j \in D_1} I[a^T x_i, a^T x_j]$$

$$\text{with } I[a^T x_i, a^T x_j] = \begin{cases} 1, & \text{if } a^T x_i < a^T x_j \\ 0.5, & \text{if } a^T x_i = a^T x_j \\ 0 & \text{otherwise} \end{cases}$$

where D_0 and D_1 are the sets of control and case observations, respectively and n_i is the sample size $(i = 0, 1)$. Then we can derive the test statistic as follows:

$$z = \frac{(eAUC_p - eAUC_{p-k} - (AUC_p - AUC_{p-k}))}{\sqrt{(1, -1)S(1, -1)^T}} \dot{\sim} N(0, 1) \tag{3.8}$$

p is the original number of features and k is the number of features removed. S is the covariance matrix of $(eAUC_p, eAUC_{p-k})$. When the $p-value$ of the test is smaller than 0.05, we reject the null hypothesis, which states that there is no difference between the two $AUCs$.

3.1.3 Wrapper Selection

To further reduce the number of features, we apply the leave-one-covariate-out (LOCO) method to test the significance of each remaining feature from embedded selection. LOCO approach has the drawback of being computationally intensive when the feature size is large, but the computation burden will be largely reduced after the filter and embedded step, which is one reason why we put it in the last step. With the p_2 selected features, we fit the model and get the cross validated AUC (CV-AUC) of the training set, denoted by AUC_f. Then for each of the p_2 features, we exclude it and retrain the model to calculate the CV-AUC. It proceeds as follows, we exclude the j^{th} feature and refit the model with the remaining features and calculate the AUC as AUC_{-j}. Then if $AUC_f > AUC_{-j}$, it means the inclusion of feature j will improve the performance thus keeping this feature and removing it otherwise. Logistic regression is used here to calculate the AUC for its computation efficiency. Figure 3.2 below shows the process of this step. It can also test the adequacy of embedded feature selection by the drop of feature size after this step, that is, if the feature size doesn't drop significantly, we can tell that embedded step has reduced the feature size sufficiently.

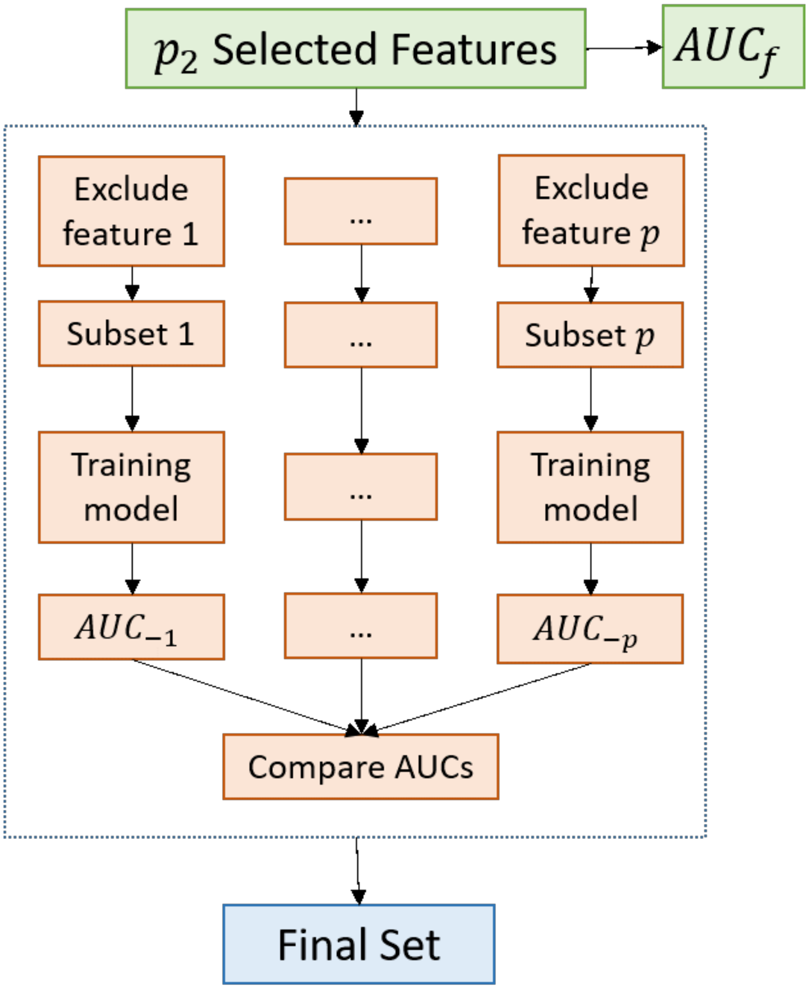

Figure 3.2. Flowchart of the wrapper selection

3.2 Time-to-event prediction

For adverse outcome prediction problem like readmission, both the readmission status and days to readmission are available. But the traditional practice in readmission prediction often ignores the days-to-readmission and only builds classification model on the readmission status, which may miss important information. Here we explore the possibility of improving prediction performance by using both the status and time information. This kind of time-to-event problem is usually treated as survival analysis problem. Suppose the time-to-readmission is denoted by T, then T is the "survival" time and the survival function is defined as $S(t) = P(T > t)$. Suppose $F(t) = 1 - S(t)$ has the density function $f(t)$ thus,

$$\begin{aligned} S(t) - S(t+\Delta t) &= P\{t \le T < t + \Delta t\} \\ &= f(t)\Delta t \text{ as } \Delta t \to 0 \end{aligned} \tag{3.9}$$

By transformation, we have

$$\begin{aligned} f(t) &= \lim_{\Delta t \to 0} \frac{S(t) - S(t+\Delta t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} -\frac{S(t+\Delta t) - S(t)}{\Delta t} \end{aligned} \tag{3.10}$$

Another important concept is the hazard function defined as:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P\{t \le T < t + \Delta t | T \ge t\}}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{S(t) - S(t+\Delta t)}{\Delta t * S(t)} \\ &= \frac{f(t)}{S(t)} \end{aligned} \tag{3.11}$$

thus $h(t) = \frac{f(t)}{S(t)}$. There are three classical survival distributions often used as the distribution of survival time T, i.e. Weibull distribution, exponential distribution and log-logistic distribution. [37] And we can derive the survival function and hazard function in Table 3.1:

Table 3.1. The survival, density and hazard function of Weibull, Exponential and Log-logistic distribution

Distribution	$S(t)$	$f(t)$	$h(t)$
Weibull(λ, p)	$\exp(-\lambda t^p)$	$\lambda p t^{p-1} \exp(-\lambda t^p)$	$\lambda p t^{p-1}$
Exponential(λ)	$\exp(-\lambda t)$	$\lambda \exp(-\lambda t)$	λ
Log-logistic(θ, κ)	$\frac{1}{1+\exp(\theta)t^k}$	$\frac{1}{1+\exp(\theta)t^k}$	$\frac{1}{1+\exp(\theta)t^k}$

The above survival processes are also called the baseline survival function since the effect of covariates on the survival status is not considered. To study the effect of covariates on the survival time, there are two classical survival models based on different assumptions.

The first type is the family of accelerated failure time (AFT) models[172], which assumes that the covariates act as acceleration factors to speed up or slow down the survival process compared with the baseline survival function. Suppose we have the feature vector $x = (x_1, x_2, \ldots, x_p)$ and the baseline survival time T_0, then the assumption is $T = \exp(\beta^T x)T_0$ and $\exp(\beta^T x)$ is also called the acceleration factor, thus the process is speed up if $\exp(\beta^T x) > 1$ and slowed down otherwise. With each of the three classical survival distributions above[173] as baseline function $S_0(t)$, we can derive three types of AFT model:

- Weibull AFT model

  Given x, we can derive conditional survival distribution of T,

$$
\begin{aligned}
S(t|x) &= P(T > t|x) \\
&= P(\exp(\beta^T x)T_0 > t|x) \\
&= P(T_0 > \exp(-\beta^T x)T_0|x) \\
&= S_0(\exp(-\beta^T x)t)
\end{aligned}
\tag{3.12}
$$

Since $S_0(t)$ is Weibull distribution with scale λ and shape $p = \frac{1}{\sigma}$, we have

$$\begin{aligned} S(t|x) &= \exp(-\lambda(\exp(-\beta^T x)t)^p) \\ &= \exp(-\lambda_1 t^{\frac{1}{\sigma}}) \\ \text{where } \lambda_1 &= \lambda \exp(-\frac{\beta^T x}{\sigma}) \end{aligned} \tag{3.13}$$

- Exponential AFT model

  Exponential distribution is a special case of Weibull distribution when $p = 1$, thus we have

$$\begin{aligned} S(t|x) &= \exp(-\lambda_1 t^{\frac{1}{\sigma}}) \\ \text{where } \lambda_1 &= \lambda \exp(-\beta^T x) \end{aligned} \tag{3.14}$$

- Log-logistic AFT model

  Same as before, let $k = \frac{1}{\sigma}$, we can derive the conditional survival distribution of T,

$$\begin{aligned} S(t|x) &= S_0(\exp(-\beta^T x)t) \\ &= \frac{1}{1 + \exp(\theta)(\exp(-\beta^T x)t)^k} \\ &= \frac{1}{1 + \exp(\theta_1)t^{\frac{1}{\sigma}}} \\ \text{where } \theta_1 &= \theta - \frac{\beta^T x}{\sigma} \end{aligned} \tag{3.15}$$

If we add an intercept term in β then, $x = (1, x_1, x_2, \ldots, x_p)$ and λ, θ in Equation 3.12 - 3.15 can be combined into β. Thus the parameters need to estimate are β and σ, which can be solved with partial likelihood method[174] and Newton Raphson (NR) algorithm.

The second type is the Proportional Hazard (PH) model, which assumes that the covariates execute their effect on the hazard rate rather than on the survival time directly, that is, $h(t) = \exp(\beta^T x)h_0(t)$. For different baseline survival distribution, we can derive the survival function in the same way as before and estimate the parameters with NR method. Cox proportional hazards model[136] is an approach that can estimate the effect parameter without considering the hazard function.

Since $h(t) = \frac{f(t)}{S(t)} = -\frac{dS(t)}{S(t)} = -d\log(S(t))$, we have $S(t) = \exp[-\int_0^t h(u)du]$. Thus,

$$\begin{aligned} S(t|x) &= \exp[-\int_0^t \exp(\beta^T x) h_0(u)du] \\ &= \exp(-\exp(\beta^T x)\int_0^t h_0(u)du) \\ &= \exp(-\exp(\beta^T x)H_0(t)) \end{aligned} \tag{3.16}$$

where $H_0(t)$ is the cumulative hazard function and can be estimated with the Breslow's method.[175]

Both of the AFT model and Cox proportional hazards (Cox PH) model assume that each subject will eventually experience the event of interest, given enough follow-up time. But for the readmission prediction problems, there may be some exceptions, like some patients may never get readmitted. In other words, there should be a "cure" fraction in the population, and the mixture cure model[176, 177] is motivated for this need. It not only involves modeling the survival distribution of the uncured patients but also the cure rate.[139] The mixture cure model can be expressed as follows:

$$S(t|x, z) = \pi(z)S(t|Y = 1, x) + (1 - \pi(z)) \tag{3.17}$$

where $S(t|x, z)$ is the conditional survival function of the entire population and Y is the indicator of whether the patient will experience the event eventually.($Y = 1$ if the patient will experience it and $Y = 0$ if not). $S(t|Y = 1, x)$ is the conditional survival function of those susceptible to experience the event and in this study we use the Cox proportional hazards model to describe their survival distribution.

$$S(t|Y = 1, x) = \exp(-\exp(\beta^T x)H_0(t)) \tag{3.18}$$

The incidence portion is often modeled using the logit link function, where z is the new covariate vector for the logistic regression:

$$\pi(z) = \frac{\exp(b'z)}{1 + \exp(b'z)} \tag{3.19}$$

Here we will use the same set of features to fit the logistic regression and survival function, that is, $z = x$. Since the information of the indicator Y is not complete, an iterative approach like the expectation maximization (EM) algorithm is used for the parameter estimation thus the mixture cure is more computation-demanding. We will explore the use of the mixture cure model on adverse outcome prediction. For comparison, we will also use Cox PH and AFT model to predict readmission and test the significance of including cure fraction. The R-package `survival`[178] is used for AFT and Cox PH model and `smcure`[139] is for mixed cure model. For this time-to-event prediction problem, we will use the 30-day hip fracture readmission prediction as an example. When evaluating or comparing the performance, we will use the AUC at the time point of 30 days as the evaluation metric.

For survival analysis, we need to define the time-to-readmission and censoring status first. Time-to-readmission is defined as the number of days from the index hospital discharge to the first readmission before the date of data collection. And cases are censored if the patients don't get readmitted before the data collection date. In this study, we aim to improve prediction performance by integrating the time-to-readmission with survival analysis.

3.3 Steps for Personalized Predictive Modeling

To better capture the unique characteristics of individual patients as well as improve the prediction accuracy, we introduce the personalized modeling to adverse outcome prediction. The key idea of personalized modeling is to identify the most similar patients and fit a separate model to make prediction for each individual patient. It involves four steps: 1) Selection of features and similarity measurement; 2) Extraction of similar patients; 3) Predictive modeling; 4) Personalized risk factor profile building. Details of each step are discussed in each sub-section as follows.

3.3.1 Selection of features and similarity measurement

We first need to select a set of features for both similarity calculation and predictive modeling. Then with the selected set of features, we explore the effect of different similarity metrics on the predictive performance.

- Select a subset of features to measure the similarity among patients

  A survey on case based reasoning discussed about the top issues in deriving individual based models in medical field.[179] They pointed out that a representative and comprehensive case library is the key to the good model performance. In our study, EHRs database can help to solve this issue with large amount of various patient information. Another limitation of person based learning is the need of expert knowledge for feature selection and weighting. Our proposed feature selection is a perfect fit for this need. To ensure that only significant features are included when measuring patient similarity and reduce the burden of computation, we will use the final feature set selected by the proposed three-stage feature selection method.

- Select a static similarity measure or learn it by supervised metric learning

  The similarity and dissimilarity measures are important in clustering and classification analysis. [35, 180, 181] Various methods of measuring similarity among individuals have been extensively used in different fields, like biology, ecology and image recognition.[158–160] Some examples of similarity measures are listed in Table 3.2. Different similarity measures can be used to identify cohort of patients from the training set that are most similar to the test patient. The similarity function is often defined as Similarity$(x, y) = -f(x, y)$ where f is the distance function and the instances are described by p attributes. For Euclidean distance, $f(x, y) = \sqrt{\frac{1}{n}\sum_{j=1}^{p}(x_i - y_i)^2}$ and is often used for numeric-valued attributes

  In different clinical scenarios, we may require different similarity metrics rather a single static measure like Euclidean distance.[35, 185] For example, patients that are similar to each other with respect to one disease, e.g. heart failure, may not be similar for a different disease such as diabetes. One remedy to this is the distance metric learning techniques, like Locally Supervised Metric Learning (LSML)[186] and Large

Table 3.2. Examples of Similarity Measures

Category	Distance/Similarity	Examples
Static Distance	General	Euclidean; Cosine; Manhattan[35]
	Binary vector	Jaccard index; Tanimoto; Roger and Tanimoto[160]
Learned Distance	Supervised	Locally Supervised Metric Learning[177]; Sparse Distance Metric Learning[182]; Large Margin Nearest Neighbor(LMNN)[183]
	Unsupervised	Covariance metric[184]

Margin Nearest Neighbor Classification (LMNN) [187], which can be trained for specific clinical condition. Many researches have shown that learning a distance metric can significantly improve the classification accuracy.[187–189] Distance metric learning is done by learning a transformation matrix W from the training data and then the distance metric between any two observations (x_i, y_i) and (x_j, y_j), is defined as follows:

$$D_{ML} = \sqrt{(x_i - x_j)^T W W^T (x_i - x_j)} \tag{3.20}$$

The trainable metric is customized for the problem setting thus satisfying the requirement that different clinical scenarios need different similarity metrics to measure how similar two patients are. For example, LMNN learns the transformation matrix as a neighborhood margin maximization problem which tries to keep the close k-nearest neighbors from the same class, while keeping examples from different classes separated by a large margin.[190] Based on this objective, W can be determined by solving the following optimization problem:

$$\min_W \sum_{i,j \in N_i} D_{ML}(x_i, x_j) + \lambda \sum_{i,j \in N_i, l, y_l \neq y_i} [D_{ML}(x_i, x_j) + 1 - D_{ML}(x_i, x_l)]_+ \tag{3.21}$$

where

$$D_{ML}(x_i, x_j) = \sqrt{(x_i - x_j)^T W W^T (x_i - x_j)} \tag{3.22}$$

and N_i is the set of exactly k different nearest neighbors with label y_i under the learned metric, $[a]_+ = \max(a, 0)$ and $W \geq 0$.

In our study, we will compare the performance of personalized modeling with the Euclidean distance, cosine distance, LMNN and random selection.

3.3.2 Extraction of similar patients

Prior to similarity calculation, all continuous predictors are normalized to fit the range between 0 and 1. For each categorical predictor, we convert them into dummy variables. For any new patient, we compute their distance to all the patients in the training set and rank the training patients by the distance in ascending order. The top K patients are selected to form the similar patient cohort. K is a tuning parameter that can be chosen by cross validation. The selection of similar patients can be done in two ways: one is not considering the outcome occurrence rate of the training patients so the resulting similar patient cohort may not maintain the case-control balance; another is to select similar case and control patients separately and then combine them into one cohort to control the case-control balance. For simplicity, here we do not consider the case-control balance.

3.3.3 Predictive modeling

Once the personalized training cohorts are identified, any supervised classification or regression method can be used to fit the model and make prediction on the test patient. We use logistic regression as the predictive model here because it is not only interpretable but also produces risk probability. With the predictions for each test patient, we can combine them into a prediction vector and calculate the performance metric AUC to evaluate the prediction performance.

3.3.4 Building personalized risk factor profile

The personalized model can reveal the specific association of each factor to patients on the individual basis. From the logistic regression model, we can get the beta coefficient for each factor and also their significance level. And the factors with large coefficient value and significant test result are important risk factors, which can be used to build risk factor profile.

With the individualized risk factor profile, individual-based care management plan can be provided to patients, thus contributing to the better interactions between physicians and patients.

3.4 Applications

In this study, we will extract the study populations from the Cerner Health Facts EHR database (Cerner Corporation, Kansas City, MO). The data include time-stamped admission, diagnosis, laboratory, surgical, and medication information and are de-identified in compliance with Health Insurance Portability and Accountability Act (HIPPA). In pursuit of our goal of improve adverse outcome prediction with EHRs data, the proposed feature selection and model development, we will apply our proposed methods to two example adverse outcome prediction problems: one is the 30-day hip fracture (HF) readmission and the other is the diabetic retinopathy(DR) prognosis.

These two datasets are chosen for three reasons: The first reason is their clinical significance. Hip fracture readmission not only increases patients mortality rate but also brings big economic burden to both hospitals and patients [4, 191] and diabetic retinopathy is the leading cause of blindness in American adults.[7] However, compared with diseases and conditions like heart failure, heart attack and chronic obstructive pulmonary disease (COPD), the prediction problems of these two diseases are less studied.[10] The second reason is because they have different data properties; The HF data has much more features than the DR data and all features in the HF data are categorical or binary while most features in DR data are numeric. Lastly, the performance of prediction models on different adverse outcomes often varies widely, and it's often better on prognosis of diseases while worse on readmission. [20] The proposed methods can be better evaluated with data of different prediction difficulty. And the studies of these two questions have different application scenarios: the prediction of hip fracture readmission is more classification-oriented and tends to be used for patient risk stratification while the prognosis of diabetic retinopathy is for the purpose of preventing DR from happening.

3.4.1 30-day Hip Fracture Readmission

Hip fracture diagnoses and surgeries were identified using the International Classification of Diseases, 9th Revision - Clinical Modification (ICD-9-CM) diagnosis and procedure codes listed in Table 3.3. A patient's index admission is defined as an admission with one or more 820.xx hip fracture diagnoses as the primary or secondary conditions (priority $\leq$ 2) and at least one procedure coded as 79.15, 79.35, and 81.52. In this study, we will focus on the patients discharged between January 2006 and August 2015. A total of 38,981 index admissions with associated diagnosis and procedure code for hip fracture surgery are eligible for this study. Based on the exclusion rules defined in Figure 3.3, we end up with 35,561 index admissions as our study population. Since only 3.1% of the patients have more than one index admissions, we assume all the index admissions are independent. A readmission is defined as a subsequent all-cause inpatient admission in the same or a different hospital within 30 days following an index admission. We labelled each index admission by "readmitted" or "non-readmitted" based on whether they were re-hospitalized for any reason within 30-days of discharge from the index admission.

ICD-9-CM code	Description
820.xx	Fracture of neck of femur (hip)
79.15	Closed reduction of fracture with internal fixation; femur
79.35	Open reduction of fracture with internal fixation; femur
81.52	Partial hip replacement

Table 3.3. ICD-9-CM codes used to identify hip fracture diagnoses and surgeries

Although hospital system characteristics, such as the bed size, also influence the readmission rate, it should not affect the quality of care[192]. Therefore, our study doesn't include such hospital characteristics. The predictor variables examined in this study can be grouped into

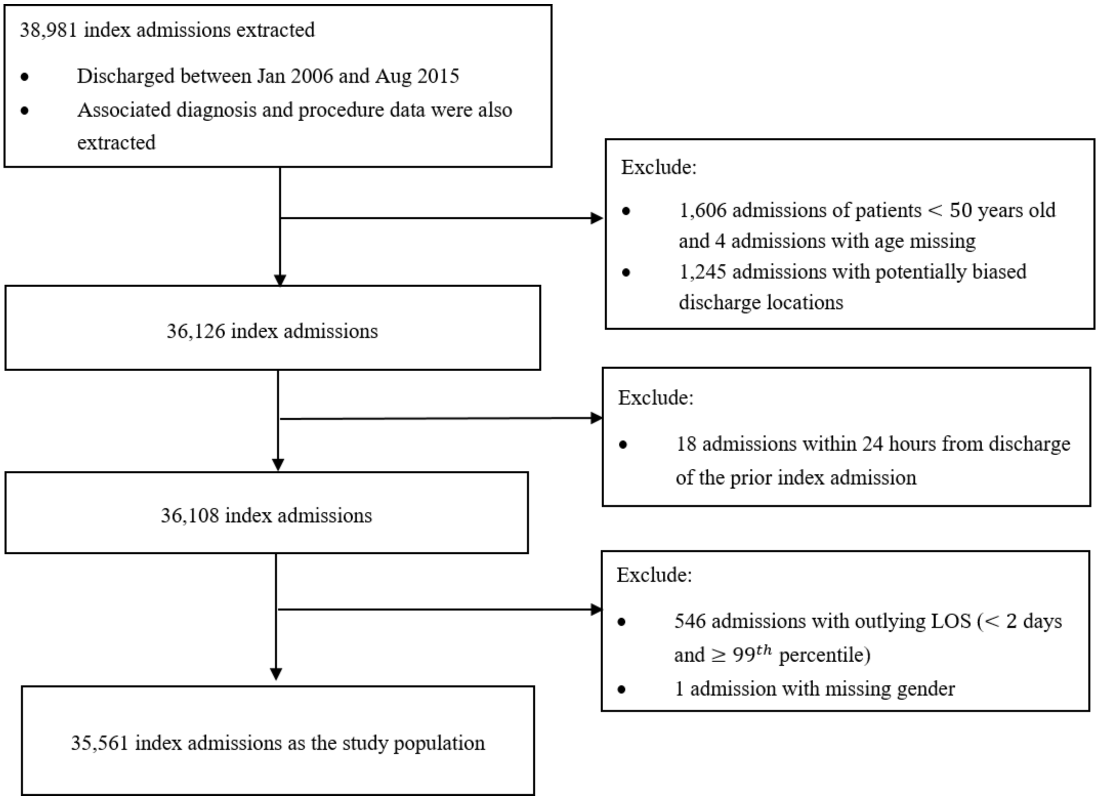
38,981 index admissions extracted
Discharged between Jan 2006 and Aug 2015
Associated diagnosis and procedure data were also extracted
Exclude:
1,606 admissions of patients < 50 years old and 4 admissions with age missing
1,245 admissions with potentially biased discharge locations
36,126 index admissions
Exclude:
18 admissions within 24 hours from discharge of the prior index admission
36,108 index admissions
Exclude:
546 admissions with outlying LOS (< 2 days and ≥ 99th percentile)
1 admission with missing gender
35,561 index admissions as the study population

Figure 3.3. Cohort derivation of 30-day hip fracture readmission

six categories:

- Demographics, e.g. age, race, gender and marital status
- Encounter-related variables, e.g. discharge location and length of stay
- Comorbidities and Charlson Comorbidity Index[193]
- Procedures and diagnosis
- Lab tests
- Hospital utilization: the number of hospital visits (inpatient or emergency room (ER)) during six months/one year prior to the index admission

The diagnoses and procedures were classified based on the Clinical Classifications Software (CCS) available from the Agency for Healthcare Research and Quality. The Charlson Comorbidity Index is calculated based on the diagnoses recorded during the stay of each admission using the algorithm presented in[194]. There is also research showing that previous admission to hospitals is a robust predictor of future hospitalization.[195] In this study, we construct hospital utilization features by aggregating patients' previous inpatient and emergency room visits. The number of prior inpatient / emergency room (ER) visits is counted based on inpatient / ER encounters during the past 3 months / 6 months / 1 year prior to the index admission. All the continuous variables are discretized into nominal variables since we check that there is no significant difference between the performance of their continuous and categorical version, but the discretized data is relatively easier to interpret from the clinical point of view. After discretization and dummy encoding, there are 534 features in total and the readmission rate is 11%.

3.4.2 Diabetic Retinopathy Prognosis

The diagnosis of diabetic retinopathy is identified by the International Classification of Diseases – Ninth Revision (ICD-9) codes as shown in Table 3.4. We firstly extract diabetic patients who have one or more 250.xx ICD-9-CM diabetes codes. Then based on the appearance of diabetic retinopathy (DR) diagnosis code 362.0x ICD-9-CM, patients were labelled by either "DR" or "non-DR". A total of 52,375 patients are selected to form the final study population, including 2,147 (4.1%) DR patients and 50,228 (95.9%) non-DR patients.

ICD-9-CM code	Description
250.xx	Diabetes
362.0x	Diabetic retinopathy

Table 3.4. ICD-9-CM codes used to identify diabetic retinopathy diagnoses

Based on the literature on diabetic retinopathy prediction, we selected features that have shown statistically significant association with the onset of diabetic retinopathy. A total of 32 predictor variables were included in this study and they were grouped into two categories:

- Demographics: age, race and gender
- Lab results: HbA1c, blood urea nitrogen (BUN), creatinine, glucose, hemoglobin, hematocrit, calcium, triglyceride, potassium, chloride, mean corpuscular hemoglobin (MCH), sodium, mean corpuscular hemoglobin concentration (MCHC), mean corpuscular volume (MCV), albumin, bilirubin, protein, anion gap, aspartate aminotransferase (AST), alanine aminotransferase (ALT), alkaline phosphatase (ALP), high-density lipoprotein (HDL), red blood cell count (RBC), white blood cell count (WBC), platelet, cholesterol, blood pressure diastolic (BPD) and blood pressure systolic (BPS)

CHAPTER IV

RESULTS

4.1 Feature selection

4.1.1 30-day Hip Fracture Readmission

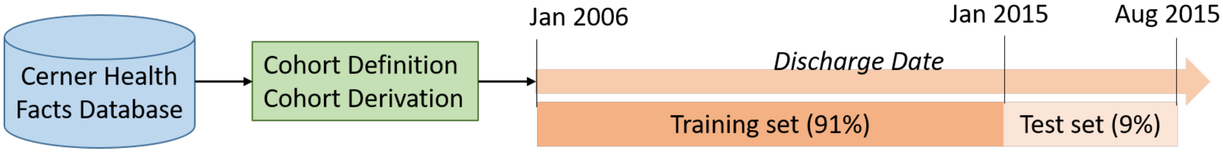

Figure 4.1. Data split for hip fracture readmission dataset

After the discretization of continuous features and dummy encoding of categorical features, we ended up with 534 features, all of which are encoded as 0/1 features. The common practice of prognostic research is training the model in the past and making predictions in the future[196], thus we split the data based on the discharge date of the index admission. The whole dataset is split into the training set and test set, with January 2015 as the cut point. 91% of the admissions have discharge date prior to January 2015 and they build the training set, and the remaining 9% form the test set. So the model is trained on the past patient records and validated with patients new to the system.[197] The readmission rate of the whole patient cohort is 11% while the rates for the training and test set are 11.16% and 9.48% respectively. Demographic characteristics of the patient population are presented in Table 4.1. We can observe some difference in the distribution of both the covariates and response variable between the training and test sets, which indicates the potential of population shift problem.

Demographic characteristics	Training (2006.01 – 2015.01)	Test (2015.01 – 2015.08)
N	32,332	3,229
Readmitted (%)	11.16	9.48
Age (%)		
[50, 65)	9.72	11.68
[65, 75)	14.76	15.17
[75, 80)	13.37	12.98
[80, 85)	20.35	18.74
[85, 90)	23.23	21.86
[90, 100]	18.57	19.57
Male (%)	27.66	29.24
Race (%)		
African American	5.41	5.45
Caucasian	89.34	88.05
Other	2.84	5.30
UNKNOWN	2.41	1.21

Table 4.1. Patient demographic characteristics of training and test sets – 30-day hip fracture readmission

We applied the proposed three-stage feature selection method on the training set to select the most predictive features. In the filter selection step, the chi-square test or Fisher exact test is used to test the univariate significance of each feature and 148 out of 534 are selected ($p - value < 0.1$). The remaining features are input into the embedded and wrapper selection step and the feature size change are depicted in Figure 4.2. On the left is the selection with Gradient Boosting Machine (GBM) as the embedded method and on the right is with Logistic regression (LR) with LASSO as the embedded method. Different color is used to denote the results from the two ranking aggregation methods: orange is for the *simple average of ranking*(denoted by ”Mean_rank” in the legend) and blue is for the *weighted average of ranking*(denoted by ”Wt_Mean_rank” in the legend).

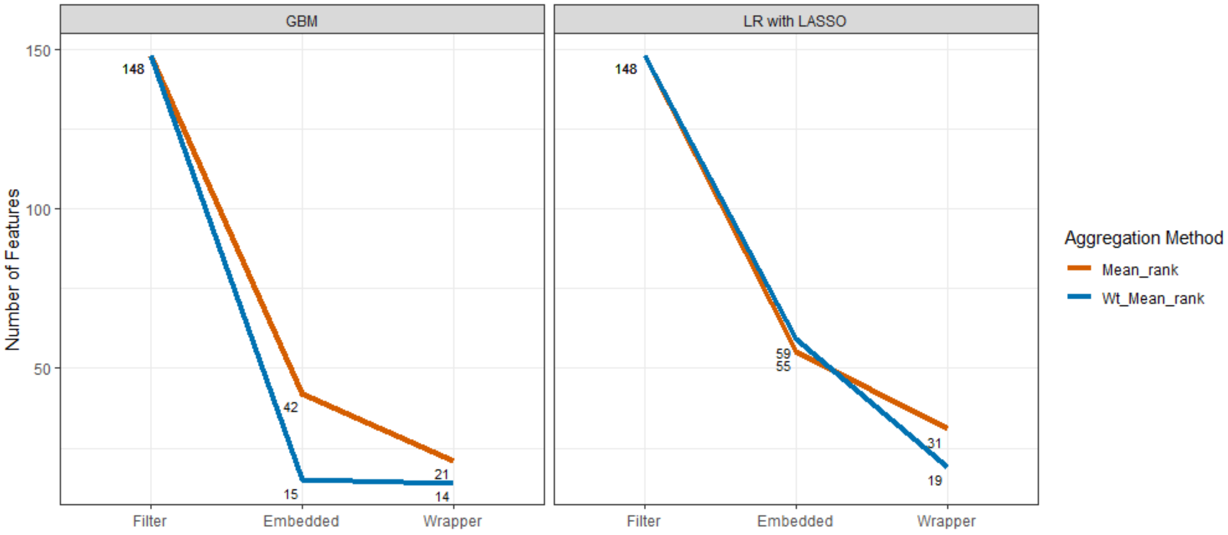

Figure 4.2. Number of selected features after each step – hip fracture readmission. On the left is the selection with Gradient Boosting Machine (GBM) as the embedded method and on the right is with Logistic regression (LR) with LASSO as the embedded method.

In both plots, there is a big drop in the feature size from the filter step to the embedded step for both ranking aggregation methods, indicating that there are still some falsely significant features remaining after the filter selection. But the filter step largely reduced the computation burden of embedded selection compared to staring with all 534 features. The two embedded methods reduce the feature size to different degrees and overall the GBM has a higher reduction in feature size than LR with LASSO. For the embedded selection with GBM, the *weighted average of ranking* ends in a much smaller set of features than the *simple average of ranking*. For LR with LASSO, the two ranking aggregation methods lead to similar feature size reduction. Features from the embedded step are then put into the wrapper step, which plays two roles: one is to reduce the feature size and the other is to detect whether the embedded selection is sufficient. For the GBM scenario, the feature set from the *simple average of ranking* is further reduced by half while that from the *weighted average of ranking* is just reduced by 1. The wrapper step seems to be more necessary for the *simple average of ranking* than the *weighted average of ranking*. This indicates that for the GBM embedded selection, the *weighted average of ranking* is more robust than the simple average. For the LR scenario, there is a big reduction in the feature size of both ranking aggregation methods, which indicates that LASSO doesn't reduce the feature size efficiently for either ranking aggregation method.

To fairly compare the performance of both methods, we should look at not only the size of selected features but also the prediction performance of the final feature set, thus we fit the logistic regression on the four selected feature sets as well as the original set to calculate the test AUC. Figure 4.3 shows the number of features and the test AUC of each feature set. The overall performance of the GBM scenario is better than the LR, which is in line with out expectations since GBM is more hypothesis-free and can capture the more complex non-linear structure in data. For the selection with GBM, compared with the simple average of ranking, the feature set selected from the *weighted average of ranking* performs better with higher test AUC and smaller size, which makes sense since it aggregated the feature rankings from 20 bootstrapped samples by taking weighted average of the rankings with the out-of-bag (OOB) AUC as weight. Compared with the 534 features in original feature set, the feature set from the *weighted average of ranking* selects only 14 features but reaches 98.37% of the original test AUC.

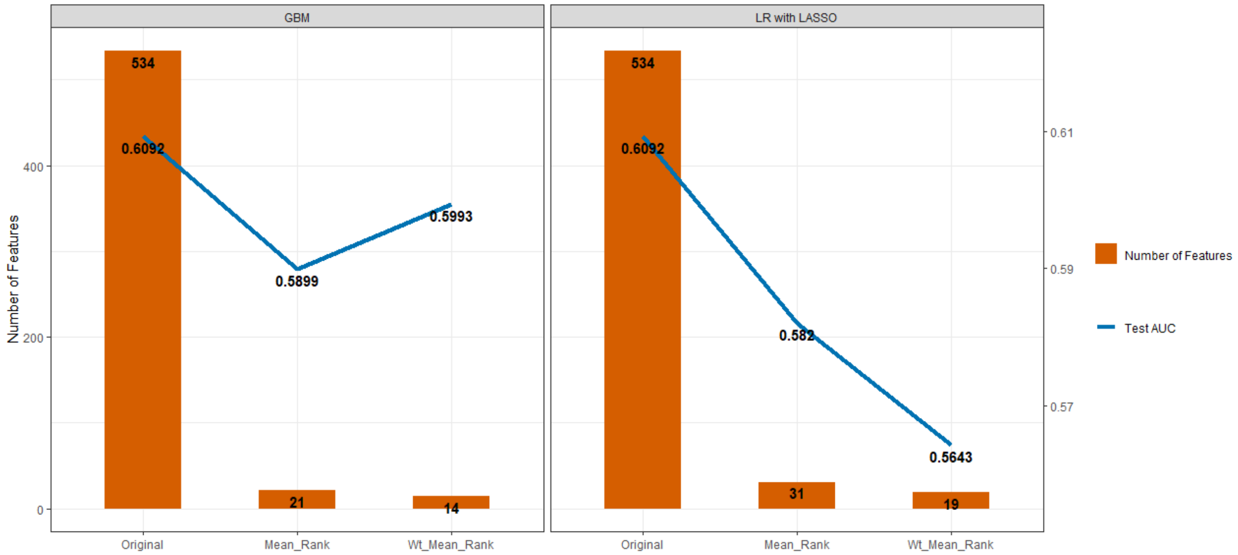

Figure 4.3. Test AUC based on different feature sets. The left plot is from the feature sets selected by GBM and the right from LR with LASSO. In both plots, the x axis represents the feature sets and the primary and secondary y axis respectively represents the number of features (right) and the test AUC (left).

GBM seems to benefit more from feature ensemble and outperforms LASSO with a higher accuracy and better stability. Among the 14 features selected by the "best" selection scenario (GBM with the *weighted average of ranking*), half of them are related to the lab tests and diagnosis. Some features are known risk factors of readmission, including age, gender and length of stay. Some less widely investigated features like previous inpatient visits and discharge location are also identified.

4.1.2 Diabetic Retinopathy

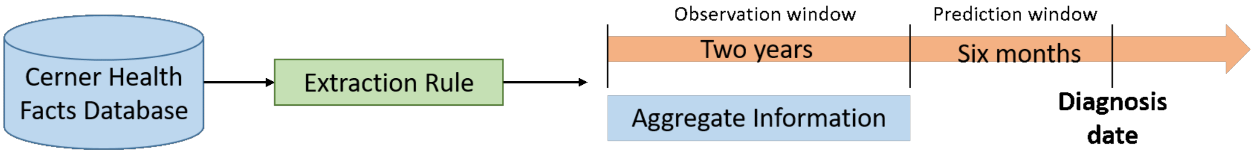

Figure 4.4. Data split for diabetic retinopathy prognosis dataset

Since diabetic retinopathy can be prevented with early treatment, we need to make early detection before the diagnosis. This means that we need to use information from much earlier stage to make prediction and suggest prevention treatment. There are two terms involved in the data collection: 1) Prediction window, which is the time window before the diagnosis that the disease is detected. 2) Observation window, which is the time window before the prediction window from which the data are collected for use.[79] In this research we chose two years for the observation window and six months for the prediction window. As shown in Figure 4.4, the response label of patients are obtained on the diagnosis date but the predictors are aggregated over the 2-year-long observation window. The whole cohort is split into training and test sets with random selection in the ratio of 7 to 3.

The lab tests selected for this research are all numerical variables, including counts and continuous measurements. It's hard to find reasonable cut points to discretize all of them. Thus, in this research, we would not discretize the numerical features but kept their original form, and the same for the age variable. Table 4.2 include the patient demographic characteristics in the training and test set and we can observe that the training and test sets have more similar distributions in both the covariates and response variable than the hip fracture readmission dataset.

Demographic characteristics	Training (70%)	Test (30%)
N	36,662	15,713
DR (%)	4.18	3.91
Mean age in years (SD)	63.43 (14.42)	63.24 (14.47)
Male (%)	45.30	45.73
Race (%)		
African American	16.65	16.45
Caucasian	74.83	75.22
Other	8.52	8.33

Table 4.2. Patient demographic characteristics of training and test sets – diabetic retinopathy prognosis

Then we applied the three-stage feature selection method on the training set to find the smallest significant feature subset. In the filter selection step, the chi-square test and t test are used to test the univariate significance of each feature and 25 out of 32 are selected ($p - value < 0.1$). We put the remaining features into the embedded and wrapper selection step and the feature size change are depicted in Figure 4.5. On the left is the selection with GBM as the embedded method and on the right is with Logistic regression (LR) with LASSO as the embedded method. Different color corresponds to the two ranking aggregation methods: orange for the *simple average of ranking* and blue for the *weighted average of ranking.* For the GBM scenario, we can observe that, after the embedded step, the two ranking aggregation methods had the same drop in the feature number. Thus we compared the selected feature set from both methods and found that although they selected the same number of features, there is one feature different between the two sets. Features from the embedded step are then put into the wrapper selection. Figure 4.5 shows that even starting with the same number of features, the two aggregation methods ended up in different number of features after the wrapper step. Feature set from the *simple average of ranking* is reduced by 2 while that from *weighted average of ranking* is reduced by 4, and it may indicate that the latter one can identify more falsely significant features from the embedded step, which is also proved by the AUCs in Figure 4.6. The selected features are put in the Appendix for reference and comparison. For the LASSO scenario, the two ranking aggregation methods show no difference in the feature reduction of embedded step and thus ending the same feature set after wrapper step. The feature size drops significantly from 25 of the filter step to 6 of the embedded step and no drop from the embedded step to the wrapper step. By comparing the two final selected feature sets in LASSO scenario, we found that the two ranking aggregation methods have selected the same set of features.

Same as before, we calculated the test AUC of logistic regression fit on the two selected feature sets and the original set. Figure 4.6 shows the number of features and test AUC of each feature set. Still, the overall performance of GBM is better than LASSO. For the former scenario, compared with the *simple average of ranking*, the features selected by the *weighted average of ranking* performed better, with higher test AUC and smaller feature size. The test AUC of features from the *weighted average of ranking* is about 98.19% of the AUC on the original feature set while its feature size is just 40.6% of the original size.

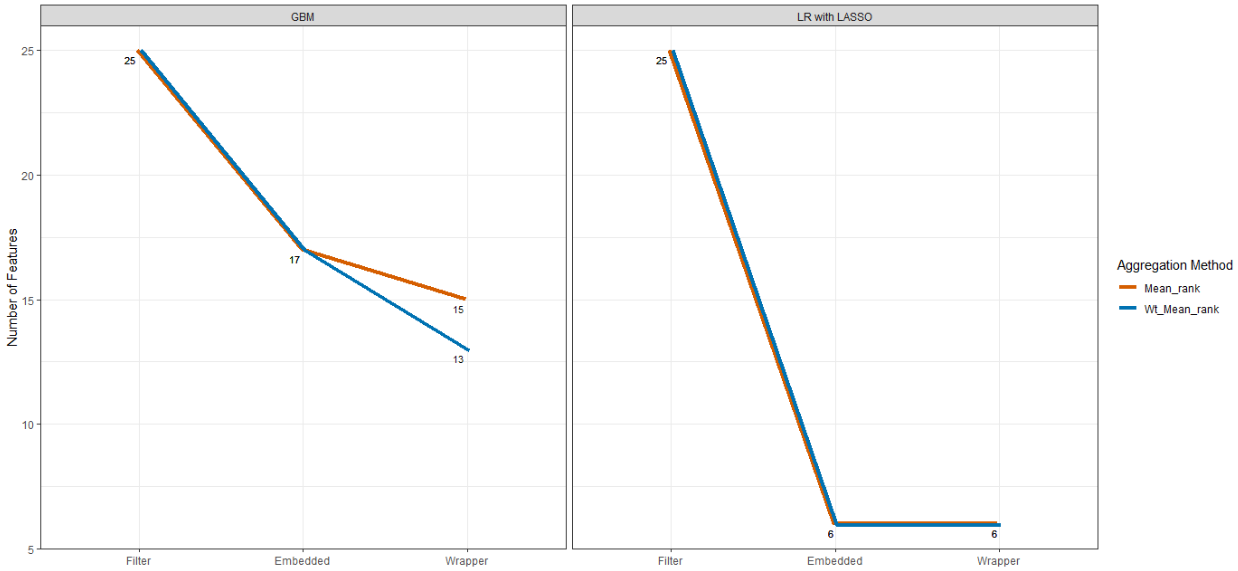

Figure 4.5. Number of features selected after each step - diabetic retinopathy prognosis. On the left is the selection with GBM as the embedded method and on the right is with Logistic regression (LR) with LASSO as the embedded method.

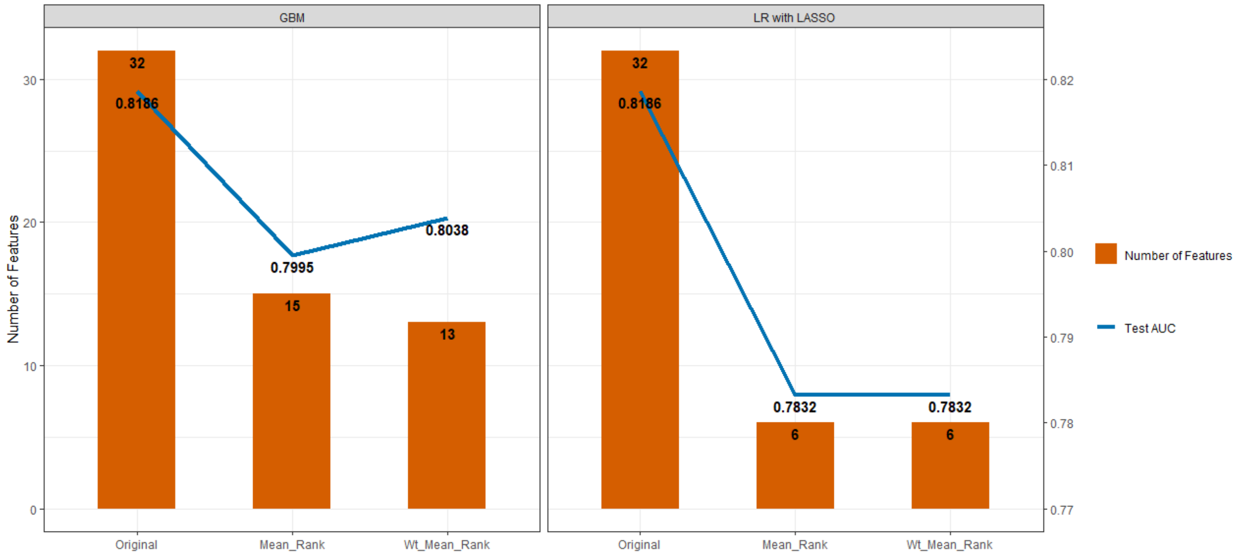

Figure 4.6. Test AUC based on different feature sets. The left plot is from the feature sets selected by GBM and the right from LR with LASSO. In both plots, the x axis represents the feature sets and the primary and secondary y axis respectively represents the number of features (right) and the test AUC (left).

By comparing the feature selection result of hip fracture readmission and diabetic retinopathy prognosis, GBM outperforms LASSO with reasonable feature size and higher prediction accuracy. We can observe that under the GBM scenario, for both datasets, the *weighted average of ranking* achieves the best performance since it selected less than half of both original feature sets but the prediction accuracy is about the same as the original feature sets. But there is also difference in their magnitude of feature reduction. Potential reasons for this include:

- The predictive power of the original features. It's easy to tell that compared with the HF data, the original DR data has a smaller number of features but there are more predictive features to the response variable. It also means that the HF data contains more insignificant features than DR data. That's why the feature size reduction of DR data is smaller than that of HF data but the overall performance of DR data is in higher level.
- The population shift problem. The way that the two datasets are split into the training and test sets is different, one is by time and the other is by random selection. This

partially explains the poor performance of the HF model on the test set.

4.2 Prediction with the time-to-event

For each admission of patients to hospital, there are two important time points, the admission date and the discharge date. If we use the readmission date to minus the discharge date of the index admission, we will get the days-to-readmission. If the days-to-readmission is less than or equal to 30 days, it also means that patients get readmitted within 30 days. Although classification models are often used to predict patients' 30-day readmission status, in this part we converted it into a time-to-event problem and use survival analysis to predict the status. Since it's hard to define a meaningful term like "days-to-readmission" for diabetic retinopathy prognosis problem, we only use the hip fracture data as example in this part.

Here we used the same training and test sets as in the feature selection experiment. The training set is used to fit the model and the test set is used for evaluation. Since the ubiquitous evaluation metric in readmission prediction modeling is AUC. And AUC can be computed for survival analysis at any timepoint of the survival curve.[198] In this research, we calculated AUC for the test set at day 30, day 60 and day 90, as listed in Table 4.3.[199]

Method/AUC	Mean rank			Weighted mean rank		
	30	60	90	30	60	90
AFT Weibull	0.5929	0.5908	0.5897	0.6011	0.5986	0.5982
AFT Exponential	0.5995	0.5971	0.5967	0.5993	0.5973	0.5962
AFT Log-logistic	0.5932	0.5910	0.5998	0.6009	0.5985	0.5981
Cox PH	0.5900	0.5881	0.5867	0.6006	0.5981	0.5981
Mixture cure	0.5908	0.5888	0.5873	0.6007	0.5982	0.5981

Table 4.3. Test AUC for different methods, at three timepoints, day 30/60/90

Since our goal in this part is to compare the performance of classification model and time-to-event prediction model and the computation would be much slower with the original set of features, we trained each model with the two selected final feature sets from the GBM scenario. We can observe that overall, features selected by the *weighted average of ranking* outperformed features from the *simple average of ranking.* For each set of features, the family of AFT models performs slightly better than others but there is no significant difference among different models. And it's unexpected that the mixture cure model doesn't outperform the others.

Although survival analysis doesn't significantly outperform the logistic regression, it does provide flexible and dynamic analysis on patient risk status. With survival analysis, we can predict the probability of an adverse outcome at any time point rather than at a certain time point. So we are able to extend the classification problem to any time point we are interested in. For example, for readmission prediction problem, the time window used to define readmission has long been a controversial topic, like 15-day, 30-day, 60-day and 90-day.[130] With the time-to-event model, we can make readmission prediction at any time point with one single model rather than fitting one classification model for each of the time windows of interest.

4.3 Personalized Modeling (PM)

In this part, for the sake of computation, we used the features selected by the "best" combination, i.e. GBM with the *weighted average of ranking* aggregation method from previous result part. To study the effect of the number of nearest neighbor training patients (K) and the distance metrics on the prediction performance, we chose a set of candidate values for each of them and repeated the same personalized modeling steps. For the choice of K, we need to set it big enough to make sure the inclusion of observations from both the majority and minority class. Since the degree of imbalance problem and total observation number of the two datasets are different, the selected candidate values for K are a little different for them. For the distance metrics, we chose four candidate metrics, including 1) two static distance metrics – Euclidean distance, cosine distance; 2) two metrics from metric learning – LMNN + Euclidean distance; LMNN + cosine distance. Below are the detailed setting and results on the two datasets.

4.3.1 30-day Hip Fracture Readmission

We selected nine values for K, including 100, 500, 1000, 3000, 5000, 7000, 9000, 10000 and 15000 and plotted the performance of the personalized logistic regression as a function of K as shown in Figure 4.7. And there are five different distance configurations denoted by curves of different colors and labels. Additionally, we put the performance of global logistic regression model (∘) in the same plot as reference and need to note that it's trained with the whole training set and does not change with the size K. Firstly, as a baseline, K training patients are randomly selected to train the personalized model (∗). Its performance keeps increasing as the number of training patients increases and almost levels off after K=3000. With slightly increase, it reaches the performance of global model around K=8000. The steady increase before K=3000 is reasonable since the training of logistic regression requires enough data. And when the data size is big enough, more training observations won't

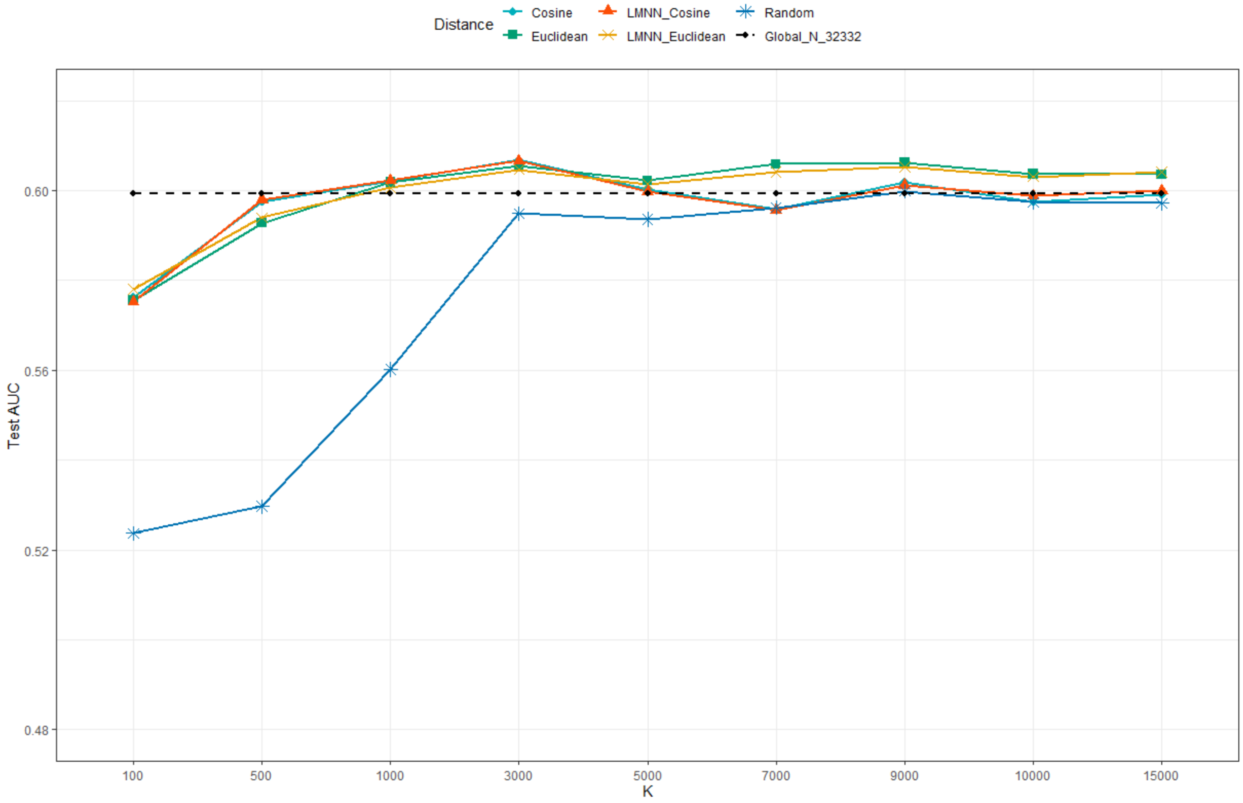

Figure 4.7. Performance of the personalized logistic regression model in terms of AUC as a function of K – hip fracture readmission

contribute much to its performance, and that's why there is just minor increase after this point. Then, instead of selecting the patients randomly, we used static distance metrics, the Euclidean distance (□) and the cosine distance (⋄), to select the K most similar patients for model training. For both of them, the performance is consistently better than the random selection, especially when K is small. This indicates the gain of using the more similar patients compared with the randomly selected patients. Their performance keeps increasing but starts to level off after K=3000, which implies that the inclusion of more dissimilar patients won't improve the performance. For cosine distance, there is even a minor decrease with more dissimilar patients included into model training. Lastly, we used the LMNN version of both distance metrics to select the most similar patients ($\times$ and $\triangle$). But there is no significant difference between the static measure and the trained measure for different values of K. Overall, the Euclidean distance performs slightly better than the others.

We also analyze the characteristics and distribution of the patient-specific risk factors with clustering analysis to prove the need of personalized predictive modeling. We randomly select 200 test patients and extract the coefficients of the risk factors when K=3000. Agglomerative hierarchical clustering is used here to cluster the risk factors and patients respectively and the cluster tree is shown in Figure 4.8.The rows are the risk factors and are clustered along the vertical axis. The columns stand for the risk factor profile of each patient and are clustered along the horizontal axis. The patient 30-day readmission status is also plotted as a horizontal bar on the top; green stands for non-readmitted and red stands for readmitted. Different colors are used to represent the coefficient values of the risk factors in the heat map: red means high and blue means low. Based on the risk factor profile (i.e. columns), similar patients are clustered together while those with very different risk factor profiles are put into clusters that are further away from each other. Patients with the same readmission status can have very different risk factor profile. For example, the patients with 30-day readmission are scattered in different clusters. For different patients, the same risk factor may exert effect on them in different directions, which the global model may fail to capture.

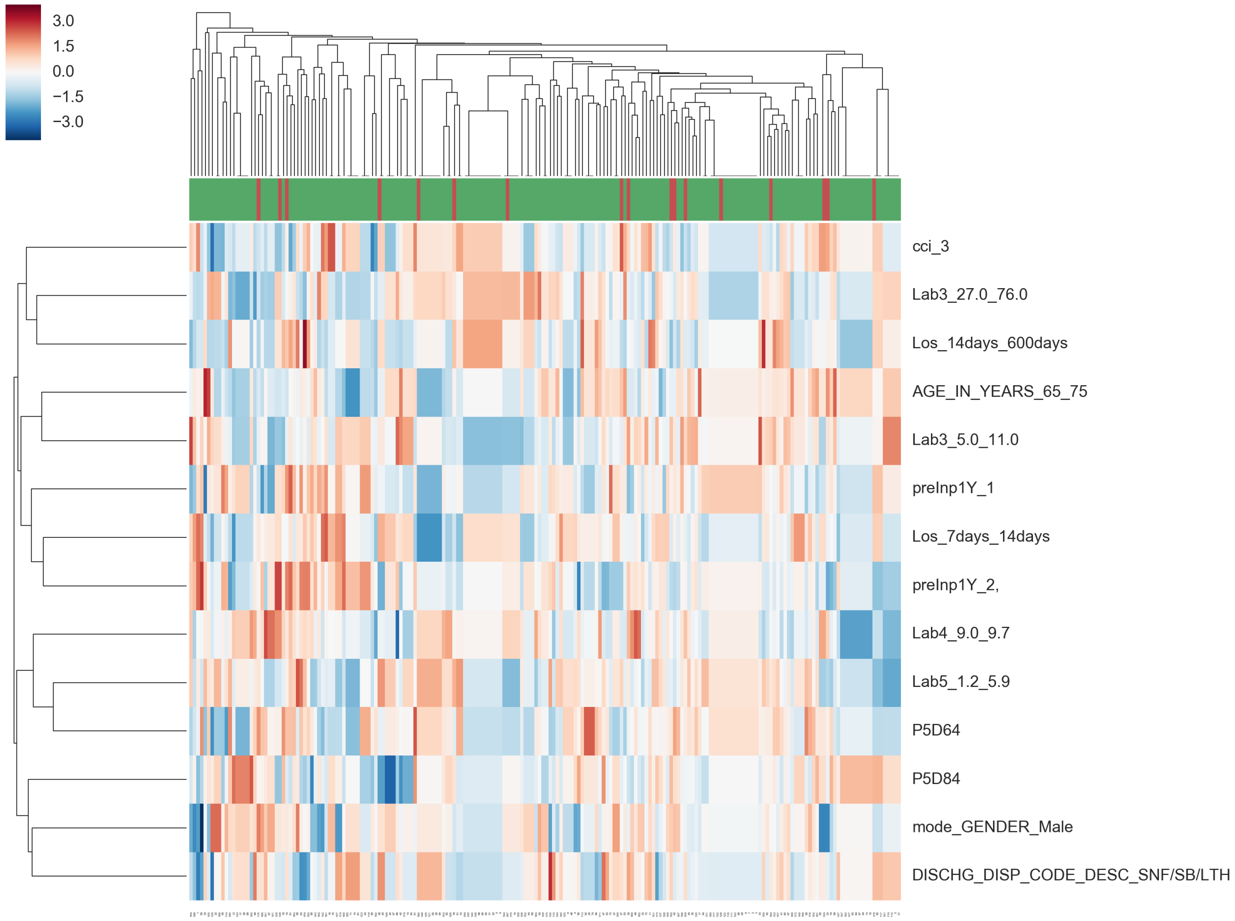

Figure 4.8. Hierarchical heat map plot of the risk factors from the personalized predictive models for 200 randomly selected test patients - Hip fracture readmission. The rows are the risk factors and are clustered along the vertical axis. Each column stands for a patient and the red-green horizontal bar on the top stands for their 30-day readmission status.(Green: Non-readmitted; Red: Readmitted.)

4.3.2 Diabetic Retinopathy prognosis

Compared with the hip fracture readmission data, diabetic retinopathy has a much lower occurrence rate. When we set K to be 100 or 500, sometimes the selected nearest neighbor training patient cohort doesn't contain any positive cases. Thus, we need to set K to be big enough and here we selected eight values for K, including 1000, 3000, 5000, 7000, 9000, 10000, 15000 and 20000. The performance of the personalized logistic regression as a function of K is shown in Figure 4.9. Other configurations are the same as in the previous experiment. Here we can observe the same patterns of how the performance change with the increase of K as before. But one big difference is the significantly better performance of personalized modeling over the global model and random selection for all values of K, compared with result of the hip fracture readmission. This is because, here K starts from 1000 and we are not able to see the performance when K is smaller as before. And for this dataset, the cosine distance and its trained distance perform better than the Euclidean distance.

Same as before, we also analyze the characteristics and distribution of the patient-specific risk factors with clustering analysis and plot the heat map of the coefficient values of risk factors in Figure 4.10. Here we randomly select 200 test patients and extract the coefficients of the risk factors when K=7000. The horizontal bar on the top stands for whether the patient is diagnosed of diabetic retinopathy on the diagnosis date; green stands for no and red stands for yes. Patients with the same diagnosis status can have very different risk factor profile and are scattered in different clusters, indicating the need for personalized predictive modeling.

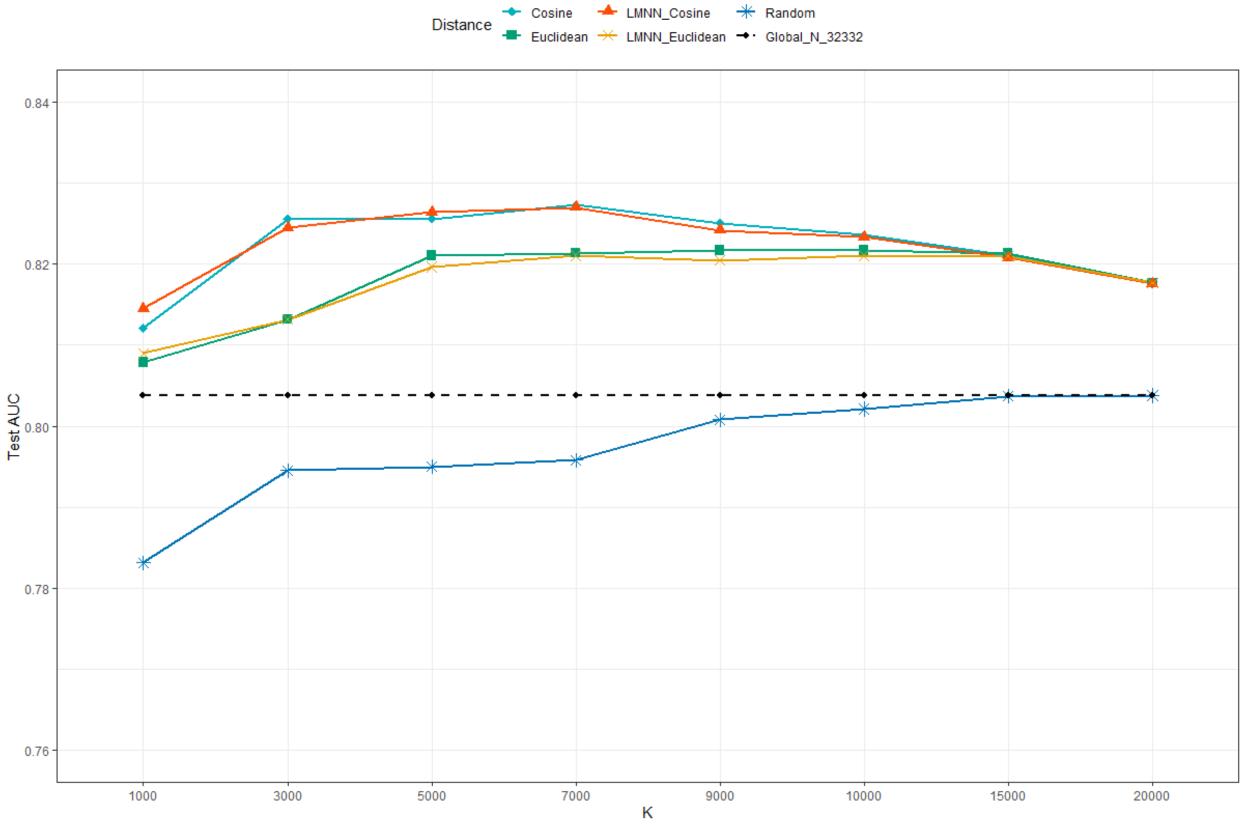

Figure 4.9. Performance of the personalized logistic regression model in terms of AUC as a function of K – diabetic retinopathy prognosis

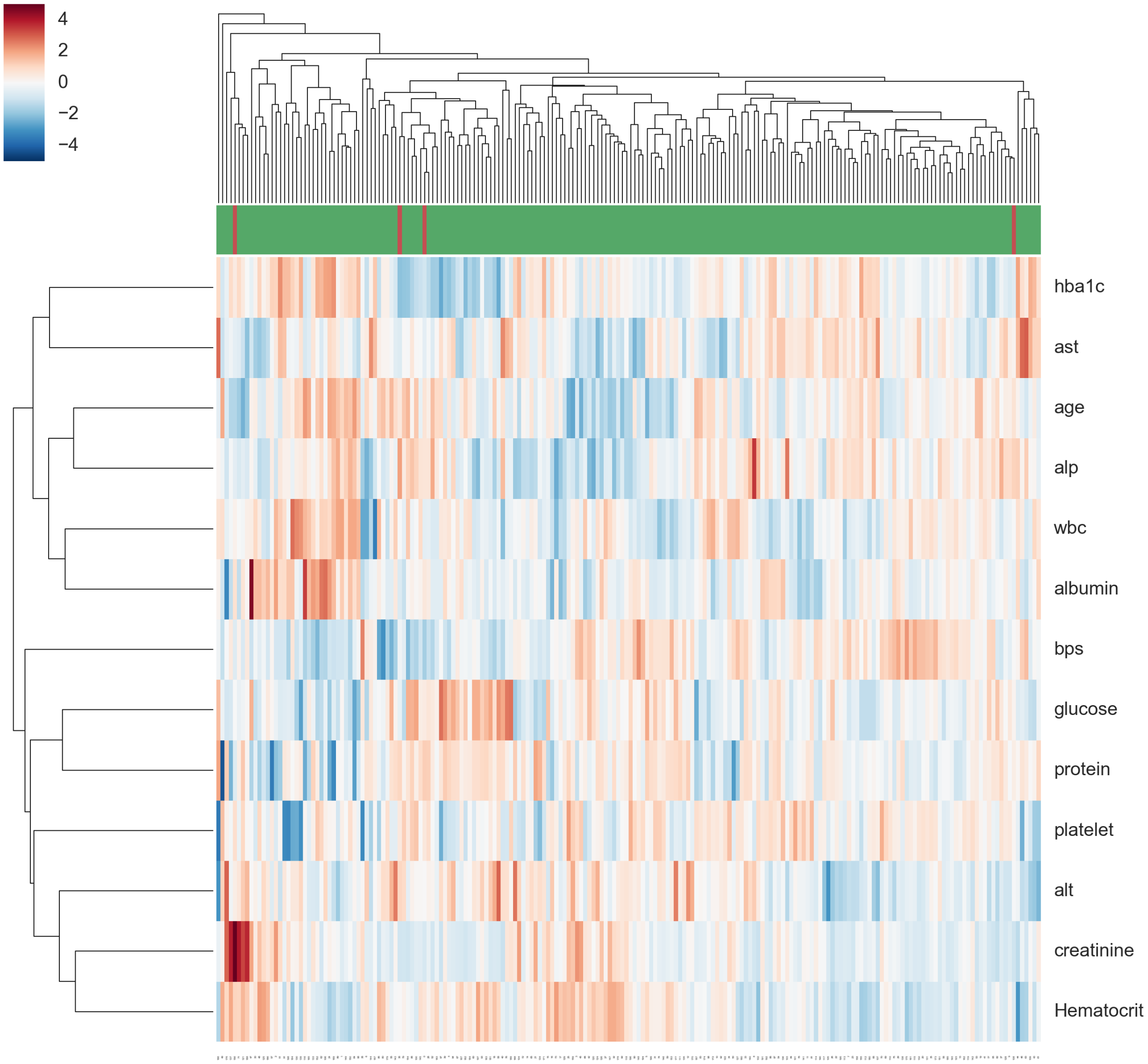

Figure 4.10. Hierarchical heat map plot of the risk factors from the personalized predictive models for 200 randomly selected test patients - Diabetic retinopathy prognosis. The rows are the risk factors and they are clustered along the vertical axis. Each column stands for a patient and the red-green horizontal bar on the top stands for their diagnosis of diabetic retinopathy(DR).(Green: No DR; Red: Have DR.)

CHAPTER V

CONCLUSION AND DISCUSSION

In this thesis, we try to improve the adverse outcome prediction from two aspects, the feature selection and model development. We propose an ensemble feature selection method and explore the use of survival analysis methods and personalized modeling on adverse outcome prediction. While we exemplify the application of our methods with specific research questions, they can be applied to various disease targets and outcome types.

5.1 Feature selection

We have developed a stable feature selection framework that can learn from the data and select features automatically. The criteria of defining a good feature selection method should include both accuracy and stability. Accuracy measures the predictive power and reliability of the final feature set in clinical applications and stability ensures the insensitivity to over-fitting of the result. To improve prediction performance, we include the embedded and wrapper methods in the selection process and use accuracy as the updating criterion in the search algorithm. We strengthen the stability of the selection method by combining different feature selection techniques[126] and using bootstrapping to aggregate the selected feature subsets[41] to reduce the variance when selecting features on different data subsets. GBM seems to outperform LASSO in selecting and ranking the features with more accuracy-related importance values. Feature ranking aggregation is crucial to improving accuracy and stability of the selection result and the result shows that the weighted aggregation method performs better than the simple aggregation, with fewer features and higher accuracy.

For hip fracture readmission, the best feature set has 14 features, most of which are lab tests and diagnoses. Some features are the known predictors of hip fracture readmission, like age, gender and length of stay. For diabetic retinopathy prognosis, we have identified 12 vital signs and age as the most important risk factors. We need to note that the prediction performance of hip fracture readmission is pretty modest compared with that of the diabetic retinopathy. Since the data types of features in these two datasets are different, we are interested in studying the effect of data types on predictive power. As the stability of feature selection technique is gaining more importance, a consensus stability measure also needs to be developed for fair comparison of feature selection methods.

5.2 Time-to-event Prediction

In this study, we investigate the performance of several survival analysis methods on the 30-day hip fracture readmission prediction. It's shown that the family of AFT models all perform well and slightly outperforms the other methods. The mixture cure model does not seem to meet the expectations, although it has the advantage of not assuming the survival function will go to zero when time goes to infinity. Compared with binary classifiers, the time-to-event prediction achieves similar performance but provides more flexibility on the dynamic analysis of patients' status. With only one survival analysis model, we are able to predict the patients' readmission status over different time frames.

The performance of the mixture cure model is sensitive to the cure fraction and previous simulation study showed that the convergence of cure model gets more difficult as the cure fraction decreases to zero.[200] This is a possible reason for the modest performance of mixture model in comparison with AFT and Cox regression models. Further investigations on other patient data sets with higher cure fraction are necessary for uncovering the reason behind this.

In the medical context, people are concerned about more than one type of adverse outcomes. For example, a patient may have the risk of developing new diseases, readmission and mortality at the same time. In the future, we can extend the one-type adverse outcome prediction to multiple events modeling to have a more comprehensive evaluation of patients' risk.[201]

5.3 Personalized Modeling

In this part, we use empirical study to show the potential of personalizing patients' status prediction by identifying and analyzing their past similar patients. By exploiting similarity analytics, personalized predictive model trained on a smaller set of clinically similar patients can perform better than the global model fit on all available data. Personalized models can also identify the important risk factors on an individual basis and provide customized decision support for doctors. Our experiment also characterizes the trade-off between the training set size and the degree of similarity between the training cohort and the test patient for whom we are making the prediction, as concluded in [40]. And the performance of personalized modeling also depends on the data types and the actual predictive power of the features.

There are some directions for future work. Firstly, the approach need to be validated on additional research questions and more datasets. Secondly, the accurate identification of clinically similar patients is a crucial step to the prediction performance thus selecting an appropriate similarity measure is of great importance. We can investigate the relationship between data types and similarity measures to choose the most suitable one based on the data types, disease target and so on. We can also replace the logistic regression with more sophisticated classification algorithms as the prediction model. Lastly, since both of the adverse outcome problems discussed in this study have binary response variable, we may extend the personalized modeling to continuous response variables to test its effectiveness.

REFERENCES

[1] Lian Leng Low, Kheng Hock Lee, Hock Ong, Marcus Eng, Sijia Wang, Shu Yun Tan, Julian Thumboo, and Nan Liu. Predicting 30-day readmissions: performance of the lace index compared with a regression model among general medicine patients in singapore. *BioMed research international*, 2015, 2015.

[2] Gregory M Garrison, Paul M Robelia, Jennifer L Pecina, and Nancy L Dawson. Comparing performance of 30-day readmission risk classifiers among hospitalized primary care patients. *Journal of evaluation in clinical practice*, 23(3):524–529, 2017.

[3] LeeAnna Spiva, Marti Hand, Lewis VanBrackle, and Frank McVay. Validation of a predictive model to identify patients at high risk for hospital readmission. *Journal for Healthcare Quality*, 2014.

[4] Dustin D French, Elizabeth Bass, Douglas D Bradham, Robert R Campbell, Rubenstein, and Laurence Z. Rehospitalization after hip fracture: predictors and prognosis from a national veterans study. *Journal of the American Geriatrics Society*, 56(4):705–710, 2008.

[5] Ray Marks. Hip fracture epidemiological trends, outcomes, and risk factors, 1970–2009. *International journal of general medicine*, 3:1, 2010.

[6] Karen E Joynt and Ashish K Jha. Characteristics of hospitals receiving penalties under the hospital readmissions reduction program. *Jama*, 309(4):342–343, 2013.

[7] Jason S Ng, Marcus A Bearse, Marilyn E Schneck, Shirin Barez, and Anthony J Adams. Local diabetic retinopathy prediction by multifocal erg delays over 3 years. *Investigative ophthalmology & visual science*, 49(4):1622–1628, 2008.

[8] Ralph B D'Agostino, Scott Grundy, Lisa M Sullivan, and Peter Wilson. Validation of the framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *Jama*, 286(2):180–187, 2001.

[9] Jyoti Soni, Ujma Ansari, Dipesh Sharma, and Sunita Soni. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, 17(8):43–48, 2011.

[10] Devan Kansagara, Honora Englander, Amanda Salanitro, David Kagen, Cecelia Theobald, Michele Freeman, and Sunil Kripalani. Risk prediction models for hospital readmission: a systematic review. *Jama*, 306(15):1688–1698, 2011.

[11] Jacques Donzé, Drahomir Aujesky, Deborah Williams, and Jeffrey L Schnipper. Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. *JAMA internal medicine*, 173(8):632–638, 2013.

[12] Shipeng Yu, Faisal Farooq, Alexander Van Esbroeck, Glenn Fung, Vikram Anand, and Balaji Krishnapuram. Predicting readmission risk with institution-specific prediction models. *Artificial intelligence in medicine*, 65(2):89–96, 2015.

[13] Manoranjan Dash and Huan Liu. Feature selection for classification. *Intelligent data analysis*, 1(3):131–156, 1997.

[14] Thomas Abeel, Thibault Helleputte, Yves Van de Peer, Pierre Dupont, and Yvan Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–398, 2010.

[15] Anne-Claire Haury, Pierre Gestraud, and Jean-Philippe Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS one*, 6(12), 2011.

[16] Max Kuhn and Kjell Johnson. *Applied predictive modeling*, volume 26. Springer, 2013.

[17] Enric Junqué de Fortuny, David Martens, and Foster Provost. Predictive modeling with big data: is bigger really better? *Big Data*, 1(4):215–226, 2013.

[18] Akbar K Waljee, Peter DR Higgins, and Amit G Singal. A primer on predictive models. *Clinical and translational gastroenterology*, 5(1):e44, 2014.

[19] L Nelson Sanchez-Pinto, Laura Ruth Venable, John Fahrenbach, and Matthew M Churpek. Comparison of variable selection methods for clinical predictive modeling. *International journal of medical informatics*, 116:10–17, 2018.

[20] Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1):198–208, 2017.

[21] Jay R Desai, Pingsheng Wu, Greg A Nichols, Tracy A Lieu, and Patrick J O'Connor. Diabetes and asthma case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness and epidemiologic research. *Medical care*, 50:S30, 2012.

[22] Chunhua Weng, Paul Appelbaum, George Hripcsak, Ian Kronish, Linda Busacca, Karina W Davidson, and J Thomas Bigger. Using ehrs to integrate research with patient care: promises and challenges. *Journal of the American Medical Informatics Association*, 19(5):684–687, 2012.

[23] Nicole Gray Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151, 2013.

[24] Shamsul Huda, John Yearwood, Herbert F Jelinek, Mohammad Mehedi Hassan, Giancarlo Fortino, and Michael Buckland. A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis. *IEEE access*, 4:9145–9154, 2016.

[25] Carla S Alvarado, Kathleen Zook, and J Henry. Electronic health record adoption and interoperability among us skilled nursing facilities in 2016. *ONC Data Brief*, (39), 2017.

[26] Noura AlNuaimi, Mohammad M Masud, and Farhan Mohammed. Examining the effect of feature selection on improving patient deterioration prediction. *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol*, 5, 2015.

[27] K Strand and H Flaatten. Severity scoring in the icu: a review. *Acta Anaesthesiologica Scandinavica*, 52(4):467–478, 2008.

[28] Carl van Walraven, Irfan A Dhalla, Chaim Bell, Edward Etchells, Ian G Stiell, Kelly Zarnke, Peter C Austin, and Alan J Forster. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Cmaj*, 182(6):551–557, 2010.

[29] Ying-Jui Chang, Min-Li Yeh, Yu-Chuan Li, Chien-Yeh Hsu, Chao-Cheng Lin, Meng-Shiuan Hsu, and Wen-Ta Chiu. Predicting hospital-acquired infections by scoring system with simple parameters. *PloS one*, 6(8), 2011.

[30] Huan Liu and Hiroshi Motoda. *Computational methods of feature selection*. CRC Press, 2007.

[31] Irene L Vegting, Marlou van Beneden, Mark HH Kramer, Abel Thijs, Piet J Kostense, and Prabath WB Nanayakkara. How to save costs by reducing unnecessary testing: lean thinking in clinical practice. *European journal of internal medicine*, 23(1):70–75, 2012.

[32] Federico Cismondi, Leo A Celi, André S Fialho, Susana M Vieira, Shane R Reti, Joao MC Sousa, and Stan N Finkelstein. Reducing unnecessary lab testing in the icu with artificial intelligence. *International journal of medical informatics*, 82(5):345–358, 2013.

[33] Joon Lee and David M Maslove. Using information theory to identify redundancy in common laboratory tests in the intensive care unit. *BMC medical informatics and decision making*, 15(1):59, 2015.

[34] Timothy Schmutte, Christine L Dunn, and William H Sledge. Predicting time to readmission in patients with recent histories of recurrent psychiatric hospitalization: a matched-control survival analysis. *The Journal of nervous and mental disease*, 198(12):860–863, 2010.

[35] Kenney Ng, Jimeng Sun, Jianying Hu, and Fei Wang. Personalized predictive modeling and risk factor identification using patient similarity. *AMIA Summits on Translational Science Proceedings*, 2015:132, 2015.

[36] Scott Powers, Trevor Hastie, and Robert Tibshirani. Customized training with an application to mass spectrometric imaging of cancer tissue. *The annals of applied statistics*, 9(4):1709, 2015.

[37] Lore Dirick, Gerda Claeskens, and Bart Baesens. Time to default in credit scoring using survival analysis: a benchmark study. *Journal of the Operational Research Society*, 68(6):652–665, 2017.

[38] David Collett. *Modelling survival data in medical research*. CRC press, 2015.

[39] Nitesh V Chawla and Darcy A Davis. Bringing big data to personalized healthcare: a patient-centered framework. *Journal of general internal medicine*, 28(3):660–665, 2013.

[40] Joon Lee, David M Maslove, and Joel A Dubin. Personalized mortality prediction driven by electronic medical data and a patient similarity metric. *PloS one*, 10(5), 2015.

[41] Xing Song, Lemuel R Waitman, Yong Hu, Alan SL Yu, David Robins, and Mei Liu. Robust clinical marker identification for diabetic kidney disease with ensemble feature selection. *Journal of the American Medical Informatics Association*, 26(3):242–253, 2019.

[42] S Geisser. An introduction to predictive inference, 1993.

[43] CV Apte, Se June Hong, Ramesh Natarajan, Edwin PD Pednault, FA Tipu, and Sholom M Weiss. Data-intensive analytics for predictive modeling. *IBM Journal of Research and Development*, 47(1):17–23, 2003.

[44] Christopher Westphal. *Data Mining for Intelligence, Fraud & Criminal Detection: Advanced Analytics & Information Sharing Technologies*. CRC Press, 2008.

[45] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. Recommender systems survey. *Knowledge-based systems*, 46:109–132, 2013.

[46] Se June Hong and Sholom M Weiss. Advances in predictive models for data mining. *Pattern Recognition Letters*, 22(1):55–61, 2001.

[47] David J Sheskin. *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2003.

[48] Anju Verma, Maurizio Fiasché, Maria Cuzzola, Pasquale Iacopino, Francesco C Morabito, and Nikola Kasabov. Ontology based personalized modeling for type 2 diabetes risk analysis: an integrated approach. In *International Conference on Neural Information Processing*, pages 360–366. Springer, 2009.

[49] El-Sayed M El-Alfy. Discovering classification rules for email spam filtering with an ant colony optimization algorithm. In *2009 IEEE Congress on Evolutionary Computation*, pages 1778–1783. IEEE, 2009.

[50] Joaquín Abellán and Carlos J Mantas. Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, 41(8):3825–3830, 2014.

[51] Chih-Fong Tsai and Jhen-Wei Wu. Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert systems with applications*, 34(4):2639–2649, 2008.

[52] Leo Guelman. Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, 39(3):3659–3667, 2012.

[53] SS Thakur and JK Sing. Mining customer's data for vehicle insurance prediction system using k-means clustering-an application. *International journal of computer Applications in Engineering sciences*, 3(4):148, 2013.

[54] Linda Miner, Pat Bolding, Joseph Hilbe, Mitchell Goldstein, Thomas Hill, Robert Nisbet, Nephi Walton, and Gary Miner. *Practical predictive analytics and decisioning systems for medicine: Informatics accuracy and cost-effectiveness for healthcare administration and delivery including medical research*. Academic Press, 2014.

[55] Ivo D Dinov, Ben Heavner, Ming Tang, Gustavo Glusman, Kyle Chard, Mike Darcy, Ravi Madduri, Judy Pa, Cathie Spino, Carl Kesselman, et al. Predictive big data analytics: a study of parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. *PloS one*, 11(8):e0157077, 2016.

[56] Eric WT Ngai, Yong Hu, Yiu Hing Wong, Yijun Chen, and Xin Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems*, 50(3):559–569, 2011.

[57] Anuj Sharma and Prabin Kumar Panigrahi. A review of financial accounting fraud detection based on data mining techniques. *arXiv preprint arXiv:1309.3944*, 2013.

[58] Devendra Kumar Tayal, Arti Jain, Surbhi Arora, Surbhi Agarwal, Tushar Gupta, and Nikhil Tyagi. Crime detection and criminal identification in india using data mining techniques. *AI & society*, 30(1):117–127, 2015.

[59] Kenney Ng, Amol Ghoting, Steven R Steinhubl, Walter F Stewart, Bradley Malin, and Jimeng Sun. Paramo: A parallel predictive modeling platform for healthcare analytic research using electronic health records. *Journal of biomedical informatics*, 48:160–170, 2014.

[60] J Kievit, M Krukerink, and PJ Marang-van de Mheen. Surgical adverse outcome reporting as part of routine clinical care. *Qual Saf Health Care*, 19(6):e20–e20, 2010.

[61] Susan Mallett, Patrick Royston, Rachel Waters, Susan Dutton, and Douglas G Altman. Reporting performance of prognostic models in cancer: a review. *BMC medicine*, 8(1):21, 2010.

[62] Gary S Collins, Susan Mallett, Omar Omar, and Ly-Mee Yu. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC medicine*, 9(1):103, 2011.

[63] Sharath Cholleti, Andrew Post, Jingjing Gao, Xia Lin, William Bornstein, Dedra Cantrell, and Joel Saltz. Leveraging derived data elements in data analytic models for understanding and predicting hospital readmissions. In *AMIA Annual Symposium Proceedings*, volume 2012, page 103. American Medical Informatics Association, 2012.

[64] Adria E Navarro, Susan Enguídanos, and Kathleen H Wilber. Identifying risk of hospital readmission among medicare aged patients: an approach using routinely collected data. *Home health care services quarterly*, 31(2):181–195, 2012.

[65] Charles A Baillie, Christine VanZandbergen, Gordon Tait, Asaf Hanish, Brian Leas, Benjamin French, C William Hanson, Maryam Behta, and Craig A Umscheid. The readmission risk flag: Using the electronic health record to automatically identify patients at risk for 30-day readmission. *Journal of hospital medicine*, 8(12):689–695, 2013.

[66] Ying P Tabak, Xiaowu Sun, Linda Hyde, Ayla Yaitanes, Karen Derby, and Richard S Johannes. Using enriched observational data to develop and validate age-specific mortality risk adjustment models for hospitalized pediatric patients. *Medical care*, pages 437–445, 2013.

[67] Rajeev Ayyagari, Francis Vekeman, Patrick Lefebvre, Siew Hwa Ong, Elizabeth Faust, Alex Trahey, Gerardo Machnicki, and Mei Sheng Duh. Pulse pressure and stroke risk: development and validation of a new stroke risk model. *Current medical research and opinion*, 30(12):2453–2460, 2014.

[68] Santu Rana, Truyen Tran, Wei Luo, Dinh Phung, Richard L Kennedy, and Svetha Venkatesh. Predicting unplanned readmission after myocardial infarction from routinely collected administrative hospital data. *Australian Health Review*, 38(4):377–382, 2014.

[69] Alan J Forster, Heather D Clark, Alex Menard, Natalie Dupuis, Robert Chernish, Natasha Chandok, Asmat Khan, and Carl van Walraven. Adverse events among medical patients after discharge from hospital. *Cmaj*, 170(3):345–349, 2004.

[70] Jin Huang and Charles X Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310, 2005.

[71] John Concato, Alvan R Feinstein, and Theodore R Holford. The risk of determining risk with multivariable models. *Annals of internal medicine*, 118(3):201–210, 1993.

[72] Steven C Bagley, Halbert White, and Beatrice A Golomb. Logistic regression in the medical literature:: Standards for use and reporting, with particular attention to one medical domain. *Journal of clinical epidemiology*, 54(10):979–985, 2001.

[73] Janus Christian Jakobsen, Christian Gluud, Jørn Wetterslev, and Per Winkel. When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. *BMC medical research methodology*, 17(1):162, 2017.

[74] Yvonne Vergouwe, Patrick Royston, Karel GM Moons, and Douglas G Altman. Development and validation of a prediction model with missing predictor data: a practical approach. *Journal of clinical epidemiology*, 63(2):205–214, 2010.

[75] Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

[76] Jing Zhao, Aron Henriksson, Lars Asker, and Henrik Boström. Predictive modeling of structured electronic health records for adverse drug event detection. *BMC medical informatics and decision making*, 15(S4):S1, 2015.

[77] Hans-Ulrich Prokosch and Thomas Ganslandt. Perspectives for medical informatics. *Methods of information in medicine*, 48(01):38–44, 2009.

[78] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[79] Kenney Ng, Steven R Steinhubl, Christopher deFilippi, Sanjoy Dey, and Walter F Stewart. Early detection of heart failure using electronic health records: practical implications for time before diagnosis, data diversity, data quantity, and data density. *Circulation: Cardiovascular Quality and Outcomes*, 9(6):649–658, 2016.

[80] Kristiina Häyrinen, Kaija Saranto, and Pirkko Nykänen. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics*, 77(5):291–304, 2008.

[81] Pascal Coorevits, Mats Sundgren, Gunnar O Klein, Anne Bahr, B Claerhout, C Daniel, Martin Dugas, D Dupont, A Schmidt, P Singleton, et al. Electronic health records: new opportunities for clinical research. *Journal of internal medicine*, 274(6):547–560, 2013.

[82] Sebastian Pölsterl, Sailesh Conjeti, Nassir Navab, and Amin Katouzian. Survival analysis for high-dimensional, heterogeneous medical data: Exploring feature extraction as an alternative to feature selection. *Artificial intelligence in medicine*, 72:1–11, 2016.

[83] David Gans, John Kralewski, Terry Hammons, and Bryan Dowd. Medical groups' adoption of electronic health records and information systems. *Health affairs*, 24(5):1323–1333, 2005.

[84] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[85] Tian Xia, Dacheng Tao, Tao Mei, and Yongdong Zhang. Multiview spectral embedding. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(6):1438–1446, 2010.

[86] Wojciech Siedlecki and Jack Sklansky. On automatic feature selection. In *Handbook of Pattern Recognition and Computer Vision*, pages 63–87. World Scientific, 1993.

[87] Richard Butterworth, Gregory Piatetsky-Shapiro, and Dan A Simovici. On feature selection through clustering. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 4–pp. IEEE, 2005.

[88] Kenji Kira and Larry A Rendell. A practical approach to feature selection. In *Machine Learning Proceedings 1992*, pages 249–256. Elsevier, 1992.

[89] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[90] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[91] Samuel H Huang. Supervised feature selection: A tutorial. *Artif. Intell. Research*, 4(2):22–37, 2015.

[92] Lei Yu and Huan Liu. Redundancy based feature selection for microarray data. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 737–742, 2004.

[93] Iman Kamkar, Sunil Kumar Gupta, Dinh Phung, and Svetha Venkatesh. Stable feature selection for clinical prediction: Exploiting icd tree structure using tree-lasso. *Journal of biomedical informatics*, 53:277–290, 2015.

[94] Mark Bower, Brian Gazzard, Sundhiya Mandalia, Tom Newsom-Davis, Christina Thirlwell, Tony Dhillon, Anne Marie Young, Tom Powles, Andrew Gaya, Mark Nelson, et al. A prognostic index for systemic aids-related non-hodgkin lymphoma treated in the era of highly active antiretroviral therapy. *Annals of internal medicine*, 143(4):265–273, 2005.

[95] Hyung L Kim, David Seligson, Xueli Liu, Nicolette Janzen, Matthew HT Bui, Hong Yu, Tao Shi, Arie S Belldegrun, Steve Horvath, and Robert A Figlin. Using tumor markers to predict the survival of patients with metastatic renal cell carcinoma. *The Journal of urology*, 173(5):1496–1501, 2005.

[96] Nynke Halbesma, Desiree F Jansen, Martijn W Heymans, Ronald P Stolk, Paul E de Jong, Ronald T Gansevoort, PREVEND Study Group, et al. Development and validation of a general population renal risk score. *Clinical journal of the American Society of Nephrology*, 6(7):1731–1738, 2011.

[97] Allan B Massie, Joseph Leanza, Lara M Fahmy, Eric KH Chow, Niraj M Desai, Xun Luo, Elizabeth A King, Mary G Bowring, and Dorry L Segev. A risk index for living donor kidney transplantation. *American Journal of Transplantation*, 16(7):2077–2084, 2016.

[98] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

[99] Tsang-Hsiang Cheng, Chih-Ping Wei, and Vincent S Tseng. Feature selection for medical data mining: Comparisons of expert judgment and automatic approaches. In *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, pages 165–170. IEEE, 2006.

[100] Carlos A Alvarez, Christopher A Clark, Song Zhang, Ethan A Halm, John J Shannon, Carlos E Girod, Lauren Cooper, and Ruben Amarasingham. Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data. *BMC medical informatics and decision making*, 13(1):28, 2013.

[101] Ruben Amarasingham, Billy J Moore, Ying P Tabak, Mark H Drazner, Christopher A Clark, Song Zhang, W Gary Reed, Timothy S Swanson, Ying Ma, and Ethan A Halm. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Medical care*, pages 981–988, 2010.

[102] Amir Dembo, Thomas M Cover, and Joy A Thomas. Information theoretic inequalities. *IEEE Transactions on Information theory*, 37(6):1501–1518, 1991.

[103] Huan Liu and Rudy Setiono. Chi2: Feature selection and discretization of numeric attributes. In *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*, pages 388–391. IEEE, 1995.

[104] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *Advances in neural information processing systems*, pages 507–514, 2006.

[105] Edward F Philbin and Thomas G DiSalvo. Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data. *Journal of the American College of Cardiology*, 33(6):1560–1566, 1999.

[106] Hao Wang, Richard D Robinson, Carlos Johnson, Nestor R Zenarosa, Rani D Jayswal, Joshua Keithley, and Kathleen A Delaney. Using the lace index to predict hospital readmissions in congestive heart failure patients. *BMC cardiovascular disorders*, 14(1):97, 2014.

[107] Christie M Atchison, Shilpa Arlikar, Ernest Amankwah, Irmel Ayala, Laurie Barrett, Brian R Branchford, Michael Streiff, Clifford Takemoto, and Neil A Goldenberg. Development of a new risk score for hospital-associated venous thromboembolism in noncritically ill children: findings from a large single-institutional case-control study. *The Journal of pediatrics*, 165(4):793–798, 2014.

[108] Michael E Egger, Malcolm H Squires III, David A Kooby, Shishir K Maithel, Clifford S Cho, Sharon M Weber, Emily R Winslow, Robert CG Martin II, Kelly M McMasters, and Charles R Scoggins. Risk stratification for readmission after major hepatectomy: development of a readmission risk score. *Journal of the American College of Surgeons*, 220(4):640–648, 2015.

[109] Alisa B Busch, Brian Neelon, Katya Zelevinsky, Yulei He, and Sharon-Lise T Normand. Accurately predicting bipolar disorder mood outcomes—implications for the use of electronic databases. *Medical care*, 50(4):311, 2012.

[110] Peiyao Cheng, Britta Neugaard, Philip Foulis, and Paul R Conlin. Hemoglobin a1c as a predictor of incident diabetes. *Diabetes care*, 34(3):610–615, 2011.

[111] Iffat A Gheyas and Leslie S Smith. Feature subset selection in large dimensionality domains. *Pattern recognition*, 43(1):5–13, 2010.

[112] Christian C Apfel, Esa Läärä, Merja Koivuranta, Clemens-A Greim, and Norbert Roewer. A simplified risk score for predicting postoperative nausea and vomiting conclusions from cross-validations between two centers. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 91(3):693–693, 1999.

[113] Sarah J Crane, Ericka E Tung, Gregory J Hanson, Stephen Cha, Rajeev Chaudhry, and Paul Y Takahashi. Use of an electronic administrative database to identify older community dwelling adults at high-risk for hospitalization or emergency department visits: the elders risk assessment index. *BMC health services research*, 10(1):338, 2010.

[114] Shahid A Choudhry, Jing Li, Darcy Davis, Cole Erdmann, Rishi Sikka, and Bharat Sutariya. A public-private partnership develops and externally validates a 30-day hospital readmission risk prediction model. *Online journal of public health informatics*, 5(2):219, 2013.

[115] Taghi M Khoshgoftaar, Alireza Fazelpour, Huanjing Wang, and Randall Wald. A survey of stability analysis of feature subset selection techniques. In *2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)*, pages 424–431. IEEE, 2013.

[116] Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *Icml*, volume 97, page 35, 1997.

[117] Mark Andrew Hall. Correlation-based feature selection for machine learning. 1999.

[118] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[119] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[120] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[121] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[122] Pablo Bermejo, Jose A Gámez, and Jose M Puerta. A grasp algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets. *Pattern Recognition Letters*, 32(5):701–711, 2011.

[123] Waad Bouaguel and Mohamed Limam. A new way for combining filter feature selection methods. In *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*, pages 411–419. Springer, 2016.

[124] Eugene Tuv, Alexander Borisov, George Runger, and Kari Torkkola. Feature selection with ensembles, artificial variables, and redundancy elimination. *Journal of Machine Learning Research*, 10(Jul):1341–1366, 2009.

[125] Verónica Bolón-Canedo, Noelia Sánchez-Marono, Amparo Alonso-Betanzos, José Manuel Benítez, and Francisco Herrera. A review of microarray datasets and applied feature selection methods. *Information Sciences*, 282:111–135, 2014.

[126] Yvan Saeys, Thomas Abeel, and Yves Van de Peer. Robust feature selection using ensemble feature selection techniques. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 313–325. Springer, 2008.

[127] Wael Awada, Taghi M Khoshgoftaar, David Dittman, Randall Wald, and Amri Napolitano. A review of the stability of feature selection techniques for bioinformatics data. In *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*, pages 356–363. IEEE, 2012.

[128] Donghai Guan, Weiwei Yuan, Young-Koo Lee, Kamran Najeebullah, and Mostofa Kamal Rasel. A review of ensemble learning based feature selection. *IETE Technical Review*, 31(3):190–198, 2014.

[129] Barbara Pes, Nicoletta Dessì, and Marta Angioni. Exploiting the ensemble paradigm for stable feature selection: a case study on high-dimensional genomic data. *Information Fusion*, 35:132–147, 2017.

[130] Kathryn R Fingar, Marguerite L Barrett, and H Joanna Jiang. A comparison of all-cause 7-day and 30-day readmissions, 2014: Statistical brief# 230. *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Rockville, MD: Agency for Healthcare Research and Quality*, 2006.

[131] Jeph Herrin, Justin St. Andre, Kevin Kenward, Maulik S Joshi, Anne-Marie J Audet, and Stephen C Hines. Community factors and hospital readmission rates. *Health services research*, 50(1):20–39, 2015.

[132] Chenyang Zhong and Robert Tibshirani. Survival analysis as a classification problem. *arXiv preprint arXiv:1909.11171*, 2019.

[133] Mert R Sabuncu. A bayesian algorithm for image-based time-to-event prediction. In *International Workshop on Machine Learning in Medical Imaging*, pages 74–81. Springer, 2013.

[134] Patrick S Kamath, Russell H Wiesner, Michael Malinchoc, Walter Kremers, Terry M Therneau, Catherine L Kosberg, Gennaro D'Amico, E Rolland Dickson, and W Ray Kim. A model to predict survival in patients with end-stage liver disease. *Hepatology*, 33(2):464–470, 2001.

[135] Thomas J Wang, Philimon Gona, Martin G Larson, Geoffrey H Tofler, Daniel Levy, Christopher Newton-Cheh, Paul F Jacques, Nader Rifai, Jacob Selhub, Sander J Robins, et al. Multiple biomarkers for the prediction of first major cardiovascular events and death. *New England Journal of Medicine*, 355(25):2631–2639, 2006.

[136] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[137] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

[138] Tianxi Cai, Lu Tian, and LJ Wei. Semiparametric box–cox power transformation models for censored survival observations. *Biometrika*, 92(3):619–632, 2005.

[139] Chao Cai, Yubo Zou, Yingwei Peng, and Jiajia Zhang. smcure: An r-package for estimating semiparametric mixture cure models. *Computer methods and programs in biomedicine*, 108(3):1255–1260, 2012.

[140] Anis Sharafoddini, Joel A Dubin, and Joon Lee. Patient similarity in prediction models based on health data: a scoping review. *JMIR medical informatics*, 5(1):e7, 2017.

[141] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[142] Claire Cardie. Using decision trees to improve case-based learning. In *Proceedings of the tenth international conference on machine learning*, pages 25–32, 1993.

[143] Femke Kirschner, Fred Paas, and Paul A Kirschner. Individual and group-based learning from complex cognitive tasks: Effects on retention and transfer efficiency. *Computers in Human Behavior*, 25(2):306–314, 2009.

[144] Nitin Bhatia et al. Survey of nearest neighbor techniques. *arXiv preprint arXiv:1007.0085*, 2010.

[145] Upendra Shardanand and Pattie Maes. Social information filtering: algorithms for automating "word of mouth". In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 210–217, 1995.

[146] Will Hill, Larry Stead, Mark Rosenstein, and George Furnas. Recommending and evaluating choices in a virtual community of use. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 194–201, 1995.

[147] Kelly Peterson, Ognjen Rudovic, Ricardo Guerrero, and Rosalind W Picard. Personalized gaussian processes for future prediction of alzheimer's disease progression. *arXiv preprint arXiv:1712.00181*, 2017.

[148] Leigh Ann Simmons, Michaela Ann Dinan, Timothy John Robinson, and Ralph Snyderman. Personalized medicine is more than genomic medicine: confusion over terminology impedes progress towards personalized healthcare. *Personalized medicine*, 9(1):85–91, 2012.

[149] Ralph Snyderman. Personalized health care: from theory to practice. *Biotechnology journal*, 7(8):973–979, 2012.

[150] Jennifer Pittman, Erich Huang, Holly Dressman, Cheng-Fang Horng, Skye H Cheng, Mei-Hua Tsou, Chii-Ming Chen, Andrea Bild, Edwin S Iversen, Andrew T Huang, et al. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences*, 101(22):8431–8436, 2004.

[151] Ping Zhang, Fei Wang, Jianying Hu, and Robert Sorrentino. Towards personalized medicine: leveraging patient similarity and drug similarity analytics. *AMIA Summits on Translational Science Proceedings*, 2014:132, 2014.

[152] Nikola Kasabov. Global, local and personalised modeling and pattern discovery in bioinformatics: An integrated approach. *Pattern Recognition Letters*, 28(6):673–685, 2007.

[153] David W Aha, Dennis Kibler, and Marc K Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991.

[154] Kimberly L Elmore and Michael B Richman. Euclidean distance as a similarity metric for principal component analysis. *Monthly weather review*, 129(3):540–549, 2001.

[155] Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4), 2014.

[156] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):881–892, 2002.

[157] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.

[158] Thrasyvoulos N Pappas and Nikil S Jayant. An adaptive clustering algorithm for image segmentation. In *International Conference on Acoustics, Speech, and Signal Processing,*, pages 1667–1670. IEEE, 1989.

[159] Graham Hepworth and Ian Gordon. Assessing similarity of dna profiles. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(1):125–133, 2011.

[160] Sony Hartono Wijaya, Farit Mochamad Afendi, Irmanida Batubara, Latifah K Darusman, Md Altaf-Ul-Amin, and Shigehiko Kanaya. Finding an appropriate equation to measure similarity between binary vectors: case studies on indonesian and japanese herbal medicines. *BMC bioinformatics*, 17(1):520, 2016.

[161] Raymond Austin Jarvis and Edward A Patrick. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on computers*, 100(11):1025–1034, 1973.

[162] Ka Yee Yeung, David R. Haynor, and Walter L. Ruzzo. Validating clustering for gene expression data. *Bioinformatics*, 17(4):309–318, 2001.

[163] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee. A similarity measure for text classification and clustering. *IEEE transactions on knowledge and data engineering*, 26(7):1575–1590, 2013.

[164] Shyam Visweswaran, Derek C Angus, Margaret Hsieh, Lisa Weissfeld, Donald Yealy, and Gregory F Cooper. Learning patient-specific predictive models from clinical data. *Journal of biomedical informatics*, 43(5):669–685, 2010.

[165] Oya Ekin, Peter L Hammer, Alexander Kogan, and Pawel Winter. Distance-based classification methods. *INFOR: Information Systems and Operational Research*, 37(3):337–352, 1999.

[166] Ayça Çakmak Pehlivanlı. A novel feature selection scheme for high-dimensional data sets: four-staged feature selection. *Journal of Applied Statistics*, 43(6):1140–1154, 2016.

[167] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[168] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[169] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

[170] Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.

[171] Randall Wald, Taghi M Khoshgoftaar, David Dittman, Wael Awada, and Amri Napolitano. An extensive comparison of feature ranking aggregation techniques in bioinformatics. In *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*, pages 377–384. IEEE, 2012.

[172] Jonathan Buckley and Ian James. Linear regression with censored data. *Biometrika*, 66(3):429–436, 1979.

[173] Norman Lloyd Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. Continuous univariate distributions. 1994.

[174] Wing Hung Wong et al. Theory of partial likelihood. *The Annals of statistics*, 14(1):88–123, 1986.

[175] John P Klein and Melvin L Moeschberger. Semiparametric proportional hazards regression with fixed covariates. *Survival analysis: techniques for censored and truncated data*, pages 243–293, 2003.

[176] Yu Yakovlev Andrei, B Asselain, et al. *Stochastic models of tumor latency and their biostatistical applications*, volume 1. World Scientific, 1996.

[177] Lu Wang, Pang Du, and Hua Liang. Two-component mixture cure rate model with spline estimated nonparametric components. *Biometrics*, 68(3):726–735, 2012.

[178] Terry M Therneau and Thomas Lumley. Package 'survival'. *Survival analysis Published on CRAN*, 2:3, 2014.

[179] Shahina Begum, Mobyen Uddin Ahmed, Peter Funk, Ning Xiong, and Mia Folke. Case-based reasoning systems in the health sciences: a survey of recent trends and developments. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(4):421–434, 2010.

[180] Chun Zeng, Chun-Xiao Xing, Li-Zhu Zhou, and Xiao-Hui Zheng. Similarity measure and instance selection for collaborative filtering. *International Journal of Electronic Commerce*, 8(4):115–129, 2004.

[181] Tommy Hielscher, Myra Spiliopoulou, Henry Völzke, and Jens-Peter Kühn. Using participant similarity for the classification of epidemiological data on hepatic steatosis. In *2014 IEEE 27th International Symposium on Computer-Based Medical Systems*, pages 1–7. IEEE, 2014.

[182] Fei Wang and Jimeng Sun. Survey on distance metric learning and dimensionality reduction in data mining. *Data mining and knowledge discovery*, 29(2):534–564, 2015.

[183] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480, 2006.

[184] Bo Wang, Jiayan Jiang, Wei Wang, Zhi-Hua Zhou, and Zhuowen Tu. Unsupervised metric fusion by cross diffusion. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2997–3004. IEEE, 2012.

[185] Liu Yang and Rong Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2(2):4, 2006.

[186] Fei Wang, Jimeng Sun, Tao Li, and Nikos Anerousis. Two heads better than one: Metric+ active learning and its applications for it service classification. In *2009 Ninth IEEE International Conference on Data Mining*, pages 1022–1027. IEEE, 2009.

[187] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.

[188] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[189] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[190] Kilian Q Weinberger and Lawrence K Saul. Fast solvers and efficient implementations for distance metric learning. In *Proceedings of the 25th international conference on Machine learning*, pages 1160–1167, 2008.

[191] Russell A Reeves, William W Schairer, and David S Jevsevar. The national burden of periprosthetic hip fractures in the us: costs and risk factors for hospital readmission. *HIP International*, 29(5):550–557, 2019.

[192] E Anne Nelson, Mark E Maruish, and Joel L Axler. Effects of discharge planning and compliance with outpatient appointments on readmission rates. *Psychiatric services*, 51(7):885–889, 2000.

[193] Mary Charlson, Ted P Szatrowski, Janey Peterson, and Jeffrey Gold. Validation of a combined comorbidity index. *Journal of clinical epidemiology*, 47(11):1245–1251, 1994.

[194] Hude Quan, Vijaya Sundararajan, Patricia Halfon, Andrew Fong, Bernard Burnand, Jean-Christophe Luthi, L Duncan Saunders, Cynthia A Beck, Thomas E Feasby, and William A Ghali. Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data. *Medical care*, pages 1130–1139, 2005.

[195] Harlan M Krumholz, Ya-Ting Chen, Yun Wang, Viola Vaccarino, Martha J Radford, and Ralph I Horwitz. Predictors of readmission among elderly survivors of admission with heart failure. *American heart journal*, 139(1):72–77, 2000.

[196] Douglas G Altman, Yvonne Vergouwe, Patrick Royston, and Karel GM Moons. Prognosis and prognostic research: validating a prognostic model. *Bmj*, 338:b605, 2009.

[197] Shivapratap Gopakumar, Truyen Tran, Tu Dinh Nguyen, Dinh Phung, and Svetha Venkatesh. Stabilizing high-dimensional prediction models using feature graphs. *IEEE journal of biomedical and health informatics*, 19(3):1044–1052, 2014.

[198] Matthias Schmid, Hans A Kestler, and Sergej Potapov. On the validity of time-dependent auc estimators. *Briefings in Bioinformatics*, 16(1):153–168, 2015.

[199] Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and LJ Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

[200] Sven Stringer, Damiaan Denys, René S Kahn, and Eske M Derks. What cure models can teach us about genome-wide survival analysis. *Behavior genetics*, 46(2):269–280, 2016.

[201] John GT Watkins, Andrey L Vasnev, and Richard Gerlach. Multiple event incidence and duration analysis for credit data incorporating non-stochastic loan maturity. *Journal of Applied Econometrics*, 29(4):627–648, 2014.

APPENDICES

Below list the selected features after each step under the GBM scenario and all the names are given by the authors rather than the original names appear in the EHRs system.

1. 30-day Hip fracture readmission

    - Simple average of ranking

        - After the embedded step (42)

          Lab3_[27.0, 76.0), Los_[7days, 14days), preInp1Y_[2,), Lab5_[1.2, 5.9), Los_[14days, 600days), Gender, Discharge to SNF/SB/LTH, P5D84, Age[65, 75), P5D64, cci_3, Lab3_[5.0, 11.0), Lab4_[9.0, 9.7), preInp1Y_1, P5D8, P157, P5D95, cci_2, Lab4_[11.1, 15.1), Age_[90, 100), Lab1_[140.0, 147.0), Lab3_[15.0, 19.0), P12, P5D70, Lab5_[0.6, 0.76), P5D66, P5D114, P5D238, Lab2_[268.0, 513.0), preER1Y_2,, P5D53, Lab3_[19.0, 27.0), P5D56, Lab0_[4.4, 5.6), Lab6_[26.9, 28.8), preInp3M_1, Discharge to Rehab or STH, epn_3, preInp6M_1, P72, P5D259, P5D78

        - After the wrapper step (21)

          Lab3_[27.0, 76.0), Los_[7days, 14days), preInp1Y_[2,), Los_[14days, 600days), Discharge to SNF/SB/LTH, P5D64, cci_3, Lab3_[5.0, 11.0), preInp1Y_1, P157, P5D95, Age_[90, 100), P12, P5D66, P5D114, P5D238, Lab2_[268.0, 513.0), Lab3_[19.0, 27.0), Discharge to Rehab/STH, preInp6M_1, P72

    - Weighted average of ranking

        - After the embedded step (15)

          Lab3_[27.0, 76.0), Los_[7days, 14days), preInp1Y_[2,), Lab5_[1.2, 5.9), Los_[14days, 600days), Gender, Discharge to SNF/SB/LTH, P5D84, Age_[65, 75), P5D64, cci_3, Lab3_[5.0, 11.0), Lab4_[9.0, 9.7), preInp1Y_1, P5D8

    - After the wrapper step (14)
      Lab3_[27.0, 76.0), Los_[7days, 14days), preInp1Y_2, Lab5_[1.2, 5.9), Los_[14days, 600days), Gender, Discharge to SNF/SB/LTH, P5D84, Age_[65, 75), P5D64, cci_3, Lab3_[5.0, 11.0), Lab4_[9.0, 9.7), preInp1Y_1

2. Diabetic retinopathy prognosis

    - Simple average of ranking
        - After the embedded step (17)
          creatinine, HbA1c, WBC, age, BUN, platelet, glucose, albumin, ALP, AST, BPD, BPS, potassium, protein, Hematocrit, race_AfricanAmerican, ALT
        - After the wrapper step (15)
          creatinine, HbA1c, WBC, age, BUN, glucose, albumin, platelet, ALP, AST, BPD, BPS, potassium, protein, Hematocrit
    - Weighted average of ranking
        - After the embedded step (17)
          creatinine, HbA1c, WBC, age, BUN, glucose, albumin, platelet, AST, ALP, potassium, BPS, BPD, Hematocrit, protein, calcium, ALT
        - After the wrapper step (13)
          creatinine, HbA1c, WBC, age, glucose, albumin, platelet, AST, ALP, BPS, Hematocrit, protein, ALT

VITA

Qingqing Dai

Candidate for the Degree of:

Doctor of Philosophy

Dissertation: FEATURE SELECTION AND PERSONALIZED MODELING ON MEDICAL ADVERSE OUTCOME PREDICTION

Major Field: Statistics

Biographical:

Education:

Completed the requirements for Doctor of Philosophy in Statistics at Oklahoma State University, Stillwater Oklahoma in May, 2020.

Completed the requirements for Bachelor of Economics in Statistics at Central University of Finance and Economics, Beijing China in 2011.

Experience:

Employed by Center for Health Systems Innovation, Oklahoma State University in the position of Research Assistant in Stillwater, Oklahoma from January 2017 to May 2020.

Employed by Department of Statistics, Oklahoma State University in the position of Teaching Assistant in Stillwater, Oklahoma from August 2015 to December 2016.