Daehr: A Discriminant Analysis Framework for Electronic Health Record Data and an Application to Early Detection of Mental Health Disorders

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

Electronic health records (EHR) provide a rich source of temporal data that present a unique opportunity to characterize disease patterns and risk of imminent disease. 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 for early disease diagnosis when too few patients are known to have the target disease. Furthermore, EHR data are heterogeneous with significant noise. In such cases, the covariance matrices used in LDA are usually singular and estimated with a large variance.

This article presents *Daehr*, an extension of the LDA framework using electronic health record data 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 variance of parameter (covariance matrices) estimation for LDA models when only a few patients' EHR are available for training. To achieve these two goals, we designed an iterative algorithm to improve the covariance matrix estimation with embedded data-noise/parameter-variance reduction for LDA. We evaluated *Daehr* extensively using the College Health Surveillance Network, a large, real-world EHR dataset. Specifically, our experiments compared the performance of LDA to three baselines (i.e., LDA and its derivatives) in identifying college students at high risk for mental health disorders from 23 U.S. universities. Experimental results demonstrate *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: Data Mining, Algorithms, Performance

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

ACM Reference Format:
Haoyi Xiong, Jinghe Zhang, Yu Huang, Kevin Leach, and Laura E. Barnes. 2017. *Daehr*: A discriminant analysis framework for electronic health record data and an application to early detection of mental health disorders. ACM Trans. Intell. Syst. Technol. 8, 3, Article 47 (February 2017), 21 pages.
DOI: http://dx.doi.org/10.1145/3007195

This work was supported in part by the University of Virginia Hobby Postdoctoral and Predoctoral Fellowships in Computational Science.
Authors' addresses: H. Xiong, Department of Computer Science, Missouri University of Science and Technology, 324, Computer Science Bldg., Rolla, Missouri, 65409, USA; email: xiongha@mst.edu; J. Zhang and L. E. Barnes, Department of System and Information Engineering, University of Virginia, 151 Engineer's Way, Charlottesville, VA 22904, USA; emails: {jz4kg, lb3dp}@virginia.edu; Y. Huang and K. Leach, Department of Computer science, Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia, 151 Engineer's Way, Charlottesville, VA 22904, USA; emails: {yh3cf, kjl2y}@virginia.edu.

DOI: http://dx.doi.org/10.1145/3007195

1. INTRODUCTION

Electronic health records (EHRs) are digital versions of a patient's medical history, maintained over time by health care providers, that contain information relevant to a patient's care, including to demographics, diagnoses, medical procedures, medications, vital signs, immunizations, laboratory results, and radiology images [cms 2012]. With the widespread adoption of patient EHR systems, new methodologies have emerged for assisting with patient diagnosis and studying patterns of care. Because patient EHR data reflect the longitudinal nature of patient care, patients' sequences of diagnoses and treatments have the potential to be utilized to build models to predict future disease state. Effective mining of this data is crucial to gaining actionable clinical insights.

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

1.1. Motivation and Background

EHR data are the most comprehensive, accessible standard and frequently used across health care organizations and, thus, provide the most promising opportunity for data-driven healthcare research. Recently, there has been a great deal of work on predicting future disease state and adverse events from patients' EHR data [Gil-Herrera et al. 2015; Ng et al. 2015a; Amarasingham et al. 2010; Pittman et al. 2004; Jensen et al. 2012]. For example, retrospective clinical data have been used to build models that predict both diagnosis and severity of depression [Huang et al. 2014a], sepsis in hospitalized patients [Mitchell et al. 2016], and coronary heart disease [D'Agostino et al. 2001].

Figure 1 depicts an individual patient's EHR data, which consist of a temporal record of their past diagnoses and treatments by visit. Diagnoses, $DX1 \ldots DXn$, characterize the patients disease state while procedures ($Proc1 \ldots Procn$) and medications ($Med1 \ldots Medn$) characterize the patient treatments. These data are commonly coded using International Classification of Diseases version 9 (ICD-9) codes [Dubberke et al. 2006].

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 (i.e., whether each patient is diagnosed with the target disease or not) in a supervised manner. Further, given each new patient's EHR data, these models then predict if the given patient will develop the targeted disease in the near future.

EHR Data Representation for Early Detection of Diseases. There are many existing approaches for representation of EHR data, including the use of diagnosisfrequencies [Sun et al. 2012; Wang and Sun 2015; Ng et al. 2015a], pairwise diagnosis transitions [Zhang et al. 2015; Jensen et al. 2001], and graph representations of diagnosis sequences [Liu et al. 2015]. Among these approaches, the diagnosis-frequency is the most common way to represent EHR data. Given each patient's EHR data, which consist of the patient's demographic information and a sequence of past visits, existing methods first retrieve the diagnosis codes recorded during each visit. Next, each occurrence of each diagnosis appearing in all past visits are counted, followed by subsequent transformation of the frequency of each diagnosis into a vector of frequencies (e.g., $\langle 1, 0, \ldots, 3 \rangle$, where the 0 indicates the second diagnosis does not appear in all past

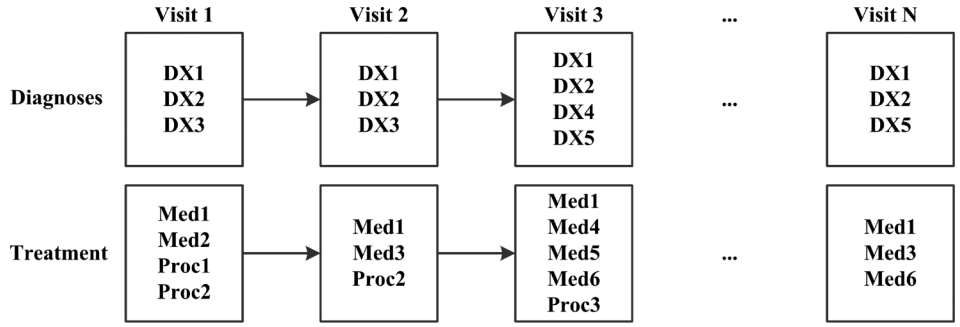


Fig. 1. An Example of a Patient's EHR Data.

visits). By using this structure, each patient's diagnoses can be represented with a fixed-length data vector despite differing numbers of visits and diagnoses among patients. These fixed-length vectors are easily used in common machine-learning algorithms.

Using the diagnosis-frequency representation of EHR data, the problem is extremely high dimensional and sparse: For example, there are more than 14,000 standard ICD-9 codes, and, thus the diagnosis-frequency vector using raw ICD-9 codes contains thousands of dimensions with many zero values [HCUP 2014].

One way to reduce the dimensionality of the data is to group codes by meaningful clinical groupings [HCUP 2014]. For example, using the Agency for Healthcare Research and Quality (AHRQ) Clinical Classification System (CCS), each ICD-9 diagnosis code can map to 1 of 295 groups [Zhang et al. 2015], allowing us to compress each raw diagnosis-frequency vector to roughly 200 dimensions.

Supervised Learning for Early Detection of Diseases. Given EHR data and a target disease for early detection, existing methods first select patients both with and without the disease and then use an appropriate representation of their EHR data to form a training set. To train an accurate predictive model, many machine-learning methods such as Support Vector Machine (SVM), Random Forest, Bayesian Network, Gaussian Process, and LDA have been adopted [Sun et al. 2012; Wang and Sun 2015; Ng et al. 2015a; Zhang et al. 2015; Jensen et al. 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; Zhang et al. 2015; Kalina et al. 2013; Karlsson and Bostrom 2014; Wang et al. 2014] because it effectively reduces dimensionality.

For example, when using diagnosis-frequency vectors for representing EHR data, an 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 an LDA model can be improved when more samples are available for training. This is because the expected loss [Berger 2013] and the variance of parameter estimation for an 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 a few training samples (especially when the number of training samples is less than the number of dimensions), the model cannot produce any valid prediction results.

Because LDA needs to use the *inverse of the covariance matrices* to make a prediction, in such cases, the covariance matrices estimated in LDA are not invertible or namely singular [Huang et al. 2002; Gao and Davis 2006].

With this background in mind, we are motivated to enhance supervised learning methodologies by building on high-dimensional EHR data to improve predictive models. Specifically, we study the LDA model using the diagnosis-frequency features because of the clinical relevance and application of such methods to early disease detection.

1.2. Research Assumptions and Objectives

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

Observation 1. Variation in EHR Coding Practices – EHR data are input by clinicians without a unified coding scheme. Coding variation can result from a variety of factors, including EHR vendor platform, organizational culture, and policies (e.g., insurance) [Nobles et al. 2015]. One such example is differing numbers of diagnosis records for repeated visits for the same condition. Some physicians will code the diagnosis at each repeated visits while others will only code it in the first visit. Other typical coding variations include using different codes for the same diagnosis.

Assumption I. Non-negative Noise in Diagnosis-Frequency Vector Data – Based on 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 or organizational coding practices (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 college that wants to identify the students at high-risk of mental health disorders after their recent adoption of an EHR system. The clinic first separates all students into two groups (i.e., with/without mental health disorders diagnosed). Next, they select a subset of students from each group as training samples. However, the EHR system was installed only 1 year ago, and thus the available training samples that include at least one type of mental health disorders are too few (e.g., 100–500 students) in the school.

Assumption II. Variance of Parameter (Covariance Matrix) Estimation for LDA-Based Early Detection of Diseases – Considering the dimension p of diagnosis-frequency vectors (e.g., $p \geq 200$ using diagnosis groups), we assume that the size of positive samples for LDA training is relatively small (i.e., $0 < m \ll 2^p$, where m refers to the number of positive training samples). When $0 < m < p$, the trained LDA model cannot produce any valid prediction results, since the estimated covariance matrix is singular/non-invertible; when $p \leq m \ll 2^p$, 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 these 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 the problem of early detection of mental health disorders as the "target disease" in evaluation and experimental design with respect to *Assumption II*.

1.3. Technical Contributions

In this article, we make following technical contributions:

—In this work, we study the problem of improving the existing LDA framework for early detection of diseases based on the aforementioned assumptions. To the best of our knowledge, this article is the first work for LDA-based early detection of diseases utilizing EHR data addressing the issues of coding variation and small sample size for for training.

—In order to address the technical challenges, we propose *Daehr*—an extending LDA framework. We propose a novel approach to eliminate noise and lower the variance of parameter (covariance matrices) estimation for the 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. According to the theory of minimax-risk covariance estimation [Cai and Zhou 2012], under certain assumptions, the maximal expected loss of *Daehr* for parameter estimation is minimal among all possible solutions.
—We evaluate *Daehr* using data from the College Health Surveillance Network (CHSN) [Turner and Keller 2015], electronic health record data from student health centers representing more than 300,000 students from 23 U.S. universities. We designed a set of experiments based on CHSN for large-scale early detection of mental health 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 article is structured as follows: Section 2 introduces the background and problem formulation of our study. Section 3 first presents the *Daehr* framework to solve the problem and then describes two core algorithms used in *Daehr*. Section 4 describes the data used in this research, the experimental design, results, and analyses. Section 5 discusses the previous studies that have been done in the data-mining approaches to early detection of disease and LDA extensions. Finally, the summary of this work, future work, and clinical context are discussed in Section 6.

2. *DAEHR* SYSTEM MODEL

In this section, we first introduce relevant background material in this area and then formulate the research problem.

2.1. Diagnosis-Frequency Vector and Sample Covariance Matrix Estimation

Given EHR data of m patients (both with and without the targeted disease), we can extract m diagnosis-frequency vectors $X_0, X_1 \ldots X_{m-1}$. Each vector (e.g., $X_i = <1, 0, \ldots, 3>$) 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:

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

With diagnosis-frequency vectors of a group of m patients (i.e., $X_0, \ldots X_{m-1}$), considering the sparsity of the vectors, the sample diagnosis-to-diagnosis covariance matrix is estimated Σ as follows:

$$\begin{aligned}\Sigma &= \frac{1}{m}\sum_{i=0}^{m-1} X_i X_i^T, \\ &= \frac{1}{m}\sum_{i=0}^{m-1}(\hat{X}_i + E_i)(\hat{X}_i + E_i)^T \\ &= \hat{\Sigma} + \frac{1}{m}\sum_{i=0}^{m-1}(2\hat{X}_i E_i^T + E_i E_i^T).\end{aligned} \tag{2}$$

As both $\hat{X}_i$ and E_i for all $0 \leq i \leq m-1$ are non-negative vectors (i.e., all elements are non-negative), we can conclude $\Sigma - \hat{\Sigma} = \frac{1}{m}\sum_{i=0}^{m-1}(2\hat{X}_i E_i^T + E_i E_i^T)$ is a non-negative matrix, and thus $\hat{\Sigma}$ should be a sparse estimation of Σ.

2.2. Minimax-Risk Covariance Matrix Estimator

Previous work [Cai and Zhou 2012; Xue et al. 2012] showed that it is possible to estimate covariance matrix "precisely" from a few samples using a *minimax-risk covariance matrix estimator* [Xue et al. 2012]. To achieve this goal, given the sample covariance estimation Σ, we estimate the minimax-risk estimation of the covariance matrix $\hat{\Sigma}$ as the ℓ_1-*norm-minimal approximation* of Σ:

$$\textbf{min.}\ |\hat{\Sigma}|_1\ \textbf{s.t.}\ ||\hat{\Sigma} - \Sigma||_F^2 \leq \epsilon,\ \textbf{and}\ \hat{\Sigma} \in I^+, \tag{3}$$

where I^+ refers to the overall set of positive semidefinite matrices—that is, $\forall \Sigma\ m \times m$ positive semidefinite matrix and $\forall X \in \mathbb{R}^m$, there exists $X^T \Sigma^{-1} X \geq 0$.

2.3. Fisher's Linear Discriminant Analysis Model for Disease Prediction

According to the common implementation of a Generalized Two-class Fisher's Discriminant Analysis (FDA) classifier [Ziegel 2003], given m training samples as well as the labels (i.e., $(X_0, l_0) \ldots (X_{m-1}, l_{m-1})$, where $l_i \in \{-1, +1\}$ such that $l_i = +1$, indicates patient i has been diagnosed with the target disease, $l_i = -1$ indicates otherwise), a two-class FDA model first sorts each sample into two groups according to the label and estimates covariance matrix/mean vector of the two classes, that is, (Σ_+, μ_+) and (Σ_-, μ_-), using the positive samples and negative samples, respectively. Then, generalized two-class FDA determines if a new patient (X') would develop the targeted disease, using

$$\begin{aligned}&(X' - \mu_-)^T \Sigma_-^{-1}(X' - \mu_-) + ln|\Sigma_-| - \\ &(X' - \mu_+)^T \Sigma_+^{-1}(X' - \mu_+) - ln|\Sigma_+| < T,\end{aligned} \tag{4}$$

where T is an optimal threshold based on the training samples. The model in Equation (4) becomes an LDA classifier and can achieve the Bayes optimal solution when $\Sigma_- = \Sigma_+$. When $\Sigma_- \neq \Sigma_+$, the classifier in Equation (4) is referred to as Quadratic Discriminant Analysis. In our study, we explicitly force $\Sigma_- = \Sigma_+$ through proportionally updating the covariance matrix, that is, $\Sigma_+, \Sigma_- \leftarrow \frac{m_+}{m}\Sigma_+ + \frac{m_-}{m}\Sigma_-$, where m_+ and m_- respectively refer to the quantities of positive and negative samples in the m samples and $m_- + m_+ = m$.

Given the two estimated matrices Σ_+ and Σ_- as well as the training samples, an LDA model works as follows:

(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 (4). In this case, *Daehr* uses an LDA model as $(\Sigma_+^*, \mu_+, \Sigma_-^*, \mu_-, T^*)$.
(2) **LDA-Based New Patient Prediction—**Given a new patient's EHR data, *Daehr* first converts his/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 (4).

2.4. Problem Formulation

According to the above definitions and preliminary work, this article 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

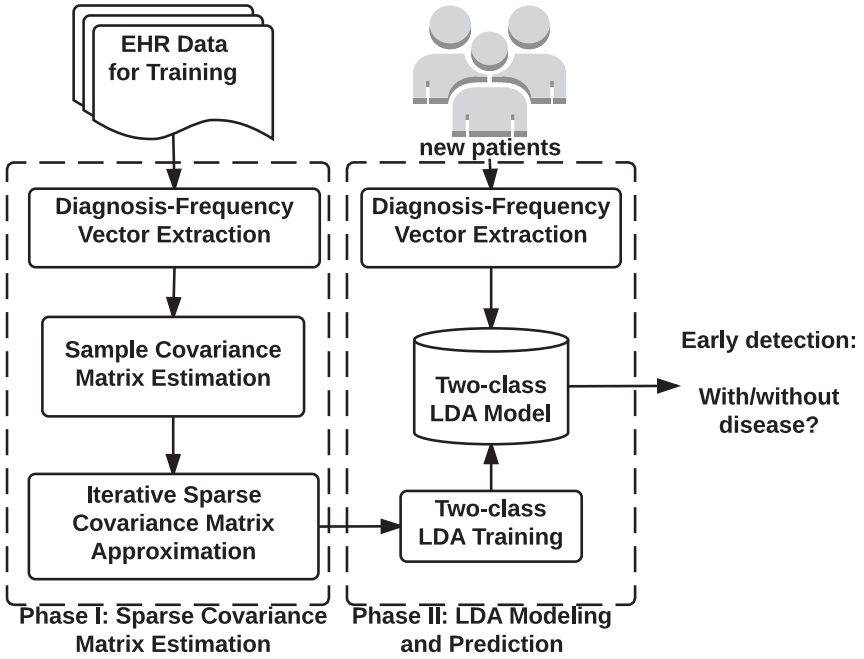


Fig. 2. *Daehr* framework.

detection of disease. Hereby, we define our research problem whereby, given m diagnosis-frequency vectors $X_0, X_1 \ldots X_{m-1}$, our problem is to find $\tilde{\Sigma}$,

$$\min. |\tilde{\Sigma}|_1 \text{ s.t. } ||\tilde{\Sigma} - \hat{\Sigma}||_F^2 \leq \epsilon \text{ and } \tilde{\Sigma} \in I^+. \tag{5}$$

Please note that $\hat{\Sigma}$ is not foreknown due to the unknown data noise. 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. *DAEHR* FRAMEWORK AND ALGORITHMS

In this section, we first introduce the design of the *Daehr* framework and then address several key algorithms used in *Daehr*. Then, we present the theoretical analysis of our algorithms.

3.1. *Daehr* Framework

In this section, we introduce the *Daehr* framework (Figure 2). *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 $(X_0, l_0) \ldots (X_{m-1}, l_{m-1})$, where $l_i \in \{-1, +1\}$ is the label of the ith 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 (2).

(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 terminates 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 that is, Σ_+ or Σ_-
Result: Σ_{t+1}—the positive definite ℓ_1-penalized estimation of Σ_0
begin
  while $\Delta' \geq tol,\ or\ 0 \leq t \leq maxit'$ **do**
    $\Sigma_{t+\frac{1}{2}} \leftarrow \ell_1$-penalized sparse estimation of Σ_t
    $\Sigma_{t+1} \leftarrow$ the nearest positive semidefinite approximation to $\Sigma_{t+\frac{1}{2}}$
  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, all based on the state-of-the-art of LDA introduced in Section 2.3.

3.2. 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:

$$\min \frac{1}{2} ||\Sigma_{t+\frac{1}{2}} - \Sigma_t||_F^2 + \lambda |\Sigma_{t+\frac{1}{2}}|_1, \tag{6}$$

where λ is a Lagrange multiplier [Wu 2009]. When $\lambda \geq 0$, Equation (6) 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.

3.3. 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 (7),

$$min.\ ||\Sigma_{t+1} - \Sigma_{t+\frac{1}{2}}||_F^2\ s.t.\ \Sigma_{t+1} \in I^+. \tag{7}$$

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 $\lambda_+ = \langle min\{\lambda_0, 0\}, min\{\lambda_1, 0\}\dots\rangle$, 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 on 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 3.4.

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 = 1/2(Σ_{t+1/2} + Σ^T_{t+1/2}), k = 1, I_{mod_0} = 0, Δ = 1;
    while Δ'' ≥ tol, or 0 ≤ k ≤ 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
    Σ_{t+1} = H_{k+1}
    return Σ_{t+1}
end
```

3.4. Algorithm Analysis

To understand theoretical properties of the *iterative sparse covariance matrix approximation* process, we first analyze the core algorithms used in the process, and 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 (6) is convex, the proximal gradient descent algorithm can approximate the optimal solution of Equation (6) 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 $k \to +\infty$, 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,

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

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 that 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 tth iteration) of the process, the algorithm first projects the matrix Σ_t to its ℓ_1-penalized sparse estimation $\Sigma_{t+\frac{1}{2}} \in G$, and 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] and Bregman [1967], when $t \to +\infty$, the *iterative sparse covariance matrix approximation* process converges (i.e., $||\Sigma_{t+1} - \Sigma_t|| \to 0$[1]).

Specifically, when $D \cap G \neq \emptyset$ and $k \to +\infty$, we can find $||\Sigma_{t+1} - \Sigma_t|| = ||\Sigma_t - \Sigma_{t+\frac{1}{2}}|| \to 0$ 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 High Dimension Low Sample Size (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.

4. EVALUATION

In this section, we introduce the experimental design for our evaluation. Next, 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 models. Finally, we compare the computational time of *Daehr* with other modeling methodologies.

4.1. Experimental Design

We first present the data used for our evaluation and then introduce the problem formulation for early detection of mental health disorders. We also specify the settings and context for early detection of mental health disorders.

[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.

Dataset for Evaluation—In this study, to evaluate *Daehr*, we use the de-identified EHR data from the 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 of the participating schools. The CHSN database includes ICD-9 diagnostic codes, Current Procedural Terminology (CPT) procedure codes, and limited demographic information. The selected 10 schools include over 200,00 enrolled students representing all geographic regions of the United States. The demography of enrolled students (sex, race/ethnicity, age, undergraduate/graduate status) closely matched the demography for the population of 108 Carnegie Research Universities/Very High classification.

Target Disease Early Detection—To evaluate our proposed approach, we select the most common mental health disorders in CHSN, *anxiety and depression disorders*, as the target disease for early detection. Specifically, we plan to evaluate *Daehr* using the early detection of mental health disorders in *college students*, considering the following motivation:

(1) *Increasing prevalence of mental health disorders*—Psychiatric disorders in the college student population have increased in frequency and severity in the United States with 18.6% of adults suffering from at least one active mental health disorder [nim 2015]. According to the Spring 2014 American College Health Association's National College Health Assessment report, almost 50% of college students have had the feeling of hopeless and overwhelming anxiety [American College Health Association 2014].
(2) *Difficulty of recognizing mental health disorders in primary care*—Mental health disorders are often unrecognized in primary care settings such as in student health centers [Wittchen et al. 2003]. This oversight leads to adverse outcomes and higher costs when patients with anxiety/depression cannot receive proper treatment in a timely manner. Thus, this work on the early detection of anxiety/depression could potentially aid student health centers in identifying high-risk patients in advance and referring them to behavioral health services.
(3) *Limitations of existing approaches for assessment and early detection of mental health disorders*—In primary care, physicians require sophisticated consulting skills to differentiate a wide range of symptoms [Tylee and Gandhi 2005]. However, symptoms of mental health disorders such as depression are dominated by somatic symptoms so physicians become preoccupied investigating possible underlying organic disease rather than considering a mental health diagnosis. Standard instruments, such as The Patient Health Questionnaire (PHQ)-9 [Kroenke and Spitzer 2002], also called psychological screening, are most commonly used to assess a patient's risk of suffering from mental health disorders. However, these methods are not generally applicable in primary care settings and the evaluation data are not widely accessible.

We are motivated to use EHR data for the early detection of mental health disorders, considering its wide accessibility and standardized use in primary care settings. Further, we are especially interested in identifying those high-risk students who have not yet received diagnosis or treatment for a mental health disorder. We thus design our experiments as follows.

Early Detection Settings—From the CHSN data, 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 their diagnosis-frequency vector based on the AHRQ CCS groups with patients. This minimizes potential noise caused by variable coding practices both within and between organizations.

If a patient has any of the ICD-9 diagnosis codes for anxiety or depression in their EHR, then they are a part of our target class. It is important to note that we are not

attempting to predict anxiety and depression disorders independently, as they frequently co-occur [Kendler et al. 2003]. Our model further assumes that, for each patient, at least two visits are required to provide enough patient history for early detection. Thus, patients with fewer than two visits from both the control and target groups were excluded from the analysis. Furthermore, for the target group, there must be at least two visits in the month prior to a patient's first diagnosis of anxiety/depression. Patients with mental health diagnosis other than anxiety and depression are removed from our analysis.

Notably, the visit data and corresponding diagnosis information from within 1 month[3] of the first diagnosis of anxiety/depression in the target group is excluded for the aim of early detection at least 1 month prior to diagnosis.

The diagnosis-frequency vectors used as predictors in our experiment only include primary care visit data (non-mental health visits); all mental health related visit information has been removed. In this case, our experiment is equivalent to predicting whether a patient would develop a mental health disorder from their primary care visit data.

4.2. 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 (4). This algorithm uses the pseudo-inverse [Ye et al. 2004] to replace the matrix inverse in Equation (4) 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 (5).
—**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 and 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 1,000 unselected patients (not included in the training set) from the target and control groups as the testing set; here, the testing set is also balanced.

[3]All experimental results presented in this article are based on a 1-month window. One month was chosen due to the seasonal patterns of student health center utilization (e.g., academic semesters, summer). In the appendix, we also provide the evaluation results of the early detection of mental health disorders based on both 2-month and 3-month windows.
[4]Java-Implementation of the Linear Discriminant Analysis, Institute for Psychological Diagnosis, http://www.psychometrica.de/lda.html.

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

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + TN + FP + FN}, \\ \text{F1-score} &= \frac{2 * TP}{2 * TP + FP + FN}, \end{aligned} \tag{8}$$

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 as having a mental health disorder. The F1-score is a weighted average of the precision and recall of the algorithm.

Figure 3 presents 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.

4.3. Comparison to Other Predictive Models

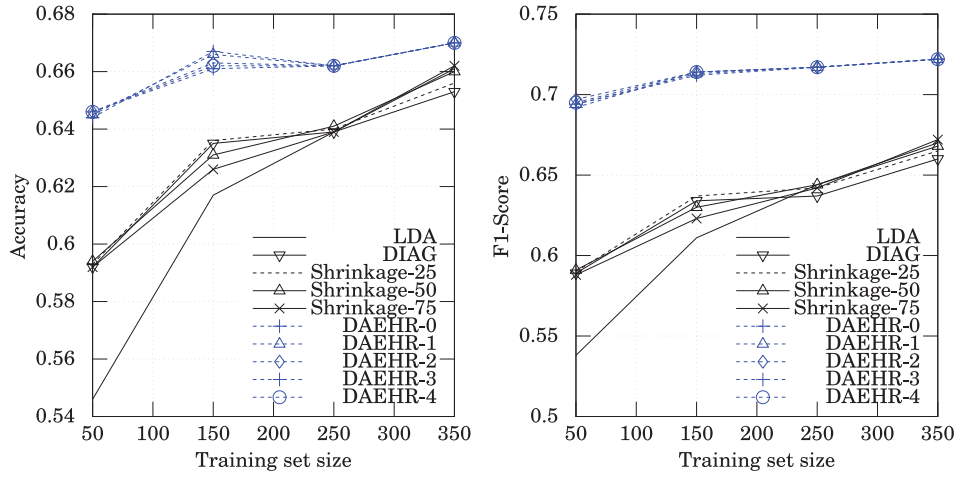
In order to understand the performance of *Daehr*, we compare it to other predictive models frequently used for early detection of diseases. Specifically, we consider to use following algorithms for the comparison:

—*SVM*—As in similar, studies [Sun et al. 2012; Ng et al. 2015a; Zhang et al. 2015], we build a linear binary SVM classifier with fine-tuned parameters.
—*Logistic Regression (Logit. Reg.)*—Additionally, we build a logistic regression classifier similar to recent work focused on prediction of both diagnosis and severity of depression from EHR data [Huang et al. 2014a].
—*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.

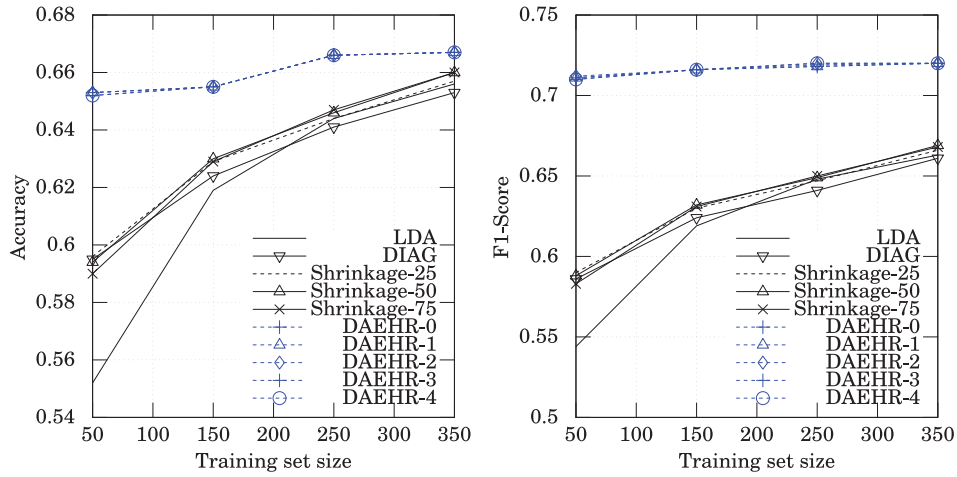
Combined with LDA and *Daehr* ($\lambda = 0.005 * 0.5^2$), we evaluate these six algorithms using the experiment settings introduced in Section 4.2. The comparison results are shown in Table I.[5]

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

[5]Please note that the results of LDA and *Daehr* in Table I slightly differ from those in Figure 3, since we conduct the two sets of experiments separately.



(a) Results with 2 × 200 Testing Samples



(b) Results with 2 × 1000 Testing Samples

Fig. 3. Performance comparison between *Daehr* and LDA baselines (Shrinkage-25 refers to Shrinkage Estimator with $\beta = 25\%$; Shrinkage-50 refers to Shrinkage Estimator with $\beta = 50\%$; Shrinkage-75 refers to Shrinkage Estimator with $\beta = 75\%$; DAEHR-0 refers to *Daehr* with $\lambda = 0.005 \times 0.5^0$; DAEHR-1 refers to *Daehr* with $\lambda = 0.005 \times 0.5^1$; DAEHR-2 refers to *Daehr* with $\lambda = 0.005 \times 0.5^2$; DAEHR-3 refers to *Daehr* with $\lambda = 0.005 \times 0.5^3$; and DAEHR-4 refers to *Daehr* with $\lambda = 0.005 \times 0.5^4$).

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 also that *Daehr* is a leading predictive model for early detection of mental health disorders.

4.4. Two Case Studies

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

Table I. 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±0.001	0.547±0.001	0.639±0.001	0.640±0.001
Logit. Reg.	0.614±0.004	0.471±0.056	0.615±0.003	0.468±0.035
SVM	0.614±0.001	0.607±0.001	0.660±0.001	0.669±0.001
AdaBoost-10	0.643±0.001	0.589±0.007	0.629±0.001	0.523±0.011
AdaBoost-50	0.633±0.002	0.548±0.019	0.633±0.001	0.538±0.008
Daehr	0.653±0.036	0.711±0.013	0.666±0.022	0.720±0.013

Table II. 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

Computational Time Analysis—We measure computational time of the six algorithms in the experiments introduced in Section 4. 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 II 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 longer to train—however, the average time to train a model with $250 \times 2 = 500$ samples is less than 12s, 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 six algorithms to classify patients is quite fast (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.

Covariance Matrix Estimation Analysis—We assume *Daehr* improves the LDA 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 23 U.S. universities); (2) randomly select 10,000 patients from them to estimate covariance matrix Σ_{+l}, (3) randomly select another 50 or 250 samples to train LDA and *Daehr*; and (4) further compare Σ_{+l} to the covariance matrices estimated in LDA and *Daehr* separately through measuring the error of matrices. We repeat steps (1) through (4) for a total of 30 trials to obtain the average error between the covariance matrices. We similarly compare the matrices estimated using negative samples (i.e., patients without mental disorders). Table III presents the average error between covariance matrices in ℓ^1/Frobenius-norm. The results show that, compared to LDA, the covariance matrix estimated in *Daehr* using small samples is *more closed* to the covariance matrix estimated using large samples. In this case, we conclude that *Daehr* can accurately estimate the covariance

Table III. Covariance Matrix Comparison—Training Set Size: 50 × 2

Algorithm	$\lvert\Sigma_+ - \Sigma_{+l}\rvert_1$	$\lvert\lvert\Sigma_+ - \Sigma_{+l}\rvert\rvert_F^2$	$\lvert\Sigma_- - \Sigma_{-l}\rvert_1$	$\lvert\lvert\Sigma_- - \Sigma_{-l}\rvert\rvert_F^2$
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

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 extended evaluation results and experimental insights.

5. RELATED WORK AND DISCUSSION

In this section, we first summarize previous studies related to this article from two aspects: *data-mining approaches to early detection of diseases* and *extensions to LDA learning*. Then we compare our work to the most relevant work. Further, we discuss several open issues of our study.

5.1. Data-Driven Approaches to Early Detection of Disease

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. 2014b]. In this section, we will discuss previous work in two areas: *predictive modeling* and *data representation* approaches.

5.1.1. 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 earlier, 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 studies on the early detection of mental disorders, a Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression model has been applied to predict the depression severity to help personalize treatment for high-risk patients [Huang et al. 2014b].

5.1.2. 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. 2015b; Huang et al. 2014b] 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 [Zhang et al. 2015; Wang et al. 2012a; Liu et al. 2015; Gotz et al. 2014; Perer and Wang 2014; Perer et al. 2015]. Specifically, Zhang et al. [2015] intend 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], and Perer et al. [2015] consider the frequencies of transitions crossing multiple diagnoses using a hyper-graph.

5.2. Extensions to LDA Model

We introduce several LDA extensions in HDLSS settings such as those challenges presented by EHR data. As discussed above, when LDA works in HDLSS, there 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 expected loss [Berger 2013] is inherited from the variance of small samples, through classical LDA training. To handle the singular (non-invertible) covariance matrix issue, 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 risk via regularizing the estimated covariance matrices.

5.3. Comparing *Daehr* to Existing Work

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], and Yoo et al. [2011]), *Daehr* is the first work that focuses on improving the performance of the LDA model by addressing noisy data and small target training sample size.

Second, our contribution is complementary with the work in EHR data representation [Wang et al. 2012a, 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], and 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 mathematically differ from ours with different objectives and assumptions.

5.4. Discussion and Open Issues

5.4.1. Electronic Health Record (EHR) Data. In this article, we use EHR data as predictors for the early detection of mental health disorders. Specifically, we leveraged the diagnosis records. There exists a limited body of related work using both diagnosis and treatment records in EHR data [Wang et al. 2012a; Liu et al. 2015; Gotz et al. 2014]. We plan to integrate procedures in *Daehr* to capture patient treatment patterns in future work and hypothesize that these additional data could improve model performance.

In this article, rather than using the raw ICD-9 codes, we also used higher-level groups of codes from AHRQ CCS [HCUP 2014] to represent the diagnosis records. The grouping scheme maps 15,000 ICD-9 codes to 295 groups, where each ICD-9 code corresponds to a single group. Further, we assume to use diagnosis-frequency vectors to

represent the diagnosis records in EHR data. Compared to using raw ICD-9 codes, using the AHRQ CCS groups reduces the dimensionality and noise of diagnosis-frequency vectors. However, such dimensionality reduction can also lead to information loss and is not necessarily optimal. In our future work, we will study other data representations with *Daehr*.

While there are many other sources of medical data such as personal health data, surveys, claims data, and disease registries to names a few, EHR data are the most standardized and accessible to study patient populations on a large scale. While this data creates great opportunity for predictive modeling, the sparsity, noisiness, heterogeneity, and bias [Goldstein et al. 2016] leave many challenges and open areas for investigation.

5.4.2. Dimensionality Reduction and Other Approaches to HDLSS Problems. In this article, we demonstrated that the performance of traditional LDA might be bottlenecked due to 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.

5.4.3. Performance Metrics for Early Detection of Diseases. In this article, we evaluate *Daehr* against other baselines using accuracy and F1-score. We also report sensitivity and specificity of the methods in the Appendix. On average, compared to classic LDA, *Daehr* achieves 15%–25% higher sensitivity while scarifying no more than 8%–18% specificity. In the context of early detection of mental health disorders in a student health setting, this performance is quite good for two reasons. First, identifying patients with the disorder and referring them to early treatment is the most important consideration. Second, the workload required to implement such a model in clinical practice may increase additional screening required for false positive patients. However, in the long term, these models have the potential to lead to more rapid referral and intervention of students with mental health disorders to behavioral health services.

6. CONCLUSIONS

In this article, 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 risk of the parameter estimation for 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 a quasi-optimal solution in terms of LDA-based early disease detection. The experimental results using a 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 the Appendix.

REFERENCES

2012. CMS: Electronic Health Records. Retrieved from https://www.cms.gov/Medicare/E-health/EHealth Records/index.html.

2015. Any Mental Illness (AMI) Among Adults. NIH National Institute of Mental Health. Retrieved from http://www.nimh.nih.gov/.

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. *Med. 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.* Retrieved from http://www.ijme.net/archive/2/communication-training-and-perceived-patient-similarity/.

James O. Berger. 2013. *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.

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 Comput. Math. Math. Phys.* 7, 3 (1967), 200–217.

T. Tony Cai and Harrison H. Zhou. 2012. Minimax estimation of large covariance matrices under l1 norm. *Stat. Sin.* 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, 137–144.

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

Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll. 2011. Sparse discriminant analysis. *Technometrics* 53, 4 (2011), 406–413.

Ralph B. D'Agostino, Scott Grundy, Lisa M. Sullivan, and Peter Wilson. 2001. Validation of the Framingham coronary heart disease prediction scores: Results of a multiple ethnic groups investigation. *J. Am. Med. Am.* 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. *Emerg. Infect. Dis.* 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. *Ann. Eugen.* 7, 2 (1936), 179–188.

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

E. Gil-Herrera, G. Aden-Buie, A. Yalcin, A. Tsalatsanis, L. E. Barnes, and B. Djulbegovic. 2015. Rough set theory based prognostic classification models for hospice referral. *BMC Med. Inform. Dec. Making* 15, 1 (2015), 98.

Benjamin A. Goldstein, Ann Marie Navar, Michael J. Pencina, and John P. A. Ioannidis. 2016. Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *J. Am. Med. Inform. Assoc.* (2016). DOI:http://dx.doi.org/10.1093/jamia/ocw042

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. *J. Biomed. Inform.* 48 (April 2014), 148–159. DOI:http://dx.doi.org/10.1016/j.jbi.2014.01.007

HCUP. 2014. Appendix A - Clinical Classification Software-DIAGNOSES (January 1980 through September 2014). Retrieved from https://www.hcup-us.ahrq.gov/toolssoftware/ccs/AppendixASingleDX.txt.

Nicholas J. Higham. 2002. Computing the nearest correlation matrixa problem from finance. *IMA J. Numer. Anal.* 22, 3 (2002), 329–343.

Pao-Lu Hsu and Herbert Robbins. 1947. Complete convergence and the law of large numbers. *Proc. Natl. Acad. Sci. U.S.A.* 33, 2 (1947), 25.

Rui Huang, Qingshan Liu, Hanqing Lu, and Songde Ma. 2002. Solving the small sample size problem of LDA. In *Proceedings of the 16th International Conference on Pattern Recognition, 2002*, 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. *J. Am. Med. Inform. Assoc.* 21, 6 (2014), 1069–1075.

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. *J. Am. Med. Inform. Assoc.* 21, 6 (Dec. 2014), 1069–1075. DOI:http://dx.doi.org/10.1136/amiajnl-2014-002733

Peter B. Jensen, Lars J. Jensen, and Søren Brunak. 2012. Mining electronic health records: Towards better research applications and clinical care. *Nat. Rev. Genet.* 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*.

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. *Eur. J. Biomed. Inform.* 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 *Proceedings of the 2014 IEEE International Conference on Healthcare Informatics (ICHI)*. 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. *Arch. Gen. Psychiat.* 60, 8 (2003), 789–796.

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

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, 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. *IEEE Trans. Neur. Netw.* 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 Res. 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.

S. Mitchell, K. Schinkel, Y. Song, Y. Wang, J. Ainsworth, T. Halbert, S. Strong, J. Zhang, C. C. Moore, and L. E. Barnes. 2016. Optimization of sepsis risk assessment for ward patients. In *Proceedings of the IEEE Systems and Information Engineering Design Symposium (SIEDS)*. 107–112. DOI:http://dx.doi.org/10.1109/SIEDS.2016.7489280

Yurii Nesterov. 2004. *Introductory Lectures on Convex Optimization*. Vol. 87. Springer Science & Business Media.

Kenney Ng, Jimeng Sun, Jianying Hu, and Fei Wang. 2015a. Personalized predictive modeling and risk factor identification using patient similarity. *AMIA Summit on Clinical Research Informatics (CRI)* (2015), 132.

Kenney Ng, Jimeng Sun, Jianying Hu, and Fei Wang. 2015b. Personalized predictive modeling and risk factor identification using patient similarity. *AMIA Summits on Translational Science Proceedings* 2015 (March 2015), 132–136.

Alicia Nobles, Ketki Vilankar, Hao Wu, and Laura Barnes. 2015. Evaluation of data quality of multisite electronic health record data for secondary analysis. In *Proceedings of the 2015 International Conference on Big Data (Workshop)*. IEEE.

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.

Adam Perer, Fei Wang, and Jianying Hu. 2015. Mining and exploring care pathways from electronic medical records with visual analytics. *J. Biomed. Inform.* 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. *Proc. Natl Acad. Sci. U.S.A.* 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. *Ann. Stat.* 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. *Br. Med. J.* 344 (2012).

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

Barbara G. Tabachnick, Linda S. Fidell, and others. 2001. Using multivariate statistics. (2001). 530–538.

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. *J. Am. College Health: J. ACH* (June 2015). DOI:http://dx.doi.org/10.1080/07448481.2015.1055567

A. Tylee and P. Gandhi. 2005. The importance of somatic symptoms in depression in primary care. *Prim. Care Compan. J. Clin. Psychiatr.* 7, 4 (2005), 167–176.

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.

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. 272–285.

Fei Wang and Jimeng Sun. 2015. PSF: A unified patient similarity evaluation framework through metric learning with weak supervision. *IEEE J. Biomed. Health Inform.* 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.

H.-U. Wittchen, S. Mhlig, and Beesdo K. 2003. Mental disorders in primary care. *Dialog. Clin. Neurosci.* 5, 2 (2003), 115–128.

Hsien-Chung Wu. 2009. The Karush–Kuhn–Tucker optimality conditions in multiobjective programming problems with interval-valued objective functions. *Eur. J. Operat. Res.* 196, 1 (2009), 49–60.

Lingzhou Xue, Shiqian Ma, and Hui Zou. 2012. Positive-definite 1-penalized estimation of large covariance matrices. *J. Am. 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. *IEEE Trans. Pattern Anal. Mach. Intell.* 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. *J. Med. Syst.* 36, 4 (May 2011), 2431–2448. DOI:http://dx.doi.org/10.1007/s10916-011-9710-5

Jinghe Zhang, Haoyi Xiong, Yu Huang, Hao Wu, Kevin Leach, and Laura E. Barnes. 2015. MSEQ: Early detection of anxiety and depression via temporal orders of diagnoses in electronic health data. In *2015 International Conference on Big Data (Workshop)*. IEEE.

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 Syst. Appl.* 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.

Received December 2015; revised July 2016; accepted October 2016