A

Daehr: a Discriminant Analysis Framework for Electronic Health Record Data with its Application to Early Detection of Mental Health Disorders

HAOYI XIONG, University of Virginia
JINGHE ZHANG, University of Virginia
YU HUANG, University of Virginia
KEVIN LEACH, University of Virginia
LAURA E. BARNES, University of Virginia

Electronic Health Records (EHR) containing a massive number of patient diagnosis records have been used to predict patients' future or potential diseases according to their past diagnoses. While many data mining tools have been adopted for EHR-based disease early detection, Linear Discriminant Analysis (LDA) is one of the most commonly used statistical methods. However, it is difficult to train an accurate LDA model to detect specific diseases early when too few patients are known to have the target disease and the EHR data are manually coded with noise. In such cases, the covariance matrices used in LDA are usually singular and estimated with a large variance.

This paper presents *Daehr*, an extending LDA framework using Electronic Health Records, to address these issues. Beyond existing LDA analyzers, we propose *Daehr* to 1) eliminate the data noise caused by the manual encoding of EHR data, and 2) lower the decision risk of LDA model with finely-estimated parameters when only a few patients' EHR are given for training. To achieve these two goals, we designed an iterative algorithm to improve the covariance matrix estimation with embedded data-noise/decision-risk reduction for LDA. We evaluated *Daehr* extensively using a large-scale real-world EHR dataset, CHSN. Specifically, our experiments compared the performance of LDA to three baselines (i.e., LDA and its derivatives) in terms of identifying college students at high risk for mental health disorders from 23 US universities. Experimental results show *Daehr* significantly outperforms the three baselines by achieving 1.4%–19.4% higher accuracy, and a 7.5%–43.5% higher F1-score.

Categories and Subject Descriptors: J.3 [**Applied computing**]: Health care information systems

General Terms: Human Factors, Algorithms, Performance

Additional Key Words and Phrases: predictive models, early detection, anxiety/depression, temporal order, electronic health data

1. INTRODUCTION

With the rapid development of medical big data, forecasting future or potential diseases based on patients' past medical records has emerged as a promising approach towards preventing high-risk diseases. Rather than individualizing patients (e.g., screening or counseling), a medical informatics system can predict each patient's potential diseases using his or her past diagnoses as well as diagnoses collected from many other patients. In this way, the medical system can identify high-risk patients from a large corpus of patients with low cost. These high-risk patients can then receive targeted care to employ disease prevention techniques in advance. Naturally, the accuracy of such early detection is crucial to improving the efficiency of high-risk patient identification and disease prevention.

In this paper, we present *Daehr*—an extended linear discriminant analysis (LDA) [Fisher 1936; McLachlan 2004] framework for early detection of diseases using Electronic Health Records (EHR), which can improve the prediction accuracy of the standard LDA model by reducing the noise in EHR data and regularizing the estimated covariance matrices. We first discuss the motivations and background of this research, then we formulate a new research problem based on our observations and assumptions. We elaborate the technical challenges of the proposed research. Finally, we summarize our technical contributions.

1.1. Motivations and Backgrounds

To predict patients' potential diseases according to their past medical records, a variety of predictive models utilizing heterogeneous medical data have been studied [Soni et al. 2011; Palaniap-

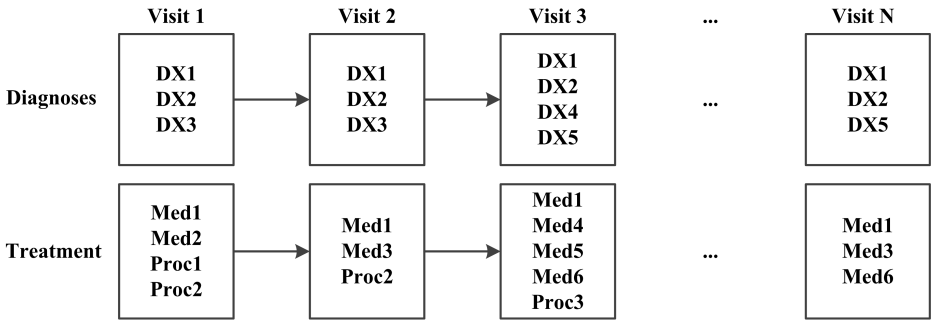

Fig. 1: An Example of a Patient's EHR Data

pan and Awang 2008; Kumari and Godara 2011]. For example, chest imaging has been used for early detection of chest cancer, questionnaire-based assessment (e.g., PHQ-9 [Kroenke and Spitzer 2002]) data for predicting mental health disorders, and screening data for predicting heart disease [D'Agostino Sr et al. 2001]. Among these medical data, Electronic Health Records (EHR) consisting of the diagnosis records from patients' visits are used as a general purpose data source that enables early detection of diseases based on the previous diagnoses at a massive scale. Further more, this data is more accessible to clinicians and researchers, and holds comprehensive information of patients' medical history especially within the primary care setting. Thus, EHR data provides a promising opportunity for early detection of diseases due to its generality, accessibility and standardized use and features.

As shown in Fig. 1, a patient's EHR data includes all his/her past diagnosis and treatment records, where the diagnosis record includes a sequence of visits, and each visit consists of multiple diagnoses. Note that all diagnoses are recorded using ICD-9 codes [Dubberke et al. 2006], where each evidence of diagnosis corresponds to a specific ICD-9 code. With diagnosis records in the EHR data, several methods [Kenney Ng 2015; Amarasingham et al. 2010; Pittman et al. 2004; Jensen et al. 2012] have been studied to predict the disease of patients. Given a disease as the prediction target (e.g., anxiety/depression) as well as the EHR data of a large population with or without the target disease, most existing methods first represent each given patient's EHR data using a set of features, and then train a predictive model using features and labels (if each patient is diagnosed with the targeted disease) in a supervised manner. Further, given each new patient's EHR data, these models predict if the given patient will develop the targeted disease in near future using the trained predictive model.

EHR Data Representation for Early Detection. In terms of representing EHR data, existing approaches include using diagnosis-frequencies [Sun et al. 2012; Wang and Sun 2015; Kenney Ng 2015], pairwise diagnosis transitions [Jinghe Zhang 2015; Jensen and SPSS 2001], graph representations of diagnosis sequences [Liu et al. 2015], and so on. Among these approaches, the diagnosis-frequency is a common way to represent EHR data. Given each patient's EHR data, which consists of the patient's demographic information and a sequence of past visits, existing methods first retrieve the diagnosis codes recorded during each visit. Next, the frequency of each diagnosis appearing in all past visits are counted, followed by further transformation on the frequency of each diagnosis into a vector of frequencies (e.g., $\langle 1, 0, \ldots, 3 \rangle$, where 0 means the second diagnosis does not exist in all past visits). In this way, each patient having different number of visits and each visit consisting of multiple diagnoses is represented as a fixed-length data vector, which can be handled by common machine learning algorithms.

Note that the diagnosis-frequency representation of EHR data is usually with ultra-high dimensions; for example, there are more than 15,000 ICD-9 codes in EHR scheme, thus the diagnosis-frequency vector using raw ICD-9 codes contains thousands of dimensions. To reduce the dimensinality, clinical professionals may suggest using clustered code set, where each ICD-9 code can map

to one of the 295 clustered codes. Thus, each raw diagnosis-frequency vector can be compressed to a vector of around 200 dimensions using clustered codes.

Supervised Learning for Early Detection. Given an EHR database and a target disease for early detection, existing methods first select patients both with and without the disease, then use an appropriate representation of their EHR data to form a training set. To train an accurate predictive model with the training set, many machine learning methods such as Support Vector Machine (SVM), Random Forest (RF), Bayesian Network, Gaussian Process and Linear Discriminant Analysis (LDA) have been adopted [Sun et al. 2012; Wang and Sun 2015; Kenney Ng 2015; Jinghe Zhang 2015; Jensen and SPSS 2001; Liu et al. 2015; Cazzanti and Gupta 2007]. Among these machine learning methods, LDA is frequently used as one of the common performance benchmarks in a series of studies [Cazzanti and Gupta 2007; Jinghe Zhang 2015; Kalina et al. 2013; Karlsson and Bostrom 2014; Wang et al. 2014], because it effectively reduces dimensionality. For example, when using diagnosis-frequency vector as the representation of EHR data, a LDA model learns a linear combination of diagnoses (from the all diagnoses) that can optimally separate patients into the two groups (i.e., with/without the disease). Then LDA predicts whether new patients will develop the targeted disease by separating their vectors into the two groups using the linear combination.

Like many other statistical learning models, the accuracy of a LDA model can be improved, when more samples are given for training. This is because the decision risk of a LDA model is inherited from the variance of its training samples, while *increasing the sample size lowers the sample variance* [Hsu and Robbins 1947; Qiao et al. 2008]. In contrast, when there are few training samples, the model cannot produce any valid prediction results. Because LDA needs to use the *inverse of the covariance matrices* to make prediction. In such case the covariance matrices estimated in LDA are singular or namely *non-invertible* [Huang et al. 2002; Gao and Davis 2006].

We are motivated to enhance the supervised learning methods building upon EHR data so as to improve the prediction accuracy for early detection of diseases. Specifically, we study the LDA model using the diagnosis-frequency features, because of the relevance of such settings in clinical practices.

1.2. Research Assumptions and Objectives

Our research is based on following two observations and two assumptions about EHR data and early detection settings:

Observation 1. EHR Encoding Variation – In terms of encoding EHR data, the diagnosis records are usually inputted manually by clinicians without a unified encoding scheme. Our previous work [Nobles et al. 2015] finds that, for a single patient, there may be a higher number of diagnosis records for one disease than the number of times that the disease as been diagnosed. For example, consider three clinicians Ann, Bob, and Carl, all working in the same clinic. A single patient has been diagnosed with *upper respiratory infection* (ICD-9 code: 465.9). Ann may leave only the record of code 465.9 for the first visit in which the disease is diagnosed. However, Bob may leave the record in the first visit as well as all of the patient's returning visits to receive screening or treatment for upper respiratory infection. Carl may leave a record in the first visit and in some of the returning visits at his discretion.

Assumption I. Non-negative Noise in Diagnosis-Frequency Vector Data – Based upon the first observation, we assume that each diagnosis is recorded at least one time in the EHR and that the number of records might differ due to clinician encoding style (i.e., *frequency of record* $\geq$ *frequency of diagnosis* for each specific disease). We further assume the encoding variation of EHR data may cause certain unknown *non-negative data noise* in the diagnosis-frequency vectors.

Observation 2. Limited Positive Training Samples – We find that the total number of patients with a specific disease (*positive samples*) might be too few to train a predictive model for early detection of the disease. For example, consider a historically black college that wants to identify the high-risk students in terms of mental health disorders using all students' EHR data in the college clinics. The clinics first separate all students in to two groups (i.e., with/without mental health disorders diagnosed). Then it selects a subset of students from each group as training samples. However,

psychiatric clinics are typically underutilized by African American [Thompson et al. 2004], and thus the available training samples that include at least one type of mental health disorders are two few (e.g., 100-500 students) in the school.

Assumption II. Decision Risk of LDA Model for Early Detection of Diseases – Considering the dimension p of diagnosis-frequency vectors (e.g., $p \geq 200$ using clustered code set), we assume that the size of positive samples for LDA training is relatively small i.e., $0 < n \lll 2^p$, where n refers to the number of positive training samples. When $0 < n < p$, the trained LDA model cannot produce any valid prediction results, since the estimated covariance matrix is singular/non-invertible; when $p \leq n \lll 2^p$ the trained LDA model might be able to produce a valid prediction, but with large decision risk inherited from the variance of small training samples.

With above two assumptions in mind, our work attempts to reduce the effect of noise while lowering the decision risk of the LDA model for early detection of diseases. Specifically we use mental health disorders as the "target disease" in evaluation and experiment design, with respect to *Assumption II*.

1.3. Technical Issues and Contributions

In order to improve LDA with respect to the two assumptions, we address the following three technical issues:

(1) ***Eliminating the data noise in diagnosis-frequency vectors caused by encoding variation –*** Given the frequency-diagnosis vectors for training, LDA first estimates diagnosis-to-diagnosis covariance matrices using sample covariance matrix estimator such as *Intrinsic Estimator* or *Maximized Likelihood Estimator (MLE)*, then builds the predictive models using estimated covariance matrices. However, our later analysis shows that the non-negative data noise in the vectors might make the estimated covariance matrices more dense than the noise-free (ideal) one. In this way, we might need a method to *sparsify* the covariance matrices in order to reduce the effect of data noise to LDA.

(2) ***Lowering the decision risk of LDA while guaranteeing non-singularity and positive definiteness of the estimated covariance matrices -*** To lower the decision risk associated with LDA, one possible solution is to use the ℓ_1-penalized estimation of the covariance matrices [Cai and Zhou 2012; Xue et al. 2012]. However, any modifications (including ℓ_1-penalty and sparse approximation) to a coavriance matrix might result in loss of its positive definiteness—we cannot use such modified matrix in the statistics model. We need an algorithm to obtain the ℓ_1-penalized estimation of the sparsifed covariance matrix while ensuring the estimation is non-singular and positive semidefinite.

(3) ***Incorporating the newly-estimated covariance matrices for improving the performance of EHR-based LDA –*** Given the non-singular/positive-definite ℓ_1-penalized sparse estimations of the covariance matrices, we might need use them to replace the covariance matrices originally used in LDA. Thus, we need a generic framework to extend the original LDA through incorporating the aforementioned covariance matrix estimation algorithms. More important, we must make sure that, compared to LDA, the new framework should provide better overall accuracy and F1-score for early detection of the diseases. While most existing predictive models consider overall accuracy as the primary performance metrics, F1-score, characterizing both correctness and completeness to identify high-risk patients from large testing samples, is yet another important metrics of our problem. An early detection framework with higher F1-score usually can identify more high-risk patients with fewer false alarms. Note that a predictive model with improved overall accuracy is not necessarily to have better F1-score. Thus, though it might be hard to achieve the two goals together, we need such a framework that can improve both accuracy and F1-score.

With the aforementioned research challenges in mind, we make following technical contributions in this study:

— In this work, we studied the problem of improving the existing Linear Discriminant Analysis (LDA) for early detection of diseases based on our two assumptions. To the best of our knowledge, this paper is the first work for LDA-based early detection of diseases built upon EHR data, by addressing the issues of encoding variation and low training sample size.
— In order to address these technical challenges, we proposed *Daehr*—an extending LDA framework. It takes a novel approach to eliminate the affect of data noise and lower the decision risk of LDA models through estimating sparse and non-singular diagnosis-to-diagnosis covariance matrices from diagnosis-frequency vectors. Theoretical analysis shows that, with low computational complexity, the proposed algorithm can approximate the ℓ_1-penalized near-sparsest estimation of the diagnosis-to-diagnosis covariance matrices with non-singularity and positive semi-definiteness guaranteed, even when a very limited number of diagnosis-frequency vectors are given for LDA training.
— We evaluated *Daehr* using a real-world dataset CHSN, which contains more than 300,000 students' EHR records collected from 23 US universities in past three years. We designed a set of experiments based on CHSN for large-scale early detection of mental disorders. The experimental results show *Daehr* significantly outperforms three baselines (i.e., LDA and its derivatives) by achieving 1.4%–19.4% higher prediction accuracy, and 7.5%–43.5% higher F1-score.

The paper is structured as follows: Section 2 discusses the previous studies that have been done in the data mining approaches to early detection of disease and LDA extensions. Section 3 introduces the problem formulation of our study and introduces the *Daehr* framework to solved the problem. Section 4 describes two core algorithms used in *Daehr*. Section 5 describes the data used in this research, the experimental design, and the experimental results and analyses. Finally, the summary of this work, future work, and clinical context are discussed in Section 6.

2. RELATED WORK

In this section, we summarize previous studies related to this paper from two aspects: *data mining approaches to early detection of diseases* and *extensions to LDA learning*.

2.1. Big Data Approaches to Disease Early Detection

Various analytical methods have been used to study the causes, prevention, progression, and interventions of diseases. Among these methods, machine learning emerged as a promising technique in the prediction of diseases [Maroco et al. 2011; Huang et al. 2014a]. In this section, we will discuss previous work in two areas: *predictive modeling* and *data representation* approaches.

2.1.1. Predictive Models for Early Detection of Disease. Predictive models have become popular in the early detection of diseases, such as breast cancer, type II diabetes and cardiovascular disease [Lindstrom and Tuomilehto 2003; Siontis et al. 2012; Zheng et al. 2015; Yoo et al. 2011]. The outcomes of the predictive models are beneficial to both care providers and patients. Accurate prediction of diseases can assist clinicians in identifying high-risk patients in an early stage, ultimately leading to more timely diagnoses and more focused delivery of effective treatments to those patients. In essence, the early detection of diseases can be viewed as a classification problem so that well-established classifiers can be used to perform the task. Among the studies on the early detection of mental disorders, a LASSO logistic regression model has been applied to predict the depression severity to help personalize treatment for high-risk patients [Huang et al. 2014a].

2.1.2. EHR Data Representation for Predictive Models. . Many data representation approaches have been developed to preserve useful information from the raw EHR data. Diagnosis-frequency vectors have been proposed [Ng et al. 2015; Huang et al. 2014a] to convert sequences of diagnoses with different lengths into fixed-length data vectors. This approach associates each patient with an intuitive notion of an "intensity" of each disease with which a patient has been diagnosed. Because such vectors can be easily handled by common predictive models without further data representation. Some novel representation methods have been proposed to characterize the temporal

order information in patients' diagnosis sequences using the frequencies of transitions between diagnoses [Jinghe Zhang 2015; Wang et al. 2012a; Liu et al. 2015; Gotz et al. 2014; Perer and Wang 2014; Perer et al. 2015]. Specifically, [Jinghe Zhang 2015] intends to project the frequency of transition between each two diagnoses onto a fully-connected graph, while [Liu et al. 2015] preserves the frequencies of important transitions using sparse graph representation and penalty. While previous work usually considers the frequency of pairwise transitions between each two diagnoses, [Gotz et al. 2014; Perer and Wang 2014; Perer et al. 2015] consider the frequencies of transitions crossing multiple diagnoses using a hyper-graph.

2.2. Extensions to LDA Model

Regarding the application of LDA to EHR-based early detection of diseases, here we mainly introduce several LDA extensions in High Dimension Low Sample Size (HDLSS) settings. As discussed above, when LDA works in HDLSS, there might exist two major technical issues: 1) LDA requires inverting covariance matrices for classification, but these covariance matrices estimated from small number of samples are usually singular (non-invertible), and 2) large decision risk is inherited from the variance of small samples, through classical LDA training. To handle the singular (non-invertible) covariance matrix issue, Ye et al. [Ye et al. 2004] uses the Pseudo-inverse of the singular covariance matrix, while Direct LDA [Lu et al. 2003; Gao and Davis 2006] uses the *simultaneous diagonalization* of covariance matrices, which are non-singular, to replace the original covariance matrices. On the other hand, several works [Clemmensen et al. 2011; Qiao et al. 2008; Shao et al. 2011] have proposed lowering the decision risk via regularizing the estimated covariance matrices.

Daehr is distinct in three ways. First, compared to other data mining approaches to early detection of diseases (e.g., [Lindstrom and Tuomilehto 2003; Siontis et al. 2012; Zheng et al. 2015; Yoo et al. 2011]), *Daehr* is the first work that intends to improve the performance of LDA model by addressing data noise and small positive training sample size issues. Second our contribution is complementary with these works in EHR data representation [Wang et al. 2012a; Wang et al. 2012b; Liu et al. 2015] and we can further improve *Daehr* by incorporating advanced EHR data representation methods. Third, when compared to existing LDA extensions, *Daehr* re-estimates the covariance matrices to (1) eliminate the effect of data noise to LDA model, (2) lower the decision risk inherited from small positive training samples, and (3) guarantee the non-singularity of covariance matrices, while [Ye et al. 2004; Lu et al. 2003; Gao and Davis 2006; Clemmensen et al. 2011; Qiao et al. 2008; Shao et al. 2011] all focus on regularizing the covariance matrices to enable LDA in a general HDLSS setting. Thus, the estimation/optimization problems considered in any single one of the previous studies are mathematically different from ours with different objectives and assumptions.

3. *DAEHR* SYSTEM MODEL

In this section, we first formulate the research problem of our study; then we propose *Daehr* framework to solve the formulated problem.

3.1. Problem Formulations

According to our research assumptions, in this section, we make two definitions and introduce several preliminary studies used in our studies. Further, we formulate our research problem based on all above definitions and preliminaries.

Definition I. *Diagnosis-frequency Vector and Non-negative Noise Vector* – Given EHR data of m patients (both with and without the targeted disease), we can extract m diagnosis-frequency vectors $X_0, X_1 \dots X_{m-1}$. Each vector e.g., $X_i =< 1, 0, \dots, 3 >$ consists of two parts: $\hat{X}_1$ the vector of true diagnosis frequencies (not diagnosis record frequencies) and E_i the non-negative noise vector:

$$X_i = \hat{X}_i + E_i \tag{1}$$

Preliminary I. *Generalized Two-class LDA and Covariance Matrices* – According to the common implementation of a LDA classifier [Ziegel 2003], given m training samples as well as the

labels i.e., $(X_0, l_0) \dots (X_{m-1}, l_{m-1})$ where $l_i \in \{-1, +1\}$ refers to whether the patient i has been diagnosed with the target disease (i.e., positive sample or negative sample), a two-class LDA model first sorts each sample into two groups according to the label, and estimates covariance matrix/mean vector of the two classes, i.e., (Σ_+, μ_+) and (Σ_-,μ_-), using the positive samples and negative samples respectively. Then generalized two-class LDA determine if a new patient (X') would develop to the targeted disease, using

$$\begin{aligned}(X' - \mu_-)^T \Sigma_-^{-1} (X' - \mu_-) + ln|\Sigma_-| - \\ (X' - \mu_+)^T \Sigma_+^{-1} (X' - \mu_+) - ln|\Sigma_+| < T,\end{aligned} \tag{2}$$

where T is an optimal threshold based on the training samples. However, as illustrated in our observation 2, when positive sample size is relatively small e.g., for the rare disease in the database, $Rank(\Sigma_+) < p$, Σ_+ is singular and Σ_+^{-1} doesn't exist. In this case, Equation 2 might not work.

Please note that in the rest of paper we name the both Σ_+ and Σ_- as a *covariance matrix* simply, since they are considered equally in our problem formulation and solution design; in contrast, the covariance matrix may refer to the either Σ_+ or Σ_-.

Definition II. *Sample Diagnosis-to-Diagnosis Covariance Matrix Estimation and Disturbance of Non-negative Noise* – With above settings in mind, we further define Σ as the sample diagnosis-to-diagnosis covariance matrix based on noisy data, $\hat{\Sigma}$ as the sample covariance matrix based on "noisy-free" vectors, and $\Delta = \Sigma - \hat{\Sigma}$ as the disturbance of non-negative noise to covariance estimation.

$$\begin{aligned}\Sigma &= \frac{1}{n} \sum_{i=0}^{n-1} X_i X_i^T = \frac{1}{n} \sum_{i=0}^{n-1} (\hat{X}_i + E_i)(\hat{X}_i + E_i)^T \\ &= \hat{\Sigma} + \Delta\end{aligned} \tag{3}$$

As the sample covariance matrix estimation shown in 3, the disturbance should be:

$$\Delta = \frac{1}{n} \sum_{i=0}^{n-1} (2\hat{X}_i E_i^T + E_i E_i^T).$$

According to our definition $\hat{X}_i$ and E_i are both non-negative matrices, it is not hard to find that $\Delta = \Sigma - \hat{\Sigma} \geq \mathbf{0}$ is a non-negative matrix and $||\Sigma|| \geq ||\hat{\Sigma}||$. Thus, we can roughly conclude that $\hat{\Sigma}$ might be a sparse estimation of Σ.

Preliminary II. *Minimax risk estimation of the covariance matrix in HDLSS settings* – Previous work [Cai and Zhou 2012; Xue et al. 2012] showed that it is possible to achieve *minimax risk* covariance matrix estimation from a few samples, using the *minimal ℓ_1-normal estimation* of the original sample covariance matrix. In this case, in terms of lowering variance of LDA, we can assume that the optimal [Cai and Zhou 2012] covariance matrix $\tilde{\Sigma}$ should be a ℓ_1-penalized sparse estimation of $\hat{\Sigma}$.

Problem Formulation. According to above definitions and preliminaries, this paper considers a problem of finding the positive-definite sparse estimation of $\hat{\Sigma}$-the noisy-free diagnosis-to-diagnosis covariance matrices, to improve the performance of LDA for early detection of disease. Hereby, we define our research problem that, Given n diagnosis-frequency vectors $X_0, X_1 \dots X_{n-1}$, our problem is to estimate $\tilde{\Sigma}$:

$$\text{min. } |\tilde{\Sigma}|_1 \text{ s.t. } ||\tilde{\Sigma} - \hat{\Sigma}||_F^2 \leq \epsilon \text{ and } \tilde{\Sigma} \in I^+ \tag{4}$$

where I^+ refers to the overall set of positive semidefinite matrices. Please note that $\hat{\Sigma}$ is not foreknown due to the unknown data noise.

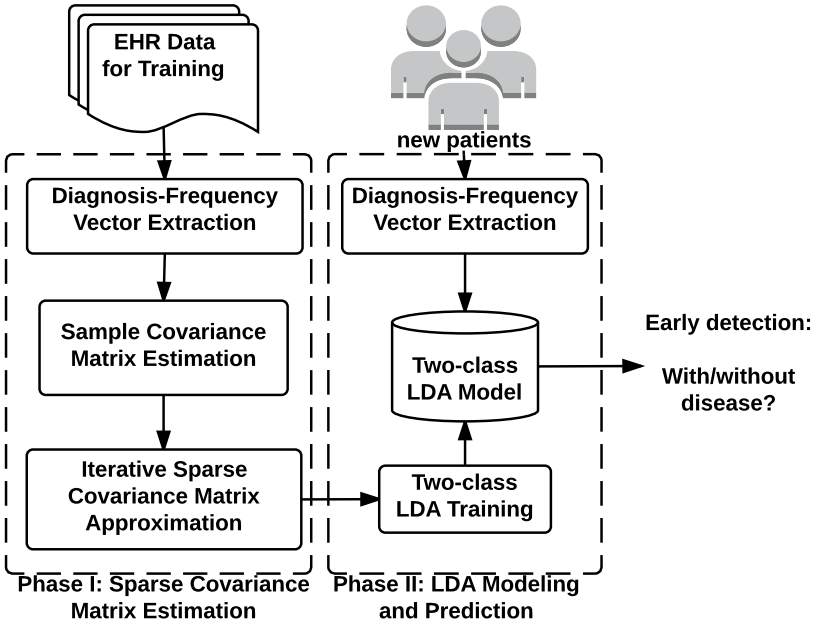

Fig. 2: *Daehr* Framework

Intuitively, it is possible to solve the formulated problem through sparsifying and regularizing the sample diagnosis-to-diagnosis covariance matrix Σ subject to the positive semidefinite and non-singularity constraint.

3.2. *Daehr* Framework

In this section, we introduce the framework design of *Daehr*. *Daehr* consists of two phases. First, we use the EHR data for training to estimate the covariance matrices used in LDA with respect to our problem formulation. Next, we adopt LDA with newly estimated parameters to predict whether the new patient will develop the targeted disease.

Phase I: Sparse Covariance Matrix Estimation — Given the patients' EHR data as a training set, this phase estimates the sparse covariance matrices for two classes of patients with following two steps:

(1) **Diagnosis-frequency Vector Extraction and Sample Covariance Matrix Estimation —** *Daehr* first converts each patient's EHR data to a diagnosis-frequency vector and combines it with his/her label (indicating whether the patient has been diagnosed with the targeted disease). Specifically, we acquire $(X_0, l_0) \dots (X_{m-1}, l_{m-1})$, where $l_i \in \{-1, +1\}$ is the label of the i^{th} patient. With the vectors corresponding to each of the two classes, *Daehr* then estimates the sample covariance matrices for the two classes Σ_+ and Σ_- using Equation 3.

(2) **Iterative Sparse Covariance Matrix Approximation —** Given sample covariance matrices Σ_+ and Σ_-, *Daehr* estimates the positive-definite ℓ_1-penalized estimation of both Σ_+ and Σ_- using a unified iterative approximation process, where *Daehr* treats Σ_+ and Σ_- equally. As shown in Algorithm 1, given an input sample covariance matrix $\Sigma_0 = \Sigma_+$ or Σ_-, the process iteratively approximates to the positive definite ℓ_1-penalized estimation of Σ_0 through alternating between two algorithms—ℓ_1-*penalized Sparse Matrix Estimation* and *Nearest Positive Semidefinite Matrix Approximation* in each iteration. In Algorithm 1, $\Delta' = \frac{||\Sigma_{t+1} - \Sigma_t||_\infty}{||\Sigma_t||_\infty}$ and

tol is a threshold characterizing the tolerance of convergence. Specifically, in each (i.e., the t^{th}, $t \geq 0$) iteration, the process obtains an improved result Σ_{t+1} using the previous result Σ_t. With the result improved each iteration, the algorithm stops only when the predefined convergence is achieved ($\Delta'' < tol$) or after iterating $maxit'$ times (i.e., $t > maxit'$).

Note that the covariance matrices for the two classes of patients are estimated in this phase through a unified process. We denote the new covariance matrices as Σ_+^* and Σ_-^* for the positive and negative classes, respectively.

Algorithm 1: Iterative Approximation Process for Sparse Covariance Matrix Estimation

Data: Σ_0 — the sample covariance matrix i.e., Σ_+ or Σ_-
Result: Σ_{t+1} — the positive definite ℓ_1-penalized estimation of Σ_0

```
begin
    while Δ' ≥ tol, or 0 ≤ t ≤ maxit' do
        Σ_{t+1/2} ← ℓ1-penalized sparse estimation of Σ_t
        Σ_{t+1} ← the nearest positive semidefinite approximation to Σ_{t+1/2}
    end
    return Σ_{t+1}
end
```

Phase II: LDA Modelling and Prediction — Given the two estimated matrices Σ_+ and Σ_- as well as the training samples, this phase first trains the optimal model for LDA prediction. Then, it uses the LDA model for new patient prediction. This phase consists of following two steps:

(1) **LDA Model Training —** Given the two estimated covariance matrices Σ_+^* and Σ_-^* as well as training samples $(X_0, l_0) \ldots (X_{m-1}, l_{m-1})$, *Daehr* searches for the optimal threshold T^* that can maximally classify the two classes of samples using Equation 2. In this case, *Daehr* uses a LDA model as $(\Sigma_+^*, \mu_+, \Sigma_-^*, \mu_-, T^*)$.
(2) **LDA-based new Patient Prediction —** Given a new patient's EHR data, *Daehr* first converts her data to a diagnosis-frequency vector (e.g., X'). Combined with the LDA model described as $(\Sigma_+^*, \mu_+, \Sigma_-^*, \mu_-, T^*)$, *Daehr* predicts whether the patient will develop the targeted disease using the criterion in Equation 2.

After the above two phases terminate, *Daehr* will have (1) learned a LDA model with advanced covariance matrix estimation, and (2) adopted the LDA model to enable the early detection of targeted disease. Though the architecture of the framework is discussed here, the design of the aforementioned ℓ_1*-penalized Sparse Matrix Estimation* and *Nearest Positive Semi-Definite Matrix Approximation* algorithms are discussed in following sections.

4. *DAEHR* CORE ALGORITHMS

In this section, we first introduce the two core algorithm used in *Daehr*, then analyzes the performance of the proposed algorithms.

4.1. ℓ_1-penalized Sparse Matrix Estimation

Given the covariance matrix estimated in the previous iteration Σ_t, this algorithm estimates $\Sigma_{t+\frac{1}{2}}$ — the ℓ_1-penalized sparse estimation of Σ_t, using the Proximal Gradient Descent algorithm [Nesterov 2004] with following objective function:

$$\textit{min.}\ \frac{1}{2}||\Sigma_{t+\frac{1}{2}} - \Sigma_t||_F^2 + \lambda|\Sigma_{t+\frac{1}{2}}|_1, \tag{5}$$

where λ is a Lagrange multiplier [Wu 2009]. When $\lambda \geq 0$, the Eq. 5 is a *convex function with sparse input*, which can be optimally converged using proximal gradient descent [Nesterov 2004]. Note that $\Sigma_{t+\frac{1}{2}}$ is neither symmetric nor positive semidefinite.

4.2. Nearest Positive Semidefinite Matrix Approximation

Given the sparse matrix $\Sigma_{t+\frac{1}{2}}$, we intend to approximate its nearest positive-definite matrix Σ_t (the output of the t^{th} iteration) as Equation 6.

$$\textit{min. } ||\Sigma_{t+1} - \Sigma_{t+\frac{1}{2}}||_F^2 \textit{ s.t. } \Sigma_{t+1} \in I^+ \tag{6}$$

To achieve the goal, we use the Nearest Correlation Matrix Approximation Algorithm [Higham 2002] shown in Algorithm 2. Specifically, the projection $P_S(A) = \frac{1}{2}(V\lambda_+V^T + (V\lambda_+V^T)^T)$ and $\lambda_+ = \langle max\{\lambda_0, 0\}, max\{\lambda_1, 0\} \dots \rangle$, where V, λ_i is the eigenvalue decomposition of A; the projection $P_U(A) = A'$, where $A'_{i,j} = 1$ when $i = j$, and $A'_{i,j} = A_{i,j}$ when $i \neq j$; the stopping criterion $\Delta'' = max\{\frac{||H_{k+1}-H_k||_\infty}{||H_k||_\infty}, \frac{||H^*_{k+1}-H^*_k||_\infty}{||H^*_k||_\infty}, \frac{||H^*_{k+1}-H^*_k||_\infty}{||H_k||_\infty}\}$.

The algorithm terminates upon predefined convergence (i.e., $\Delta'' < tol$) or when the maximal number of iterations is reached ($k = maxit''$). Note that when the algorithm stops at any $k > 0$, the output Σ_{t+1} must be a positive semidefinite matrix. A detailed analysis is discussed in Section 4.3.

Algorithm 2: Nearest Positive Definite Matrix Approximation

Data: $\Sigma_{t+\frac{1}{2}}$ — the ℓ_1-penalized sparse estimation of Σ_t, tol — the tolerance of convergence
Result: Σ_{t+1} — the nearest positive definite approximation to $\Sigma_{t+\frac{1}{2}}$

begin
  initialization:
  $H_0 = \frac{1}{2}(\Sigma_{t+\frac{1}{2}} + \Sigma^T_{t+\frac{1}{2}})$, $k = 1$, $I_{mod_0} = 0$, $\Delta = 1$;
  while $\Delta'' \geq tol$, *or* $0 \leq k \leq maxit''$ **do**
    $R_{k+1} = H_k - I_{mod_k}$,
    $H^*_{k+1} = P_S(R_{k+1})$;
    $I_{mod_{k+1}} = H^*_{k+1} - R_{k+1}$;
    $H_{k+1} = P_U(H^*_{k+1})$;
  end
  $\Sigma_{t+1} = H_{k+1}$
  return Σ_{t+1}
end

4.3. Algorithm Analysis

Daehr consists of two phases: *Phase I: Sparse Covariance Matrix Estimation* and *Phase II: LDA Modeling and Prediction*. According to [Hamsici and Martinez 2008], under certain assumptions, given the parameters (μ_+,Σ_+,μ_-,Σ_-) estimated in *Phase I*, the prediction result of *Phase II* (i.e., Equation 2) is considered to be the Bayes optimal solution based on the estimated parameters. In this case, the overall performance highly depends on the way that *Daehr* estimates parameters in *Phase I* i.e., *iterative sparse covariance matrix approximation* process shown in Algorithm 1. We first discuss the performance of Algorithm 1 in the rest of this section; then we discuss the assumptions when Phase II is optimal in the Discussion section.

To understand theoretical properties of the *iterative sparse covariance matrix approximation* process, we first analyze the core algorithms used in the process, then we conclude the overall performance of the whole approximation process. In each iteration of the process, there are two major steps:

— **ℓ_1-penalized sparse matrix estimation.** As discussed, when $\lambda \geq 0$, the objective function in Equation 5 is convex, the proximal gradient descent algorithm can approximate the optimal solution of Equation 5 when the algorithm converges. Further, we conclude that the result $\Sigma_{t+\frac{1}{2}} \in G$, where G is a convex set.

— **Nearest positive semidefinite matrix approximation.** According to [Higham 2002], when $k \rightarrow +\infty$, the output Σ_{t+1} could converge to the nearest correlation matrix of the symmetric matrix H_0, while $H_0 = \frac{1}{2}(\Sigma_{t+\frac{1}{2}} + \Sigma^T_{t+\frac{1}{2}})$ is the nearest symmetric matrix of $\Sigma_{t+\frac{1}{2}}$ in terms of the Frobenius norm. That is,

$$H_o = arg\min_{H} ||H - \Sigma_{t+\frac{1}{2}}||^2_F \textit{ s.t. } H = H^T.$$

In this case, we can conclude that, given the sparse estimation $\Sigma_{t+\frac{1}{2}}$, Algorithm 2 outputs Σ_{t+1}, the nearest correlation matrix of $\Sigma_{t+\frac{1}{2}}$. Note that the correlation matrix is a positive semidefinite matrix and can be used for linear discriminant analysis after appropriate training (e.g., Phase II of *Daehr*) [Tabachnick et al. 2001]. Further, as both projections P_U and P_S are on convex sets [Higham 2002], we can conclude $\Sigma_{t+1} \in D$, where D is a convex set.

Until now, we have shown that each step of the *iterative sparse covariance matrix approximation* process can obtain the optimal solutions of the corresponding optimization problems (which are the two sub-problems of our original problem); the optimization results of the two steps are located in two convex sets, namely G and D.

With optimality of the two steps in mind, we now analyze the *iterative sparse covariance matrix approximation* process which combines the two steps. Indeed, this process is similar to a process of Alternating Projections [Von Neumann 1951; Escalante and Raydan 2011]. In each iteration (e.g., the t^{th} iteration) of the process, the algorithm first projects the matrix Σ_t to its ℓ_1-penalized sparse estimation $\Sigma_{t+\frac{1}{2}} \in G$, then the algorithm projects $\Sigma_{t+\frac{1}{2}}$ to its nearest correlation matrix (positive semidefinite estimation) $\Sigma_{t+1} \in D$. The algorithm alternatively repeats these two projections until meeting the stopping criterion. According to [Cheney and Goldstein 1959; Bregman 1967], when $t \rightarrow +\infty$, the *iterative sparse covariance matrix approximation* process converges (i.e., $||\Sigma_{t+1} - \Sigma_t|| \rightarrow 0$[1]).

Specifically, when $D \cap G \neq \emptyset$ and $k \rightarrow +\infty$, we can find $||\Sigma_{t+1} - \Sigma_t|| = ||\Sigma_t - \Sigma_{t+\frac{1}{2}}|| \rightarrow 0$ and the iterative process converges at the optimal solution of the positive-semidefinite ℓ_1-penalized sparse estimation of the correlation matrices. Note that the positive definite ℓ_1-penalized sparse estimation of covariance matrices is considered to be the *minimax-risk* solution for covariance matrix estimation in HDLSS [Cai and Zhou 2012; Xue et al. 2012]. Thus, our *iterative sparse covariance matrix approximation* can achieve the optimal correlation matrices in terms of *minimax-risk*, when $D \cap G \neq \emptyset$. However, when $D \cap G = \emptyset$, the iterative process converges at a stationary point (non-optimal). In this case, the estimated covariance matrices satisfy positive-semidefinite constraint and the ℓ_1-norms of these matrices are low.[2] Therefore, we conclude the *Daehr* framework is a quasi-optimal solution to the proposed research problem in this study.

5. EVALUATION

In this section, we introduce the experimental design of our evaluation. Then, we present the experimental results, including the performance comparison between the *Daehr* framework and original LDA baselines. Additionally, we present performance comparisons between *Daehr* and other predictive models. Finally, we compare the time consumed by *Daehr* with other models.

[1]Please refer to [Bregman 1967] for the proof of convergence when $D \cap G \neq \emptyset$, and see [Cheney and Goldstein 1959] for the case when $D \cap G = \emptyset$. To better understand the performance of alternating projections, readers are encouraged to refer to [Escalante and Raydan 2011].

[2]The scope of convex sets D and G highly depends on the given data and cannot be determined in advanced.

5.1. Experimental Design

We first present the datasets used for our evaluation, then introduce the targeted diseases for the early detection. We also specify the settings of early detection.

Dataset for Evaluation — In this study, to evaluate *Daehr*, we plan use the de-identified EHR data from the College Health Surveillance Network (CHSN), which contains over 1 million patients and 6 million visits from 31 student health centers across the United States [Turner and Keller 2015]. In the experiments, we use the EHR data from 10 participating schools. The available information includes ICD-9 diagnostic codes, CPT procedural codes, and limited demographic information. There are over 200,000 enrolled students in those 10 schools representing all geographic regions of the US. The demography of enrolled students (sex, race/ethnicity, age, undergraduate/graduate status) closely matched the demography for the population of US universities.

Targeted Disease for Early Detection — Among all diseases recorded in CHSN, we choose mental health disorders, including *anxiety disorders, mood disorders, depression disorders, and other related disorders*, as the targeted disease for early detection. Specifically, we plan to evaluate *Daehr* using the early detection of mental health disorders in *college students*, considering following issues:

(1) *Emergence of early detection of mental health disorders —* Mental health disorders have become a severe problem in the United States and many other countries that 18.6% adults have at least one mental disorder. According to the Spring 2014 American College Health Association's National College Health Assessment report, approximately half of the college students have had the feeling of hopeless and overwhelming anxiety [American College Health Association 2014].
(2) *Difficulty of recognizing mental health disorders in early stages —* Mental health disorders are frequently unrecognized in primary care. Untimely treatment results in emotional, physical, economic, and social burdens to patients and others.
(3) *Limitations of common approaches for early detection of mental health disorders —* Questionnaires are commonly used to detect mental health disorders. Specific questionnaires, interviews, or standard measurements are often designed by researchers to collect patients' behavioral information targeting a particular psychiatric disorder. In particular, psychological screening, PHQ-9, is used to evaluate a patient's risk of mental health disorders [Kroenke and Spitzer 2002]. However, these approaches are not generally applicable in primary care thus cannot detect mental disorders at an early stage.

We are motivated to use EHR data for the early detection of mental health disorders, considering the accessibility and information contained in EHR data. We are especially interested in whether we can predict the mental health disorders for those students who have not yet received any psychiatric consulting or diagnoses and who do not have any diagnosis records related to mental health in their EHR data. We thus design our experiments as follows.

Early Detection Settings — From the CHSN datasets, we select 21,097 patients with anxiety/depression in the target group and 327,198 patients without any mental health disorder in the control group. We represent each patient using his/her diagnosis-frequency vector based on the clustered codeset, where four clustered codes (i.e., 651, 657, 658, 662) are considered to represent the diagnoses of mental health disorders. Specifically, if a patient has any of these four codes in his/her EHR, we say that he/she has been diagnosed with mental health disorders as ground truth. Note that in our research, we do not intend to predict these four types of mental disorders separately, as these four disorders are usually correlated and heavily overlapped in clinical practices [Kendler et al. 2003]. Further, patients with less than two visits from the control group were excluded from the analysis. Likewise, for target groups, there must be at least two visits in the one month before a patient's first diagnosis of anxiety/depression. Notably, the diagnosis information from within one month[3] of the first diagnosis of anxiety/depression in the target group is excluded for the aim of early

[3]All experiment results presented in this paper are based on the 1-month setting. In the appendix, we also provide the evaluation results of the early detection of mental health disorders based 2-months and 3-months settings.

Table I: Performance Comparison between *Daehr* and LDA Baselines (Testing Sample Size =200 $\times$ 2), where "ACC." refers to accuracy and "F1." refers to F1-Score

		Training Set $\times 2$							
		50		150		250		350	
Algorithm	Parameters	ACC.	F1.	ACC.	F1.	ACC.	F1.	ACC.	F1.
LDA	N/A	0.547	0.539	0.617	0.612	0.639	0.644	0.661	0.670
DIAG	N/A	0.592	0.591	0.635	0.635	0.639	0.639	0.653	0.660
Shrinkage(β)	0.25	0.593	0.592	0.636	0.638	0.640	0.643	0.656	0.665
	0.50	0.594	0.592	0.630	0.630	0.641	0.645	0.660	0.669
	0.75	0.592	0.590	0.626	0.624	0.639	0.643	0.662	0.672
Daehr(λ)	$0.005 * 0.5^0$	0.644	0.692	**0.667**	0.714	**0.662**	**0.716**	**0.670**	**0.722**
	$0.005 * 0.5^1$	0.645	0.694	0.666	0.713	**0.662**	**0.716**	**0.670**	**0.722**
	$0.005 * 0.5^2$	**0.646**	**0.697**	0.663	0.714	**0.662**	**0.716**	**0.670**	**0.722**
	$0.005 * 0.5^3$	**0.646**	0.694	0.661	0.712	**0.662**	**0.716**	**0.670**	**0.722**
	$0.005 * 0.5^4$	**0.646**	0.696	0.662	**0.715**	**0.662**	**0.716**	**0.670**	**0.722**

Table II: Performance Comparison between *Daehr* and LDA Baselines (Testing Sample Size =1000 $\times$ 2), where "ACC." refers to accuracy and "F1." refers to F1-Score

		Training Set $\times 2$							
		50		150		250		350	
Algorithm	Parameters	ACC.	F1.	ACC.	F1.	ACC.	F1.	ACC.	F1.
LDA	N/A	0.552	0.545	0.619	0.620	0.644	0.648	0.656	0.663
DIAG	N/A	0.595	0.588	0.624	0.625	0.641	0.642	0.653	0.662
Shrinkage(β)	0.25	0.596	0.592	0.629	0.631	0.644	0.648	0.657	0.667
	0.50	0.594	0.589	0.630	0.633	0.646	0.649	0.660	0.670
	0.75	0.590	0.584	0.629	0.632	0.647	0.650	0.660	0.668
Daehr(τ)	$0.005 * 0.5^0$	0.653	0.711	**0.655**	**0.716**	0.666	0.718	**0.667**	**0.720**
	$0.005 * 0.5^1$	0.653	0.711	**0.655**	**0.716**	0.666	0.718	**0.667**	**0.720**
	$0.005 * 0.5^2$	**0.653**	**0.712**	**0.655**	**0.716**	0.666	**0.720**	**0.667**	**0.720**
	$0.005 * 0.5^3$	0.652	0.710	**0.655**	**0.716**	0.666	0.719	**0.667**	**0.720**
	$0.005 * 0.5^4$	0.652	0.710	**0.655**	**0.716**	**0.667**	**0.720**	**0.667**	**0.720**

detection. Until now, the diagnosis-frequency vectors used as predictors in our experiment only include the diagnosis frequency of physical health disorders and all mental health related information has been removed. In this case, our experiment is equivalent to predicting whether a patient would develop to mental health disorders according to his/her past diagnoses of physical disorders.

5.2. Comparison to LDA Baselines

To understand the performance impact of *Daehr* beyond classic LDA, we first propose three LDA baseline approaches to compare against *Daehr*:

— **LDA** — This algorithm is based on the common implementation of generalized linear discriminant analysis using sample covariance matrix estimation and Equation 2. This algorithm uses the pseudo-inverse [Ye et al. 2004] to replace the matrix inverse in Equation 2 when the sample covariance matrix is singular.
— **Shrinkage —** This algorithm is based on the aforementioned **LDA** implementation (using pseudo-inverse). However, rather than using the sample covariance matrix, this algorithm adopts the sparse estimation of the covariance matrix $\Sigma^* = \beta * \Sigma + (1 - \beta) * diag(\Sigma)$, where Σ refers to the given sample covariance matrix, $diag(\Sigma)$ refers to a $p \times p$ matrix preserving the diagonal elements of Σ only, and $\beta \geq 0$ is a tuning parameter. The Shrinkage algorithm can be considered as a heuristic approach to the optimization problem addressed in Equation 4.
— **DIAG —** This algorithm is based on the Shrinkage approach with $\beta = 0.0$, which means the sparse estimation of the covariance matrix $\Sigma^* = diag(\Sigma)$ used in LDA only includes the diagonal information of the sample covariance matrix.

Note that the implementation of *Daehr* as well as above baselines are derived from the Java implementation of LDA released by Psychometrica[4].

With the four algorithms, we perform experiments with following settings:

— **Training Samples —** we randomly select 50, 100, 150, 200, 250, 300, 350, and 400 patients from the target group as the positive training samples, then randomly select the same number of patients from the control group as negative training samples; here, the training set of the two classes of patients is balanced.
— **Testing Samples —** we randomly select 200 and 1000 unselected patients (not included in the training set) from the target group as well as the same number of unselected patients from the control group as the testing set; here, the testing set is also balanced.

For each setting, we execute the four algorithms and repeat 30 times. In particular, we are interested in measuring following metrics:

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + TN + FP + FN}, \\ \text{F1-score} &= \frac{2 * TP}{2 * TP + FP + FN} \end{aligned} \tag{7}$$

where TP, TN, FP, and FN refer to the true-positive, true-negative, false-positive, and false-negative classification samples in early detection of mental health disorders respectively. Specifically, the Accuracy metric characterizes the proportion of patients who are accurately classified in the early detection of mental disorders. The F1-Score measures both correctness and completeness of the early detection.

Table I and Table II present part of the comparison results. The results show that under all settings, *Daehr* outperforms the three baseline algorithms in terms of overall accuracy and F1-score. Compared to LDA, *Daehr* achieves 1.4%–18.3% higher accuracy and 7.6%–29.3% higher F1-score. Compared to Shrinkage and DIAG, *Daehr* achieves 1.5%–9.7% higher accuracy and 7.9%–21.1% higher F1-score.

Further, it is clear that decreasing the quantity of training samples results in a larger improvement in accuracy and F1-score. In this case, we can conclude that *Daehr* significantly improves the accuracy and F1-score from the classic LDA, especially when the training sample size is small. *Daehr* outperforms all other baselines derived from LDA in terms of accuracy and F1-score.

[4]Java-Implementation of the Linear Discriminant Analysis, Institute for Psychological Diagnosis, http://www.psychometrica.de/lda.html

Table III: Performance Comparison between *Daehr* and other Predictive Models, where "ACC." refers to accuracy and "F1." refers to F1-Score.

	Training Set $\times 2$			
	50		250	
Algorithm	ACC.	F1.	ACC.	F1.
LDA	0.551	0.549	0.639	0.641
Logit. Reg.	0.614	0.521	0.615	0.501
SVM	0.614	0.608	0.660	0.669
AdaBoost-10	0.643	0.599	0.629	0.538
AdaBoost-50	0.633	0.568	0.633	0.550
Daehr	**0.658**	**0.695**	**0.684**	**0.719**

5.3. Comparison to other predictive models

In order to understand the performance of *Daehr*, we compare it to other predictive models frequently used for early detection of diseases. Specifically, we consider to use following algorithms for the comparison:

— *Support Vector Machine (SVM)* — Inspired by the previous studies [Sun et al. 2012; Kenney Ng 2015; Jinghe Zhang 2015], we use a linear binary SVM classifier with fine-tuned parameters.
— *Logistic Regression (Logit. Reg.)* — Inspired by the recent progress in depression prediction [Huang et al. 2014b], we use a Logistic Regression classifier.
— *AdaBoost-10* and *AdaBoost-50* — To compare an ensemble of learning methods, we use AdaBoost to ensemble multiple Logistic Regression classifiers, where AdaBoost-10 refers to the AdaBoost classifier based on 10 Logistic Regression instances and AdaBoost-50 refers to the one with 50 Logistic Regression instances.

Combined with LDA and *Daehr* ($\lambda = 0.005 * 0.5^2$), we evaluate these six algorithms using the experiment settings introduced in Section 5.2. The comparison results are shown in Table III.[5]

Compared to LDA, SVM, Logistic Regression and AdaBoost can achieve 11.4%–16.7% higher accuracy and 3.5%–10.8% higher F1-score (the only exception is the F1-score of Logistic Regression, which is 5% lower than LDA) with a relatively small training set (Training Set = 50). On a large training set (Training set = 250), SVM still attains better performance than LDA. The performance of LDA is nearly equal to Logistic Regression and AdaBoost in terms of accuracy, while achieving a better F1-score. Compared to SVM, Logistic Regression, and AdaBoost, *Daehr* can achieve 2.3%–19.4% higher accuracy and 7.5%–43.5% higher F1-score. In this case, we can conclude that the classic LDA model cannot perform as well as many other predictive models such as SVM and AdaBoost. However, *Daehr* significantly outperforms all five baseline algorithms in all settings. These results indicate that *Daehr* not only improves LDA, but that *Daehr* is also a leading predictive model for early detection of mental health disorders.

5.4. Two Case Studies

In order to further understand the performance of *Daehr*, we present two case studies to show the time consumption of *Daehr*, then analyze the reason how *Daehr* can outperform LDA baselines.

[5]Please note that the results of LDA and *Daehr* in Table III are slightly different from those in Table I and Table II, since we conduct the two sets of experiments separately.

Table IV: Computation Time Comparison (in Milliseconds, Training Samples: 250×2), "AB "refers to AdaBoost

	LDA	*Daehr*	SVM	Logit. Reg.	AB-10	AB-50
Training	249.1	11076.3	830.97	44.97	484.2	2631.0
Testing	0.098	0.098	0.001	0.002	0.016	0.077

Table V: Covariance Matrix Comparison - Training Set Size =50×2.

Algorithm	$\|\Sigma_+ - \Sigma_{+l}\|_1$	$\|\|\Sigma_+ - \Sigma_{+l}\|\|_F^2$	$\|\Sigma_- - \Sigma_{-l}\|_1$	$\|\|\Sigma_- - \Sigma_{-l}\|\|_F^2$
LDA	94493.04	187413.61	94350.30	187278.89
DIAG	93596.18	186885.30	93599.03	186881.08
Shrinkage-25	93766.30	186912.14	93726.33	186901.50
Shrinkage-50	93999.36	187009.13	93924.81	186974.58
Shrinkage-75	94243.82	187176.29	94135.50	187100.38
Daehr	25773.97	49477.79	27062.96	50418.86

Computational Time Analysis — We measure computational time consumption of the six algorithms in the experiments introduced in Section 5. We carried out the experiments using a laptop with an Intel Core i7-2630QM Quad-Core CPU and 8GB memory. All algorithms used in our experiments were implemented with the Java SE platform on a Java HotSpot(TM) 64-Bit Server VM. Table IV shows the computational time comparison between *Daehr* and the rest of methods, where the *"Training"* row refers to the average time consumption of the six algorithms to train a model. The average time consumption to classify each patient of the testing set is shown in the *"Testing"* row. Among these six algorithms, *Daehr* takes the longest time to train—however, the average time consumption to train a model with $250 \times 2 = 500$ samples is less than 12 seconds, which is acceptable. On the other hand, the average time consumption to classify a patient using *Daehr* is similar to LDA, as these two algorithms are equivalent in terms of prediction. In any case, the time consumption of all these six algorithms to classify patients is quite tolerable (i.e., thousands patients per second). We conclude that all of the algorithms described here, including *Daehr*, are computationally efficient, in terms of model training and early detection of diseases.

Correlation Matrix Estimation Analysis — We assume *Daehr* improves LDA the model because the sparse covariance matrix used in *Daehr* is more "accurate" than the sample covariance matrix used in LDA when the training sample size is limited. In order to verify our hypothesis, we (1) gather the EHR data of all 21,097 patients with mental health disorders from CHSN (4 years EHR of 22 US Universities); (2) randomly select 10,000 patients from them to estimate correlation matrix Σ_{+l}, (3) randomly select another 50 or 250 samples to train LDA and *Daehr*; and (4) further compare Σ_{+l} to the correlation matrices estimated in LDA and *Daehr* separately through measuring the error of matrices. We repeat steps 1 through 4 for a total 30 trials so as to obtain the average error between the correlation matrices. We similarly compare the matrices estimated using negative samples (i.e., patients without mental disorders). Table V presents the average error between correlation matrices in ℓ^1/Frobenius-norm. The results show that, compared to LDA, the correlation matrix estimated in *Daehr* using small samples is **more closed** to the correlation matrix estimated using large samples. In this case, we conclude that *Daehr* can accurately estimate the covariance matrix for linear discriminant analysis, even when a small number of samples are given for model training.

Note that in our experiment, we simulate a training set with a relatively large sample size (i.e., 10,000). However, for realistic predictive model training, such a large number of samples is usually not available.

To conserve space, some results are not reported here. Readers are encouraged to see the Appendix for additional details, including the evaluation results under more evaluation settings and more experimental insights.

6. DISCUSSION

Due to space limitation, many important issues have been elided from this paper. In this section, we discuss the following issues:

— **Leveraging both Diagnosis and Treatment Records.** In this paper, we use only diagnosis records as predictors for the early detection of diseases, while there exists some other papers [Wang et al. 2012a; Liu et al. 2015; Gotz et al. 2014] using both diagnosis and treatment records in EHR data. It is reasonable to assume that using both diagnosis and treatment records can further improve the performance of *Daehr*. On the other hand, EHR uses CPT codes to encode treatment records; these treatment records are frequently represented using treatment-frequency vectors for machine learning. In this way, we can conclude that after certain extension, *Daehr* can predict further/potential diseases using both diagnosis and treatment records.

— **ICD-9/Clustered Codes and Other Data Representations.** In this paper, rather than using the raw ICD-9 codes, we adopt clustered codes to represent the diagnosis records. The clustered code set maps 15000 ICD-9 codes to 295 codes, where each ICD-9 code corresponds to a single clustered code. Further, we assume to use diagnosis-frequency vectors to represent the diagnosis records in EHR data. Compared to ICD-9 codes, using the clustered codes reduces the dimensionality of diagnosis-frequency vectors. Apparently, such practical dimensionality reduction causes information loss and is not optimal. In our future work, we will study other data representations with *Daehr*.

— **Dimensionality Reduction and Other Approaches to HDLSS Problems.** In this paper, we have shown that the performance of traditional LDA might be bottlenecked due to High Dimension Low Sample Size (HDLSS) settings. In this case, we proposed to use sparse covariance matrix estimation to lower the decision risk of LDA caused by HDLSS. There are several alternative approaches to tackling HDLSS challenges such as feature extraction, representation learning, and kernel methods. We believe *Daehr* can be further improved when combining with these approaches and further our contribution in *Daehr* is complementary to these studies.

— **Bayes Optimality and Gaussian Distribution Assumptions.** *Phase II* of *Daehr* is Bayes optimal, only when the diagnosis frequency data is normally distributed and follows the multivariate Gaussian distribution with the estimated parameters [Hamsici and Martinez 2008]. However, the data extracted from real-world EHR might be non-Gaussian (e.g., Poisson-alike). Empirically, even though the assumptions of Gaussian distribution are often violated, LDA frequently achieves good performance in many classification tasks. *Daehr* can be further improved when working with normally distributed data. In future work, we will study the way to re-scale the frequency data with respect to the Gaussian distribution assumptions.

— **Sensitivity, Specificity and Other Performance Metrics.** In this paper, we present the comparison results of accuracy and F1-score between *Daehr* and other baselines. Actually, many other performance metrics including sensitivity and specificity are frequently used in other early detection work. To conserve space, our main paper presents only the accuracy and F1-score results. Please refer the comparison of specificity and sensitivity in our Appendix. Indeed, *Daehr* also achieves acceptable sensitivity and specificity. On average, compared to classic LDA, *Daehr* achieves 30% higher sensitivity while scarifying no more than 10% specificity. We conclude the overall performance improvement satisfies our expectations for early detection of mental health diseases for two reasons. 1). For early detection of mental health disorders, sensitivity is more important than specificity. Compared to the cost of recognizing a patient with mental health

disorders as negative (i.e., false-negative), the cost of false-positive is relatively low: all patients predicted to be high-risk are suggested to use psychiatric services to receive further psychological diagnoses/intervention, while patients predicted to be low-risk might never go for consultation due to the low utilization rate of psychiatric services; 2) A lower specificity is expected with respect to the real positive/negative-patient distribution in ground truth. When testing our algorithms, many negative testing samples are in fact patients with mental health disorders but have not yet been diagnosed. Since mental health disorders such as anxiety and depression are not with significant/obvious symptoms or physiological changes, many patients with anxiety/depression cannot be recognized if they do not visit psychological departments for diagnoses.

7. CONCLUSIONS

In this paper, we proposed *Daehr* — a novel discriminant analysis framework for early detection of diseases, based on electronic health record data. *Daehr* is designed to 1) reduce the effect of EHR data noise to LDA model training, and 2) lower the decision risk of LDA prediction through regularizing the covariance matrix estimation. To improve the performance of LDA model by achieving these two goals, *Daehr* leverages the process of alternating projections with ℓ_1*-penalized sparse matrix estimation* and *nearest positive-definite matrix approximation* to train the LDA model. Theoretical analysis shows that *Daehr* can achieve quasi-optimal solution in terms of LDA-based early disease detection. The experimental results using real-world EHR dataset CHSN showed *Daehr* significantly outperformed three baselines by achieving 1.4%–19.4% higher prediction accuracy and 7.5% – 43.5% higher F1-score. Further experimental results and discussion details are addressed in Appendix.

REFERENCES

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. 2010. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Medical care* 48, 11 (2010), 981–988.

American College Health Association. 2014. American College Health Association National College Health Assessment. *Spring 2014 Reference Group Executive Summary* (2014). http://www.ijme.net/archive/2/communication-training-and-perceived-patient-similarity/

Lev M Bregman. 1967. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics* 7, 3 (1967), 200–217.

T Tony Cai and Harrison H Zhou. 2012. Minimax estimation of large covariance matrices under l1 norm. *Statistica Sinica* 22, 4 (2012), 1319–1378.

Luca Cazzanti and Maya R. Gupta. 2007. Local Similarity Discriminant Analysis. In *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*. ACM, New York, NY, USA, 137–144.

Ward Cheney and Allen A Goldstein. 1959. Proximity maps for convex sets. *Proc. Amer. Math. Soc.* 10, 3 (1959), 448–450.

Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll. 2011. Sparse discriminant analysis. *Technometrics* 53, 4 (2011).

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

Erik R Dubberke, Kimberly A Reske, L Clifford McDonald, and Victoria J Fraser. 2006. ICD-9 codes and surveillance for Clostridium difficile–associated disease. *Emerging infectious diseases* 12, 10 (2006), 1576.

René Escalante and Marcos Raydan. 2011. *Alternating projection methods*. Vol. 8. SIAM.

Ronald A Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics* 7, 2 (1936), 179–188.

Hui Gao and James W Davis. 2006. Why direct LDA is not equivalent to LDA. *Pattern Recognition* 39, 5 (2006), 1002–1006.

David Gotz, Fei Wang, and Adam Perer. 2014. A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *Journal of Biomedical Informatics* 48 (April 2014), 148–159. DOI: http://dx.doi.org/10.1016/j.jbi.2014.01.007

Onur C Hamsici and Aleix M Martinez. 2008. Bayes optimality in linear discriminant analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 30, 4 (2008), 647–657.

Nicholas J Higham. 2002. Computing the nearest correlation matrixa problem from finance. *IMA journal of Numerical Analysis* 22, 3 (2002), 329–343.

Pao-Lu Hsu and Herbert Robbins. 1947. Complete convergence and the law of large numbers. *Proceedings of the National Academy of Sciences of the United States of America* 33, 2 (1947), 25.

Rui Huang, Qingshan Liu, Hanqing Lu, and Songde Ma. 2002. Solving the small sample size problem of LDA. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, Vol. 3. IEEE, 29–32.

Sandy H. Huang, Paea LePendu, Srinivasan V. Iyer, Ming Tai-Seale, David Carrell, and Nigam H. Shah. 2014a. Toward personalizing treatment for depression: predicting diagnosis and severity. *Journal of the American Medical Informatics Association: JAMIA* 21, 6 (Dec. 2014), 1069–1075. DOI: http://dx.doi.org/10.1136/amiajnl-2014-002733

Sandy H Huang, Paea LePendu, Srinivasan V Iyer, Ming Tai-Seale, David Carrell, and Nigam H Shah. 2014b. Toward personalizing treatment for depression: predicting diagnosis and severity. *Journal of the American Medical Informatics Association* 21, 6 (2014), 1069–1075.

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

Susan Jensen and UK SPSS. 2001. Mining medical data for predictive and sequential patterns: PKDD 2001. In *Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases*.

Yu Huang Hao Wu Kevin Leach Laura E. Barnes Jinghe Zhang, Haoyi Xiong. 2015. MSEQ: Early Detection of Anxiety and Depression via Temporal Orders of Diagnoses in Electronic Health Data. In *Big Data (Workshop), 2015 International Conference on*. IEEE.

Jan Kalina, Libor Seidl, Karel Zvára, Hana Grünfeldová, Dalibor Slovák, and Jana Zvárová. 2013. Selecting relevant information for medical decision support with application to cardiology. *European Journal for Biomedical Informatics* 9, 1 (2013), 2–6.

Isak Karlsson and Henrik Bostrom. 2014. Handling Sparsity with Random Forests When Predicting Adverse Drug Events from Electronic Health Records. In *Healthcare Informatics (ICHI), 2014 IEEE International Conference on*. IEEE, 17–22.

Kenneth S Kendler, John M Hettema, Frank Butera, Charles O Gardner, and Carol A Prescott. 2003. Life event dimensions of loss, humiliation, entrapment, and danger in the prediction of onsets of major depression and generalized anxiety. *Archives of general psychiatry* 60, 8 (2003), 789–796.

Jianying Hu Fei Wang Kenney Ng, Jimeng Sun. 2015. Personalized Predictive Modeling and Risk Factor Identification using Patient Similarity. *AMIA Summit on Clinical Research Informatics (CRI)* (2015).

Kurt Kroenke and Robert L Spitzer. 2002. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatr Ann* 32, 9 (2002), 1–7.

Milan Kumari and Sunila Godara. 2011. Comparative study of data mining classification methods in cardiovascular disease prediction 1. (2011).

Jaana Lindstrom and Jaakko Tuomilehto. 2003. The Diabetes Risk Score: A practical tool to predict type 2 diabetes risk. *Diabetes Care* 26, 3 (2003), 725–731.

Chuanren Liu, Fei Wang, Jianying Hu, and Hui Xiong. 2015. Temporal Phenotyping from Longitudinal Electronic Health Records: A Graph Based Framework. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 705–714. DOI: http://dx.doi.org/10.1145/2783258.2783352

Juwei Lu, Kostantinos N Plataniotis, and Anastasios N Venetsanopoulos. 2003. Face recognition using LDA-based algorithms. *Neural Networks, IEEE Transactions on* 14, 1 (2003), 195–200.

Joo Maroco, Dina Silva, Ana Rodrigues, Manuela Guerreiro, Isabel Santana, and Alexandre de Mendona. 2011. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Research Notes* 4, 1 (Aug. 2011), 299. DOI: http://dx.doi.org/10.1186/1756-0500-4-299

Geoffrey McLachlan. 2004. *Discriminant analysis and statistical pattern recognition*. Vol. 544. John Wiley & Sons.

Yurii Nesterov. 2004. *Introductory lectures on convex optimization*. Vol. 87. Springer Science & Business Media.

Kenney Ng, Jimeng Sun, Jianying Hu, and Fei Wang. 2015. Personalized Predictive Modeling and Risk Factor Identification using Patient Similarity. *AMIA Summits on Translational Science Proceedings* 2015 (March 2015), 132–136. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4525240/

Alicia Nobles, Ketki Vilankar, Hao Wu, and Laura Barnes. 2015. Evaluation of Data Quality of Multisite Electronic Health Record Data for Secondary Analysis. In *Big Data (Workshop), 2015 International Conference on*. IEEE.

Sellappan Palaniappan and Rafiah Awang. 2008. Intelligent heart disease prediction system using data mining techniques. In *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on*. IEEE, 108–115.

Adam Perer and Fei Wang. 2014. Frequence: interactive mining and visualization of temporal frequent event sequences. In *Proceedings of the 19th international conference on Intelligent User Interfaces*. ACM, 153–162. http://dl.acm.org/citation.cfm?id=2557508

Adam Perer, Fei Wang, and Jianying Hu. 2015. Mining and exploring care pathways from electronic medical records with visual analytics. *Journal of Biomedical Informatics* 56 (Aug. 2015), 369–378. DOI: http://dx.doi.org/10.1016/j.jbi.2015.06.020

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, and others. 2004. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences of the United States of America* 101, 22 (2004), 8431–8436.

Zhihua Qiao, Lan Zhou, and Jianhua Z Huang. 2008. Effective linear discriminant analysis for high dimensional, low sample size data. In *Proceeding of the World Congress on Engineering*, Vol. 2. Citeseer, 2–4.

Jun Shao, Yazhen Wang, Xinwei Deng, Sijian Wang, and others. 2011. Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of statistics* 39, 2 (2011), 1241–1265.

George C M Siontis, Ioanna Tzoulaki, Konstantinos C Siontis, and John P A Ioannidis. 2012. Comparisons of established risk prediction models for cardiovascular disease: systematic review. *BMJ* 344 (2012).

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

Jimeng Sun, Fei Wang, Jianying Hu, and Shahram Edabollahi. 2012. Supervised patient similarity measure of heterogeneous patient records. *ACM SIGKDD Explorations Newsletter* 14, 1 (2012), 16–24.

Barbara G Tabachnick, Linda S Fidell, and others. 2001. Using multivariate statistics. (2001).

Vetta L Sanders Thompson, Anita Bazile, and Maysa Akbar. 2004. African Americans' perceptions of psychotherapy and psychotherapists. *Professional psychology: Research and practice* 35, 1 (2004), 19.

James C. Turner and Adrienne Keller. 2015. College Health Surveillance Network: Epidemiology and Health Care Utilization of College Students at U.S. 4-Year Universities. *Journal of American college health: J of ACH* (June 2015), 0. DOI: http://dx.doi.org/10.1080/07448481.2015.1055567

John Von Neumann. 1951. *Functional operators: The geometry of orthogonal spaces*. Princeton University Press.

Fei Wang, Noah Lee, Jianying Hu, Jimeng Sun, and Shahram Ebadollahi. 2012a. Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 453–461. http://dl.acm.org/citation.cfm?id=2339605

Fei Wang, Noah Lee, Jianying Hu, Jimeng Sun, Shahram Ebadollahi, and A. Laine. 2012b. A Framework for Mining Signatures from Event Sequences and Its Applications in Healthcare Data. (2012). http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6200289

Fei Wang and Jimeng Sun. 2015. PSF: A Unified Patient Similarity Evaluation Framework Through Metric Learning With Weak Supervision. *Biomedical and Health Informatics, IEEE Journal of* 19, 3 (May 2015), 1053–1060. DOI: http://dx.doi.org/10.1109/JBHI.2015.2425365

Fei Wang, Ping Zhang, Xiang Wang, and Jianying Hu. 2014. Clinical Risk Prediction by Exploring High-Order Feature Correlations. In *AMIA Annual Symposium Proceedings*, Vol. 2014. American Medical Informatics Association, 1170.

Hsien-Chung Wu. 2009. The Karush–Kuhn–Tucker optimality conditions in multiobjective programming problems with interval-valued objective functions. *European Journal of Operational Research* 196, 1 (2009), 49–60.

Lingzhou Xue, Shiqian Ma, and Hui Zou. 2012. Positive-definite 1-penalized estimation of large covariance matrices. *J. Amer. Statist. Assoc.* 107, 500 (2012), 1480–1491.

Jieping Ye, Ravi Janardan, Cheong Hee Park, and Haesun Park. 2004. An optimization criterion for generalized discriminant analysis on undersampled problems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 26, 8 (2004), 982–994.

Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, and Lei Hua. 2011. Data Mining in Healthcare and Biomedicine: A Survey of the Literature. *Journal of Medical Systems* 36, 4 (May 2011), 2431–2448. DOI: http://dx.doi.org/10.1007/s10916-011-9710-5

Bichen Zheng, Jinghe Zhang, Sang Won Yoon, Sarah S. Lam, Mohammad Khasawneh, and Srikanth Poranki. 2015. Predictive modeling of hospital readmissions using metaheuristics and data mining. *Expert Systems with Applications* 42, 20 (Nov. 2015), 7110–7120. DOI: http://dx.doi.org/10.1016/j.eswa.2015.04.066

Eric R Ziegel. 2003. Modern Applied Statistics With S. *Technometrics* 45, 1 (2003), 111.