© Copyright 2017

Ying Lin

Large-Scale Personalized Health Surveillance by Collaborative Learning and Selective Sensing

Ying Lin

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Shuai Huang, Chair

Shan Liu, Co-Chair

W. Art Chaovalitwongse

Xiaohua (Andrew) Zhou

Program Authorized to Offer Degree:

Industrial & Systems Engineering

University of Washington

Abstract

Large-Scale Personalized Health Surveillance by Collaborative Learning and Selective Sensing

Ying Lin

Chair of the Supervisory Committee:

Assistant Professor Shuai Huang

Assistant Professor Shan Liu

Industrial & Systems Engineering

Recent advance in sensing and information technology provides abundance of risk predictive data, leading to the development of personalized health surveillance. However, effective use of the sensing technology is prohibited by the complexity of disease progression, heterogeneity in a large population, and the lack of cost-effective monitoring strategy. To scale up personalized health surveillance for a large population, we developed innovative methodologies for individual prognostic and personalized monitoring strategy design in this thesis. We proposed a statistical framework, collaborative learning, for characterizing the individual disease progression trajectories from sparse and irregular data. We then developed a decision support algorithm, selective sensing, to adaptively allocate limited monitoring resources to high-risk individuals. We further proposed a rule-based method and a prognostic-based monitoring framework to translate the sensing data into risk-predictive patterns for individual prognostic and identify the cost-effective monitoring strategies for disease prevention. We applied the proposed methods to real-world applications, including cognitive monitoring in Alzheimer's Disease (AD), and follow-up monitoring in depression treatment.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deepest and sincerest gratitude to the people who made my past three years at the University of Washington the most memorable and intellectual stimulating time in my life. Without the generous support, endless patience and encouragement and insightful guidance from my advisors, mentors, friends and families, this thesis would not have been possible.

First of all, my deepest gratitude goes to my advisors Dr. Shuai Huang and Dr. Shan Liu. Dr. Huang brought me to the fantastic world of applying data analytics techniques to solve quality engineering problems. His rigorous statistical thinking, incisive opinions and thoughtful insights guided me to see the beauty and power of statistics and the opportunities in the interdisciplinary research. Dr. Liu further opened the door of medical decision making for me. Her rich experience and in-depth domain knowledge helps me underin decision analysis. The experiences of working with them shaped my interest of the multidisciplinary research area lying at the interaction of data analytics and medical decision making. Their insightful guidance helped me to solve the challenging problems in my thesis. Solving research problems was like climbing mountains, in which rich skills, sincere attitudes and thoughtful insights were necessaries for successfully reaching the top. These skills and experiences learned from my advisors lay a solid foundation to this thesis and will be continuously invaluable for my future career. I also appreciate the endless patience and encouragement from my advisors to support me and their high expectations to push me to where I'm today.

In addition, I would like to thank my other committee members, professor W. Art Chaovalitwongse and professor Xiaohua (Andrew) Zhou. They provided thoughtful suggestions and insightful comments that greatly enhanced the quality of my thesis and inspire me to extend

my work for solving new problems in the future. I would also like to thank our collaborator, Dr. Gregory E. Simon, in the Group Health Research Institute the wonderful opportunity to work on the depression treatment population and learn how to collaborate with the health professionals and psychiatrists. My gratitude also goes to Dr. Linda Ng Boyle, Dr. Archis Ghate and Dr. Youngjun Choe for their valuable suggestions and comments for my future career.

Thirdly, I would like to thank all the people, such as Zach Chen, Boxiang Dong, Yan Jin, Vivi Shang, who have helped me during my Ph.D study.

Last but not least, my gratitude goes to my parents. I really appreciate their endless support and understanding in the past years. Their unconditional love is like sunshine brightening my drizzle days.

Chapter 1. INTRODUCTION

1.1 MOTIVATION

The rapid advances in sensing and information technology have provided unprecedented measurement capabilities and an abundance of risk predictive data, leading to breakthroughs in many disease research with better risk screening and monitoring procedures being developed (Bu et al.,2007; Ortiz et al., 2003). Despite the risk prediction capability currently available to researchers and clinical practitioners, it is still a challenge to make individual-level prediction and design the patient-specific monitoring strategy (Swan, 2012; Srikanta et al., 1984). One challenge is that existing risk prediction methods more focus on revealing the associations between risk factors and disease progression on population-level (Sosenko et al., 2008), which assumes the homogeneity between individuals, while personalized risk prediction methods are inadequately developed due to lack of understanding of the disease etiology or lack of sensitive and specific biomarker measurement technologies. Another challenge is that, although the existing risk prediction model could provide stratification capability, how to use the risk stratification for patient-specific monitoring strategy design is inadequately studied. This thesis concentrates on developing novel statistical methods for mitigating these challenges. In the following, we illustrate the objectives and organization of this thesis.

1.2 RESEARCH OBJECTIVES

The objectives of this research are:

1) Develop novel statistical methods and their associated computational algorithms for translating the sensing data into the understanding of disease progression and personalized

prognostics by drawing on recent theoretical developments in statistics and machine learning.

2) Develop decision support algorithms for patient-specific monitoring strategy design, and optimally allocate the limited sensing resources to monitor a heterogeneous population.
3) Compare the patient-specific monitoring strategies with existing one-size-fits-all monitoring guideline in the Cost-Effectiveness frontier to inform better decision making in clinical practice.

By achieving these goals, we will help build a solid foundation for the future development of personalized health surveillance.

1.3 ORGANIZATION OF THESIS

The following outlines the organization of this thesis: the first part of this thesis focuses on the trajectory based disease prediction and management. We develop a collaborative modeling framework to effectively characterize a population of heterogeneous trajectories by exploiting the latent structure and similarity between individuals in Chapter 2. We further develop a selective sensing framework that integrates the Markov models, collaborative modeling, and sensing resource allocation to efficiently and economically monitor a large number of individuals by exploiting the similarities between them in Chapter 3. The second part of this thesis presents a rule-based approach for individual risk prediction and monitoring strategy design. We first propose a Rule-based prognostic model in Chapter 4 to characterize heterogeneous disease progression as a set of patterns of risk factors (rules) and dissect out individual's disease risk using these patterns. We further develop a prognostic-based monitoring framework in Chapter 5 to compare different

prognostic models and monitoring strategies by integrating the individual prognostic, monitoring strategy design, and cost-effectiveness analysis into a computational framework.

Chapter 2. COLLABORATIVE LEARNING FRAMEWORK FOR INDIVIDUALIZED MONITORING

2.1 INTRODUCTION

This chapter concerns the problem of estimating many individualized regression models in a heterogeneous population where each individual has a distinct regression model. Regression models have been popular tools in a wide range of reliability and prognostic tasks. A classic example is to model the relationship between demographical and/or intellectual variables with students' academic achievements in a school district where a number of regression models may be needed since each school may require a different model. The heterogeneous population is also observed in many healthcare applications. It has been widely known that significant heterogeneities exist among patients, calling for individualized prognostic models. Examples include the prognostic modeling for Alzheimer's disease (AD) to predict the cognitive status over time of each at-risk individual for early detection of AD and trajectory modeling for depression patients to monitor the depression progression. These individualized prognostic models are critical for the development of adaptive monitoring strategy, treatment therapy design and healthcare resources management.

One approach to model heterogeneous individuals is to estimate the regression model separately for each individual. This approach, however, could be less effective because the information from others is not exploited. Moreover, in many healthcare studies, it is difficult to collect an abundance of longitudinal data that can cover the whole spectrum of disease progression for each individual

(Hedeker and Gibbons 1997; Little 1995). Consequently, measurements from an individual may not be sufficient to build an accurate prognostic model. For example, the cognitive data of each individual is oftentimes sparse in the AD study and the depression symptoms collected from sensing and information technology of each individual are irregular. Considering the significant complexity of the progression trajectory of the disease status reported in the literature (Reynolds 2002; Mcgue and Christensen 2002), the individual modeling approach will result in significant bias when predicting in the area with sparse or no training data. Another common approach is to merge all the individual data together and estimate a group-level prediction model (Zhou et al. 2013). However, this approach will only characterize the average effect, failing to capture the personalized variations.

A more advanced treatment to mitigate the limitations of these two approaches is to use the mixed effects model (MEM) (Galecki and Burzykowski 2013), also known as hierarchical models (Bryk and Raudenbush 1987) and multilevel models (Goldstein 1987). The MEM assumes that the regression parameters of these regression models are sampled from a distribution which is usually a multivariate normal distribution, where the mean vector characterizes the average tendency of the regression parameters and the covariance matrix captures the dispersion of the parameters. The MEM has been widely used to address the individual-to-individual (or unit-to-unit) variations in various applications (Hedeker and Gibbons 1997; Rasmusson et al. 1996; Baayen et al. 2008; Penny et al. 2003).

The MEM approach could be effective when there is a central tendency and the heterogeneous patterns of individuals (i.e., the deviations from the central tendency) are randomly distributed. It, however, will be less effective where a mixture of central tendency exists that presents a complex heterogeneous structure. For instance, it has been discovered in the AD research community that

roughly, there are three latent phenotypes in the AD population (Hertzog et al. 2003; Jack et al. 2010; Royall et al. 2005; Petersen et al. 1999): the diseased group, the normal aging group (NC), and the mild cognitive impairment group (MCI). Previous studies demonstrated that these three groups have exhibited distinct progression patterns on cognitive deterioration (Petersen et al. 1999), while the individuals in the same group experience relatively similar disease progression. Recent research findings have further revealed that there are more subgroups in the MCI group (Jack et al. 2010; Albert et al. 2011; Lopez et al. 2003), suggesting a more complicated phenotype structure within the heterogeneous population.

In a heterogeneous population, although the regression model for each individual should differ from each other, the models can be characterized by a low-dimensional structure, e.g., considering the few types of degradation mechanisms of the phenotypes in a population. Individuals' models may be variants of these typical degradation mechanisms. To address the population heterogeneity and data challenges such as sparse measurements, this study proposes a collaborative learning framework that provides a generic methodology for estimating regression models for heterogeneous individuals. Specifically, we exploit the idea of canonical models and model regularization. While the framework is generic, we illustrate its utility using two real-world case studies on AD and depression in subsequent discussions.

The basic idea of the proposed collaborative learning framework is to use a set of canonical models to represent the heterogeneous population characteristics. For example, in the AD application, there are K representative disease progression mechanisms that form a heterogeneous population. The proposed collaborative model (CM) assumes that there are K canonical models, in which each progression mechanism can be represented by a canonical model. In practice, it is usually unknown a prior which progression mechanism an individual may follow, and some

individuals could follow progression pattern between multiple canonical models. Thus, mathematically, these canonical models span the modeling space for the individuals and provide a basis to characterize the individuals' variations on their own progression mechanisms. This can be done by creating a membership vector (with K elements) for each individual, denoted as $\boldsymbol{c}_i$ for individual i, where c_{ij} represents the degree to which the model of individual i resembles the canonical model j. We assume that, although each individual has a distinct model, the model can be approximated by a weighted combination of the K canonical models.

Utilizing the flexible modeling structure in the proposed collaborative learning framework, we further extend the CM formulation to incorporate the similarity information between individuals to enhance the estimation of the models. The similarity between two individuals can be quantified by comparing their profiles on covariates such as the demographic, social-economical, clinical and/or genetic and imaging information, which are commonly available in many healthcare applications. This extension is named similarity-regularized CM (SCM). In summary, by using the canonical models to characterize the heterogeneity structure in a heterogeneous population, and further incorporating the similarity information to strengthen the learning of the models, the proposed collaborative learning framework can effectively integrate the sparse data of multiple individuals.

While the collaborative learning framework provides intuitive interpretations, it presents a challenging constrained optimization problem. Inspired by non-negative matrix factorization methods (Ding et al. 2006; Cai et al. 2011), we develop an efficient computational algorithm to solve the optimization problem and prove that the proposed algorithm can guarantee convergence to a stationary solution. Moreover, we conduct a theoretical analysis that makes the connection

between the proposed models with MEM and provide theoretical justifications on why the proposed methods could be superior when the low-dimensional heterogeneity structure exists.

This chapter is organized as follows. Section 2.2 review the related methods. Section 2.3 provides the details of the proposed collaborative learning framework. Section 2.4 discusses the relationship between proposed collaborative framework with the mixed effects model. Section 2.5 focuses on deriving the computational algorithm that solves the optimization problems. Section 2.6 conducts comprehensive simulation studies to demonstrate the efficacy and superiority of the proposed method over alternative methods. The proposed method is applied to model cognitive degradations of Alzheimer's Disease (AD) patients and depression trajectories using two real-world datasets on AD and depression respectively. The results of these real case studies are presented in Section 2.7 and 2.8. Section 2.9 draws the conclusion and discussion of this chapter.

2.2 RELATED WORK

As we have described in Section 2.1, the proposed collaborative learning framework is different from the existing methods such as MEM models (Galecki and Burzykowski 2013; Bryk and Raudenbush 1987; Goldstein 1987). It is also different from the finite mixture regression models that have been developed in (McLachlan and Peel, 2004; Leisch, 2004). The finite mixture regression models are also closely related to mixture regression model (Hoshikawa, 2013), latent class regression model (Vermunt, 2002), and clusterwise linear regression (DeSarbo, 1988). The finite mixture regression models assume that there are latent clusters, which is a different assumption from our proposed canonical structure. Our proposed collaborative learning approach is fundamentally different from these mixture regression models since our goal is to learn a regression model for each individual, rather than assigning individuals into clusters. Specifically, in those mixture regression models, a similar parameter as c_{ij} was also defined. However, in

mixture regression models, it is assumed that the underlying model has each individual following one-and-only-one of the K canonical models, so the c_{ij} represents the probability that individual i follows the model j. However, here, a fundamental difference is that we assume that each individual i truly follows a linear combination of the K canonical models rather than one-and-only-one of the K canonical models. The model formulation is quite different, while the associated computational challenges are also different. The proposed method is also different from the reduced-rank regression models (Anderson, 1995; Izenman, 1975; Reinsel and Velu, 1998) that have been proposed for restricting the rank of the regression coefficient matrix of a regression model where multivariate outcomes are concerned simultaneously, and the dynamic weighted ensemble models (Shen and Kong 2004; Liu et al 2015) which are developed to enhance the conventional framework of ensemble learning where a set of models are combined.

While many of the abovementioned methods could be shown as equivalent to the MEM model (but differ in the forms of the basis used in these regression models), the proposed collaborative learning method target a different type of applications. The central assumption is that, in a heterogeneous population, although the regression model for each individual should differ from each other, the models can be characterized by a low-dimensional structure, e.g., considering the few types of degradation mechanisms of the phenotypes in a population. Individuals' models may be variants of these typical degradation mechanisms. Another unique merit of the proposed method is that it can address the data challenges such as sparse and irregular measurements, and explicitly utilize similarity information for modeling.

2.3 COLLABORATIVE MODELING (CM)

We use the AD application as the context to illustrate the development of the collaborative learning framework. Specifically, we focus on the prognostics of the AD disease progression process which calls for accurate degradation models of the cognitive health of the individuals. There has been a range of technical approaches to acquire measurements that reflect cognitive health of individuals including several neuroimaging markers, clinical biomarkers, and neuropsychological instruments. The degradation models can be used to predict the cognitive status over time using variables such as age, genetic factors, risk factors, and/or other biological factors (Jack et al. 2010; Regan et al. 2006; Rasmusson et al. 1996; Suh et al. 2004; Zhou et al. 2013; Sliwinski et al. 2003; Wilkosz et al. 2010). These cognitive degradation models are critical for operationalizing the idea of cognitive monitoring and screening in primary care or community setting. For instance, accurate prediction of cognitive status of healthy subjects may enable timely detection of the subjects that are about to experience incipient cognitive decline. Also, accurate prediction model can help health care providers to prioritize screening efforts on the high-risk individuals.

Mathematically, the degradation models aim to capture the predictive relationships between p predictors and the cognitive status. For each subject i, $i = 1, \cdots, N$, we assume that there are longitudinal measurements at n_i time points that include the longitudinal cognitive status measurements, denoted as $\boldsymbol{y}_i = \left[y_{i1}, \dots, y_{in_i}\right]^{\mathrm{T}} \in \mathbb{R}^{n_i \times 1}$, and the longitudinal measurements of p predictors, denoted as $\mathbf{X}_i = \left[\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{in_i}\right]^{\mathrm{T}} \in \mathbb{R}^{n_i \times p}$.

We demonstrate the collaborative learning method using linear models in this study, i.e., we assume that the cognitive degradation model of individual i, denoted as $f_i(\boldsymbol{x}), i = 1, \cdots, N$, is a linear model on $\boldsymbol{x}$ as $f_i(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}_i$, where $\boldsymbol{\beta}_i$ is the coefficient vector. Linear models have been

found successful in characterizing the dynamic progression of cognitive status (Jack et al. 2010; Regan et al. 2006; Rasmusson et al. 1996; Zhou et al. 2013). However, the proposed approach can incorporate nonlinear models. Many nonlinear models such as Gaussian processes or kernel models can be represented as linear models using nonlinear basis functions or kernel tricks that map the original variables $\boldsymbol{x}(t)$ into the reproducing kernel Hilbert space defined by a certain kernel function (Smola et al. 2007; Scholkopf and Smola 2002). Here, we use the Gaussian process as an example for explanation.

Assuming the degradation model $f_i(\boldsymbol{x})$ takes the form as a Gaussian process which can be defined as (Rasmussen 2006; Doksum and Normand 1995):

$$f_i(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}_i + g(\boldsymbol{x}), \text{ where } g(\boldsymbol{x}) \sim GP(\mathbf{0}, k(\boldsymbol{x}, \boldsymbol{x}')).$$

It consists of a mean structure, represented by $\boldsymbol{x}\boldsymbol{\beta}_i$, and a Gaussian process structure, $g(\boldsymbol{x})$**,** with mean as zero and covariance function, $k(\boldsymbol{x}, \boldsymbol{x}')$. It can be further written as:

$$f_i(\boldsymbol{x}) \sim GP(\boldsymbol{x}\boldsymbol{\beta}_i, k(\boldsymbol{x}, \boldsymbol{x}')).$$

Given the observational data that includes a vector of measurements, $\boldsymbol{y}_i$, and the predictors, $\mathbf{X}_i$, the conditional distribution of y^* conditioning on $\boldsymbol{x}^*$ is also Gaussian:

$$E(y^*|\boldsymbol{x}^*) = \boldsymbol{x}^*\boldsymbol{\beta}_i + \sum_{j=1}^{n_i} \alpha_{ij} k\left(\boldsymbol{x}^*, \boldsymbol{x}_{ij}\right),$$

where $\boldsymbol{x}^*$ represents the new observations of predictors and $\boldsymbol{\alpha}_i = [\alpha_{i1}, \dots, \alpha_{in_i}]^T = \left(K_{\mathbf{y}_i,\mathbf{y}_i} + \sigma_e \mathbf{I}\right)^{-1}(\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i)$. $K_{\mathbf{y}_i,\mathbf{y}_i} = \left[k\left(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'}\right)\right]_{n_i \times n_i}$ is the covariance matrix of observational data on the predictors. Using a set of covariance functions, $\{k\left(\boldsymbol{x}^*, \boldsymbol{x}_{ij}\right), \mathrm{j} = 1, \dots, n_i\}$, the mean prediction for y^*can be expressed as the linear regression form. Therefore, a linear model provides a flexible framework for encompassing a wide range of models and can be easily extended to capture the nonlinear patterns.

Therefore, a linear model provides a flexible framework for encompassing a range of models and can be easily extended to capture the nonlinear dynamics as demonstrated in (Ashford and Schmitt 2001; Ito et al. 2010).

In subsequent sections, we propose the CM and SCM to effectively learn $f_i(\boldsymbol{x})$, $i = 1, \cdots, N$. We include the detailed proofs of the theorems in the Appendix A.

2.3.1 *Model formulation*

Let $g_k(\boldsymbol{x})$, $k = 1, \cdots, K$, be the degradation model of the k^{th} canonical model such that $g_k(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{q}_k$, where $\boldsymbol{q}_k$ is the corresponding regression parameter vector. We assign a membership vector $\boldsymbol{c}_i = [c_{i1}, \dots, c_{iK}]^{\mathrm{T}}$to each subject i, where c_{ik} represents the degree to which the subject i resembles the canonical model k. To derive prediction for subject i, we use the weighted combination of the K canonical degradation models as $f_i(\boldsymbol{x}) = \sum_k c_{ik} g_k(\boldsymbol{x})$. By assumption, the more similar between the model of subject i and canonical model k (i.e., larger c_{ik}), the more importance the degradation model $g_k(\boldsymbol{x})$ will have in determining $f_i(\boldsymbol{x})$. With linear models, the coefficient vector $\boldsymbol{\beta}_i$ of $f_i(\boldsymbol{x})$ can be represented as a linear combination of the coefficient vectors of the K canonical models. That is, when $f_i(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}_i$, $g_k(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{q}_k$, then it is easy to derive that $\boldsymbol{\beta}_i = \sum_k c_{ik}\boldsymbol{q}_k = \mathbf{Q}\boldsymbol{c}_i$ while $\mathbf{Q} = [\boldsymbol{q}_1, \dots, \boldsymbol{q}_K]$.

Based on the canonical models, we use the least square loss function to measure the goodness-of-fit of the individuals' models and consider the constraint that the membership vector is normalized and has nonnegative elements, leading to the following optimization formulation of CM:

$$\min_{\boldsymbol{c}_i, \mathbf{Q}} \sum_i \|\boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i\|^2, \tag{2.1}$$

$$\text{subject to } c_{ik} \geq 0, \quad \sum_k c_{ik} = 1, \quad \mathbf{X}_i\mathbf{Q} \geq \mathbf{0}, \forall\, i = 1, \dots, N \text{ and } k = 1, \dots, K.$$

Here, the last inequality $\mathbf{X}_i\mathbf{Q} \geq \mathbf{0}$ is imposed due to the constraint that the predicted cognitive status should stay nonnegative. The other two constraints, $c_{ik} \geq 0$ and $\sum_k c_{ik} = 1$, are imposed on $\boldsymbol{c}_i$ due to its definition as a membership vector. Then, by solving this optimization formulation, the K canonical models, encoded in $\mathbf{Q}$, and the membership vector for each individual, encoded in $\boldsymbol{c}_i$, can be estimated. Next, individuals' degradation models can be obtained by using $\boldsymbol{\beta}_i = \mathbf{Q}\boldsymbol{c}_i$.

The proposed collaborative learning framework is flexible and capable of fusing data and information from multiple sources. For instance, we could further extend the CM to incorporate the similarity information, denoted as w_{jl} for the similarity between individuals j and l. The similarity w_{jl} is commonly available in many healthcare applications and reflects how likely that the degradation models of the two individuals could be similar. Therefore, it is reasonable to assume that $\boldsymbol{c}_j$ and $\boldsymbol{c}_l$ are more similar with each other when w_{jl} is larger. w_{jl} can be obtained by various approaches that will be discussed in more details in Section 2.5.

To incorporate the similarity knowledge in the model formulation of CM, we add a regularization term, $\sum_{j,l} ||\boldsymbol{c}_j - \boldsymbol{c}_l||^2 w_{jl}$, into the objective function of (2.1), leading to the following SCM formulation:

$$\min_{\boldsymbol{c}_i, \mathbf{Q}} \sum_i \|\boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i\|^2 + \lambda \sum_{j,l} ||\boldsymbol{c}_j - \boldsymbol{c}_l||^2 w_{jl}, \tag{2.2}$$

$$\text{subject to } c_{ik} \geq 0, \quad \sum_k c_{ik} = 1, \quad \mathbf{X}_i\mathbf{Q} \geq \mathbf{0},$$

$$\forall\, i = 1, \ldots, N \text{ and } k = 1, \ldots, K.$$

Here, λ is the tuning parameter to control the degree of how much the regularization term affects the parameter estimation. The larger λ, the more influence the regularization term will impose on the estimation of the parameters.

The proposed SCM is a constrained optimization problem with non-convex objective function, which has no closed-form solution and thus, it is difficult to solve by regular gradient-based algorithms. We will present the details of our proposed algorithm in Section 2.5.

2.4 THEORETICAL ANALYSIS ON THE RELATIONSHIP OF CM WITH MIXED EFFECTS MODEL (MEM)

This section presents a connection between the proposed SCM with MEM. For linear models, MEM assumes that $\{\boldsymbol{\beta}_i, i = 1,2, \dots, N\}$ are i.i.d., sampled from a multivariate normal distribution, i.e., $\boldsymbol{\beta}_i \sim N(\mathbf{0}, \boldsymbol{G})$, where $\boldsymbol{G}$ denotes a covariance matrix. It can be shown that the objective function in Eq. (2.2) becomes equivalent to MEM under the specific conditions where $w_{jl} = 1$ for all j and l and $\boldsymbol{G} = \mathbf{Q}\mathbf{Q}^T$. Note that, here, $w_{jl} = 1$ for all j and l corresponds to the fact that MEM actually treats the individual models as identical samples from a distribution model.

Theorem 2-1 *The objective function of the optimizing problem (2.2) is equivalent to the objective function of MEM when* $\boldsymbol{W}$ *is a matrix with all the elements being one and* $\boldsymbol{G} = \mathbf{Q}\mathbf{Q}^T$

Theorem 2-1 provides a useful insight into the proposed collaborative learning approach's flexibility and unique capability of studying the heterogeneous models at a more fundamental and detailed level than MEM: (***a***) the proposed SCM provides more flexibility of incorporating information sources (w_{jl}) for capturing the similarity among individuals. However, the MEM is limited to $w_{jl} = 1$ for all j and l. These results suggest that the proposed approach should be more general than MEM; (***b***) the proposed model can characterize individual's heterogeneity by allocating different membership vectors, while the MEM treats individuals as identically distributed; (***c***) further, unlike the MEM that encapsulates the population heterogeneity into a

variance-covariance matrix of random effects (e.g., encoded in **G**), the proposed method can model the heterogeneity of the population by explicitly learning multiple canonical models in **Q**.

Remark 1: While Theorem 2-1 suggests that the objective function of MEM can be considered as a special case of the objective function of SCM, it does not imply that MEM is a strictly special case of SCM. Because SCM employs the constraints in (2.2), it presents a more constrained version of MEM. In addition, the number of canonical models K plays an interesting role in defining an upper bound of the rank of the covariance matrix **Q**. As such, SCM can be considered as a knowledge-driven MEM with an extra capability to incorporate the canonical structure of the random effects. On the other hand, SCM has more flexibility than MEM by further incorporating the similarity information between individuals as Theorem 2-1 indicates.

Remark 2: In the literature, a number of transfer learning and multitask learning approaches (Zhou et al. 2012, 2013; Pan and Yang 2010) have been proposed to jointly learn multiple regression models by treating these models as related. One distinct difference between the proposed methods with them is that the proposed methods explicitly exploit the low-dimensional canonical structure embedded in many applications. By allowing the prediction model of each individual to be a weighted combination of multiple canonical models, the proposed methods essentially enable automatic determination of the relatedness of individual regression model to the canonical models. Finally, as a byproduct, the proposed methods can reveal the low-dimensional canonical structure of the underlying problem (i.e., by using model selection methods to identify the number K) and produce the subgroup-level canonical models, leading to valuable domain insights which are not available in many existing multitask learning methods.

2.5 PROPOSED COMPUTATIONAL ALGORITHM FOR SOLVING CM

2.5.1 *Derivation of the computational algorithm*

This section presents the algorithm for estimating the parameters $\mathbf{Q}$ and $\mathbf{C}$ of SCM (CM is a special case of SCM with $\lambda = 0$). First, we rewrite the formulation in (2.2) in a matrix form:

$$\min_{\mathbf{C},\mathbf{Q}} \sum_i \|\boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i\|^2 + \lambda \mathrm{Tr}(\mathbf{C}^{\mathrm{T}}\mathbf{L}\mathbf{C}) \tag{2.3}$$

$$\text{subject to } c_{ik} \geq 0, \quad \sum_k c_{ik} = 1, \quad \mathbf{X}_i\mathbf{Q} \geq \mathbf{0},$$

$$\forall\, i = 1, \dots, N \text{ and } k = 1, \dots, K.$$

Here, we use the fact that $\frac{1}{2}\sum_{j,l} ||\boldsymbol{c}_j - \boldsymbol{c}_l||^2 w_{jl} = \sum_{j=1}^{N} (\boldsymbol{c}_j)^{\mathrm{T}} \boldsymbol{c}_j d_{jj} - \sum_{j,l=1}^{N} (\boldsymbol{c}_j)^{\mathrm{T}} \boldsymbol{c}_l w_{jl} = \mathrm{Tr}(\mathbf{C}^{\mathrm{T}}\mathbf{L}\mathbf{C})$, where $d_{jj} = \sum_l w_{jl}$, $\mathbf{D}$ is a diagonal matrix with entries $\{d_{jj}, j = 1,2, \dots, N\}$, and $\mathbf{L} = \mathbf{D} - \mathbf{W}$.

We observe that by decoupling the estimation of $\mathbf{Q}$ and $\mathbf{C}$ as two sub-optimization problems, we can derive an iterative algorithm. This is due to the fact that the constraints are imposed on $\mathbf{C}$ and $\mathbf{Q}$ separately. Within each iteration, using the latest estimation of $\mathbf{C}$, we can derive a solution of the optimizing problem with respect to $\mathbf{Q}$; while using the latest estimation of $\mathbf{Q}$, we can derive an efficient updating algorithm for $\mathbf{C}$.

1) Solve for $\boldsymbol{Q}$ at fixed parameters $\boldsymbol{C}^r$:

Given $\boldsymbol{C}^r$, the optimization problem (2.3) is reduced to

$$\min_{\mathbf{Q}} \sum_i \|\boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i^r\|^2,$$

$$\text{subject to } \mathbf{X}_i\mathbf{Q} \geq \mathbf{0}\ \forall\, i = 1, \dots, N.$$

This is essentially a constrained Least Square (LS) problem. We define $\mathbf{X}_i^*$ as

$$\mathbf{X}_i^* = \mathbf{X}_i\tilde{\mathbf{C}}_i^r,$$

where

$$\tilde{\mathbf{C}}_i^r = \begin{bmatrix} (\boldsymbol{c}_i^r)^{\mathrm{T}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & (\boldsymbol{c}_i^r)^{\mathrm{T}} \end{bmatrix}_{(p\times Kp)}. \tag{2.4}$$

Define $\mathbf{B}_{i_{(Kp\times Kp)}} = diag(\mathbf{X}_i, \dots, \mathbf{X}_i)$ and $\boldsymbol{q}$ as a $Kp \times 1$ vector that is generated by concatenating the columns of the matrix $\mathbf{Q}$. Then, the objective function of $\mathbf{Q}$ is

$\min_{\mathbf{Q}} \sum_i \|\boldsymbol{y}_i - \mathbf{X}_i^* \boldsymbol{q}\|^2$,

The constraint $\mathbf{X}_i\mathbf{Q} \geq \mathbf{0}$ can be written as $\mathbf{B}_i\boldsymbol{q} \geq \mathbf{0}$. In this way, the problem formulation in (2.3) can be rewritten as

$$\min_{\mathbf{Q}} \sum_i \|\boldsymbol{y}_i - \mathbf{X}_i^* \boldsymbol{q}\|^2, \tag{2.5}$$

$$\text{subject to } \mathbf{B}_i\boldsymbol{q} \geq \mathbf{0}, \ \forall\, i = 1, \dots, N.$$

This quadratic programming problem can be solved by existing algorithms such as NNLS and FNNLS (Lawson and Hanson 1947; Bro and Jong 1997).

2) Solve for $\boldsymbol{C}$ at fixed $\boldsymbol{Q}^r$:

Given $\mathbf{Q}^r$, the objective function in (2.3) becomes

$$\sum_i \|\boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i\|^2 + \lambda \mathrm{Tr}(\mathbf{C}^{\mathrm{T}}\mathbf{L}\mathbf{C}).$$

By introducing the Lagrange multiplier μ_i for the constraint $(\boldsymbol{c}_i)^{\mathrm{T}}\mathbf{1} = 1$, the Lagrangian is

$$L = \sum_{i=1}^N (\boldsymbol{y}_i)^{\mathrm{T}}\boldsymbol{y}_i - 2\sum_{i=1}^N (\boldsymbol{y}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i + \sum_{i=1}^N (\boldsymbol{c}_i)^{\mathrm{T}}(\mathbf{Q}^r)^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i +$$

$$\lambda \mathrm{Tr}(\mathbf{C}^{\mathrm{T}}\mathbf{L}\mathbf{C}) + \sum_{i=1}^N \mu_i[(\boldsymbol{c}_i)^{\mathrm{T}}\mathbf{1} - 1].$$

The partial derivative of L with respect to $\boldsymbol{c}_i$ is

$$\frac{\partial L}{\partial \boldsymbol{c}_i} = -2(\mathbf{Q}^r)^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i + 2(\mathbf{Q}^r)^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i + 2(\lambda\mathbf{L}\mathbf{C})_i + \mu_i\mathbf{1}.$$

Using the complementarity condition to enforce the nonnegativity of c_{ik}, we get the following equation for c_{ik},

$$-[(\mathbf{Q}^r)^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]_k c_{ik} + [(\mathbf{Q}^r)^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i]_k c_{ik} + (\lambda\mathbf{L}\mathbf{C})_{ik}c_{ik} + \frac{1}{2}\mu_i c_{ik} = 0. \tag{2.6}$$

Summing over i and using the primal feasibility $(\boldsymbol{c}_i)^{\mathrm{T}}\mathbf{1} = 1$ and $\mathbf{L} = \mathbf{D} - \mathbf{W}$, we have

$$\frac{1}{2}\mu_i = [(\mathbf{Q}^r)^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]^{\mathrm{T}}\boldsymbol{c}_i + \lambda[(\mathbf{WC})_i]^{\mathrm{T}}\boldsymbol{c}_i - [(\mathbf{Q}^r)^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i]^{\mathrm{T}}\boldsymbol{c}_i - \lambda[(\mathbf{DC})_i]^{\mathrm{T}}\boldsymbol{c}_i. \tag{2.7}$$

Using the expression of multiplier μ_i in (2.7), (2.6) can be written as

$$\{[(\mathbf{Q}^r)^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i]_k + \lambda(\mathbf{DC})_{ik} + [(\mathbf{Q}^r)^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]^{\mathrm{T}}\boldsymbol{c}_i + \lambda[(\mathbf{WC})_i]^{\mathrm{T}}\boldsymbol{c}_i\}c_{ik} -$$

$$\{[(\mathbf{Q}^r)^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]_k + [(\mathbf{Q}^r)^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i]^{\mathrm{T}}\boldsymbol{c}_i + \lambda(\mathbf{WC})_{ik} + \lambda[(\mathbf{DC})_i]^{\mathrm{T}}\boldsymbol{c}_i\}c_{ik} = 0.$$

This equation leads to the following updating rule:

$$\boldsymbol{c}_{ik}^{m+1} = \boldsymbol{c}_{ik}^{m}\frac{[(\mathbf{Q}^r)^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i + (\lambda\mathbf{WC}^m)_i]_k + [(\mathbf{Q}^r)^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i^m]^{\mathrm{T}}\boldsymbol{c}_i^m + \lambda[(\mathbf{DC}^m)_i]^{\mathrm{T}}\boldsymbol{c}_i^m}{[(\mathbf{Q}^r)^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i^m + (\lambda\mathbf{DC}^m)_i]_k + [(\mathbf{Q}^r)^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]^{\mathrm{T}}\boldsymbol{c}_i^m + \lambda[(\mathbf{WC}^m)_i]^{\mathrm{T}}\boldsymbol{c}_i^m}. \tag{2.8}$$

We derive an algorithm that iteratively optimize for $\mathbf{C}$ and $\mathbf{Q}$. Figure 2-1 provides a summary of the overall algorithm. Our numerical studies in Section 2.5-2.7 suggest that the proposed algorithm is efficient and easy to converge.

Input: measurements for risk factors and cognitive status on each subject, $\mathbf{X}_i$ and $\boldsymbol{y}_i$, $i = 1, \dots, N$; initial values for the parameters, $\mathbf{C}^{(0)}$ and $\mathbf{Q}^{(0)}$; similarity matrix, W; tuning parameter, λ; maximal iteration number, $MaxIter$

For $r = 0,1, \dots, MaxIter$

1. Transform $\boldsymbol{c}_i^r$ to $\tilde{\mathbf{C}}_i^r$ using (2.4).
2. Let $\mathbf{X}_i^* = \mathbf{X}_i\tilde{\mathbf{C}}_i^r, \mathbf{B}_i = diag(\mathbf{X}_i, \dots, \mathbf{X}_i)$. Calculate $\boldsymbol{q}^{r+1}$ by solving the quadratic programming problem in (2.5).
3. Transform $\boldsymbol{q}^{r+1}$ to $\mathbf{Q}^{r+1}$ by partitioning the $Kp \times 1$ vector to the $p \times K$ matrix.
4. Calculate $\mathbf{C}^{r+1}$ by the updating rules in (2.8).

End for

Output: $\{\mathbf{Q}^{(MaxIter+1)}, \mathbf{C}^{(MaxIter+1)}\}$.

Figure 2-1: Procedure of the proposed algorithm for solving SCM.

2.5.2 *Convergence properties of the proposed algorithm*

Theorem 2-2 below shows that by using the proposed algorithm, the objective function is non-increasing, converging to a stationary point.

Theorem 2-2. *The solution converges to a stationary point using the iterative algorithm in Figure 2-1.*

Note that, stationary point is a necessary condition for optimality, but not a sufficient condition. Thus, Theorem 2-2 only implies that the algorithm will converge to a stationary point but not necessary the optimal point. However, empirically, we observe that it is usually the case that the algorithm will converge to the local optimal as well.

2.5.3 *Empirical issues of implementing the algorithm*

Note that we need to define a similarity matrix **W** in SCM. Sometimes it can be readily available through querying prior knowledge or expert opinion. There have been many methods developed in the literature (Wang et al. 2011; Sun et al. 2010) that can extract the patient similarity from domain knowledge or medical records. We could also quantify the similarity between individuals based on some covariates that reflect the characteristics of the individuals, e.g., in AD, the ApoE genotype and some other biomarkers could be used to define the patient similarity. Even when the covariates are not available, a heuristic approach could be used that treats the regression parameters of the individuals as the covariates. For instance, the MEM method will be used to learn the regression models of the individuals. The regression parameters estimated by the MEM represent the individual-to-individual variations. If the underlying low-dimensional canonical structure exists in the heterogeneous population, the SCM can further improve the estimation by extracting the similarity information from the regression parameters. On the other hand, to convert the covariates into similarity, we adopt some existing approaches that have been found effective

in the literature (Belkin and Niyogi 2001), including the 0-1 weighting, Heat Kernel Weighting and Dot-Product Weighting. In our numerical studies on both synthetic datasets and real-world dataset, we note that the heat kernel weighting method consistently lead to satisfactory results. One possible reason is that comparing with the 0-1 weighting and dot-product weighting, the heat kernel weighting method allows an optimal tuning of the similarity by introducing a scaling parameter σ^2 in defining w_{ij} as $w_{ij} = exp\left(-\left\|x_i - x_j\right\|^2/\sigma^2\right)$ for individuals i and j. As such, the scaling parameter can be automatically determined by model selection methods such as cross-validation.

Other implementation issues include how to obtain initial values of $\mathbf{C}^{(0)}$ and $\mathbf{Q}^{(0)}$. Empirical evidence in the simulation studies (Sec 2.6) and the real-world data analysis (Sec 2.7, 2.8) suggests that the solutions from MEM (Galecki and Burzykowski 2013) can provide good initial values. For instance, clustering algorithms such as k-means can be applied on the regression parameters that are learned by the MEM method. Then, the centroid vectors of the clusters that are learned by the k-means algorithm can be the initial values of $\mathbf{Q}^{(0)}$, and the similarity between the regression parameters of the individuals with the centroid vectors of the clusters can be used as the initial values of $\mathbf{C}^{(0)}$.

2.6 SIMULATION STUDY

In this section, we conduct simulation studies to compare the performances of the proposed CM (i.e., Eq. (2.1)), the SCM (i.e., Eq. (2.2), the mixed effect models (MEM) (i.e., MEM assumes that $\boldsymbol{\beta}_i$ comes from a multivariate normal distribution), and the trivial method that builds the model for each individual separately (IGM) (i.e., estimate $\boldsymbol{\beta}_i$ independently).

We compare these models across various settings of the following parameters, the number of canonical models (e.g., here we use $K = 3$ and $K = 5$), the sparsity of the samples (e.g., sparse sampling and dense sampling, that will be described later), and the types of regression models (e.g., Type 1 model and Type 2 model, that will be described later). Here, the sparsity of the samples controls how many samples we can usually collect for each individual as training data. Throughout our simulation studies, for each individual, we simulate longitudinal observations on 25 consecutive time points regardless of any other factors such as the number of canonical models or the type of degradation model. The last 5 observations of each individual are used as testing data. To generate the training data of each individual, in the "dense sampling" scenario, we randomly select M observations (i.e., $M \sim \text{Unif}(15,20)$) from the first 20 observations as the training data for each individual; while in the "sparse sampling" scenario, we randomly select M observations (i.e., $M \sim \text{Unif}(4,8)$) from the first 20 observations as the training data for each individual. For the types of degradation models, many degradation models have been used in AD for predicting cognitive decline (Pearson et al. 2005; Duchesne et al. 2009) and most of them can be formulated as linear regression models, depending on the selected predictors. Here, we focus on two common types of degradation models. One is commonly referred as the disease trajectory model (Type 1 model), which uses the age (or polynomial basis functions of the variable age) as the predictor (or predictors) (Head et al. 2004; Bartzokis et al. 2004; Raz 2000). In our simulation, we adopt the 2^{nd} order polynomial model. Another type of degradation models (Type 2 model) uses risk factors as the predictors, such as the Disease Progression Score (DPS) (Jedynak et al. 2012) that is adopted here to simulate the Type 2 model.

For any given K and the type of degradation model, we can generate the underlying model for each individual by the following procedure. Specifically, the underlying model for the individual

i at time point t is $z_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta}_{\mathrm{i}} + \varepsilon_{it}$, where $\boldsymbol{x}_{it} = [1, t, t^2]$ for Type 1 model, and $\boldsymbol{x}_{it} = [x_{i1t}, x_{i2t}, \ldots, x_{ipt}]$ for Type 2 model where x_{ijt} is the measurement of the j^{th} biomarker of subject i at time t. We simulated the $\boldsymbol{x}_{it}$ from the standard multivariate normal distribution. We randomly generate the matrix $\mathbf{Q}$ and $\boldsymbol{c}_i$ to obtain $\boldsymbol{\beta}_{\mathrm{i}}$ as $\boldsymbol{\beta}_{\mathrm{i}} = \mathbf{Q}\boldsymbol{c}_i$. Specifically, to encourage the low-dimensional canonical structure, we generate the membership vector $\boldsymbol{c}_i$ by the following procedure. For instance, considering the case that there are three canonical models. We design three multivariate normal distributions as below:

$$F_1(\boldsymbol{c}) \sim N\left(0, \begin{bmatrix} v^2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right),\ F_2(\boldsymbol{c}) \sim N\left(0, \begin{bmatrix} 1 & 0 & 0 \\ 0 & v^2 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right),\ F_3(\boldsymbol{c}) \sim N\left(0, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & v^2 \end{bmatrix}\right).$$

For generating $\boldsymbol{c}_i$, we first randomly select one multivariate normal distribution among the three distributions and generate a random sample from the selected distribution. The resulting random sample is further normalized to obtain $\boldsymbol{c}_i$. Evidently, the larger the magnitude of v^2 in F_i, the more dominant the i^{th} element in $\boldsymbol{c}_i$. Note that, here, v^2 controls the significance of the low-dimensional canonical structure, i.e., when v^2 is small, the difference between the canonical models becomes less significant. Thus, it is anticipated that when v^2 is large, SCM should outperform MEM; when v^2 is small, SCM and MEM should perform similarly. Our implementation results appear to be robust in a wide range of v^2 as long as the low-dimensional canonical structure is significant. In the sequel, we use $v^2 = 100$.

After generating the data, we determine the optimal K, which is unknown in practice by using the Akaike information criterion (AIC) (Akaike 1974). To implement the SCM, here, the similarity between any two individuals is calculated based on the regression parameters estimated by MEM using the heat kernel function. We evaluate the performances of these models based on two criteria, the parameter estimation and prediction accuracy. For parameter estimation, we calculate the

difference between the estimated coefficients with the true coefficients, e.g., $\sum_{i,j}(\widehat{\beta_{ij}} - \beta_{ij})^2/(Np)$. For prediction accuracy, we compare the normalized mean square error (nMSE) on the testing set, e.g., $nMSE(\boldsymbol{Z}, \widehat{\boldsymbol{Z}}) = [\sum_{t=1}^{T}\|\boldsymbol{z}_t - \hat{\boldsymbol{z}}_t\|_2^2/\sigma(\boldsymbol{z}_t)]/(\sum_{t=1}^{T} n_t)$, where β_{ij} is the true coefficient of risk factor j and subject i, $\widehat{\beta_{ij}}$ is estimated value of β_{ij}, $\boldsymbol{z}_t$ are the true measurements (including all the subjects in the testing dataset) at a single time point, $\hat{\boldsymbol{z}}_t$ are the predicted values of $\boldsymbol{z}_t$, and n_t is the number of observations at that time point. Besides nMSE, we also use the weighted correlation coefficient (wR), defined as $wR(\boldsymbol{z}, \hat{\boldsymbol{z}}) = [\sum_{t=1}^{T} Corr(\boldsymbol{z}_t, \widehat{\boldsymbol{z}_t})n_t]/(\sum_{t=1}^{T} n_t)$, as another criterion for evaluating the prediction accuracy.

Table 2-1 summarizes the results, which correspond to $K = 3$. Note that, in Table 2-1, rMSE1-step means that the model is used to predict the degradation in the next time point; rMSE3-step means that the model is used to predict the degradation in the 3rd time point. Our overall observations include; 1) When the low-dimensional canonical structure is significant (v^2 is large), the proposed CM and SCM is effective on exploiting the low-dimensional canonical structure as demonstrated by its better performance than IGM in terms of both parameter estimation and model prediction. 2) Overall SCM outperforms other models, which demonstrates that SCM can effectively incorporate the similarity information between subjects to further enhance the model estimation. 3) The advantage of SCM is generally larger in the sparse sampling scenario, indicating that the incorporation of the structure of the heterogeneity of the population will be more preferred when there is a lack of observations.

Table 2-1:Comparison of CM, MEM, SCM, and IGM on simulated dataset when $K = 3$.

	Type 1 Model				Type 2 Model			
	IGM	CM	MEM	SCM	IGM	CM	MEM	SCM
Dense Sampling								
Running time	**0.060**	28.260	75.000	95.580	**0.670**	203.74	248.05	260.12
MSE	8.143	3.933	3.795	**3.221**	58.050	**37.647**	43.337	40.705
nMSE	0.029	0.028	0.020	**0.017**	0.996	0.064	0.131	**0.045**
wR	0.986	0.987	0.990	**0.992**	0.669	0.967	0.933	**0.976**
rMSE1-step	26.182	28.559	24.506	**23.453**	37.542	10.745	12.518	**10.044**
rMSE3-step	38.827	40.772	35.386	**32.293**	32.810	13.939	22.015	**12.235**
rMSE5-step	50.086	45.987	35.613	**33.667**	51.582	10.748	15.226	**7.720**
Sparse Sampling								
Running time	**0.450**	16.830	72.970	71.200	**0.590**	109.32	227.71	244.81
MSE	54.969	5.857	6.088	**5.391**	85.192	**57.607**	66.621	60.221
nMSE	20.979	0.233	0.348	**0.181**	2.841	0.696	0.626	**0.385**
wR	0.112	0.896	0.851	**0.917**	0.320	0.705	0.665	**0.809**
rMSE1-step	768.53	87.975	103.979	**78.835**	76.511	50.310	41.597	**33.147**
rMSE3-step	1028.0	111.641	137.560	**96.633**	87.744	42.914	42.914	**31.407**
rMSE5-step	1327.1	131.729	162.059	**115.804**	68.253	28.968	37.220	**18.193**

Comparison of CM, MEM, SCM, and IGM on simulated dataset when $K = 5$.

	Type 1 Model				Type 2 Model			
	IGM	CM	MEM	SCM	IGM	CM	MEM	SCM
Dense Sampling								
Running time	**0.070**	34.210	76.290	76.750	**0.090**	285.09	364.56	381.90
MSE	11.753	6.589	7.758	**5.905**	52.606	40.004	47.166	**37.654**
nMSE	0.046	0.022	0.026	**0.018**	1.584	0.161	0.265	**0.154**
wR	0.978	0.991	0.988	**0.991**	0.580	0.921	0.768	**0.925**
rMSE1-step	39.828	30.620	35.149	**28.423**	56.393	15.408	21.029	**15.222**
rMSE3-step	56.066	38.667	41.008	**35.439**	50.239	16.521	18.817	**16.229**
rMSE5-step	74.589	50.243	52.345	**42.701**	58.343	16.865	21.341	**16.534**
Sparse Sampling								
Running time	**0.240**	18.610	73.450	75.860	**0.510**	141.13	139.15	166.40
MSE	90.045	14.903	24.104	**10.784**	69.710	52.691	52.369	**52.088**
nMSE	9.473	0.358	0.114	**0.060**	1.703	0.569	0.543	**0.490**
wR	0.395	0.887	0.941	**0.971**	0.562	0.715	0.699	**0.752**
rMSE1-step	596.00	129.890	72.373	**54.688**	52.248	29.893	28.954	**28.551**
rMSE3-step	803.87	154.972	86.974	**64.305**	54.842	34.917	33.268	**31.080**
rMSE5-step	1050.2	191.821	108.631	**75.960**	68.883	31.403	32.560	**26.214**

2.7 APPLICATION IN LEARNING COGNITIVE DEGRADING TRAJECTORIES

This section demonstrates the performance of our proposed methods on a real-world dataset that is collected by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller 2005). In the dataset, we identified a set of 478 subjects whose longitudinal measurements of Mini-Mental State Estimation (MMSE) were collected at baseline, 12th month, 24th month, 36th month, 48th month and 60th month. These 478 subjects include 104, 261 and 113 individuals in the NC, MCI and AD groups, respectively. The underlying canonical structure of these subjects is revealed by the cognitive degradation patterns: the MMSE measurements of the NC subjects at different time points maintain at a high level with small fluctuations, while the MMSE measurements of AD patients degrade dramatically, and the degradation of the MMSE measurements of MCI patients is faster than NC but slower than AD. Among these subjects, 21, 156 and 244 individuals have only 3, 4, and 5 observations, respectively, presenting a typical dataset that has sparse measurements.

We remove the group information before implementing the proposed methods to demonstrate that the CM and SCM can effectively recover the low-dimensional canonical structure. For each individual, we use the measurements in the 48^{th} month and 60^{th} month as testing data, and others as training data. We use the ApoE genotypes, the baseline MMSE score, and the baseline regional brain volume measurements extracted from MRI via FreeSurfer (Jack et al. 2008) to derive the similarity of individuals by using the heat kernel weighting.

We use the polynomial model as the degradation model of each individual i (Biesanz et al. 2004; Sliwinski et al. 2003):

$$f_i(t) = \sum_{\mathrm{k}} c_{ik}\, q_{k0} + \sum_k c_{ik} q_{k1}\, t + \sum_k c_{ik} q_{k2}\, t^2 + \varepsilon_{it},$$

where $f_i(t)$ is the MMSE measurement of subject i at time t.

We first investigate if the proposed methods (CM and SCM) can automatically identify the low-dimensional canonical structure. AIC is used to automatically select the best K (together with other parameters such as λ and the scaling parameter in the heat kernel function) in the CM and SCM model. The results show that both methods can identify the low-dimensional canonical structure, e.g., the AIC results of SCM in Figure 2-2 clearly show that the AIC value reaches minima when $K = 3$, which is consistent with our prior knowledge of the AD dataset. Figure 2-3 also shows that the algorithm for SCM converges quickly within less than 20 iterations.

We also investigate the canonical models discovered in Alzheimer's Disease population, and show their cognitive degradation patterns in Figure 2-4. It can be observed that three patterns of cognitive decline are discovered. These patterns represent the cognitive degradation trajectories of NC, MCI, and AD patients respectively.

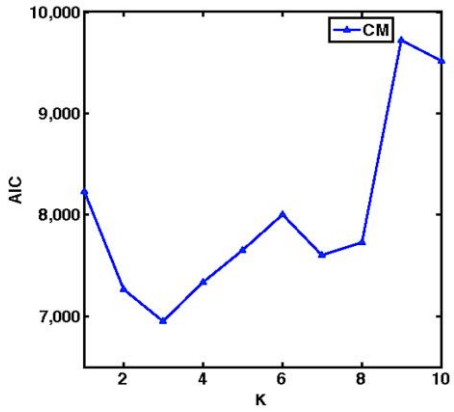


Figure 2-2: The AIC values versus K for CM.

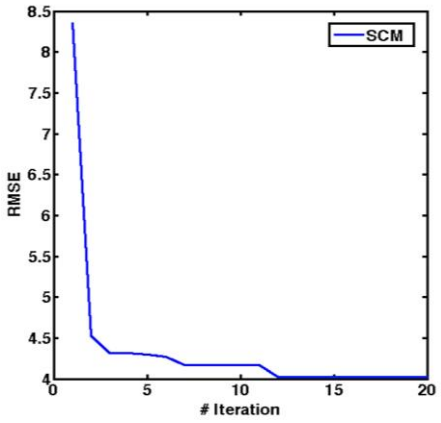


Figure 2-3: Convergence performance of the computational algorithm for SCM.

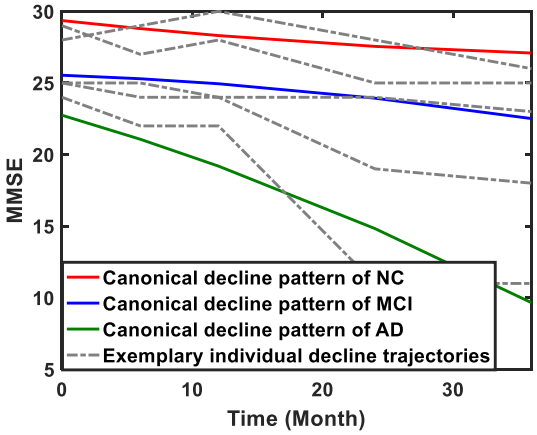


Figure 2-4: Three canonical models discovered in Alzheimer's Disease population using SCM.

Table 2-2 summarizes the prediction results. Clearly, SCM outperforms other methods on all the performance metrics. We can also observe that CM is better than MEM and IGM, indicating that the CM is indeed effective on utilizing the low-dimensional canonical structure embedded in the AD dataset. As a result, explicitly exploiting the heterogeneity of the population coupled with the low-dimensional canonical structure in CM and SCM appears to be better than only considering the variations among individuals in MEM. It is also clear that the performance deterioration of CM and SCM on predicting the 60th month (a long-term prediction) than the 48th

month is much smaller than the other methods. This result indicates that the proposed methods are particularly advantageous if long-term prediction/monitoring is required.

Table 2-2: The performances of the models (IGM, CM, MEM, and SCM) in terms of normalized mean square error (nMSE) and weighted correlation coefficient (wR).

	IGM	CM	MEM	SCM
Target:MMSE				
Running time (sec)	0.160	6.220	8.050	92.090
nMSE	1.799	0.936	0.755	**0.531**
wR	0.580	0.618	0.660	**0.716**
M48 rMSE	4.874	4.330	3.705	**3.651**
M60 rMSE	8.326	5.458	5.040	**3.777**

Overall, the results suggest that the proposed collaborative learning framework has a superior prediction capability over other alternative methods. The gain of the prediction accuracy has the potential of being translated into higher detection rate of early AD, which will lead to more powerful and cost-effective routine cognitive screening and monitoring programs that are needed in many primary health care settings.

2.8 APPLICATION IN LEARNING DEPRESSION TRAJECTORIES

In this section, we further applied the proposed method for learning depression trajectories. Depression is a common, complex and dynamic mental disorder characterized by sad mood, loss of interest in activities, weight gain or loss, psychomotor agitation or retardation, fatigue, inappropriate guilt, difficulties concentrating, and recurrent thoughts of death (CDC, 2012). Due to potential side effects of the medication, the Food and Drug Administration (FDA) emphasizes that patients taking antidepressants should be closely monitored (NIMH). While depression is essentially a heterogeneous dynamic process, monitoring needs a quantitative understanding of the progression patterns. Understanding trajectories of depression is essential when health care

providers and health systems want to allocate attention and resources to those who need them most. We need more sophisticated approaches for monitoring outcomes and identifying those likely to need more intensive treatment. Hence, empirically-based monitoring strategies are critical for diagnosis, prognosis and evaluation of treatment outcome of depression. The emerging use of electronic health record (EHR) in health care systems provides an abundance of data that contains the disease trajectories of many individuals. Thus, statistical analysis of these trajectories could lead to a quantitative understanding of the progression patterns, providing useful knowledge for guiding the clinical actions for depression monitoring. However, knowledge of how depression progress over time is still limited. Much previous research on depressive symptom trajectories has examined average effects. For example, longitudinal study of depressive trajectory across adult lifespan has found a quadratic pattern (U-shaped), in which depression is highest among young and older adults (Sutin et al., 2013). Research regarding individual trajectories is more limited. One investigation found a five class trajectory patterns in a cohort of Australian family practice (Gunn et al., 2013). To the best of our knowledge, there is no research on individual depression trajectories for the intermediate timeframe (2-5 years) using large U.S. EHR data. Understanding depression progression in the 2-5 years' time window is clinically relevant for designing monitoring and treatment follow-up strategies.

Thus we applied the proposed method on a dataset from the Mental Health Research Network (MHRN), which provides one of the largest depression dataset in the U.S. The MHRN dataset contains depression monitoring outcome data with approximately 1.2 million observations from a diverse and representative sample of outpatients in five states (California, Colorado, Minnesota, Washington and Idaho). It includes personal-level longitudinal depression measures using the Patient Health Questionnaire (PHQ)-9, a self-administered questionnaire that includes 9

multiple-choice questions (Kroenke et al., 2001). between years 2007 and 2012. Our study sample was limited to 3,159 frequently measured individuals who are receiving ongoing treatment and have at least six biweekly measurements.

Due to the nonlinearity nature of depression progression, we extent the linear based collaborative modeling method to characterize the nonlinear dynamics by using B-spline model. In the B-spline model, irregular time interval is firstly transformed to the B-splines bases, then the trajectory signals are represented as linear combination of these bases. B-splines bases are constructed from polynomial pieces, joined at certain time points, which are denoted as the knots. Once the knots are given, the B-splines bases can be computed recursively for any desired degree of the polynomial using the algorithms in (Boor, 1978). Assume subject i is repeatedly measured N_i times and the measurement times are denoted as $\boldsymbol{t_i} = [t_1, \dots, t_{N_i}]$, and the corresponding measurements are denoted as $\boldsymbol{y_i} = [y_{it_1}, \dots, y_{\boldsymbol{it_{Ni}}}]$, we use the cubic B-splines to compute the bases for each time. Let $B_j(t)$ denotes the j^{th} basis at time t, M denotes the number of bases. The trajectory of measurements on subject i, denoted as f_i, can be fitted by a linear combination of these bases,

$$f_i(\boldsymbol{t_i}) = \sum_{j=1}^{M} \beta_{ij} B_j(\boldsymbol{t_i}) .$$

where β_{ij} is the coefficient relating to the j^{th} basis of subject i.

To choose the number of bases in B-spline model, we apply the leave-one-out cross-validation (LOOCV) technique on the PHQ-9 measurements within subject and use the square root of the mean square error (rMSE) to evaluate the prediction error across all measurements of this subject. As revealed in Figure 2-5, B-spline models with 5 bases give the best prediction performance when the number of basis is limited at low level, we specify the number of bases at 5. To measure the similarity between individuals, we consider both risk markers including the age,

gender, and Charlson Comorbidity Scores, and the exploratory features from the individual measurements such as the statistical summary of measurements and the B-spline coefficients. We further apply the clustering algorithm (K-means) on the similarity informations to obtain the initial values of number of canonical models and model parameters. We find a five progression patterns from the dataset, as shown in Figure 2-6. The five-subgroup pattern is also consisted with what have been known in the literature using a smaller sample size of observational data collected in one year (Gunn et al.,2013).

We compare the predictive performance of IGM, MEM, CM and SCM on the last two measurements of each individual. The prediction accuracy measured by rMSE is summarized in Table 2-3. It can be observed that incorporating the canonical structure and similarity between individuals will greatly improve the prediction accuracy of individual depression trajectories, which may enable the development of patient-specific depression management.

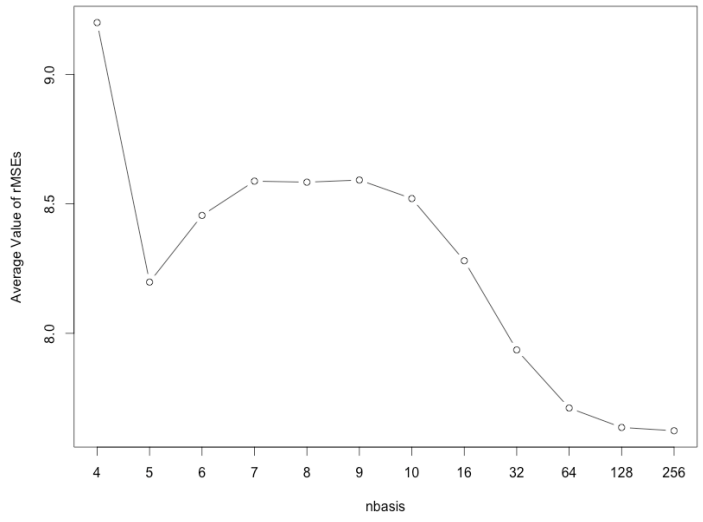


Figure 2-5: Choosing the optimal number of bases for the B-spline model

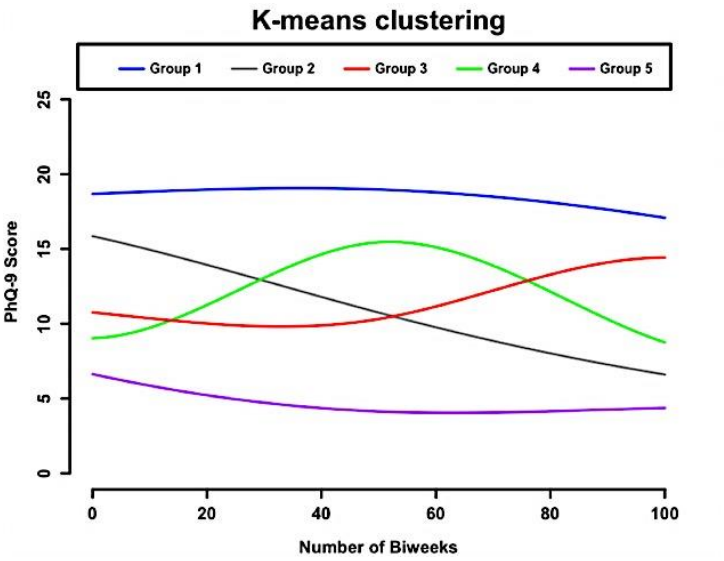


Figure 2-6: Five progression patterns found by applying K-means clustering algorithm on B-spline coefficients.

Table 2-3: Prediction performance of four methods on depression data.

Method	IGM	MEM	CM	SCM
rMSE	12.5335	5.9131	5.1780	3.2098

2.9 CONCLUSION

In this chapter, we propose a novel collaborative learning framework to estimate a heterogeneous population of regression models. It is motivated by the fact that existing models, such as the MEM, impose the homogeneity assumption that assumes the parameters of these prediction models are sampled from a distribution which is usually a multivariate normal distribution, so the mean vector can characterize the mean tendency of these parameters and the covariance matrix can characterize the dispersion of the models. While MEM provides an effective solution for estimating "mixed effects", however, it is not opted for applications in a heterogeneous population, where considerable heterogeneity of the models exists. To mitigate such heterogeneity, we propose the collaborative learning framework by exploiting the idea of "canonical models" and

model regularization. First, to characterize the heterogeneity of the population, it uses a set of canonical models to represent the population characteristics. Then, the model of each individual resembles these canonical models in a probabilistic way, in which the uncertainty is characterized by a membership vector that can be learned from data. To enhance the estimation of the individual models, the similarity between the individuals can also be used by imposing a network-regularized term in the learning framework. Such a collaborative learning framework is applied in the contexts of degradation modeling in AD and nonlinear progression modeling in depression trajectory, which lead to the development of CM and SCM. We develop an efficient computational algorithm to estimate the model parameters, and we further provide theoretical results that ensure that the proposed algorithm can guarantee the convergence property. Both simulation studies, AD cognitive data analysis, and depression data analysis show that the proposed methodology can lead to significantly better performance on both parameter estimation and prediction over existing methods. In addition, theoretical analysis is conducted to reveal the connection between the proposed methods with MEM.

It is important to point out that the proposed method is valuable for improving the current prognostic practices of chronic conditions in AD and depression, particularly considering the feasibility and cost of acquiring measurements for these diseases prognostics and diagnosis. The identification of the pre-symptomatic individuals who are going to experience abnormal cognitive decline or depression progression can provide an evidence-based platform to better deliver existing clinically validated assessment tools of AD (e.g. biomarker measurement technologies, such as the MRI images, the PiB (Pittsburgh Compound B) PET scan, and cerebral spinal fluid (CSF) measurements) and depression (e.g. PHQ-9) to targeted population.

Chapter 3. SELECTIVE SENSING OF HETEROGENEOUS POPULATION

3.1 INTRODUCTION

In this chapter, we aim to answer this question: given a heterogeneous population of individuals where each individual's health is characterized as a dynamic process, how could we optimally allocate limited sensing resources to maximize the likelihood of identifying those individuals who will move into high-risk health status? This problem is ubiquitous in a range of applications including healthcare (e.g., where many depressed patients are monitored) (Löwe et al. 2004), manufacturing (e.g., where a fleet of robots are employed in an automation process) (Robinson et al. 1999), and wind energy (e.g., where many wind turbines are installed in a wind farm) (Wiser and Bolinger, 2008). Accurately detecting the high-risk (abnormal) individuals where problems (e.g. failure of engines and disease onset) are most likely occurring holds great promise for initializing timely actions including preventive care and early treatment (maintenance) to prevent problems occurrence (Selkoe and Schenk 2003; Hameed et al. 2009). In the cognitive monitoring of at-risk individuals of dementia, for example, early treatment can be initialized on the identified high-risk patients to prevent irreversible brain damage or mental decline (Weimer and Sager 2009). In the depression monitoring, the treatment strategies can be adjusted for the high-risk individuals who tend to resist the current treatment plan (Simon, et al. 2000). However, due to the resource constraints (e.g. limited transmission and processing capabilities in sensor system, or limited number of staffs and appointment slots in healthcare system), To answer this question, it is desirable to develop an individual-specific adaptive monitoring strategy, based on partially observed data at each observation epoch, which can dynamically allocate sensing resources in the

next observation epoch for maximum detection of anomalies. Specifically, in what follows we present a technical description of the problem. Consider a large-scale population that consists of N units (in healthcare application, each individual corresponds to a unit), denote the measurements of these units at each sensing epoch as $\boldsymbol{x}_t = [x_{1t}, \dots, x_{Nt}]$, where each measurement is a single value indicating the health condition of each unit. We are interested in detecting the abnormal units. Due to the limited sensing resources, we can only observe C ($C \leq N$) out of N units at each sensing epoch. By introducing the binary decision variable δ_{it} for each measurement x_{it} such that $\delta_{it} = 1$ if and only if x_{it} is observed at epoch t, the sensing constraint can be expressed as $\sum_i \delta_{it} = C, \forall t$. Thus, the problem is how to choose δ_{it}'s at each sensing epoch such that the sensing constraint is satisfied and maximum abnormal units are detected.

At the first glance, this problem seems to be similar to existing efforts in problems such as ranking and selection problem (Nelson et al., 2001), adaptive monitoring (Byon et al., 2010), reinforcement learning (Barrett, 2013), multi-arm bandit (MAD) problems (Bubeck and Cesa-Bianchi; 2012) and optimal learning (Powell et al., 2012). In comparison, our problem is different since the problem we concern involves a heterogeneous population of individuals where each individual may embody a different dynamic model, whereas these methods concern one dynamic model or a set of homogenous dynamic models and sometimes they assume the dynamic models are known. Further, we concern optimal decision-making given the similarity structure among the individuals, while most of existing efforts assume independence (except a few exceptions such as (Frazier et al., 2009) that concerned dependency which are quite different from ours). Note that, the similarity between units can be quantified by the characteristics of units. For example, the demographic, social-economical, and clinical information are widely used to measure the similarity between patients for better clinical decision making (Zhang et al., 2014).

To solve this problem, we develop a decision support framework that integrates prognosis and sensing. For prognosis, a collaborative modeling approach is developed to predict the future risk of individuals by modeling the individuals' degradation models from partially observed data; for sensing, a selective sensing approach is developed to allocate limited sensing resources to monitor the units that are most likely abnormal. Particularly, to tackle the challenges such as the heterogeneity of the individuals' dynamic models that are unknown, we focus our methodological development on a type of applications where a few representative dynamic models exist that span the space of the dynamic models of the individuals. For instance, for some diseases such as Alzheimer's disease and depression (Jack et al., 2010; Sutin et al., 2013), it is known that there are typical phenotypes in a population where each phenotype follows a certain dynamic progression process and individuals could follow a phenotype or a blend of several phenotypes to progress from normal state to diseased state. This inspires us to develop a collaborative modeling (CM) framework for prognosis: the heterogeneous population characteristics are represented by a number of canonical models, whereas each unit's progression model is captured as variants of these typical models. In this CM based prognosis framework, similarity between units will be further incorporated to improve prognostic accuracy. Then, we formulate the selective sensing (SS) strategy as an optimization problem with respect to decision variables δ_{it}'s, which also integrates similarity information between units. New observations collected by the sensing strategy at next observation epoch are further incorporated in the CM based prognostic method to update the prognosis of all the units, guiding the sensing operation in the next epoch. The overall work flow is illustrated in Figure 3-1. In this chapter, we focus on applications where the disease process could be modeled as a Markov model, but the framework is generic for many other applications.

The remainder of this article is organized as follows. Section 3.2 reviews the related methods. Section 3.3 presents a detailed description of the proposed model. Section 3.4 will present a comprehensive evaluation of the proposed model using simulation studies. Section 3.5 will present a real-world application of the proposed method, and Section 3.6 will present the conclusions and suggestions for future research.

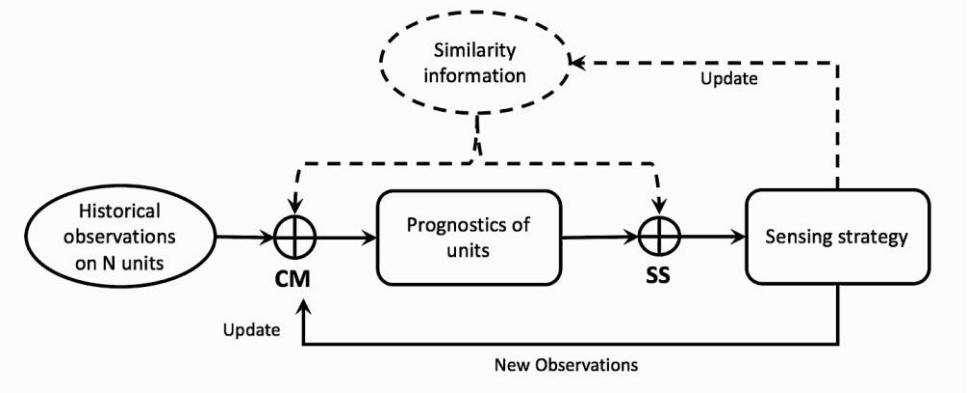


Figure 3-1: Flowchart of the integrated collaborative modeling (CM) based prognosis and selective sensing (SS) for monitoring a heterogeneous patient population.

3.2 RELATED WORK

In what follows we review some methods that are related to our problem. The sequential resource allocation problems have been extensively studied, e.g., the Markov decision process (MDP) and partially observed Markov decision process (POMDP) are widely used stochastic models for dynamically taking actions and updating information at each monitoring period. They are commonly used in adaptive monitoring (Ayer et al. 2012), and condition-based maintenance (Byon et al., 2010). However most of these methods were developed for monitoring one system. On the other hand, ranking and selection strategy is widely used in system control (Liu et al., 2015), simulation (Nelson et al., 2001) and channel sensing (Jiang et al., 2009) to allocate limited

resources among units. However, these methods concentrate on problems where beliefs about the units' progressions are independent, which may result in inaccurate sensing strategies when local prognostics are unreliable under uncertain environment (Nelson et al., 2001). Reinforcement learning, as a class of machine learning methods, has been used to find an optimal data collection policy in an uncertain environment (Baird, 1994). These reinforcement learning algorithms find optimal strategy by iteratively alternating between the exploitation process in which strategy is evaluated by the knowledge learnt from historical data, and the exploration process which improves the strategy (Sutton et al., 1998; Barrett, 2013). The trade-off between exploitation and exploration is also thoroughly analyzed in multi-arm bandit (MAD) problems where resources need to be allocated among competing processes when the distribution of rewards is initially unknown (Bubeck and Cesa-Bianchi; 2012). Optimal learning methods approach the problem in another way and consider the value of information from each measurement (Powell and Ryzhov, 2012). However, these methods impose specific assumptions on the underlying dynamic model of the unit or belief evolution or the reward mechanism, and they generally don't incorporate the similarity among the units and the canonical structure of the population.

3.3 PROBLEM FORMULATION FOR SELECTIVE SENSING

Our method consists of two major components, the prognosis method and the selective sensing method. Both methods are unique in terms of their capacity to utilize the similarity between individuals and the canonical structure of the population for model learning, optimization, and decision-making. In this section, we will introduce the prognosis method in Section 3.3.1 and the selective sensing method in Section 3.3.2.

3.3.1 *A collaborative prognostic model*

3.3.1.1 The model formulation

Formally, assume that the health state of each unit i, $i = 1, \dots, N$ at n_i epochs as $[X_{i1}, \dots, X_{in_i}] \in \mathbb{R}^{1\times n_i}$, and $X_{it} \in \{1, \dots, S\}$ $t = 1, \dots, n_i$. Each unit's health progresses from state s_1 to state s_2, $s_1, s_2 \in \{1, \dots, S\}$, according to a Markov chain with transition probability matrix $\mathbf{P}_i \in \mathbb{R}^{S\times S}$. Each element in the transition matrix $\mathbf{P}_i(s_1, s_2)$ is given by $\mathbf{P}_i(s_1, s_2) = \Pr(X_{it+1} = s_2 | X_{it} = s_1)$. The initial state of each unit is sampled from an initial distribution $\boldsymbol{\nu}_i \in \mathbb{R}^{S\times 1}$ with each element given by $\boldsymbol{\nu}_{is} = \Pr(X_{i1} = s), s \in \{1, \dots, S\}$. Given observations on N subject $\boldsymbol{x}_i = [x_{i1}, \dots, x_{in_i}] \in \mathbb{R}^{1\times n_i}, i = 1, \dots, N$, the log-likelihood function can be written as:

$$\sum_{i=1}^{N} \log\left[P\left(x_{i1}, \dots, x_{in_i}\right)\right] = \sum_{i=1}^{N}\left\{\sum_s e_{is} \log[v_{is}] + \sum_{s_1}\sum_{s_2} N_i(s_1, s_2)\log[\mathbf{P}_i(s_1, s_2)]\right\}, \quad (3.1)$$

where $e_{is} = 1$ if $x_{i1} = s$; $e_{is} = 0$ otherwise. $N_i(s_1, s_2)$ is the number of transitions from s_1 to s_2 on unit i. Parameters in the transition probability matrix $\mathbf{P}_i$ can be estimated by maximizing the log-likelihood function. However, traditional maximum likelihood estimator (MLE) of transition probability matrix may be unreliable since the transitions observed on each individual are sparse and irregular. To effectively learn the individual Markov chain on sparse observations, we adopt the collaborative modeling framework developed in (Lin et al., 2015). As mentioned in the Introduction section, collaborative modeling uses a set of canonical models to represent the heterogeneous population characteristics and assigns each unit a membership vector to represent the degree to which the individual progression resembles the canonical models. Thus, each individual progression model is characterized as a weighted combination of the canonical models. Note that the CM method developed in (Lin et al., 2015) was developed for linear models. Here, we adopt it for dynamic modeling. Formally, let $\boldsymbol{\Pi}_k \in \mathbb{R}^{S\times S}, \boldsymbol{\theta}_k \in \mathbb{R}^{S\times 1}$, $k = 1, \dots, K$, denote the transition probability matrix and initial distribution of the k^{th} canonical model, $\boldsymbol{c}_i =$

$[c_{i1}, \dots, c_{iK}] \in \mathbb{R}^{1\times K}$ represents the membership vector of subject i while each element c_{ik} represents the probability that the unit i's progression resembles the k^{th} canonical model. The transition probability matrix and initial distribution of each unit i is assumed to be:

$$\mathbf{P}_i = \sum_k c_{ik}\mathbf{\Pi}_k, \boldsymbol{\nu}_i = \sum_k c_{ik}\boldsymbol{\theta}_k, \tag{3.2}$$

Based on the canonical models, the log-likelihood function in (5.2) can be rewritten as:

$$\sum_{i=1}^{N} \log\left[P\left(x_{i1}, \dots, x_{in_i}\right)\right] = \sum_{i=1}^{N}\left\{\sum_s e_{is} \log[\sum_k c_{ik}\theta_{ks}] + \right.$$

$$\left.\sum_{s_1}\sum_{s_2} N_i(s_1, s_2)\log[\sum_k c_{ik}\mathbf{\Pi}_k\,(s_1, s_2)]\right\},$$

Considering the constraints that the membership vectors are normalized and nonnegative, and the parameters in the Markov models should come from the probability space, the MLE of parameters in the CM model is formulated as:

$$\max_{c_i,\boldsymbol{\theta}_k,\mathbf{\Pi}_k} \sum_{i=1}^{N}\left\{\sum_s e_{is} \log[\sum_k c_{ik}\theta_{ks}] + \sum_{s_1}\sum_{s_2} N_i(s_1, s_2)\log[\sum_k c_{ik}\mathbf{\Pi}_k\,(s_1, s_2)]\right\},$$

$$\text{s.j.}\quad \sum_{s_2}\mathbf{\Pi}_k(s_1, s_2) = 1, \sum_s \theta_{ks} = 1, \sum_k c_{ik} = 1\ ,$$

$$\forall s_1 = 1, \dots, S, k = 1, \dots, K,\ \ \forall i = 1, \dots, N,$$

$$\text{all parameters are nonnegative.} \tag{3.3}$$

The first equality constraint, $\sum_{s_2}\mathbf{\Pi}_k(s_1, s_2) = 1$, is imposed since the summation of each row in the transition probability matrix must be 1. The second equality, $\sum_s \theta_{ks} = 1$, is imposed due to the summation of initial probabilities must be 1. The third equality, $\sum_k c_{ik} = 1$, is imposed on c_i due to its definition as a membership vector. By solving this problem, the MLE of K canonical models, encoded in $\mathbf{\Pi}_k, \boldsymbol{\theta}_k\ , k = 1, \dots, K$, and membership vectors, encoded in $c_i, i = 1, \dots, N$, can be obtained. The individual Markov chain can be estimated by using relationships depicted in (3.2).

The formulation above could be further extended to incorporate similarity information between units as a regularization term on the membership vectors. The similarity between pair of units, i, j,

denoted as w_{ij}, reflects how likely the disease progressions of them could be similar. Adding a regularization term $\sum_{i,j} w_{ij} \left\| \boldsymbol{c_i} - \boldsymbol{c_j} \right\|^2$ into the objective function in (3.3) leads to the following problem:

$$\max_{c_i, \boldsymbol{\theta}_k, \boldsymbol{\Pi}_k} \sum_{i=1}^{N} \left\{ \sum_s e_{is} \log[\sum_k c_{ik} \theta_{ks}] + \sum_{s_1, s_2} N_i(s_1, s_2) \log[\sum_k c_{ik} \boldsymbol{\Pi}_k (s_1, s_2)] \right\} -$$

$$\frac{\lambda}{2} \sum_{i,j} w_{ij} \left\| \boldsymbol{c_i} - \boldsymbol{c_j} \right\|^2,$$

$$\text{subject to } \sum_{s_2} \boldsymbol{\Pi}_k(s_1, s_2) = 1, \sum_s \theta_{ks} = 1, \sum_k c_{ik} = 1$$

$$\forall s_1 = 1, \dots, S, k = 1, \dots, K, \ \forall i = 1, \dots, N,$$

$$\text{all parameters are nonnegative.} \tag{3.4}$$

Here, λ is a tuning parameter that controls the effect of regularization term.

3.3.1.2 Computational algorithm

The formulation in (3.4) is a constrained optimization problem with non-convex objective function, which has no closed form and is difficult to be solved by regular gradient-based algorithm. This results in a computational challenge that is different from the CM method for linear models in (Lin et al., 2015). To solve this problem, we derive multiplicative update rules (MUR) (Guan et al., 2011) to maximize the objective function by updating the canonical models and membership vectors alternatively in each iteration. By introducing the Lagrange multipliers a_k, b_{ks}, and μ_i for the constraints $\sum_{\boldsymbol{s}} \theta_{ks} = 1$, $\sum_{s_2} \boldsymbol{\Pi}_k(s, s_2) = 1$, and $\boldsymbol{c_i 1} = 1$, the Lagrangian function of (3.4) is:

$$L = \sum_{i=1}^{N} \left\{ \sum_s e_{is} \log[\sum_k c_{ik} \theta_{ks}] + \sum_{s_1, s_2} N_i(s_1, s_2) \log[\sum_k c_{ik} \boldsymbol{\Pi}_k (s_1, s_2)] \right\} - \frac{\lambda}{2} \mathrm{Tr}(\mathbf{C}^{\mathrm{T}} \mathbf{L} \mathbf{C}) -$$

$$\sum_k a_k (\sum_{\boldsymbol{s}} \theta_{ks} - 1) - \sum_{k,s} b_{ks} \left(\sum_{s_2} \boldsymbol{\Pi}_k(s, s_2) - 1 \right) - \sum_{i=1}^{N} \mu_i (\boldsymbol{c_i 1} - 1), \tag{3.5}$$

Here we use the fact that $\frac{1}{2} \sum_{i,j} w_{ij} \left\| \boldsymbol{c_i} - \boldsymbol{c_j} \right\|^2 = \sum_i \left(\sum_j w_{ij} \right) \boldsymbol{c_i c_i}^{\mathrm{T}} - \sum_{i,j} w_{ij} \boldsymbol{c_j c_i}^{\mathrm{T}} = \mathrm{Tr}(\mathbf{C}^{\mathrm{T}} \mathbf{L} \mathbf{C})$, where $\mathbf{C} = [\boldsymbol{c_1}, \dots, \boldsymbol{c_N}]^{\mathbf{T}} \in \mathbb{R}^{N \times K}$ and $\mathbf{L}$ is the Laplacian matrix of w_{ij}. It can be expressed as $\mathbf{L} =$

$\mathbf{D} - \mathbf{W}$, and $\mathbf{D}$ is a diagonal matrix with elements $d_{ii} = \sum_j w_{ij}$. Then, the optimization problem in (3.4) can be simplified as maximizing the Lagrangian L function in two steps.

Step-1. ***<u>Update canonical models with fixed c_i^*</u>***

The partial derivatives of L with respect to θ_{ks} and $\mathbf{\Pi}_k(s, s_2)$ are:

$$\frac{\partial L}{\partial \theta_{ks}} = \sum_i e_{is} \frac{c_{ik}^*}{\sum_k c_{ik}^* \theta_{ks}} - a_k,$$

$$\frac{\partial L}{\partial \mathbf{\Pi}_k(s,s_2)} = \sum_i N_i(s_1, s_2) \frac{c_{ik}^*}{\sum_k c_{ik}^* \mathbf{\Pi}_k(s,s_2)} - b_{ks}, \tag{3.6}$$

Using the complementary condition to enforce the nonnegativity of θ_{ks} and $\mathbf{\Pi}_k(s, s_2)$, we get the following equations:

$$\theta_{ks} \sum_i e_{is} \frac{c_{ik}^*}{\sum_k c_{ik}^* \theta_{ks}} - a_k \theta_{ks} = 0,$$

$$\mathbf{\Pi}_k(s, s_2) \sum_i N_i(s_1, s_2) \frac{c_{ik}^*}{\sum_k c_{ik}^* \mathbf{\Pi}_k(s,s_2)} - b_{ks} \mathbf{\Pi}_k(s, s_2) = 0, \tag{3.7}$$

Using the primal feasibilities $\sum_{\boldsymbol{s}} \theta_{ks} = 1$ and $\sum_{s_2} \mathbf{\Pi}_k(s, s_2) = 1$, the multipliers can be expressed as:

$$a_k = \sum_{i,s} e_{is} \frac{c_{ik}^* \theta_{ks}}{\sum_k c_{ik}^* \theta_{ks}},$$

$$b_{ks} = \sum_{i,s_2} N_i(s_1, s_2) \frac{c_{ik}^* \mathbf{\Pi}_k(s,s_2)}{\sum_k c_{ik}^* \mathbf{\Pi}_k(s,s_2)}, \tag{3.8}$$

Incorporating the expressions of multipliers in (3.8) to (3.7), we obtain that

$$\theta_{ks} \sum_i e_{is} \frac{c_{ik}^*}{\sum_k c_{ik}^* \theta_{ks}} - \sum_{i,s} e_{is} \frac{c_{ik}^* \theta_{ks}}{\sum_k c_{ik}^* \theta_{ks}} \theta_{ks} = 0,$$

$$\mathbf{\Pi}_k(s, s_2) \sum_i N_i(s_1, s_2) \frac{c_{ik}^*}{\sum_k c_{ik}^* \mathbf{\Pi}_k(s,s_2)} - \sum_{i,s_2} N_i(s_1, s_2) \frac{c_{ik}^* \mathbf{\Pi}_k(s,s_2)}{\sum_k c_{ik}^* \mathbf{\Pi}_k(s,s_2)} \mathbf{\Pi}_k(s, s_2) = 0,$$

which leads to the updating rules:

$$\theta_{ks}^{m+1} = \frac{\sum_i e_{is} \frac{c_{ik}^* \theta_{ks}^m}{\sum_k c_{ik}^* \theta_{ks}^m}}{\sum_{i,s} e_{is} \frac{c_{ik}^* \theta_{ks}^m}{\sum_k c_{ik}^* \theta_{ks}^m}},$$

$$\boldsymbol{\Pi}_k(s,s_2)^{m+1} = \frac{\sum_i N_i(s_1,s_2)\frac{c_{ik}^* \boldsymbol{\Pi}_k(s,s_2)^m}{\sum_k c_{ik}^* \boldsymbol{\Pi}_k(s,s_2)^m}}{\sum_{i,s_2} N_i(s_1,s_2)\frac{c_{ik}^* \boldsymbol{\Pi}_k(s,s_2)^m}{\sum_k c_{ik}^* \boldsymbol{\Pi}_k(s,s_2)^m}}. \quad (3.9)$$

Step-2. ***<u>Update membership vectors with fixed $\boldsymbol{\theta}_k^*$, $\boldsymbol{\Pi}_k^*$</u>***

The partial deviation of L with respect to $\boldsymbol{c_i}$ is:

$$\frac{\partial L}{\partial \boldsymbol{c_i}} = \sum_s e_{is}\frac{\boldsymbol{\theta}_{+s}^*}{\boldsymbol{\theta}_{+s}^* \boldsymbol{c_i}^{\mathrm{T}}} + \sum_{s_1,s_2} N_i(s_1,s_2)\frac{\boldsymbol{\Pi}_+^*(s_1,s_2)}{\boldsymbol{\Pi}_+^*(s_1,s_2)\boldsymbol{c_i}^{\mathrm{T}}} - \mu_i \mathbf{1}^{\mathrm{T}} - \lambda(\mathbf{LC})_i, \quad (3.10)$$

where $\boldsymbol{\theta}_{+s}^* = [\theta_{1s}^*, \dots, \theta_{Ks}^*] \in \mathbb{R}^{1\times K}$,$\boldsymbol{\Pi}_+^*(s_1,s_2) = [\boldsymbol{\Pi}_1^*(s_1,s_2), \dots, \boldsymbol{\Pi}_K^*(s_1,s_2)] \in \mathbb{R}^{1\times K}$.

Using the complementary condition $\frac{\partial L}{\partial \boldsymbol{c_i}} c_{ik} = \mathbf{0}$, and primal feasibility $\boldsymbol{c_i}\mathbf{1} = 1$, the multiplier can be expressed as:

$$\mu_i = \sum_s e_{is}\frac{\boldsymbol{\theta}_{+s}^* \boldsymbol{c_i}^{\mathrm{T}}}{\boldsymbol{\theta}_{+s}^* \boldsymbol{c_i}^{\mathrm{T}}} + \sum_{s_1,s_2} N_i(s_1,s_2)\frac{\boldsymbol{\Pi}_+^*(s_1,s_2)\boldsymbol{c_i}^{\mathrm{T}}}{\boldsymbol{\Pi}_+^*(s_1,s_2)\boldsymbol{c_i}^{\mathrm{T}}} - \lambda(\mathbf{LC})_i \boldsymbol{c_i}^{\mathrm{T}} = 1 + n_i - \lambda(\mathbf{LC})_i \boldsymbol{c_i}^{\mathrm{T}}, \quad (3.11)$$

Introducing the expression of multiplier to the equation that $\frac{\partial L}{\partial \boldsymbol{c_i}} c_{ki} = 0$**,** we have:

$$\sum_s e_{is}\frac{\boldsymbol{\theta}_{+s}^* c_{ik}}{\boldsymbol{\theta}_{+s}^* \boldsymbol{c_i}^{\mathrm{T}}} + \sum_{s_1,s_2} N_i(s_1,s_2)\frac{\boldsymbol{\Pi}_+^*(s_1,s_2) c_{ik}}{\boldsymbol{\Pi}_+^*(s_1,s_2)\boldsymbol{c_i}^{\mathrm{T}}} - (1+n_i)c_{ik} + \lambda[(\mathbf{D}-\mathbf{W})\mathbf{C}]_i \boldsymbol{c_i}^{\mathrm{T}} c_{ik} -$$

$$\lambda[(\mathbf{D}-\mathbf{W})\mathbf{C}]_{ik} c_{ik} = 0\ , \quad (3.12)$$

We obtain the updating rule of c_{ik}:

$$c_{ik}{}^{m+1} =$$

$$c_{ik}{}^{m}\frac{\sum_s e_{is}\frac{\boldsymbol{\theta}_{ks}^*}{\boldsymbol{\theta}_{+s}^* \boldsymbol{c_i}^{m\mathrm{T}}} + \sum_{s_1,s_2} N_i(s_1,s_2)\frac{\boldsymbol{\Pi}_k^*(s_1,s_2)}{\boldsymbol{\Pi}_+^*(s_1,s_2)\boldsymbol{c_i}^{m\mathrm{T}}} + \lambda[\mathbf{DC}^m]_i \boldsymbol{c_i}^{m\mathrm{T}} + \lambda[\mathbf{WC}^m]_{ik}}{1+n_i+\lambda[\mathbf{WC}^m]_i \boldsymbol{c_i}^{m\mathrm{T}} + \lambda[\mathbf{DC}^m]_{ik}}, \quad (3.13)$$

The overall algorithm is summarized in Figure 3-2. We have also studied the convergence performance of the algorithm and included the proof of the theorem in Appendix B.

Theorem 3-1: The solution converges to an optimal solution using the iterative updating algorithm.

Markov based Collaborative Modeling Algorithm

Input: data on N individuals, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$; initial values for the parameters, $\mathbf{C}^{(0)}$, $\boldsymbol{\theta}^{(0)}$, and $\boldsymbol{\Pi}^{(0)}$; similarity matrix, $\mathbf{W}$; tuning parameter, λ; number of canonical models, K; maximal iteration number, M

Initial: $\mathbf{C}^* = \mathbf{C}^{(0)}$, $\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(0)}$, and $\boldsymbol{\Pi}^* = \boldsymbol{\Pi}^{(0)}$

For $m = 0,1, \ldots, M$

1) Convert data $\boldsymbol{x}_i$ to the number transitions between states and initial state, $N_i(s_1, s_2)$ and e_{is}.

2) Update $\boldsymbol{\theta}^*$, $\boldsymbol{\Pi}^*$ by the updating rule in (3.9).

3) Update $\mathbf{C}^*$ by the updating rules in (3.13).

End for

Output: $\{\boldsymbol{\theta}^*, \boldsymbol{\Pi}^*, \mathbf{C}^*\}$.

Figure 3-2: Procedure of the proposed algorithm for parameter estimation.

Remark 1: The optimization problem in (3.4) is a nonconvex optimization problem with a set of constraints. It can be solved by the heuristic algorithms including the genetic algorithm (GA) and simulated annealing algorithm (SA). However, the heuristic algorithms search for the optimal solution over whole feasible region by random search which will result in large computational complexity especially for problems with a large number of continuous variables (Rylander 2001; Yang 2012). In addition, these algorithms cannot theoretically guarantee to always produce a global optimal solution, unless the number of samplings tends to (practically impossible) infinity (Xu 2003). Since the problem in (3.4) can be decomposed into two convex optimization problems by solving for the parameters in canonical models and the membership vectors alternatively, existing line search methods in convex optimization problem such as the gradient based algorithms, the quasi-Newton methods can be applied to solve each sub-problem. They start from an initial point and search for a local optimal solution along the search directions which are more efficient than exploiting all possible solutions. To avoid being triggered in a local optimal solution, the quasi-Newton methods with multiple starting points are developed (Renders and Flasse 1996).

However, these algorithms are not developed for the problems with nonnegativity constraints, leading to the possibility that the optimal solution found from these algorithms may be infeasible. A nonlinear operator $[\cdot]_\varepsilon = \max\{\cdot, \varepsilon\}$ is often used on the quasi-Newton algorithm to enforce nonnegativity of solution in each iteration (Zdunek and Cichocki 2006). However, the convergence of the quasi-Newton algorithm may not be preserved after incorporating the operator. The multiplicative update rules hold a good balance between speed and ease of implementation for solving the nonnegative optimization problems (Lee and Seung 2001). It can be viewed as a variation of the gradient based algorithm which automatically decides the step-size to guarantee the nonnegativity and normalization of parameters in each update iteration (Lee and Seung 2001). Therefore, we drive the multiplicative update rules for solving the optimization problem in (3.4) to guarantee a local optimal solution that satisfies the nonnegativity and normalization constraints can be founded.

Remark 2: Individuals may have irregular measurements under the limited sensing capacity. To apply the proposed method in (3.1) to update the individual models under irregular measurements, we further account for the effects of missing values in the collaborative learning using the EM algorithm. Based on the current estimation of each individual Markov model, we first impute the missing values by estimating the expected number of transitions between each pair of states in one monitoring epoch, $\mathrm{E}[N_i(s_1, s_2)]$. We then update the individual Markov models by using the latest estimation of number of transitions in problem (3.1). The two-step process is repeated until the transition matrices stabilize. To estimate the expected number of transitions in one monitoring epoch, we consider the real observations as well as the estimations for the missing epochs. Given the transition matrix, $\mathbf{P}_i$, the probabilities of transiting from s_1 to s_2 after t_{miss} cycles can be predicted as:

$\Pr\left(X_{it+t_{miss}} = s_2 \middle| X_{it} = s_1\right) = \sum_{x_{t+1},\ldots,x_{t+t_{miss}-1} \in S} \mathbf{P}_i(s_1, x_{t+1})\mathbf{P}_i(x_{t+1}, x_{t+2}) \cdots \mathbf{P}_i(x_{t+t_{miss}-1}, s_2)$,

Each element $\mathbf{P}_i(s_1, x_{t+1})\mathbf{P}_i(x_{t+1}, x_{t+2}) \cdots \mathbf{P}_i(x_{t+t_{miss}-1}, s_2)$ represents a possible path of health progression in the missing epochs. The expected number to have followed this path can be calculated as:

$$\mathrm{E}\left[N_i\left(s_1, x_{t+1}, \ldots, x_{t+t_{miss}-1}, s_2\right)\right] = N_i^{t_{miss}}(s_1, s_2)\frac{\mathbf{P}_i(s_1, x_{t+1})\mathbf{P}_i(x_{t+1}, x_{t+2}) \cdots \mathbf{P}_i(x_{t+t_{miss}-1}, s_2)}{\Pr\left(X_{it+t_{miss}} = s_2 \middle| X_{it} = s_1\right)}$$

where $N_i^{t_{miss}}(s_1, s_2)$ represents the number of transitions from s_1 to s_2 after t_{miss} epochs. This path involves the transitions in one epoch $(s_1, x_{t+1}), \ldots, \left(x_{t+t_{miss}-1}, s_2\right)$. Therefore, the expected number of transitions in the one monitoring epoch can be obtained by eliminating all possible paths.

3.3.2 *A selective sensing strategy driven by prognosis*

With $\mathbf{P}_i$, the risk of each unit can be derived at each epoch t as the probability of transiting from the last observed state x_{iT_i} to abnormal state S, where T_i denotes the time of last observation. Formally, denote r_{it} as the risk of unit i, it can be predicted as $r_{it} = \Pr\left(X_{it} = S \middle| X_{iT_i} = x_{iT_i}\right) = \mathbf{P}_{\boldsymbol{i}}^{(t-T_i)}(x_{iT_i}, S)$, where $\mathbf{P}_{\boldsymbol{i}}^{(t-T_i)}$ is a multi-step transition probability matrix of each unit i. Ranking and selection strategy ranks the units according to the descending order of their risks and selects the top C units at each epoch. However, the ranking and selection strategy only works when the dynamic model of each unit could be accurately determined. To be more specific, note that for any sensing choice that determines which $\delta_{it} = 1$, the total risk of the selected units is $\sum_{i=1}^{N} \delta_{it} r_{it}$, under the constraint that $\sum_{i=1}^{N} \delta_{it} = C$. An optimal sensing strategy should maximize $\sum_{i=1}^{N} \delta_{it} r_{it}$. However, ranking and selection strategy optimizes for $\sum_{i=1}^{N} \delta_{it} \hat{r}_{it}$ which may be inefficient when $\hat{r}_{it}$ is not accurately determined. This is particularly true for our problem since for each unit the training data is sparse. To mitigate this problem, the similarity information between individuals

provides an opportunity to robustify the sensing strategy. We then incorporate a regularization term on the decision variables into the sensing formulation, $\sum_{i,j=1}^{N} w_{ij}\left(\delta_{it}-\delta_{jt}\right)^2$, to encourage that the sensing decisions made on similar units should be similar as well. This leads to the following optimization problem

$$\max_{\boldsymbol{\delta_t}\in\{\mathbf{0},\mathbf{1}\}^{\boldsymbol{N}}} \sum_{i=1}^{N} \delta_{it} r_{it} - \lambda \sum_{i,j=1}^{N} w_{ij}\left(\delta_{it}-\delta_{jt}\right)^2 - \eta\|\boldsymbol{\delta_t}\|_0 \,, \tag{3.14}$$

where λ and η are two tuning parameters that control the similarity and sparsity among selections. Larger tuning parameters impose more effect of regularizations on the selections. Note that (3.14) uses l_0 norm as an equivalent expression to replace the constraint $\sum_{i=1}^{N} \delta_{it} = C$. Given C, a larger tuning parameter η will be used to force the number of selections does not exceed C. Without a given C, the tuning parameters can be chosen by minimizing the Akaike information criterion (AIC) (Akaike,1974). When $\lambda = 0$, it can be observed that the solution of (3.14) reduces to the ranking and selection strategy.

Since the decision variables are binary, the objective function in (3.14) can be written as:

$$\min_{\boldsymbol{\delta_t}\in\{\mathbf{0},\mathbf{1}\}^{\boldsymbol{N}}} \sum_{i=1}^{N} \delta_{it}(\eta - r_{it}) + \lambda \sum_{i,j=1}^{N} w_{ij}\left(\delta_{it}-\delta_{jt}\right)^2 \,, \tag{3.15}$$

To solve the problem in (3.15), we reformulate it as a s/t min-cut problem on a graph [Leighton and Rao; 1999]. Note that $\left(\delta_{it}-\delta_{jt}\right)^2$ equals to 1 if $\delta_{it} \neq \delta_{jt}$, and it is 0 otherwise. The regularization term can be represented as a cut-function on the graph defined by the similarity matrix:

$$\sum_{i,j=1}^{N} w_{ij}\left(\delta_{it}-\delta_{jt}\right)^2 = \sum_{i,j=1}^{N} w_{ij}\delta_{it}\left(1-\delta_{jt}\right) = \sum_{i\in\Omega}\sum_{j\notin\Omega} w_{ij} \,, \tag{3.16}$$

where Ω represents the set of selected units. Let's further introduce two artificial units, $s \in \Omega$ and v $\notin$ Ω, while the decision variables of them are fixed to be $\delta_{st} = 1$ and $\delta_{vt} = 0$. Define the coefficients w_{si} and w_{iv} as:

$$w_{si} = \begin{cases} r_{it} - \eta \;\; if\; r_{it} \geq \eta \\ 0 \quad otherwise \end{cases},$$

$$w_{iv} = \begin{cases} \eta - r_{it} \;\; if\; r_{it} \leq \eta \\ 0 \quad otherwise \end{cases}$$

then, the first term of the objective function in (3.15) can also be written as a cut-function:

$$\sum_{i=1}^{N} \delta_{it}(\eta - r_{it}) = \sum_{i=1}^{N} w_{sj}\delta_{st}(1 - \delta_{it}) + \sum_{i=1}^{N} w_{iv}\delta_{it}(1 - \delta_{vt}) + \sum_{\substack{i=1 \\ r_{it} \geq \eta}}^{N} (\eta - r_{it}) , \quad (3.17)$$

where $\sum_{i=1, r_{it} \geq \eta}^{N} (\eta - r_{it})$ is a constant.

Thus, the problem in (3.15) can be formulated as a s/t min-cut problem on a graph

$$\min_{\boldsymbol{\delta_t} \in \{\mathbf{0},\mathbf{1}\}^{\boldsymbol{N}}} \sum_{i=1}^{N} w_{sj}\delta_{st}(1 - \delta_{it}) + \sum_{i=1}^{N} w_{iv}\delta_{it}(1 - \delta_{vt}) + \sum_{i,j=1}^{N} w_{ij}\delta_{it}\big(1 - \delta_{jt}\big), \quad (3.18)$$

The graph structure of the corresponding s/t-min-cut problem is shown in Figure 3-3, which can be solved by applying the maximal flow algorithm (Gallo et al., 1989) on the graph.

Remark: In addition to solving the selective sensing problem under the constraint of maximum number of measurements, the proposed method is also flexible to be extended to the situations where some units are more costly to be measured than others and the total cost of measurements is constrained. Assume the cost of measuring unit i as ω_i, and the total cost of measurements in each time epoch is represented as $\sum_{i=1}^{N} \omega_i \delta_{it}$. The sensing constraint is then imposed on the total cost of measurements, i.e. $\sum_{i=1}^{N} \omega_i \delta_{it} \leq C$. Therefore, the units with high risk and low cost are more likely to be measured. The optimization problem in (2.14) can be revised by maximizing the total risk, minimizing the total cost, and encouraging the similarity between selections simultaneously, which is represented as:

$$\max_{\boldsymbol{\delta_t} \in \{0,1\}^{N}} \sum_{i=1}^{N} \delta_{it} r_{it} - \beta \sum_{i,j=1}^{N} w_{ij}\big(\delta_{it} - \delta_{jt}\big)^2 - \eta \sum_{i=1}^{N} \omega_i \delta_{it},$$

The tuning parameter, η, controls the total cost of measurements. This formulation can still be reformulated as a s-t min-cut problem by defining the similarities between artificial units and observed units as:

$$w_{u_1 i} = \begin{cases} r_{it} - \eta\omega_i \;\; if\; r_{it} \geq \eta\omega_i \\ 0 \quad otherwise \end{cases},$$

$$w_{iu_2} = \begin{cases} \eta\omega_i - r_{it} \;\; if\; r_{it} \leq \eta\omega_i \\ 0 \quad otherwise \end{cases}.$$

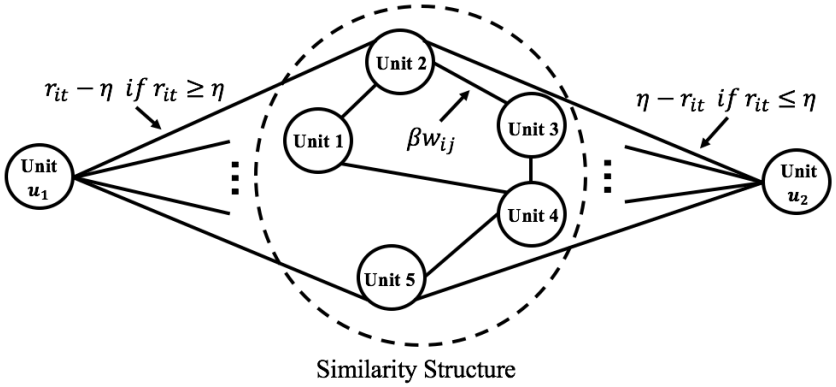


Figure 3-3: Graph of the s/t-min-cut formulation of selective sensing problem (Azencott et al., 2013).

3.3.3 *Empirical issues of implementing the algorithm*

To implement our method, we need to define the number of canonical models K and similarity matrix **W**. First, we'd like to point it out that this task doesn't involve as many free parameters as it looks like. Particularly, for similarity information, despite its appearance as a matrix, the free parameter is only one when we use the heat kernel weighting method to calculate the similarity between two patients, i.e., a scaling parameter σ^2 in defining w_{ij} as $w_{ij} = \exp(-\left\|\boldsymbol{z}_i - \boldsymbol{z}_j\right\|^2/\sigma^2)$ for individuals i and j. For the number of canonical models, sometimes they are readily available through querying prior knowledge. For example, in the two applications we will show later, three

subgroups in Alzheimer's disease population and five depression progression patterns found in literature are used as prior information to determine the number of canonical models. When the prior information is not available, we could obtain the optimal number of canonical models using model selection methods such as AIC/BIC and cross-validation (Chapter 2). It is also possible to have doctors' knowledge to help define the similarity between patients. It is not uncommon that the physician's assessment of the disease type and stage is often used as complementary information to the quantitative measurements to group similar patients together. If prior knowledge is not available in this regard, we could also quantify the similarity between units based on units' covariates, denoted as $\boldsymbol{z}_i$, that reflect the characteristics of the unit. For example, the environmental factors and operational parameters in engineering applications are often used for degradation modeling, and the demographic, social-economical, genetic and imaging information in many healthcare applications are available to identify the similarity between patients. To convert the covariates into similarity, we adopt some existing approaches (Belkin and Niyogi 2001), including the 0-1 weighting, Heat Kernel Weighting, and Dot-Product Weighting. In our numerical studies on both synthetic datasets (omitted due to space limit) and real-world dataset, we note that the heat kernel weighting method consistently leads to satisfactory results. One possible reason is that, the heat kernel weighting method allows optimal tuning by introducing a scaling parameter σ^2 in defining w_{ij} as $w_{ij} = \exp(-\left\|\boldsymbol{z}_i - \boldsymbol{z}_j\right\|^2/\sigma^2)$ for individuals i and j. Other implementation issues include how to obtain initial values of membership vectors and canonical models. Empirical evidences in the real-world data analysis (Section 3.5) suggest that the Markov models learned from a group of units can provide good initial value. The groups can be identified by applying clustering algorithms such as the K-means method on the covariates. The group label of each unit can be used to initialize its membership vector.

3.4 SIMULATION STUDY

In this section, we perform a simulation study to evaluate the effectiveness of proposed method. We first construct K canonical Markov models that characterize different progression processes, which each Markov model has five states. Without loss of generality, in what follows we present our stimulation study with $K = 3$, while the three canonical models correspond to progression processes of fast degrading units, slow degrading units, and healthy units. Corresponding to these canonical types, the transition probability matrices, $\{\boldsymbol{\Pi}_1, \boldsymbol{\Pi}_2, \boldsymbol{\Pi}_3\}$, and initial distributions, $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3\}$, for these canonical models are designed (the values of these parameters are provided in Appendix B). Then, we generate the membership vector for each unit using a Gaussian mixture distribution with three components. Specifically, the three multivariate normal distributions defining the Gaussian mixture distribution is shown in below:

$$F_1(\boldsymbol{c}) = N\left(0, \begin{bmatrix} \upsilon^2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right), F_2(\boldsymbol{c}) = N\left(0, \begin{bmatrix} 1 & 0 & 0 \\ 0 & \upsilon^2 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right), F_3(\boldsymbol{c}) = N\left(0, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \upsilon^2 \end{bmatrix}\right),$$

where υ^2 is a parameter to tune the significance of the latent structure, i.e., a larger magnitude of υ^2 in F_i corresponds to a more dominant i^{th} element in $\boldsymbol{c}$, leading to a more significant latent structure ($\upsilon = 5$ and $\upsilon = 10$ are considered in our study). For each unit, we first randomly choose one of the three distributions based on a predefined probability vector $\boldsymbol{\omega} = [\omega_1, \omega_2, \omega_3]$, e.g., $\boldsymbol{\omega} = [0.3, 0.4, 0.3]$ is used in our study. Then, the initial value of membership vector of this unit is obtained by randomly sampling from the chosen distribution. Note that, the initial value is further normalized to obtain a membership vector (i.e., as a membership vector, the sum equals to 1). Using the canonical models and membership vector, the individual Markov chain can be generated from the transition probability matrix $\mathbf{P}_i = \sum_k c_{ik}\boldsymbol{\Pi}_k$ and initial distribution $\boldsymbol{\nu}_i = \sum_k c_{ik}\boldsymbol{\theta}_k$. To simulate noise, we add random errors sampled from the uniform distribution $\mathcal{U}(0, \varepsilon)$ on each

element of the individual transition probability matrix $\mathbf{P}_i$. A larger ε will result in more noise in the simulated data ($\varepsilon = 0.05$ and $\varepsilon = 0.2$ are considered in our study). In total, we simulate data for 500 subjects over 15 consecutive observation epochs. We only use the first 5 observations of each unit as training data to estimate the prognostic model, and then, deploy a sensing method to decide which observations should be observed in the following 10 observation epochs under the sensing constraint C ($C = 30\%$ and $C = 50\%$ are considered in our study). Thus, the measurements from observation epoch 6 - 15 will give us ground truth information about the risk of the individuals, laying the foundation for performance evaluation of different sensing methods. We further simulate a set of covariates from a multivariate normal distribution to measure the similarities between individuals. To preserve the correlation between individuals, we use the correlation coefficients between membership vectors as the covariance matrix for the multivariate normal distribution.

We first evaluate the effectiveness of proposed formulation in (3.4) by comparing the statistical properties (bias and KL-divergence) of estimated Markov models from the proposed method with the maximum likelihood estimators under different canonical structures. The maximum likelihood estimators can be obtained by solving the problem in (3.4) with λ fixed at 0. As shown in Table 3-1, the incorporation of similarity information enhances the estimation of parameters leading to lower bias and divergence from the true parameters, and the advantage is consistent under different number of canonical models. Then we compare the selective sensing (SS) with the ranking and selection (RK) method under the canonical structure with 3 canonical models. Since both methods need prognosis results of the units in each epoch to derive decisions regarding which units to be monitored in the next epoch, we consider the CM based prognosis method as well as two other methods, such as the individual-specific monitoring (ISM) that estimates a prognostic model for

each patient independently, and the population-level monitoring (PSM) that aggregates individuals into homogeneous groups based on their similarities and estimates prognostic models on group level. We choose the optimal values of input parameters in CM model, including the number of canonical models, K, and the tuning parameter, λ, using the AIC and leave-one-out cross-validation (LOOCV) techniques, in which the true number of canonical models, $K = 3$, is discovered. The results are included in Appendix B. We further compare these prognosis based methods with the random selection method, which randomly chooses a subset of units to be measured in each monitoring epoch. We evaluate the performance of random selection method based on 100 repeats to avoid the randomness from single trial. As shown in Table 3-2, there are six combinations of prognosis and sensing methods. We evaluate the performance of these models based on two criteria, the prediction accuracy and the detection accuracy. For prediction accuracy, we calculate the mean square error (MSE) and Pearson's correlation coefficient between the predicted risks and real risks, e.g. $MSE(\hat{\boldsymbol{r}}, \boldsymbol{r}) = \sum_{i=1}^{N}(r_i - \hat{r}_i)^2/N$. For detection accuracy, we calculate the detection rate as the percentage of high-risk units (i.e., developed severe depression) being selected and the average real risk of the selected units.

The average performance over 10 monitoring epochs of these methods is summarized in Table 3-2. Note that we have conducted experiments on a diverse range of parameter settings. Due to page limit, here in Table 3-2 we present the results for two levels of sensing capacity $\mathbf{C}$, significance of latent structure $\boldsymbol{\upsilon}$, and effect of noise $\boldsymbol{\varepsilon}$, etc. We also calculate the improvements of performance using the proposed method (selective sensing based collaborative modeling) over the performances of other methods. Due to page limit, we present the smallest and greatest improvements of prediction and detection accuracies under each scenario in Table 3-2. It can be observed that: 1) The proposed selective sensing (SS) method is effective on detecting individuals

progressing fast towards high risk by exploiting the similarity between individuals as demonstrated by the improved detection accuracy (detection rate and average risk) of SS + ISM vs. RK + ISM, SS + PSM vs. RK + PSM, and SS + CM vs. RK + CM . 2) The proposed collaborative modeling (CM) method is effective on exploiting the low-dimensional canonical structure as demonstrated by the improved performance of RK+CM compared to RK+ISM and RK+PSM. 3) SS + CM outperforms other methods on both prediction accuracy and detection accuracy, which shows that the proposed method can efficiently incorporate the similarity information between individuals to enhance both prognostics and sensing resource allocation. 4) SS + CM is somehow robust to noise as increasing the noise in simulated data would not affect its performance as dramatically as others. 5) The advantages of SS + CM are generally more evident when the latent structure is more significant in the simulated data.

We conducted more studies to evaluate how sensitive is the result with respect to different initial values of parameters including the number of canonical models, similarity structure, and length of monitoring in Figure 3-4. To simulate different similarity structures, we incorporate random noise sampled from $N(0, v_{noise}^2)$ on the similarity matrix. The variance, v_{noise}^2, controls the degree of noise, i.e., smaller variance means better estimation of the similarity structure. The prediction accuracy and detection accuracy of the proposed method are measured by the MSE and detection rate, respectively. The prediction accuracy is sensitive to the number of canonical models and it is improved when more canonical models are used, and the detection accuracy of the proposed method is less sensitive to the number of canonical models by exploiting the similarity structure in sensing strategy design, as demonstrated in Figure 3-4 (a). As shown in Figure 3-4 (b), both prediction accuracy and detection accuracy decrease as the degree of noise increases and the performance under low sensing capacity is more sensitive to the similarity structure. This indicates

that the identification and prediction of severely degrading units rely on a good similarity structure. Since the simulation study assumes a good prior knowledge of the similarity matrix at the initial of monitoring, the performance of the proposed method is less sensitive to the length of monitoring, as demonstrated in Figure 3-4 (c).

Table 3-1: Statistical properties of CM model under different settings of parameters ($\zeta = 5$, $\varepsilon = 0.05$).

	MLE			CM		
	$K = 1$	$K = 3$	$K = 5$	$K = 1$	$K = 3$	$K = 5$
Bias	1.179	3.762	4.808	1.179	0.882	0.999
KL divergence	0.319	0.239	0.252	0.319	0.214	0.181

Table 3-2: Comparison of different sensing methods on simulated data ($K = 3, N = 500$).

M^S = 10%, $\zeta = 5$, $\varepsilon = 0.05$

	SS CM	RK CM	SS ISM	RK ISM	SS PSM	RK PSM	Random Selection
MSE	**0.096**	0.104	0.173	0.163	0.112	0.113	--
Correlation	**0.627**	0.607	0.487	0.518	0.529	0.527	--
Detection Rate	**0.257**	0.252	0.227	0.166	0.199	0.179	0.101
Average risk	**0.875**	0.835	0.856	0.778	0.858	0.794	0.399

M^S = 30%, $\zeta = 5$, $\varepsilon = 0.05$

	SS CM	RK CM	SS ISM	RK ISM	SS PSM	RK PSM	Random Selection
MSE	**0.091**	0.093	0.158	0.158	0.107	0.113	--
Correlation	**0.706**	0.705	0.557	0.556	0.657	0.647	--
Detection Rate	**0.587**	0.583	0.343	0.342	0.545	0.528	0.300
Average risk	**0.768**	0.767	0.732	0.728	0.745	0.726	0.396

M^S = 10%, $\zeta = 5$, $\varepsilon = 0.2$

	SS CM	RK CM	SS ISM	RK ISM	SS PSM	RK PSM	Random Selection
MSE	**0.107**	0.112	0.170	0.161	0.110	0.121	--
Correlation	**0.526**	0.510	0.431	0.464	0.484	0.451	--
Detection Rate	0.215	0.213	0.149	0.125	**0.221**	0.213	0.100

Average risk	0.755	0.751	0.764	0.732	**0.758**	0.748	0.357

M^S = 30%, $\zeta = 5$, $\varepsilon = 0.2$

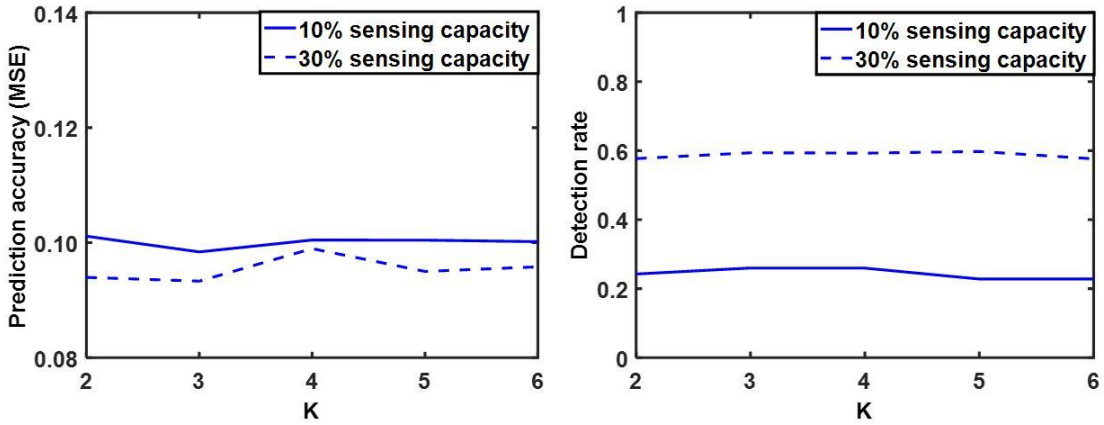
	SS CM	RK CM	SS ISM	RK ISM	SS PSM	RK PSM	Random Selection
MSE	0.102	**0.100**	0.148	0.148	0.110	0.106	--
Correlation	0.645	**0.654**	0.531	0.531	0.628	0.640	--
Detection Rate	**0.556**	0.542	0.315	0.315	0.545	0.531	0.301
Average risk	**0.645**	0.639	0.618	0.618	0.640	0.628	0.357

M^S = 10%, $\zeta = 10$, $\varepsilon = 0.2$

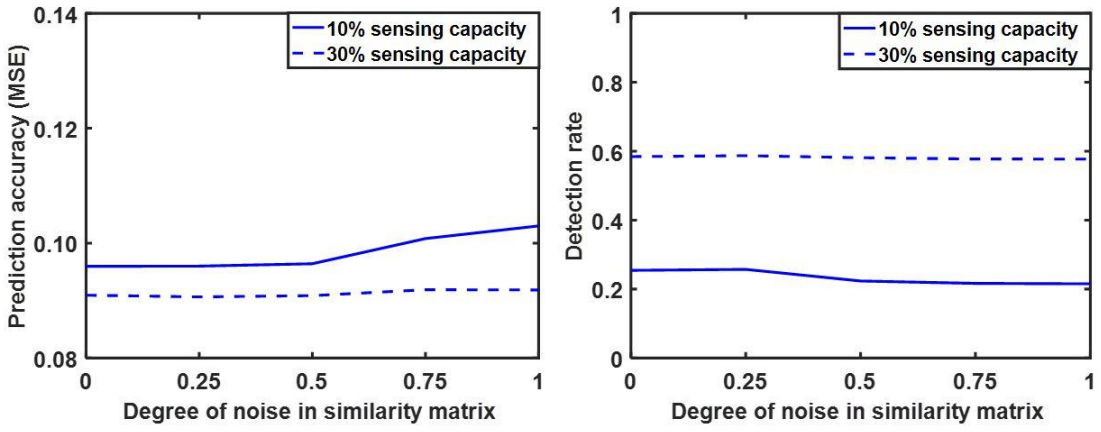
	SS CM	RK CM	SS ISM	RK ISM	SS PSM	RK PSM	Random Selection
MSE	**0.115**	0.124	0.191	0.184	0.120	0.122	--
Correlation	**0.528**	0.517	0.387	0.410	0.468	0.494	--
Detection Rate	**0.212**	0.209	0.210	0.172	0.217	0.192	0.100
Average risk	**0.835**	0.801	0.768	0.735	0.838	0.759	0.376

M^S = 30%, $\zeta = 10$, $\varepsilon = 0.2$

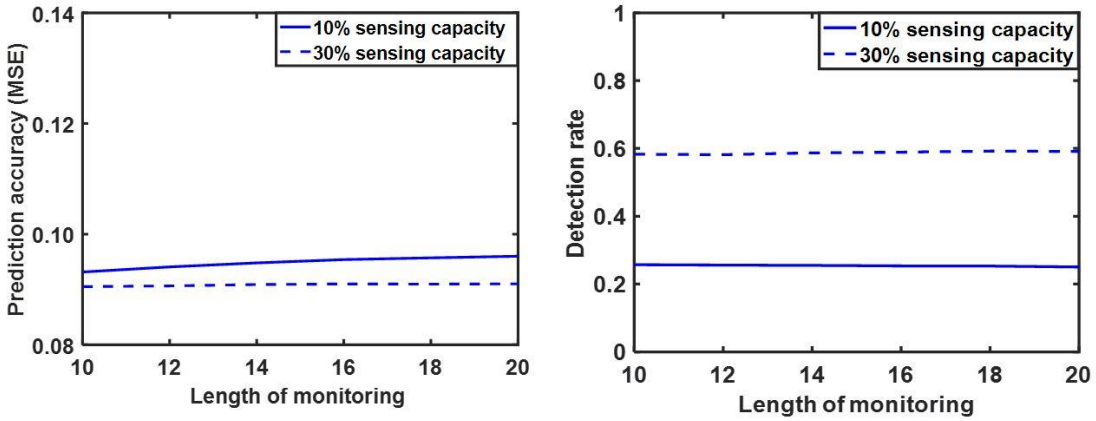
	SS CM	RK CM	SS ISM	RK ISM	SS PSM	RK PSM	Random Selection
MSE	**0.091**	0.094	0.153	0.157	0.109	0.115	--
Correlation	**0.698**	0.687	0.539	0.529	0.651	0.628	--
Detection Rate	**0.579**	0.574	0.353	0.363	0.529	0.517	0.302
Average risk	**0.724**	0.712	0.608	0.602	0.681	0.666	0.376



(a)



(b)



(c)

Figure 3-4: Prediction accuracy and detection accuracy of the proposed method under different values of parameters on the simulated data ($\zeta = 5$, $\varepsilon = 0.05$). (a) Under different number of canonical models (K); (b) Under different similarity structure; (c) Under different length of monitoring.

3.5 APPLICATION IN DEPRESSION MONITORING

In this section, we explore our method's effectiveness on monitoring a depression population that is under treatment. Depression is one of the most common mental illnesses in primary care with complex and heterogeneous progression mechanisms (CDC, 2012). Due to potential side

effects of the medication, the Food and Drug Administration (FDA) emphasizes that patients taking antidepressants should be closely monitored (NIMH). Inadequate follow-up monitoring of patient's treatment outcomes has been identified as a main challenge in depression treatment. Our depression data came from one of the largest depression dataset in U.S. provided by the Mental Health Research Network (MHRN). The dataset includes personal-level longitudinal depression measurements and corresponding measurement times in the EHR data between years 2007 and 2012. The depression treatment outcome is assessed by the score of Patient Health Questionnaire (PHQ)-9, a self-administrated questionnaire that includes 9 multiple-choice questions (Lowe et al., 2004). The PHQ-9 scores range from 0 to 27, and it stratifies depression into 5 levels including no depression (0-4), mild depression (5-9), moderate depression (10-14), moderate severe depression (15-19), and severe depression (20-27). Since the PHQ-9 scores assess patients' depression levels in the past two weeks, we monitor the patients biweekly. We consider the total monitoring time period as 20 biweeks and study a population of 610 patients that are frequently measured within this period. We first impute the missing values on each patient to ensure the measurements cover the whole monitoring window following the approach in Section 2.8. We then transform the PHQ-9 scores into 5 depression states defined above and estimate the transition probability matrix between these states. Risk is defined as the probability of progressing to severe depression (i.e., PHQ-9 is within 20-27) from the patient's current depression state.

To investigate performance of different sensing methods, we use the first 5 measurements of the individuals in this dataset to initiate the patient monitoring. The remaining measurements will give us ground truth information about the risk of the individuals, laying the foundation for performance evaluation of different sensing methods. To measure the similarities information between individuals, we extract a set of exploratory features for each patient's measurements,

including the first PHQ-9 score, last PHQ-9 score, mean, median, standard deviation, 1st and 3rd quantiles of PHQ-9 scores, and coefficients in the spline model estimated from the patient's longitudinal PHQ-9 measurements. The number of canonical models is specified at five since a five-subgroup pattern has been found in the depression populations (Gunn et al., 2013). We apply the proposed approach under different levels of sensing capacity ranging from low (10%) to high (30%), i.e., what percentage of patients we could acquire measurement in each epoch.

We compare the selective sensing (SS) based methods with the ranking and selection (RK) based methods. For detection accuracy, we calculate the detection rate as the percentage of high-risk units (i.e., developed severe depression) being selected and the average real risk of the selected units. The prediction and detection accuracy of these methods are summarized in Table 3-3. Clearly, the selective sensing (SS) method outperforms the ranking and selection (RK) method on all performance metrics under both high and low sensing capacities. And the CM based prognosis further enhanced the advantage of SS. We further evaluate how sensitive is the result with respect to different values of parameters. The prediction accuracy is sensitive to the number of canonical models and it has best performance under 5 canonical models. Different number of features are used to generate various similarity structures. The prediction accuracy is more sensitive to the number of features used to measure similarities while the detection accuracy is more robust to it. Since the similarities between high risk units (whose depression trajectories are stable high) tend to be captured by all types of features extracted from the depression measurements, these high risk units are consistently selected under different number of features, leading to the robust detection accuracy. Without a good prior knowledge of the health progression and similarity structure at the initial of monitoring, the prediction and detection accuracy are improved under longer monitoring length. The results are shown in Appendix B.

Table 3-3: Comparison of different sensing methods on depression monitoring.

10% Sensing Capacity

	SS CM	RK CM	SS ISM	RK ISM	SS PSM	RK PSM	Random Selection
MSE	**0.097**	0.103	0.124	0.123	0.109	0.111	--
Correlation	**0.488**	0.468	0.386	0.386	0.425	0.412	--
Detection Rate	**0.397**	0.393	0.350	0.321	0.396	0.384	0.100
Average risk	**0.600**	0.592	0.597	0.559	0.593	0.566	0.203

30% Sensing Capacity

	SS CM	RK CM	SS ISM	RK ISM	SS PSM	RK PSM	Random Selection
MSE	**0.095**	0.096	0.108	0.117	0.099	0.102	--
Correlation	**0.610**	0.608	0.488	0.427	0.589	0.585	--
Detection Rate	**0.827**	0.826	0.672	0.534	0.727	0.714	0.293
Average risk	**0.563**	0.562	0.420	0.322	0.552	0.541	0.200

3.6 CONCLUSION

In this chapter, we proposed a selective sensing framework to monitor a heterogeneous population of individuals under limit sensing resources. The proposed method integrates Markov models, collaborative learning, and selective sensing into a unified framework. We characterized individuals' depression progressions as Markov chains and predicted the risk to severe depression using collaborative modeling (CM) by exploiting the latent canonical structure embedded in the population and incorporating the similarity information between individuals. We formulated the selective sensing as an optimization problem to optimally allocate the sensing resources to detect a subgroup of high risk individuals at each monitoring period. Sensing results were further incorporated into the CM to update the prognosis for all individuals. Through comprehensive simulation studies, we found that the proposed method had better performance than alternative sensing methods on both risk prediction and high-risk patient detection across a wide range of settings. We also demonstrated the utility and efficiency of the proposed method on a real-world

application for depression monitoring. It is known that 30 million Americans currently use antidepressant medication. Inadequate follow-up monitoring has been identified as a main challenge in managing the depression patient population. To fill in this need, our method provides patient-specific adaptive monitoring schedules and dynamically allocates limited sensing resources to detect high risk individuals of severe depression. We applied the proposed method to monitor 610 subjects with ongoing depression treatment from the Mental Health Research Network dataset which demonstrated promising results. In the future, we plan to apply the proposed collaborative prognostic and selective sensing framework to other applications in healthcare and industry, and integrate it with other prognostic models such as the inverse Gaussian process and hidden Markov model for monitoring different populations with different dynamic health evolution processes.

Chapter 4. LONGITUDINAL PATTERN BASED PROGNOSTIC MODEL VIA RULE-BASED METHOD

4.1 INTRODUCTION

Understanding the factors associated with the risk of disease onset is an important step towards identifying the eventual healthcare needs of different individuals within a population (Greenland et al., 2004). This lays the foundation to develop and deliver appropriate resources to the right targets, called “tailored health interventions”. Evidence suggests that individuals prefer tailored care to a standardized care that is designated for the average population (Evans, 1996; Radwin and Alster, 2002; Ryan and Lauver, 2002; Whittemore, 2000). Therefore, health professionals need to identify the subgroups of individuals characterized by different patterns of risk factors.

The extended use of Electronic Health Record (EHR) provides an abundance of clinical measurements that may indicate patients' disease severities. Therefore, statistical analysis of the EHR data has the potential to identify risk predictive factors for disease progression and provide accurate prognostics for patient's outcome. But models for predicting individual's disease severity from EHR data are inadequately developed due to the following challenges. As disease progression is a complex and dynamic process, understanding its etiology requires repeated clinical measurements over time rather than relying only upon a baseline profile (Pfeiffer et al., 2015). Identifying the predictive factors for disease severity from time-varying and irregular clinical measurements poses a significant challenge. Furthermore, on top of the lack of sufficient measurement on predictive factors, the widely-reported heterogeneity on disease progression further increases this challenge. For instance, five broad trajectory patterns have been found in a depression treatment population from a U.S. EHR dataset (Lin et al., 2016). Rather than examining subpopulations with distinct disease trajectory patterns, existing research on disease prognostics focuses on identifying the risk factors that are associated with the outcome of interest, which are essentially global associations on the population level (King et al., 2008; Kendler et al., 1993; Huang et al., 2014). Logistic regression (King et al., 2008), structural equation modeling (Kendler et al., 1993) and regression analysis (Huang et al., 2014) are commonly used methods to model the association between depression progression and risk factors. Consequently, risk factors identified from these models only reflect the average effects over a population and are inadequate to be used for monitoring strategy design on the individual level.

The aim of this study is to establish a rule-based analytic framework to identify a set of risk-predictive longitudinal patterns from the EHR data for personalized disease prognostics and monitoring and apply this framework in the context of depression monitoring. Rule-based analysis

is particularly useful for identifying a set of risk patterns that segment the population into subgroups, with individuals in the same group share similar patterns that influence the outcome of interest (Lin et al., 2014). A rule describes the range of value on one or more risk factors that either indicates increased or decreased risk for disease. Rules generated on time-varying risk factors provide some natural semantics to define the risk-predictive longitudinal patterns of a subset of individuals while each rule may indicate a specific low-risk signal or warning signal for monitoring. By identifying the unknown rules from observations, a set of risk-predictive rules can be considered as a set of medical signals, providing us with a personalized risk estimation by looking into the risk patterns endorsed by each individual. By integrating the risk-predictive rules with monitoring frequencies, a set of rule-based monitoring strategies can be developed for monitoring of individuals with warning signals and less frequently monitoring of individuals with disease-free signals.

Specifically, our method integrates data transformation, rule discovery, and rule evaluation by following the steps in Figure 4-1. We first transform the EHR data of each individual to his/her disease severity assessment and measurements of a set of risk predictors. Then, we randomly split the data into training and testing data. We discover a set of rules on the training population and further investigate the risk levels of subgroups endorsing/non-endorsing these rules on both training and testing populations. Association between the identified risk-predictive rules and individual disease severities is further studied on the testing population.

This chapter is organized as follows: The data introduction and transformation is presented in Section 4.2. The rule discovery process using RuleFit is presented in Section 4.3. The rule evaluation process is described in Section 4.4. The results of applying proposed methods on the depression monitoring are analyzed in Section 4.5. Section 4.6 summarizes the conclusion and

discussion of this chapter.

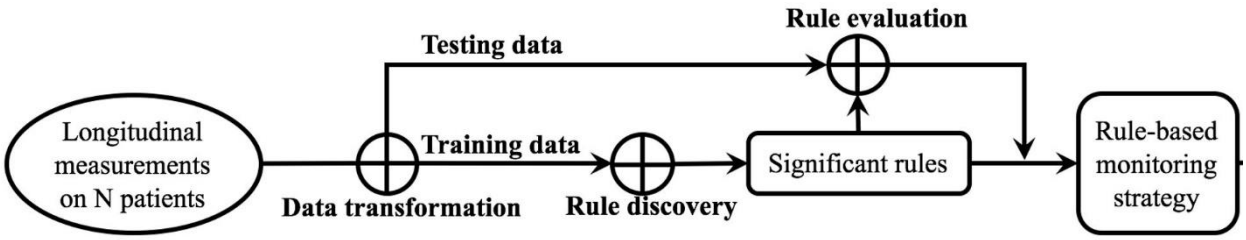


Figure 4-1: The flowchart of the rule-based analytic framework for longitudinal pattern discovery and adaptive monitoring.

4.2 DATA DESCRIPTION AND TRANSFORMATION

4.2.1 *Data description*

The Mental Health Research Network (MHRN) data are drawn from the electronic health records of four health systems participating in the Mental Health Research Network (HealthPartners, and the Colorado, Washington, and Southern California regions of Kaiser Permanente) (Simon et al., 2013). The data includes person-level longitudinal depression measures in EHR between years 2007 and 2012. The depression level is assessed by the Patient Health Questionnaire (PHQ-9), a self-administrated questionnaire with 9 multiple-choice questions (Kroenke, 2001). The data also includes relative time between PHQ-9 measures, treatment status, type of providers (primary care, specialist, mental health) where the questionnaire was conducted, individuals' age, sex, and the Charlson comorbidity score (a standard indicator of medical disease burden).

We focus our study on 9,306 individuals receiving ongoing treatment (defined as mental health diagnosis, mental health medication, or mental health specialty visit within the prior 180 days, excluding those completing the PHQ-9 for depression screening). The time window of 6 months (180 days) has been used as standard check point for depression monitoring in practice. Within the

ongoing treatment population, 1,762 individuals are regularly monitored that have at least three observations in the first 6 months and are followed up in the next 6 months. We conduct the analysis on this group by randomly separating 1,200 individuals (68% data) into a training set to build the prognostic model and identify a set of predictive rules. Then we validate the prediction capability and association with underlying disease severity of the identified rules on the remaining individuals (testing data).

4.2.2 *Data transformation*

We build a prognostic model based on measurements within a 6-month time window to predict an individual's depression severity in the following 6 months. For each individual, we translate the measurements in the first 6 months (referred as the predicting period) and the following 6 months (referred as the responding period) to predictive factors and depression severity, respectively. In total, we generate 45 factors (shown in Table C-2) to provide statistical summarizations, characterize progression trajectories, and capture non-random longitudinal patterns for the time stamped information during the predicting period. These 45 factors mainly come from three categories, e.g., statistical summarizations, progression trajectories, and non-random longitudinal patterns, which are introduced below.

4.2.2.1 Statistical summarizations

For each individual, we summarize the longitudinal PHQ-9 scores, Charlson comorbidity scores and the 9th question scores on suicide ideation in terms of their statistical measurements, which include first value, maximal value, minimal value, range, median value, 25% and 75% percentile values, average value, and volatility (i.e. standard deviation) of the longitudinal PHQ-9 scores. The PHQ-9 scores can be further segmented into five depression levels using the conventional

classification including minimum (0-4), mild (5-9), moderate (10-14), moderately severe (15-19) and severe (20-27) (Simon et al., 2013). We also summarize the percentage of PHQ-9 scores in each depression level.

4.2.2.2 Progression trajectories

As demonstrated in Section 2., changes of PHQ-9 scores can be predictive for depression progression. To extract these temporal patterns, we consider the deepest decrease and deepest increase between two consecutive PHQ-9 scores together with the volatility of the difference between nearby PHQ-9 scores. We further use the time stamp of the first observation, number of observations, observing density and the latest PHQ-9 score, to describe the irregularity and sparsity of the individual's EHR data. The observing density is defined as the number of observations divided by the time span of all observations.

In addition to the factors mentioned above, we adopt control theory to further extract some non-random longitudinal patterns in individual PHQ-9 scores. Control charts, such as the R chart, S chart and moving range chart, and control rules including the Western Electronic rules (WE) and Nelson (NC) rules are commonly used to capture the non-random patterns in system monitoring[26]. Given the sequential measurements of a dynamic process, a control chart is built to monitor the process over time. It is constructed by a central line for the average value and the upper and lower lines for the control limits which can be determined from historical data. The control rules including the WE rules and NC rules listed in Table 4-1 can be applied on the control chart to reveal the non-random patterns for the early detection of anomalies in the process. To extend the control rules for capturing non-random patterns in EHR data, we use the moving range control chart which is developed for individual measurements. Denote the measurements on an individual as $[x_1, \dots, x_T]$, the control chart is constructed by the mean value of observations $\bar{x}$, the average of

moving range $\overline{\mathrm{MR}} = \frac{\sum_{i=2}^{T}|x_i - x_{i-1}|}{T-1}$, and the estimated standard deviation of observations $\hat{\sigma} = \overline{MR}/d_2$, where d_2 is a constant that represents the expected value of the moving range of T normal observations when standard deviation equals to 1. Thus the moving range control chart is constructed by

$$\text{Central limit} = \bar{x}$$

$$\text{Upper control limit} = \bar{x} + 3\hat{\sigma}$$

$$\text{Lower control limit} = \bar{x} - 3\hat{\sigma}$$

To detect non-random patterns, control rules listed in Table 5-1 are applied on the control charts, and a set of binary factors that indicating the observations of non-random patterns on each individual is generated. The control chart on a random selected individual is shown in Figure 4-2, in which the control rules WE1, WE2, WE8, NC1 and NC2 are satisfied.

After eliminating the control rule factors that are supported by less than 1% of the population, we generate 45 risk predictive factors in total, and provide the summarization of them in Appendix C.

Table 4-4-1: The Western Electric and Nelson control rules

Control Rules	Content
Western Electronic Rule 1 (WE1)	One point falls above the upper 3σ limit.
Western Electronic Rule 2 (WE2)	Two out of three consecutive points fall above the upper 2σ limit.
Western Electronic Rule 3 (WE3)	Four out of five consecutive points fall above the upper 1σ limit.
Western Electronic Rule 4 (WE4)	Eight consecutive points fall above centerline.
Western Electronic Rule 5 (WE5)	One point falls below the lower 3σ limit
Western Electronic Rule 6 (WE6)	Two out of three consecutive points fall below the lower 2σ limit.
Western Electronic Rule 7 (WE7)	Four out of five consecutive points fall below the lower 1σ limit.
Western Electronic Rule 8 (WE8)	Eight consecutive points fall below centerline.

Western Electronic Rule 9 (WE9)	Fifteen consecutive points fall between lower and upper 1σ limits.
Western Electronic Rule 10 (WE10)	Eight consecutive points fall beyond the 1σ limits.
Nelson Rule 1 (NC1)	Night consecutive points fall on the same side of centerline
Nelson Rule 2 (NC2)	Six consecutive increasing or decreasing points.
Nelson Rule 3 (NC3)	Fourteen consecutive points alternate up and down

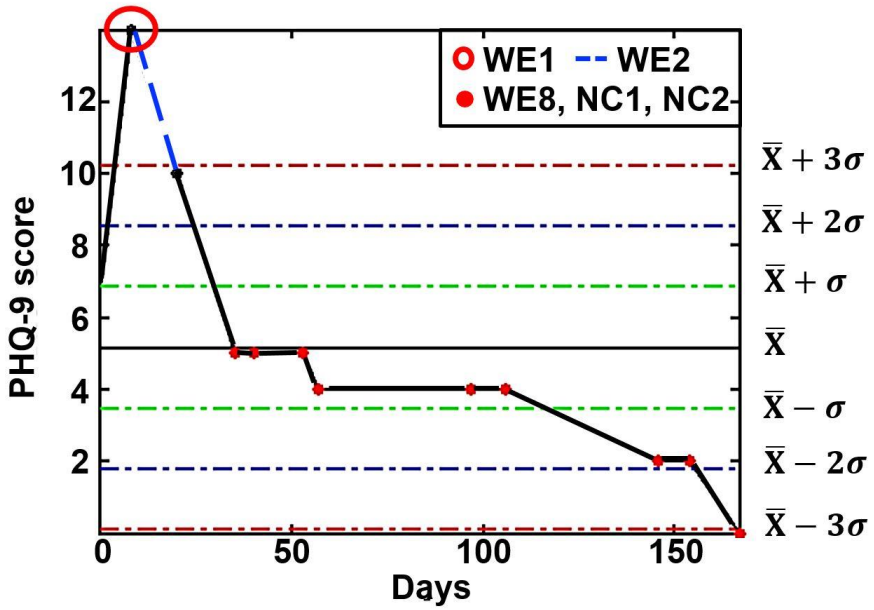


Figure 4-4-2: The moving range control chart on individual measurements. The measurements satisfy control rules: WE1, WE2, WE8, NC1, and NC2.

After eliminating the control rule factors that are supported by less than 1% of the population, we generate 45 risk predictive factors in total, and provide the summarization of them in the Appendix C. The depression severity of each patient can be assessed using the average PHQ-9 score in his/her responding period. Based on the conventional classification of PHQ-9 scores, patients with average PHQ-9 score lower than 10 can be regarded as low-risk group while the others can be regarded as depressive group (Kroenke et al., 2001). To distinguish the low-risk patients with depressive ones, we define the outcome, Y_i, as 1 if the patient i, i =1,…,N, is in low-risk group in the following 6 months and 0 otherwise.

4.3 RULE DISCOVERY USING RULEFIT

We assume individuals can be categorized into homogeneous risk groups by a set of underlying longitudinal patterns, which are characterized as rules over factors extracted from time stamped measurements. For example, individuals endorsing a rule consisting of the latest PHQ-9 score greater than 17 and the volatility of PHQ-9 score lower than 7 are predicted to have higher probability of depression onset in the next six months. We use RuleFit (Friedman and Popescu, 2008) to discover the hidden rules that may be predictive of the disease risk. RuleFit is a high-dimensional computational algorithm for rule discovery, which is capable of exhaustively searching for potential rules on a large number of candidate risk markers. It has two phases, the "rule generation phase" and "rule pruning phase": 1) **Rule generation**: At this stage, random forest (Breiman, 2001) is used to exhaustively search for candidate rules over the potential risk factors. Random forest is a high-dimensional rule discovery approach that extends traditional decision tree models. Specifically, a random forest estimates a number of trees, with each tree being estimated on a relatively homogenous subpopulation generated by bootstrapping the original dataset. Since each tree employs a set of rules to characterize a subpopulation, the random forest is actually a comprehensive collection of rules that are able to characterize the whole dataset. On the other hand, as a heuristic and exhaustive search approach, the random forest may produce a large number of less-predictive or redundant rules, which requires the following second step to refine the learning results. 2) **Rule pruning**: As the random forest will generate many rules that can be redundant or irrelevant to early withdrawal due to overfitting, the sparse regression model (Fiedman and Popescu, 2008; Tibshirani, 1996) will be applied to select a minimum set of risk-predictive rules, by using all the potential rules as predictors and the withdrawal status as the outcome. The sparse regression model is a high-dimensional variable selection model (Fiedman

and Popescu, 2008; Tibshirani, 1996). Considering each rule as a "variable", rule pruning is essentially a variable selection problem. This problem is to selecting a subset of rules out of a pool of q candidate rules, denoted as $R = \left[R_1, R_2, \ldots, R_q \right]$, which are predictive to the output variable $\mathbf{Y}$. This problem is particularly challenging in high-dimensional settings where q is large. Recently, the *Least Absolute Shrinkage Selection Operator* (LASSO) is proposed (Tibshirani, 1996), which is a sparse linear regression model that is capable to identify a subset of relevant variables out of a huge list of candidate variables. Specifically, the formulation of LASSO is

$$\min_{\beta} \| \mathbf{Y} - \mathbf{R}\boldsymbol{\beta} \|_2^2 + \lambda \| \boldsymbol{\beta} \|_1 .$$

Here, the square error term, $\| \mathbf{Y} - \mathbf{R}\boldsymbol{\beta} \|_2^2$, is used to measure the model fit. The L1-norm penalty term $\| \boldsymbol{\beta} \|_1$, defined as the sum of the absolute values of all elements in $\boldsymbol{\beta}$, is used to measure the complexity of the regression model. The user-specified penalty parameter, λ, aims to achieve an optimal balance between the model fitness and model complexity – larger λ will result in sparser estimate for $\boldsymbol{\beta}$. Efficient algorithms have been developed to solve the optimization problem (Fiedman and Popescu, 2008; Tibshirani, 1996). In our study, since the output variable $\mathbf{Y}$, i.e., the withdrawal status, is a binary variable, the sparse logistic regression (Tibshirani, 1996) is a better choice than linear regression, which can be readily implemented in the R package of RuleFit (Tibshirani, 1996). More details on RuleFit can be found in (Tibshirani, 1996).

In a summary, RuleFit is computationally efficient since efficient algorithms have been developed for both Random Forest and sparse linear regression models. Since it is an integration of random forest and LASSO, it has several important parameters to be specified, including the number of trees, the complexity of the trees that is controlled by the average number of terminal nodes, and the penalty parameter λ. According to the extensive simulation studies performed in

(Tibshirani, 1996), the default parameters values for the number of trees and the average number of terminal nodes are 333 and 4, respectively. We obtained the optimal values of these three parameters using the automated cross-validation procedure in Rulefit in a manner of grid search, which are close to these default values, e.g., the number of trees and the average number of terminal nodes are 250 and 4.5, respectively. In our experiments, we have found that the RuleFit is robust to the specification of these parameters.

A set of predictive rules are discovered where each individual rule can segment the population into subgroups with distinct disease severities. Specifically, we calculate the proportion of low-risk patients (Y_i=1) in each rule endorsing group. Then, we select the rules that have high prevalence of either high-risk patients or low-risk patients in their endorsing groups. In other words, the rules that lead to endorsing groups which are equally mixed with high-risk and low-risk patients are not predictive and thereby discarded. We then name the rules as either increasing or decreasing risk rules. Specifically, the decreasing risk rules which have high proportion of low-risk patients in their endorsing groups indicate that the patients satisfying these rules are less likely to have depression in the next 6 months. The increasing risk rules, on the other hand, have low proportion of low-risk patients in their endorsing groups and indicate higher risk of depression when they are satisfied.

4.4 RULE EVALUATION

Rules identified from the previous step are predictive of depression severity on training population. To evaluate the statistical significance of these rules on testing population, we first test whether the depression severity in the rule endorsing and un-endorsing groups are significantly different.

We apply the Mann Whitney Wilcoxon test on each rule, which is a non-parametrical statistical test for deciding whether two independent samples come from the same distribution.

We further predict individuals' disease severities based on rule endorsements by applying the item response theory (IRT) (Embreston et al., 2013). IRT is often used in psychometric problems to infer students' abilities, attitudes, or personalities by gathering evidence from questionnaire responses or tests[27]. It assigns each individual a latent variable θ denoting the underlying disease severity, and models the likelihood of endorsement of each rule as a function of both the individual's disease severity and the association between this rule and the disease severity, which corresponds to information on where the rule stands in the disease severity continuum and how predictive the rule is. Denote the probability of an endorsement of rule R_l given the disease severity θ as $P_l(R_l = 1|\theta)$, and the probability of un-endorsement as $P_l(R_l = 0|\theta)$, the probability is modeled by a monotonically increasing function called the item characteristic curve (ICC) in the following form:

$$\log \frac{P_l(R_l=1|\theta)}{P_l(R_l=0|\theta)} = a_l(\theta - b_l),$$

where b_l is the item difficulty parameter that represents the disease severity required to achieve a 50% chance of endorsement of rule R_l, i.e. $P_l(R_l = 1|\theta = b_l) = P_l(R_l = 0|\theta = b_l) = 0.5$. a_l is the item discrimination parameter for R_l that determines the amount of change in the log odds, $P_l(R_l = 1|\theta)/P_l(R_l = 0|\theta)$, for one unit change in the disease severity. A larger b_l indicates the individuals endorsing rule R_l is more likely to have higher disease severity and a larger a_l means that rule R_l is more sensitive to the small changes in the disease severity. The parameters in item characteristic curves, $\{a_l, b_l, \theta\}$, can be estimated by the Markov Chain Monte Carlo (MCMC) algorithm[28]. We use the ICCs to associate the rule endorsements with underlying disease severity.

In addition to the prediction accuracy of depression severity, we further evaluate the efficiency of using these rules for adaptively monitoring each individual. The endorsement of each risk-predictive rule provides evidence for adaptive monitoring in the following 6 months by stratifying the individual's depression severity into different levels. Specifically, individuals endorsing the decreasing risk rules are less likely to have depression in the following period and can be less frequently monitored to save cost; while individuals satisfying the increasing risk rules should be closely monitored. Therefore, each rule can be extended to a rule-based monitoring strategy in the following 6 months by applying the policies outlined in Table 4-2: for each decreasing risk rule, individuals will be monitored if the rule is not endorsed and monitored otherwise; for each increasing risk rule, individuals will be monitored if the rule is endorsed and not monitored otherwise.

To evaluate the efficiency of rule-based monitoring, we compare several monitoring strategies by estimating the number of depressive patients (PHQ-9 $\geqslant$ 10) in the next 6 months that are correctly monitored, which is the number of true positives (TP). We assume under the status quo, all patients are monitored every 6 months, which lead to unnecessary monitoring of low-risk patients. We also consider a PHQ-9 based strategy, which monitors the patients if the last-period PHQ-9 score is 10 or greater. Under rule-based monitoring, we consider both using individual rules and combining all significant rules. For the individual rule based monitoring and multiple rules based monitoring, the decisions are made using Table 4-2. In the multiple rules based monitoring, we segment the population into four groups based on the endorsements of increasing and decreasing risk rules. By considering different decisions in the group of endorsing both increasing and decreasing risk rules and the group of unendorsing all rules, we compare 4 scenarios in the multiple rules based monitoring.

Table 4-4-2: a) Individual rule based monitoring strategies in the next 6 months. b) Multiple rules based monitoring strategies in the next 6 months.

(a)

Rule endorsement / Rule type	Endorsed	Unendorsed
Increasing risk rule	Monitor	Not monitor
Decreasing risk rule	Not monitor	Monitor

(b)

	Increasing risk rule	Decreasing risk rule	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Group 1 (2.3%)	unendorsed	unendorsed	Monitor	Not monitor	Not monitor	Monitor
Group 2 (45.6%)	endorsed	unendorsed	Monitor	Monitor	Monitor	Monitor
Group 3 (41.8%)	unendorsed	endorsed	Not monitor	Not monitor	Not monitor	Not monitor
Group 4 (10.3%)	endorsed	endorsed	Monitor	Not monitor	Monitor	Not monitor

Note: the population is segmented into four groups including endorsing any of increasing risk rules and any of decreasing risk rules, endorsing any of increasing risk rules but unendorsing all decreasing risk rules, endorsing any of decreasing risk rules but unendorsing all increasing risk rules, and unendorsing all rules (percentage of patients in each group is presented in brackets).

4.5 APPLICATION IN DEPRESSION

4.5.1 *Rule discovery*

We apply the RuleFit model on the training data and identify 46 rules. By calculating the proportion of low-risk patients in each rule endorsing group, we select 12 most predictive rules listed in Table 4-3, whose proportion of low-risk patients is lower than 30% or greater than 70%. The proportions of low-risk patients in rule endorsing groups as well as the overall proportions of low-risk patients in training and testing populations are summarized in Figure 4-3. The identified rules are distinguished into 6 decreasing risk rules and 6 increasing risk rules. For example, in the

first decreasing risk rule (Rule 1), if the observations in a 6-month window of an individual have deepest increase between consecutive PHQ-9 scores smaller than 7.5 and 75 percentile of PHQ-9 scores smaller than 14.62, the individual is less likely to progress to depression in the next 6 months. It can be observed that age, sex and the latest PHQ-9 score are significant risk factors in the identified rules, which are consistent with the significant risk factors identified from the logistic model, as shown in Appendix C. However, rule-based model is more powerful in capturing the interactions between significant risk factors and their critical ranges which improve the predictability and interpretability than the logistic regression model. In addition, the identified rules include rich information for describing the depression trajectories. For example, the rule of latest PHQ-9 score greater than 17 and volatility of PHQ-9 score smaller than 7 indicates stable high PHQ-9 scores in the depression trajectory.

Table 4-4-3: 12 rules identified from the RuleFit model.

	Decreasing risk rules		**Increasing risk rules**
Rule 1	Deepest increase between consecutive PHQ9 scores < 7.50 & 75 percentile of PHQ9 score<14.62	**Rule 7**	Observing frequency > 0.03 & Minimal PHQ9 score> 8.50
Rule 2	25 percentile of PHQ9 score < 6.13 & Volatility of PHQ9 score< 9.64	**Rule 8**	Minimal PHQ9 score > 9.50 & Volatility of difference between nearby PHQ9 scores < 4.75
Rule 3	75 percentile of PHQ9 score< 15.88 & Percentage of moderate depression < 0.39	**Rule 9**	Latest PHQ9 score > 17.50 & Volatility of PHQ9 score < 7.33
Rule 4	Deepest decrease between consecutive PHQ9 scores > 2.50 & 75 percentile of PHQ9 score < 14.12	**Rule 10**	Minimal PHQ9 score > 6.50 & 75 percentile of PHQ9 score > 14.88
Rule 5	Sex is male & Mean of 9^{th} question scores < 0.71 & Percentage of moderately severe < 0.38	**Rule 11**	Age < 65 & Percentage of severe depression > 0.23

Rule 6	Latest PHQ9 score < 8.50 & Maximal PHQ9 score < 16.50	**Rule 12**	Deepest decrease between consecutive PHQ9 scores < 13.50 & Mean of PHQ9 scores > 14.73

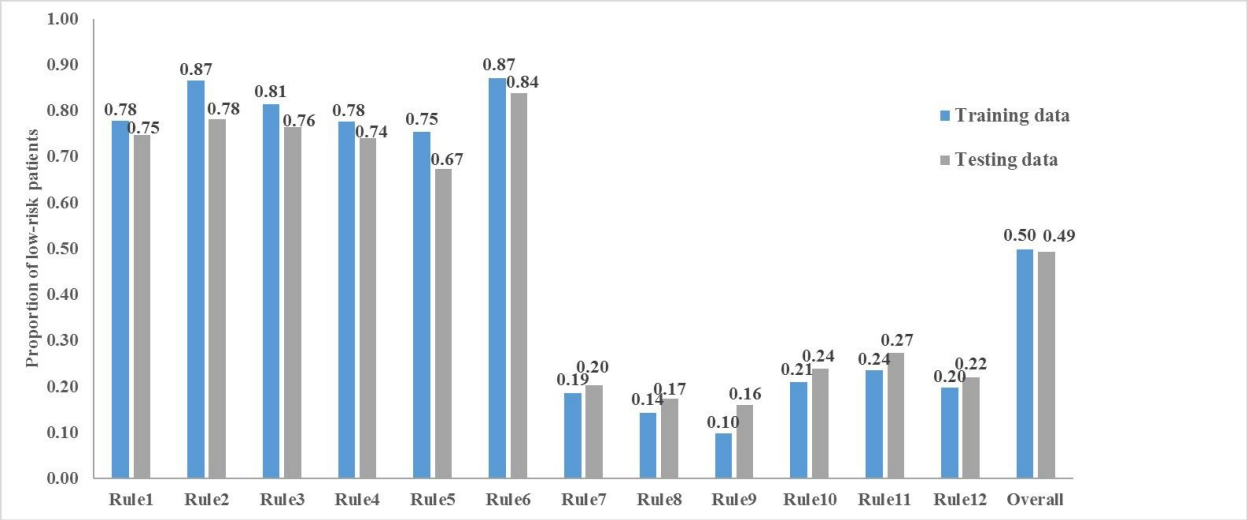


Figure 4-4-3: Proportion of low-risk patients (average PHQ-9 score < 10) in rule endorsing group.

4.5.2 *Rule evaluation*

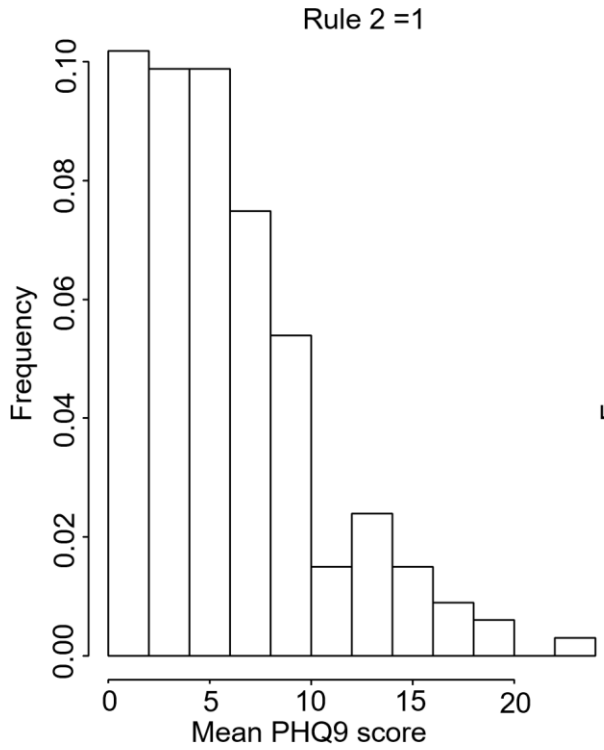
The distributions of depression severity in the rule endorsing and un-endorsing groups are presented in Figure 4-4. The distributions are plotted for a randomly selected decreasing (increasing) risk rule. It can be observed that most patients who endorse the decreasing risk rule tend to have lower depression severity than the patients who do not endorse the rule, and vice versa for the increasing risk rule. P-values of the test statistics on all rules are lower than the significant level (0.01), which indicates each rule segments the population into subgroups with significantly different depression severities.

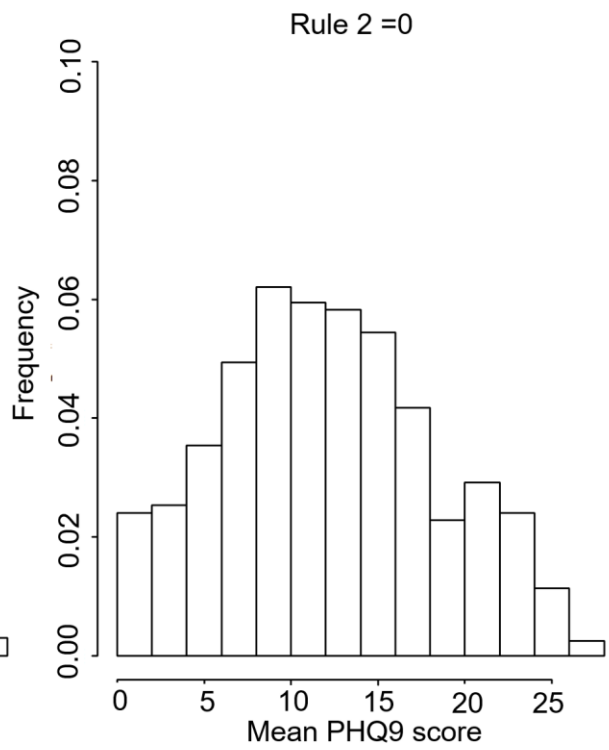
To evaluate the goodness of prediction, we compare the prediction accuracy of the rule-based prognostic model using 12 significant rules with the RuleFit, logistic regression, and Support Vector Machine (SVM) models trained on all 45 factors or the significant factors included in the

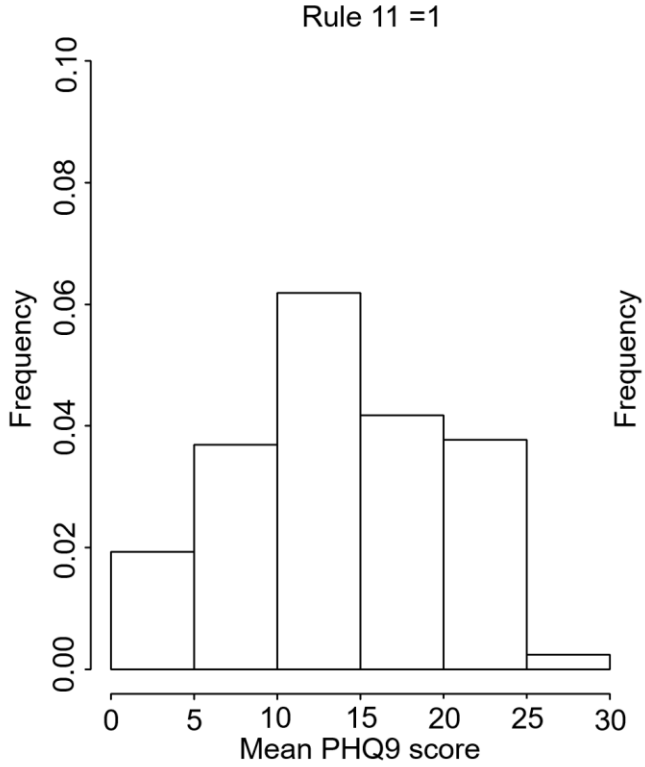
significant rules. The prediction accuracies of these models are evaluated using the average area under the curve (AUCs) on the testing population. The results are summarized in Table 4-4. The rule-based prognostic model outperforms the other methods which demonstrates the prediction capability of significant rules. We also note that the subset of factors included in the significant rules are predictive of health outcome.

We further investigate the association between the rules and depression severities by using the item characteristic curve (ICC). The probability of endorsing each rule based on the estimated disease severity is calculated. The disease severity of each patient is normalized between 0 and 1 with larger value indicating more severe depression. The associations between rule endorsements and disease severities of 12 rules are plotted in Figure 5-5; the decreasing rules are drawn by blue lines and increasing risk rules are drawn by red lines. It can be observed that the decreasing risk rules are more likely to be satisfied than the increasing risk rules when the individual has low severity of depression; while the endorsement of increasing risk rules is more likely to be seen when the individual is severely depressed. The distributions of disease severities estimated from the rule-based prognostic model in high-risk ($Y_i = 0$) and low-risk ($Y_i = 1$) groups are compared using the boxplot in Figure 4-6, where the high-risk group has higher disease severities than the low-risk group. Overall, endorsement of the increasing and decreasing risk rules are predictive of the individuals' depression severity.

The monitoring outcomes of various strategies using testing data are shown in Figure 4-7 and Table 4-5. It can be observed that the multiple-rules based monitoring strategies in scenario 1 (MultiRule 1) and scenario 3 (MultiRule 3) outperform the latest PHQ-9 score based strategy (PHQ-9 Based) by providing higher sensitivity and specificity.


Rule 2 =1
Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0
5
10
15
20
Mean PHQ9 score


Rule 2 =0
Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0
5
10
15
20
25
Mean PHQ9 score


Rule 11 =1
Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0
5
10
15
20
25
30
Mean PHQ9 score

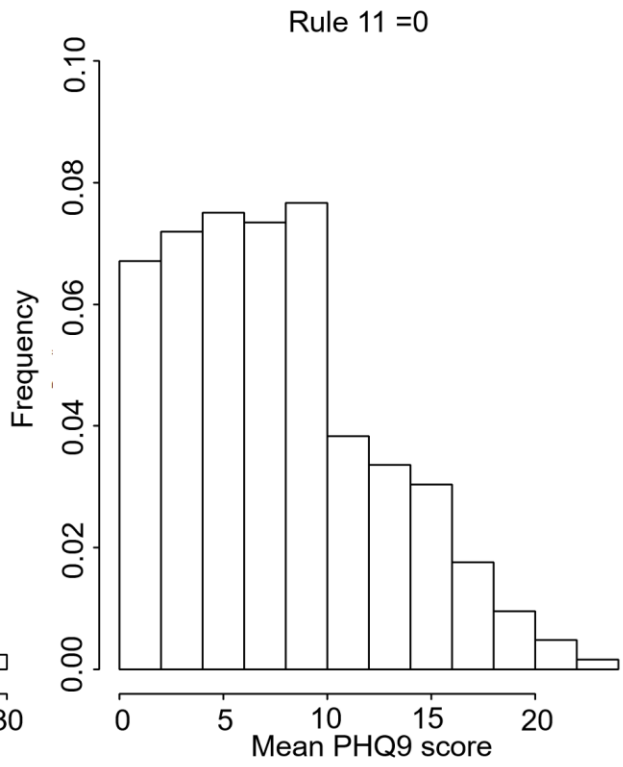

Rule 11 =0
Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0
5
10
15
20
Mean PHQ9 score

Figure 4-4: Distribution of average PHQ-9 scores in rule endorsing (=1) and un-endorsing groups (=0). Plot on randomly selected decreasing risk rule 2 (top) and increasing risk rule 11 (bottom).

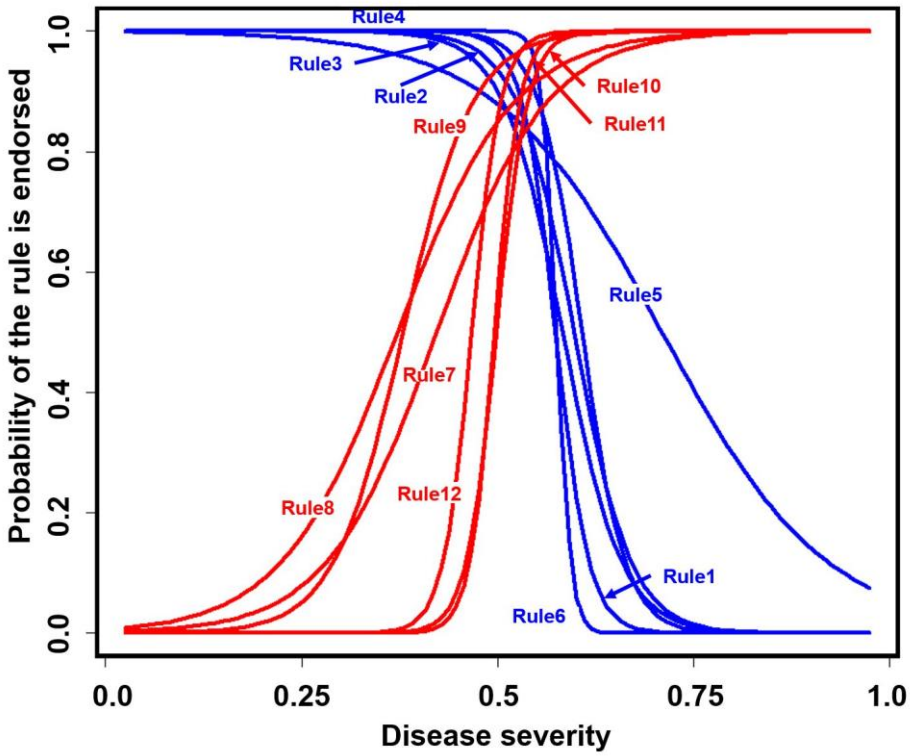


Figure 4-5: The associations between rule endorsements and depression severity of 12 rules. Each curve represents the probabilities of endorsing a rule under various depression severity. The increasing risk rules are plotted in red and the decreasing risk rules are plotted in blue.

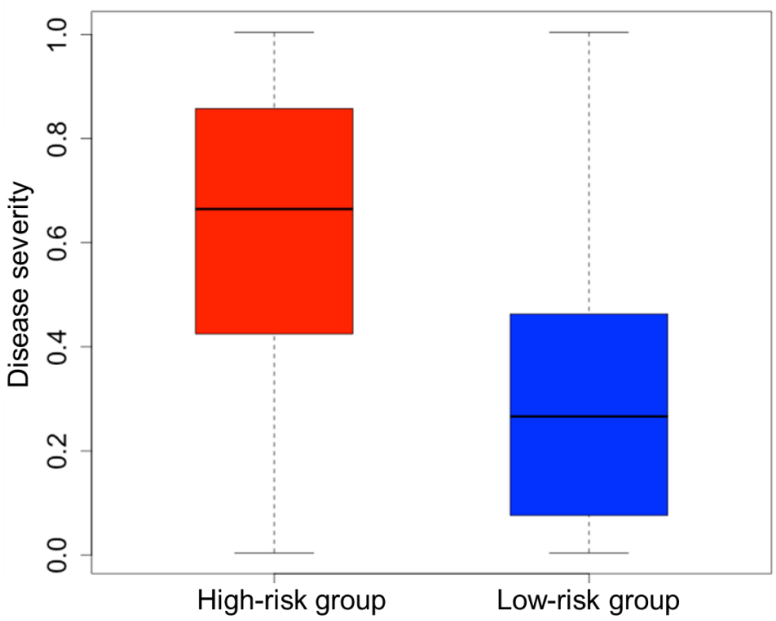


Figure 4-6: The boxplots of disease severities in high-risk and low-risk groups.

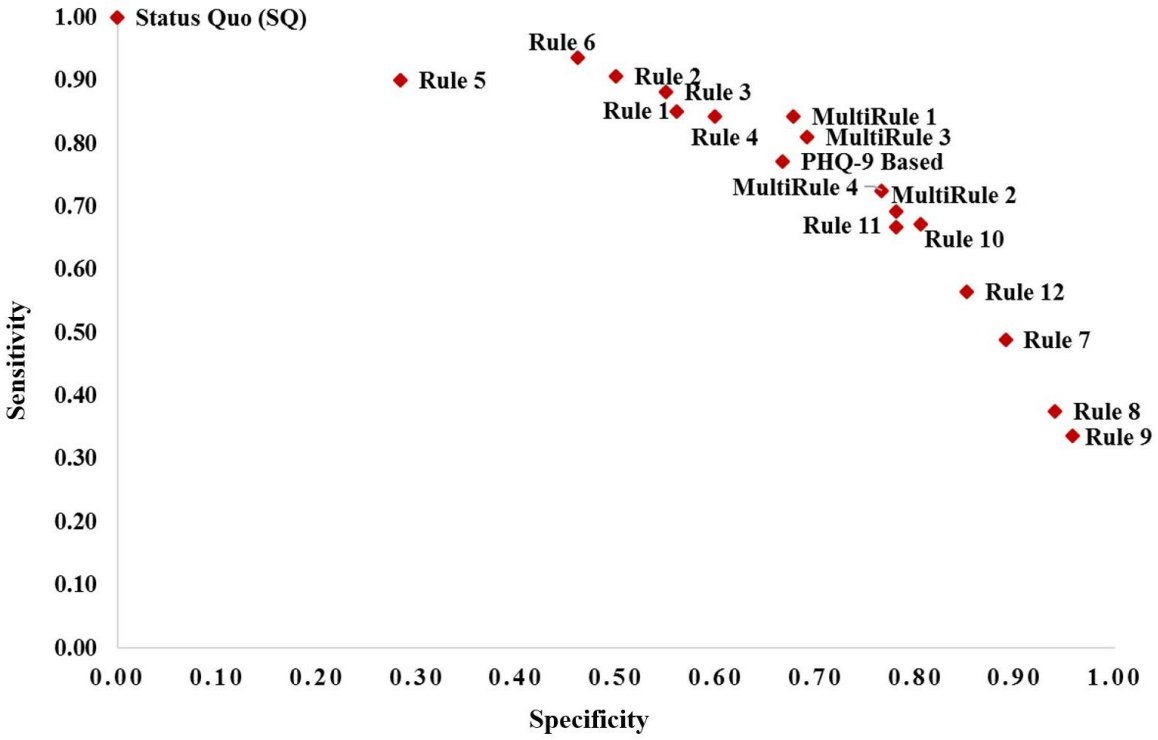


Figure 4-7: Comparison of sensitivities and specificities in different monitoring strategies.

Table 4-4-4: Prediction accuracy of several methods on testing data.

Model	Rule-based prognostic model	RuleFit		Logistic regression		SVM	
		All factors	Significant factors	All factors	Significant factors	All factors	Significant factors
AUC	0.83	0.82	0.82	0.81	0.81	0.81	0.81

Table 4-4-5: Monitoring outcomes of various strategies (N=562 patients).

Strategy	Sensitivity	Specificity
Rule 9	0.34	0.96
Rule 8	0.38	0.94
Rule 7	0.49	0.89
Rule 12	0.56	0.85
Rule 10	0.67	0.80
Rule 11	0.67	0.78
MultiRule 2	0.69	0.78
MultiRule 4	0.73	0.77
PHQ-9 Based	0.77	0.67
MultiRule 3	0.81	0.69
MultiRule 1	0.84	0.68
Rule 4	0.84	0.60
Rule 1	0.85	0.56
Rule 3	0.88	0.55
Rule 2	0.91	0.50
Rule 6	0.94	0.46
Rule 5	0.90	0.28
Status Quo (SQ)	1.00	0.00

4.6 DISCUSSION

We establish a rule-based analytic framework to identify the longitudinal patterns for predicting disease severity in a heterogeneous population and provide actionable knowledge to support the design of adaptive monitoring strategy. We apply this framework on a depression treatment population and demonstrate that the rule-based method can efficiently identify risk-predictive

longitudinal patterns from sparse and irregular measurements of depression severity (i.e. PHQ-9 score), demographic factors (age and sex), and Charlson comorbidity scores within a 6-months monitoring period. 12 longitudinal patterns are found to be predictive of depression severity in the following 6 months.

The 12 identified rules include time-varying measurements such as PHQ-9 scores as well as time invariant measurements such as age and sex. These risk factors have been identified as significant predictors for depression progression in the literature (Simon et al. 2013; Cole et al. 2013; Anstey et al. 2007; Piccinelli et al. 2000), which provide validation for our rule-based analysis. However, existing studies only reflect the average effects of risk factors over the whole population and ignore their interactions. Rule-based analysis is superior at capturing complex interactions between risk factors and further characterizes depression progression in each risk group. For instance, a two-years follow-up study on 352 patients responding to treatment of major depression (MD) had found that females were more likely than males to experience a MD in any month of the study, and marginally more likely to experience a relapse (Oquendo et al. 2013). In addition, the depression trajectories characterized by historical measurements of the PHQ-9 questionnaire have been found predictive of depression progression (Lowe et al. 2004; Simon et al. 2001). But studies considering the interactions between sex and depression trajectories are limited. Costello et al. used semi-parametric group-based modeling to explore risk and factors associated with trajectories of depressed mood from adolescence to early adulthood and found females were more likely to be classified into stable low depressed mood, early high depressed mood and late escalating depressed mood patterns versus no depressed mood (Costello et al. 2008). The associations discovered from this retrospective data analysis are inadequate to predict future depression progression. Complementary to evidence from the literature, our study finds males with

a lower than 0.71 mean score on the 9^{th} question and fewer than 38% observations being in moderately severe depression ($15 \leq$ PHQ-9 <20) are less likely to have depression, which demonstrates sex has impacts on the depression trajectory. Several studies have been conducted to examine depression onset in different age groups (Cole et al. 2013). However, there is a lack of agreement on the relationship between age and depression onset. A study on the trajectory of depression symptoms across 2,320 adults' life span has found that depressive symptoms were highest in young adulthood, decreased across middle adulthood, and increased again in older adulthood (Sutin et al. 2013). In our study, we find patients in young and middle adulthoods (Age < 65) having more than 23% severe depression measurements on their PHQ-9 records are more likely to be depressed in the future.

By evaluating the prediction capability of individual rules and the rule-based prognostic model in section 5.2, we further demonstrate that an individual rule is sufficient to segment population into groups with different risk levels and the rule-based prognostic model is capable of providing accurate personalized risk assessment by utilizing the relationships between rules and underlying disease severity. The identified rules also provide solid evidence for designing adaptive monitoring strategies in this ongoing treatment population by closely monitoring the patients with warning signals (i.e. endorse increasing risk rules) and less frequently monitoring the patients with healthy signals (i.e. endorse decreasing risk rules). By integrating the rule discovery, rule evaluation, and rule-based monitoring strategy design into a unified framework, the rule-based analytic method effectively translates the EHR data into actionable knowledge and enhances the efficient use of data driven evidence in medical decision making.

There are limitations of our research. First, our depression EHR data provide limited information on patients' socioeconomic and other clinical factors. As discovered in the literature, factors including race, education level, and family structure are likely to be associated with depression progression (Costello et al. 2008). Thus, additional data may lead to the discovery of more risk-predictive patterns and improving the accuracy of depression prognostics. Due to the lack of knowledge on treatment response scenarios and the lack of adequate depression assessments that cover the whole life-span of patients in the depression treatment population, it is challenging to estimate the long-term effectiveness of adaptive monitoring strategies.

In summary, we discovered 12 risk predictive rules from a depression treatment population that can segment individuals into risk subgroups based on their longitudinal patterns. We further developed and evaluated adaptive monitoring strategies based on these identified rules. We established a rule-based analytic framework to automatically translate the sparse, irregular and time-varying measurements in EHR data into evidence supporting the monitoring strategy design by integrating the data transformation, rule discovery and rule evaluation. The proposed method can lead to a better understanding of depression dynamics, more accurate prognostics of depression progression, and efficient monitoring of depression treatment population in clinical practice.

Chapter 5. COST-EFFECTIVENESS ANALYSIS FOR PROGNOSTIC-BASED DEPRESSION MONITORING

5.1 INTRODUCTION

Disease monitoring which is widely used to check on the progress or regress of disease and development of complications relies on the routine visit of at-risk individuals. For instance, the

Food and Drug Administration (FDA) recommends the patients taking antidepressant medications should be closely monitored every six months for preventing the side effects of antidepressant medications (NIMH). Due to the lack of considering significant heterogeneity in depression treatment outcome, the routine monitoring strategies always lead to inadequate follow-up monitoring on high-risk individuals and unnecessary monitoring of low-risk individuals (Kales et al., 2010). Therefore, the tailored health interventions which deliver appropriate resources to the right targets has the potential to enable cost-effective resources utilization in clinical practice.

Tailored health intervention is enabled by the growing availability of sensing and information technology such as the electronic heath record (EHR), and recent advances in prognostic models. The prognostic models have been applied to the massive sensing data for assessing the individual's risk of disease onset (Huang et al., 2014; Islam et al., 2013; Lasko et al., 2013; Lin et al., 2016; Sutin et al., 2013). They can be roughly classified into feature-based and trajectory based prognostic models. Feature-based prognostic models summarize the disease progression within a certain time period as a set of risk predictive features that capture the descriptive statistics, progression trajectory and abnormal progression patterns of longitudinal measurements (Huang et al., 2014; Lasko et al., 2013). The risk predictive features are further used to predict health outcome in machine learning models such as the logistic regression (Huang et al., 2014), LASSO (Huang et al., 2014) and neuron network (Lasko et al., 2013). However, these methods either inadequate to capture the heterogeneous disease progression processes or lack of interpretability. To mitigate these challenges, the rule-based method described in Chapter 4 is further developed to discover a set of longitudinal patterns from the risk predictive features and each pattern segments the population into subgroups with different risk indications. Another approach assesses the individual's risk by modeling the trajectory of disease progression (Islam et al., 2013; Lin et al.,

2016; Sutin et al., 2013). Markov process and Hidden Markov Model are widely used to explicitly describe the dynamic of disease progression. The heterogeneity in disease progression could be further captured by exploiting the latent structure in the population and similarity between the individuals, as shown in Chapter 3. The Markov models rely on the Markovian or memoryless assumption which focuses on the progression of disease between nearby time points. The feature-based model, on the other hand, exploit the long-term relationship between the current health outcome with its historical measurements. The disease progression calibrated by these statistical models are more likely to represent the average or smoothed behavior for each individual and ignore the stochastic changes of the measurements. An empiric natural history model is developed in (Alagoz et al., 2005) to capture the stochastic changes in the disease progression. It predicts the disease progression on an index patient by searching for the most similar patient in historical data and matching the disease progression on the similar patient to the index patient. Based on the simple matching process, the empiric natural history model is more sensitive to the noise in real observations.

Although the prognostic models provide assessments for individual's risk, using them for tailored health interventions still needs a seamless combination of data analysis and decision making. How to optimally allocate the monitoring resources to different risk groups is an operational challenge. Different prognostic models may lead to various monitoring strategies on the same individual. Which strategy is more efficient and whether prognostic-based monitoring strategies can lead to better disease intervention compared to the routine based monitoring are critical questions in tailored health intervention. However, existing evaluations for prognostic model mainly focus on the prediction accuracy of health outcomes or the net benefits resulted from critical events (Vickers, et al., 2006). Without considering the monitoring strategy associated with

each prognostic model as well as the cost and efficiency in long-term monitoring, existing methods are inadequate to be used for follow-up monitoring in clinical practice.

In this paper, we develop a decision support algorithm, which can automatically translate the sensing data into monitoring strategies that identify the monitoring frequencies needed for the individuals in different risk groups, leading to tailored interventions. To inform the decision making in clinical practice with consideration of monitoring cost and efficiency, the proposed method further compare different monitoring strategies together with the routine monitoring policies in the frontier of cost-effectiveness analysis. We apply the proposed method to adaptively monitor a depression treatment population and identify a set of cost-effective monitoring strategies that have potential to enable better use of healthcare resources.

5.2 METHOD

The proposed framework integrates the prognostics, monitoring strategy and cost-effectiveness analysis. As shown in Figure 5-1, the proposed method includes model training and model evaluation or cost-effectiveness analysis phases. The model training phase learns the prognostic models from a set of historical data. It includes missing value imputation, feature extraction and model learning steps. The cost-effectiveness analysis phase evaluates the efficiency of each prognostic model by iteratively stratifying the individual risks, making decision that monitor or not monitor each individual in next monitoring period and incorporating the new measurement for risk updates. In the decision making process, individuals with predicted risks higher than a threshold θ will be monitored and the individuals with lower risks are also exploited by randomly monitoring γ percentage of low-risk group (predicted risk lower than θ). By comparing the costs and effects of each monitoring strategy in the evaluation phase using a cost-effectiveness analysis

model, we identify a set of cost-effective monitoring strategies. We use the depression monitoring as an example to illustrate the methodology in each step.

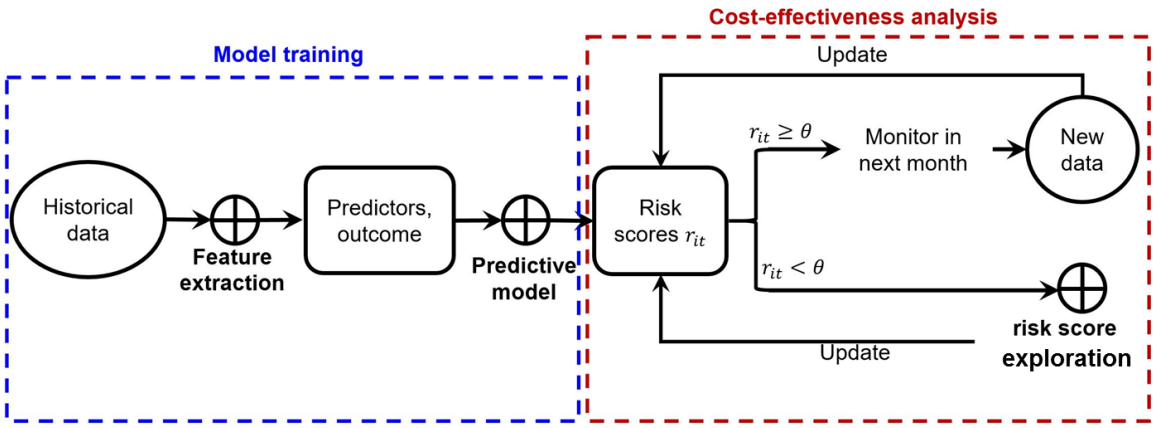


Figure 5-1: The framework of prognostic-based disease monitoring.

5.2.1 *Data description*

We study a depression treatment population of 965 individuals in the Mental Health Research Network (MHRN) data with each individual is closely monitored for one year with at least 6 measurements. In each measurement, depression severity is assessed by the Patient Health Questionnaire (PHQ-9), a self-administrated questionnaire. It ranges from 0 to 27 and indicates a high-risk state when it is greater or equal to 15. Each measurement also records the 9th question score in PHQ-9 and individual's Charlson comorbidity score (a standard indicator of medical disease burden). We conduct analysis on this group by using monthly monitoring time window and regarding the baseline to 5th month as model training phase and the remaining 7 months as cost-effectiveness analysis phase.

Since the EHR data has irregular and sparse measurements on each individual, we impute the missing value using a two-step approach. Specifically, we first impute the missing value between the initial and last measurements by fitting a smoothed B-spline model for each individual and follow the similar process described in Chapter 2. As shown in Figure 5-2, each individual may

have different length of observing spectrum. To imputed the missing value out of observing spectrum, we use the measurements from other individuals. We find the K-nearest neighborhoods of each individual that have most similar baseline features and impute the missing value after last observation using the average value of measurements from K neighbors at corresponding time points. The baseline features including patient's demographic features as well as the coefficients fitted from B-spline model. Random error from a standard normal distribution is added to simulate random noise in reality. The missing value imputation of four randomly selected individuals in Figure 5-2 demonstrate that 1) the trajectory fitted from smoothed B-spline is able to capture the individual depression trajectory and 2) imputed PHQ-9 scores follow these progression trajectories.

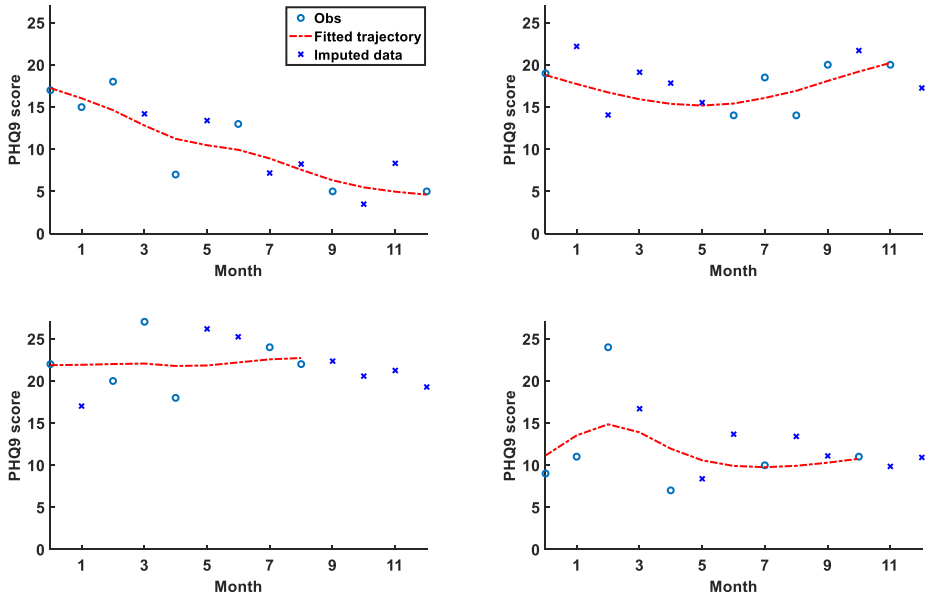


Figure 5-2: Missing value imputation on four exemplary individuals.

5.2.2 *Prognostic model*

The feature-based and trajectory-based methods, including logistic regression, rule-based method, Markov-based collaborative learning and natural history model, have been employed to predict the depression trajectories and their prediction capabilities have been demonstrated using different metrics. We consider these prognostic models in the proposed framework to identify the ones that can lead to more efficient monitoring strategy design. The detail description of each model is provided in section 5.2.2.1 – 5.2.2.3.

5.2.2.1 Feature-based prognostic model

The feature-based prognostic model uses a set of risk-predictive features, $X_1, \dots, X_p$, to describe the disease progression in a certain time window and predicts the health outcome, y in the following time point by these risk-predictive features, i.e. $y = f(X_1, \dots, X_p)$. The features characterize demographic factors of the population, statistical summarization, progression trajectories as well as the abnormal patterns for the longitudinal measurements, the detailed description of these features is provided in Chapter 4. In depression monitoring, for instance, we transform the measurements in every four months to 38 features and predicts the depression severity in 5^{th} month. The 38 features are summarized in Table D-1 in the Appendix D. The depression severity is measured by PHQ-9 score with $y = 1$ if the PHQ-9 score is no less than 15 and $y = 0$ otherwise.

The rule-based method further discovers a set of longitudinal patterns, $R_1, \dots, R_q$, to segment the population into subgroups. Individuals in each subgroup have more homogeneous risk indications, the detailed description of pattern discovery is provided in Chapter 4. 12 longitudinal patterns are discovered from the training data and summarized in Table D-2 ang Figure D-1 in the Appendix. To compare whether the rules can provide more efficient monitoring strategies

compared to the functional features, we use the logit function to link these longitudinal patterns and functional features with the health outcome. The risks in rule-based prognostic model and logistic regression model are predicted as $r_L = f_L(X_1, \dots, X_p) = \frac{1}{1+e^{-\left[\beta_0+\sum_{i=1}^{p}\beta_i X_i\right]}}$ and $r_R = f_R(R_1, \dots, R_q) = \frac{1}{1+e^{-\left[\beta_0+\sum_{i=1}^{q}\beta_i R_i\right]}}$ respectively.

5.2.2.2 Collaborative learning

Markov model is widely used to characterize disease progression by modeling the probability of transition between different health states. As demonstrated in Chapter 3, it is efficient in predicting the individual depression severity over time. To accurately estimate the Markov model for each individual, we used the similarity based collaborative model developed in Chapter 2. Collaborative learning captures the heterogeneous disease progression processes using K canonical Markov models with different initial distributions and transition matrices. Each canonical model represents a type of disease progression in the population. Each individual-level Markov model is further resembled as a weighted combination of canonical models. By assigning each individual distinct weight vector on the canonical models it captures the individual to individual variations. It further incorporates the similarity between individuals to enhance the learning of Markov models by assuming similar individuals are more likely to have similar weight vectors on the canonical models. The risk predictive features mentioned in section 5.2.2.1 can be used to measure the similarity between individuals. To initialize the canonical models and weight vectors, we cluster the individuals into groups based on the risk predictive features. The Markov models learned from measurements in the same group are used to initialize the canonical models and the cluster indexes are used to generate the weight vector of each individual. We found 3 canonical models give the best model fitting in this population, and these patterns shown in Figure D-2 in the Appendix D

represent the stable high, stable low, and moderate depression trajectories. The risk of each individual is predicted as $r_{CM} = P(S_{severe}|S_{latest})$, where S_{severe} and S_{latest} represent the severely diseased state and latest observed state of each individual respectively.

5.2.2.3 Natural history model

The natural history model starts with building a rich database from training data (Alagoz et al., 2005). To consider the progression of health condition, we regard every three sequential measurements on an individual as a triplet and build the database consist of all triplets segmented from the training data. Each individual is also associated with 38 features shown in Table D-1. Denote the predicting individual as the index individual i. We consider the index individual's PHQ-9 scores in previous period $(t-1)$ and current period (t) to determine whether the depression trajectory is improving, worsening or essentially stay the same. To predict the progression of depression in next monitoring period $(t+1)$ on index individual, Y_{it+1}, we search for K most similar individuals in database and identify a set of triplets that have closest depression trajectories from the similar individuals. When such a set of triplets is found, denoted as $\{(Z_{j1}, Z_{j2}, Z_{j3})|j \in \Omega_i\}$, the depression outcome on the index individual is predicted as a weighted average of the last measurements in triplets, i.e. $\widehat{Y_{it+1}} = \sum_{j\in\Omega_i} w_{ji} Z_{j3}$. The weight w_{ji} can be obtained from the closeness between the triples. The similarity between individuals are measured from the Euclidean distance on the 38 features. The closeness between depression trajectories of index individual and triplet is measured by the differences on previous and current PHQ-9 scores, i.e. $w_{ji} = \frac{1}{(Y_{it-1}-Z_{j1})^2+(Y_{it}-Z_{j2})^2}$. The risk score of each individual can be obtained by rescaling the predicted outcome $\widehat{Y_{it+1}}$ to a value between 0 and 1.

5.2.3 *Prognostic-based monitoring*

5.2.3.1 Monitoring policy

The prognostic models presented in section 5.2.2 stratify the individuals' risks of disease onset to different levels. Individuals with high risk tend to be closely monitored for identifying disease symptoms while the individuals with low risk will be less frequently monitored for resource saving. To decide who should be monitored in each monitoring period, we segment the population to high-risk and low-risk groups by comparing each predicted risk, r_{it}, with a predefined threshold θ. The monitoring policies of different groups are summarized in Table 5-1. All individuals in the high-risk group ($r_{it} > \theta$) are monitored to avoid missing disease symptoms. Due to uncertainty in the risk prediction, we randomly select 10% individuals from low-risk group ($r_{it} \leq \theta$) to explore better monitoring policy.

The monitoring accuracy of each policy can be measured by the rate of severe patients being monitored (sensitivity) and the rate of healthy patients not monitored (specificity). Different choice of threshold θ may result in different sensitivity and specificity in monitoring policy. For instance, the increase of threshold leads to the increase of specificity and decrease of sensitivity. To obtain the optimal monitoring policy that have both high sensitivity and high specificity, we find the optimal threshold, $\hat{\theta}$, that minimize the distance between monitoring accuracy and the perfect monitoring (sensitivity = 1 and specificity = 1), i.e.

$$\hat{\theta} = \min_{\theta}((1 - \text{sensitivity})^2 + (1 - \text{specificity})^2) \tag{5.1}$$

Note that monitoring a healthy patient and missing a severely diseased patient may result in different costs, the objective function in (5.1) could be further extended to minimize the total cost by incorporating the costs of per false positive, c_{FP}, and per false negative, c_{FN}, which results in

$$\hat{\theta} = \min_{\theta}(c_{FN}(1 - \text{sensitivity})^2 + c_{FP}(1 - \text{specificity})^2) \tag{5.2}$$

The objective function in (5.2) is more preferred when the prior knowledge about different types of monitoring costs are available, such as the monitoring of seminal vesicle invasion (SVI) prior to or during surgery (Vickers et al., 2006). Due to the lack of prior knowledge in depression monitoring, we use the objective function in (5.1) in this study.

Table 5-1: Prognostic-based monitoring policy.

	High-risk group ($r_{it} > \theta$)	**Low-risk group ($r_{it} \leq \theta$)**
Action	Monitor	Randomly select 10% of the population to be monitored

5.2.3.2 Missing value imputation

Adaptively monitoring the high-risk individuals will lead to the increased number of missing value on low-risk individuals. To guarantee an accurate estimation of the individual health condition, we impute the missing value before the risk prediction. We fit a B-spline model between the first and latest measurements to impute the missing value within monitoring period. When the new measurement is available, we update the imputation by incorporating the new measurement into B-spline model.

5.2.3.3 Risk update

Disease progression is a dynamic process, leading to the risk of each individual keeps changing over time. For example, depression progression may follow different dynamics including increasing risk, decreasing risk and fluctuant risk. To accurately monitor the high-risk individuals, we keep updating the risks on monitored individuals over time. Specifically, we update the features on monitored individuals for feature-based prognostic models by including the new measurements collected from monitoring strategy. For the natural history model, we incorporate the new measurements into database and predict the new sample for the monitored individuals. For the

Markov-based model, we update the membership vectors in collaborative learning to better estimate the individual transition matrix and update the current state on monitored individuals. The probability of progressing from current state to the abnormal state is further updated.

In addition to updating the individual prognostic over time, new measurements also provide better estimations of the sensitivity and specificity in different monitoring policies, leading to the update on optimal monitoring policy. Therefore, we integrate the new measurements and their prognostics with the historical data and update the sensitivity and specificity estimations of each prognostic model. By resolving the optimization problem in (5.1) we update optimal policy for next monitoring time.

5.3 COST-EFFECTIVENESS ANALYSIS

To evaluate the efficiency of these prognostic-based monitoring strategies, we compare them in the frontier of cost-effectiveness analysis (CEA). We conduct the cost-effectiveness analysis in the context of depression monitoring. Two commonly used guidelines in depression monitoring, the status quo and latest measurement based strategies (Unützer et al., 2002; Kroenke et al., 2001) are further considered. The status quo strategy assumes monitoring of all patients under fixed frequency, such as every month, which is limited by its low specificity which may lead to unnecessary monitoring of low-risk patients (false positives) and waste of scarce resource. This study considers different frequencies in the status quo including monitoring monthly (SQ_I), every two months (SQ_II), every three months (SQ_III), and monitoring at 1^{st}, 3^{rd} and 6^{th} following months (SQ_IV). The latest measurement based strategy, on the other hand, diagnoses the disease severity of each patient based on most recent measurements and adaptively monitor the patients with more severe measurements. In the depression monitoring, for example, patients with latest PHQ-9 scores equal to 15 or greater will be monitored in next 6 months when the other patients

will not be monitored. We also randomly select 10% of not monitored individuals to explore their health conditions. Without modeling the progression of health condition, the latest measurement based strategy may not be able to capture the patients with increasing and fluctuant risks.

To investigate which monitoring strategies can lead to more cost-effective use of monitoring, we conduct a cost-effectiveness analysis (CEA) by comparing the outcomes of monitoring costs and effects in each strategy. We estimate the cost of one-time monitoring to be $107 from the literature (Simon et al., 2001). The SQ_I strategy always correctly monitor all disease patients with the highest cost by monitoring all individuals in every period. The monitoring effect is measured by the number of high-risk patients (e.g. PHQ-9 score ≥ 15) that are correctly monitored, which is denoted as the number of true positives (TP). To identify the strategy that gives the best trade-off between cost and effect, we rank the strategies by increasing order of cost and calculate the incremental cost-effectiveness ratio (ICER) between nearby strategies. Denote the costs and effects of two strategies, strategy 0 and strategy 1, as C_0, C_1 and E_0, E_1 respectively, with $C_1 > C_0$. ICER between two strategies is measured as:

$$ICER = \frac{C_1 - C_0}{E_1 - E_0}, \tag{5.3}$$

A strategy is dominated if it has higher cost but lower effect compared to other strategies or a combination of other strategies. We identify rule-based monitoring strategies that are not dominated on the cost-effectiveness frontier.

In addition to the costs, we further consider how much the adaptive monitoring strategies can save compared to the routine monitoring by calculating the number of healthy patients being not monitored, which is denoted as true negatives (TN). The dollars each monitoring strategy saves can be calculated using $\text{Save} = \text{TN} * 107$.

5.4 RESULT

5.4.1 *Prediction accuracy*

We first compare the prediction accuracy of four prognostic models. Specifically, we train the models in first five measurements and evaluate the models using the measurement in following month (5^{th}). The prediction accuracy of four prognostic models are summarized in Table 5-2. The prediction accuracy is measured by the area under ROC curves (AUC), the correlation and the root of mean square error (rMSE) between predicted risks and the real observations. It can be observed that the prediction accuracy of natural history model is higher than the other models, but the advantage is not significant.

Table 5-2: Prediction accuracy of different prognostic model in 5^{th} month.

	Logistic	**Rule-based**	**CM**	**Lazy**
AUC	0.892	0.894	0.891	0.896
Correlation	0.734	0.733	0.733	0.780
rMSE	0.485	0.469	0.433	0.516

5.4.2 *Monitoring accuracy*

We further compare the four prognostic models in the monitoring period. The monitoring accuracy of each monitoring strategy which is measured by the average sensitivity and specificity over evaluation period is summarized in Table 5-3 and Figure 5-3. As shown in the result that feature-based model, natural history model and latest PHQ-9 score based method have similar monitoring accuracy. Rule-based method has higher sensitivity and specificity than the logistic regression which demonstrates that translating the features to a set of longitudinal patterns improve the accuracy of monitoring strategy. Latest PHQ-9 score based monitoring strategy outperforms the natural history model on both sensitivity and specificity. Markov based collaborative learning model has higher sensitivity but lower specificity compared to the Rule-based method and latest

PHQ-9 based method. The overall accuracy measured by the distance between monitoring accuracy and the perfect monitoring indicates the Markov based collaborative learning has best performance. We further evaluate the monitoring accuracy of different methods in each month in Figure A-3 (a) - (d) in the Appendix. It further demonstrates that Markov based collaborative learning model exploits more resources for monitoring leading to higher sensitivity. Other monitoring strategies have better performance on resources saving on the healthy individuals. Due to the noisy and sparse observations in follow-up monitoring, natural history model is inadequate to predict the depression progression, which is demonstrated by its lower prediction accuracy after 6^{th} month. On the other hand, the latest PHQ-9 based method is effective in the predicting the individuals with stable high and stable low depression severities, leading to its higher prediction accuracy.

Table 5-3: Average sensitivity and specificity of different strategies over evaluation period.

Method	Sensitivity	Specificity
SQ_III	0.29	0.71
SQ_II	0.43	0.57
SQ_IV	0.43	0.57
Lazy	0.58	0.84
Logistic	0.59	0.83
Latest-PHQ9	0.59	0.85
Rule-based	0.60	0.84
CM	0.73	0.75
SQ_I	1.00	0.00

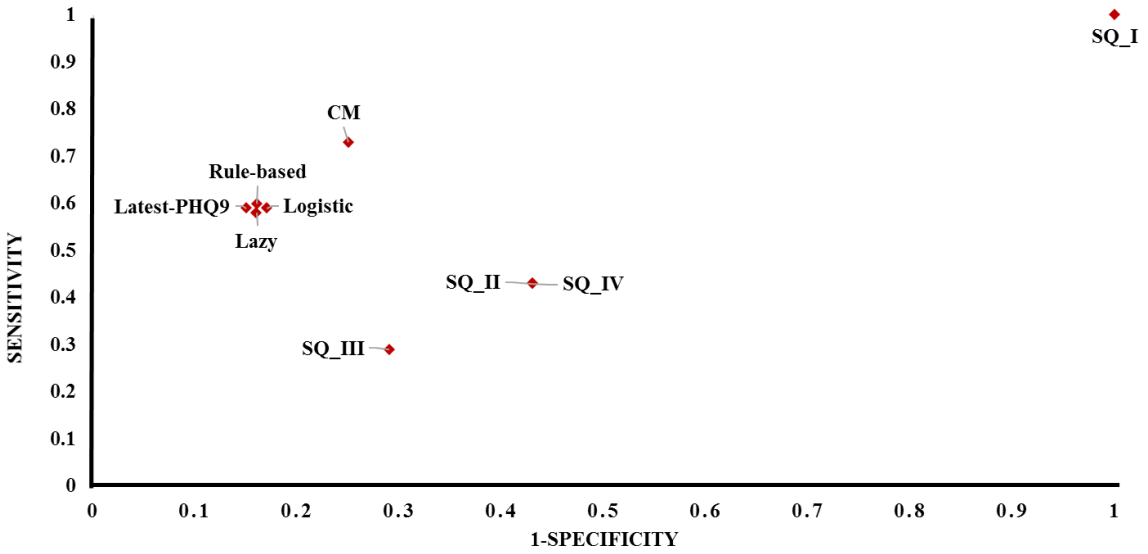


Figure 5-3: Average monitoring accuracy of different strategies over evaluation period.

5.4.3 *Cost-effectiveness analysis*

We evaluate the cost-effectiveness of adaptive monitoring strategies, including prognostic-based and latest PHQ-9 based adaptive monitoring strategies, together with the status quo strategies in the evaluation period. The results of cost-effectiveness analysis are summarized in Table 5-4. Figure 5-4 shows the cost-effectiveness frontier of various monitoring strategies; dominated strategies are plotted with blue dots and non-dominated strategies are plotted in red dots. It can be observed that natural history model and logistic regression are dominated. The latest PHQ-9 based strategy costs $5 to monitor an additional high-risk patient compared to the status quo of monitoring every 3 months. The status quos of monitoring every two months and monitoring at 1st, 3rd and 6th months are also dominated. Status Quo strategy of monitoring all patients monthly costs $746 to monitor an additional high-risk patient compared to the next-best prognostic-based strategy (CM). Using the adaptive monitoring strategies can save at most 19,581 dollars and at least 2,399 dollars than the status quo strategies. Therefore, the adaptive monitoring strategies may be preferred when the monitoring resource is constrained, while the status quo strategy of

monitoring patients monthly may be preferred if healthcare providers are willing to pay a substantial amount for additional correctly monitored patients. Among the adaptive monitoring strategies, latest PHQ-9 based monitoring, rule-based monitoring and Markov based collaborative learning (CM) are cost-effective.

Table 5-4: Cost-effectiveness analysis on the prognostic-based monitoring, latest PHQ-9 based strategy and status quos.

Method	Effect (TP)	Cost ($)	ICER ($/TP)	TN	Save ($)
SQ_III	86	29501		469	50183
Latest-PHQ9	177	29975	5	556	59492
Lazy	178	30556	Dominated	551	58957
Rule-based	185	30847	120	555	59385
Logistic	183	31580	Dominated	546	58422
CM	224	41837	278	491	52537
SQ_II	128	44252	Dominated	373	39911
SQ_IV	131	44252	Dominated	376	40232
SQ_I	307	103255	746	469	50183

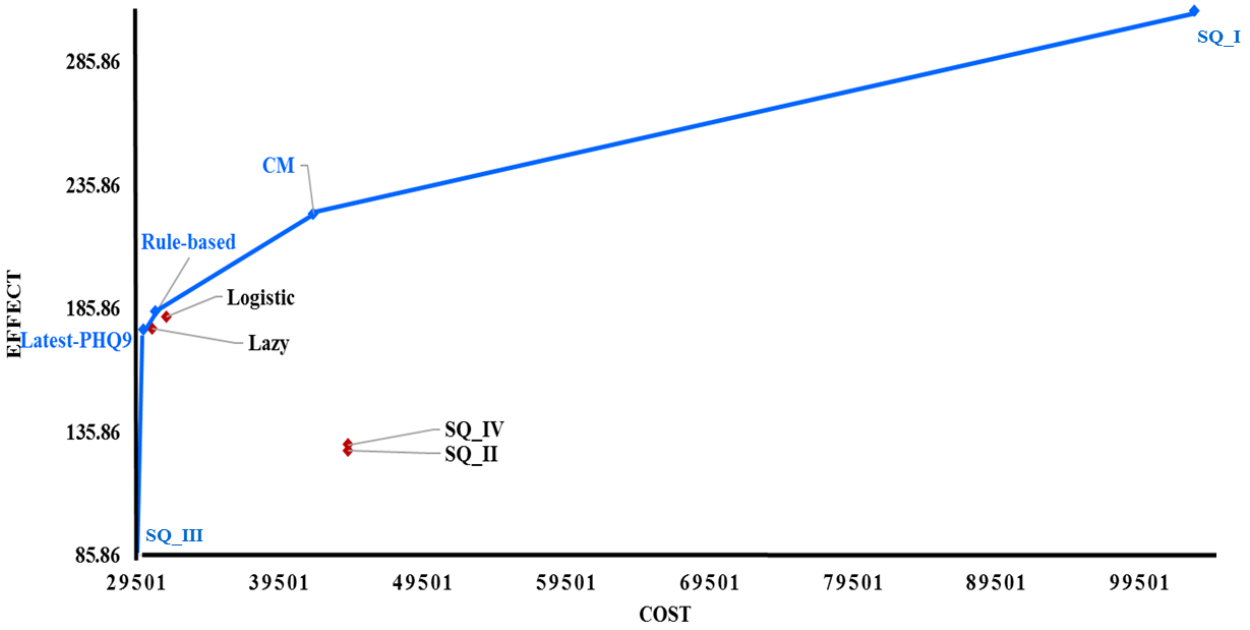
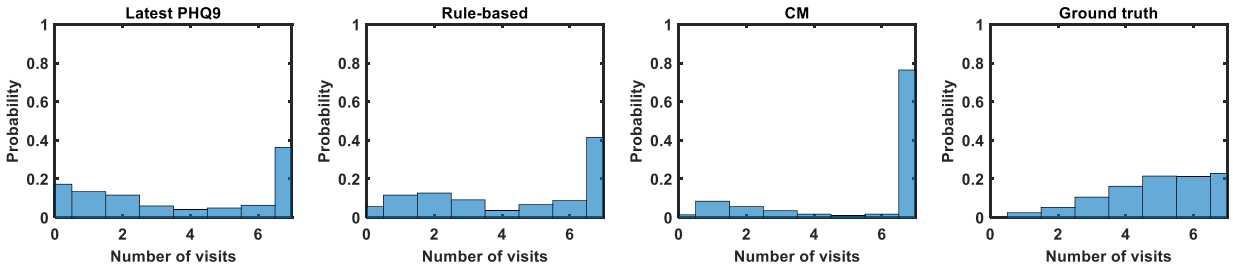


Figure 5-4: Cost-effectiveness frontier for the prognostic-based monitoring strategies.

5.5 ANALYSIS OF COST-EFFECTIVE MONITORING STRATEGIES

We further compare the cost-effective adaptive monitoring strategies on each individual by calculating the monitoring frequency (number of visits of each individual). Based on the three patterns discovered in Figure A-2 in the Appendix, we cluster the individuals to sever group which has stable high PHQ-9 scores, healthy group with stable low PHQ-9 scores and the moderate group that has PHQ-9 scores fluctuate between low and high values. The distributions of monitoring frequency under different monitoring strategies are compared in Figure 5-5. It can be observed that all adaptive monitoring strategies tend to allocate more monitoring resources to severe and moderate groups and save monitoring resources on healthy group. Specifically, CM based monitoring strategy assigns frequent monitoring to the whole severe group while the latest PHQ-9 based strategy is likely to not monitor the individuals in healthy group. The rule-based monitoring strategy has most similar distribution with the optimal strategy (ground truth) in the moderate group.



Severe Group

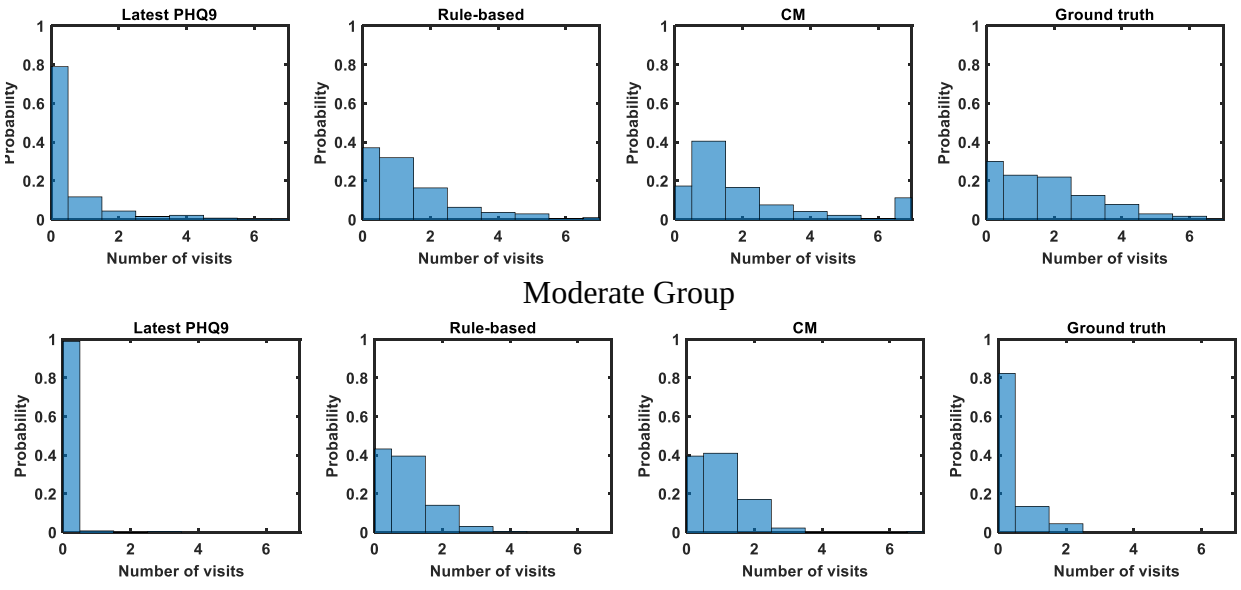


Figure 5-5: Comparison of monitoring frequencies on individuals in severe, moderate and healthy groups.

5.6 DISCUSSION

In this chapter, we establish a prognostic-based monitoring framework to translate the massive EHR data into evidence to support cost-effective monitoring strategy design. Existing methods for tailored health interventions either focus on the monitoring strategy design or risk assessment. The stochastic models including Markov decision process (MDP) and partially observed Markov decision process (POMDP) are widely used for optimal monitoring policy design when the prior knowledge of disease progression is available (Ayer et al., 2012). With increasing complexity and uncertainty in disease progression, the prognostic models such as logistic regression, survival analysis and Markov model are developed to understand the dynamic of disease progression from massive sensing data. The assessment and comparison of prognostic models usually focus on their discriminative ability of binary outcome such as area under the receiver operating characteristic (ROC) curve (AUC) and overall statistics of prediction accuracy such as R^2. Lack of incorporating the clinical consequence of a prognostic model, the use of these evaluation metrics in clinical

practice is limited. Recent developments in the decision curve analysis consider the clinical usefulness of prognostic models by using the net benefits (Vickers et al., 2006). However, the decision curve analysis focuses on the clinical consequence of a critical event, such as costs of recurrence after surgery for prostate cancer. Therefore, the proposed framework is promising to act as a supportive tool to clinical decision for disease monitoring under uncertain environment by integrating the individual prognostic, monitoring strategy design and cost-effectiveness analysis. The proposed method also has potential to enable better use of sensing and information technology for disease monitoring and lead to the quality improvement in the healthcare resources delivery.

We apply the proposed method to adaptively monitor a depression treatment population and compare four types of prognostic models including logistic regression, rule-based method, Markov based collaborative learning, natural history model. We further compared the prognostic-based monitoring strategies with the monitoring strategies in clinical practice. We identify the latest PHQ-9 based method, rule-based method and Markov based collaborative learning are potential to provide cost-effective monitoring strategies for depression monitoring. Specifically, latest PHQ-9 score is a simple approach for prognostics but provides comparable performance with the other prognostic models. Studies in the literature have consistently found that suicidal ideation was an enduring vulnerability rather than a short-term crisis, and response to single PHQ-9 measurement especially the 9^{th} question is predictive to the subsequent suicide attempts (Richardson et al., 2010; Kroenke et al., 2002). However, we discover the latest PHQ-9 based monitoring strategy has advantage in saving resources on healthy individuals but inadequate monitoring of the high-risk individuals. This may due to the fact that latest PHQ-9 based method is not able to capture the individuals with increasing and fluctuant risks. The natural history method has lowest prediction accuracy on depression trajectory which indicates the stochastic changes of PHQ-9 score is less

likely relating to the disease progression. The rule-based method which exploits the complex interactions between risk-predictive features to capture the heterogeneous disease progression processes enables more accurate risk assessment than the logistic regression model which only reflects the average effect of risk-predictive features over the population. The improvement on risk prediction leads to more cost-effective monitoring strategy being designed. On the other hand, Markov based collaborative learning is more efficient in identifying the high-risk individuals by assigning more frequent monitoring to the high-risk and moderately diseased individuals. The Markov based collaborative learning also better captures the individual-to-individual variations in the population by exploiting the heterogeneous progression patterns as well as the similarity between individuals.

One major limitation of our research is that our depression EHR data provides limited information on the treatment response scenarios and the lack of depression assessments that cover the whole life-span of patients in the depression treatment population. As discovered in the literature, the effectiveness of depression monitoring relates to the long-term health outcome of monitoring strategies such as the quality-adjusted life years gained and is also affected by the drop-off and mortality rates in depression treatment. In the future, we plan to conduct more sophisticated cost-effectiveness analysis with consideration of downstream treatment scenarios to accurately estimate long-term health outcomes (e.g. quality-adjusted life years gained). A possible approach to build a decision-analytic Markov model (Liu et al., 2012) for each individual to simulate the long-term monitoring outcomes and costs under different monitoring strategies and treatment scenarios.

In summary, we developed a decision support algorithm for cost-effectively monitoring a heterogeneous diseased population. The proposed method is adaptable to the cost-effective

monitoring strategy design from the massive sensing data such as EHR by integrating the individual prognostic, monitoring strategy design and cost-effectiveness analysis. By applying the proposed method to monitor a depression treatment population, we discover four monitoring strategies that have potential to improve the current recommendations in depression monitoring and fulfill the lack of evidence-based monitoring strategies in clinical practice.

Chapter 6. CONCLUSION AND FUTURE RESEARCH

This dissertation contributes to generic development of personalized health surveillance systems for cost-effectively monitoring the heterogeneous disease progressions in a large-scale population. To accurately predict the individual health condition, adaptively allocate the limited monitoring resources, this thesis proposes novel statistical and optimization models and computational tools, including collaborative learning, selective sensing and rule-based monitoring. It also contributes to the design of efficient monitoring guideline in clinical practice and has potential to mitigate the inadequate follow-up monitoring of chronic conditions such as depression and Alzheimer's disease.

6.1 FOR PERSONALIZED PROGNOSTICS

Recent advances in sensing and technology provide abundance of risk-predictive data that enables the prognostics of health condition in early stage. However, the heterogeneity in the population and sparse measurements on the individual pose two major challenges in disease prognostics. To effectively leverage the sparse and irregular monitoring data for modeling the heterogeneous personalized disease trajectories, I proposed a generalized statistical learning framework, **collaborative modeling**. The proposed method enables more efficient personalized model

learning by explicitly exploiting the underlying cluster structure in the population and similarity information between individuals. It has been extended to capture linear, nonlinear and stochastic dynamics in complex systems, and applied for cognitive degradation modeling in Alzheimer's disease (AD), and Depression trajectory learning.

In addition to model the disease progression based on model assumption, such as the memoryless assumption in Markov model, this thesis further develops a rule-based prognostic model to stratify the individual risk of disease onset. However, existing prognostic models only reflect average effects of risk factors, and lack interpretability, thus, they are hard to assist decision making in clinical practice. The **rule-based prognostic model** (1) identifies a set of longitudinal patterns that captures the heterogeneous disease progression processes; (2) captures complex interactions between both quantitative and qualitative factors; (3) segments population into subgroups with different risk implications; and (4) provides accurate prediction of individual risk. It has been utilized to identify the longitudinal predictive patterns for depression progression prognostics.

6.2 FOR ADAPTIVE MONITORING

Disease monitoring which refers to the repeated measurements of health condition is widely used to check the progress or regress of disease progression. The advances in sensing and information technology such as electronic health record (HER) improve the accessibility and efficiency of disease monitoring. However, under the limited monitoring resources including available appointment slots and staffs in clinical practice current monitoring guideline which relies on routine visit of at-risk individuals is insufficiently developed leading to inadequate monitoring under limit monitoring resources. To effectively allocate the limited monitoring and screening resources to a large-scale population for disease monitoring, this thesis develops a **selective**

sensing framework by integrating the personalized prognostics, optimization, and sensing strategy design into a unified framework and exploiting the correlation between individuals in each step. It has been applied in the context of depression monitoring and cognitive monitoring in AD to solve the problem of inadequate follow-up monitoring in depression treatment.

To evaluate the efficiency and cost-effectiveness of disease monitoring strategies in a heterogeneous population, this thesis further develops a **prognostic-based monitoring system** to identify the cost-effective monitoring strategies by integrating the individual prognostics, adaptive monitoring strategy design and cost-effectiveness analysis. The proposed method has been applied to a depression treatment population and identified four cost-effective monitoring strategies for improving current fixed-frequency monitoring strategy in depression management.

6.3 Future research: Disease monitoring with multimodal sensing

The use of multimodality data improves accuracy of disease diagnosis and prognosis by providing comprehensive information from heterogeneous sensor modalities, such as using neuroimaging measurements together with the cognitive measurements for early detection of Alzheimer's disease and fusing comorbidities' progressions to inform disease dynamic in diabetes. However, as the data present several challenging characteristics such as super-high-dimensionality, irregularity, longitudinality, as well as heterogeneity between sensor modalities and between individuals, existing research are only able to explore data to a limit extend. In addition, the design of sensing strategy is complicated by the need of multi-level decision making, including when each patient should be monitored as well as what measurements he/she should take. To enable effective use of multimodal sensing technology, I aim to 1) develop novel statistical models and computational algorithms to model disease trajectories from multimodality data for accurate prognosis, better understanding of disease pathology, and early detection of disease; 2) develop sophisticate

decision making algorithm to provide the adaptive monitoring strategy and multimodal sensors selection simultaneously; 3) integrate the disease prognosis, adaptive monitoring, and sensor selection into a systematic framework for assisting the development of personalized medication.

6.4 FUTURE RESEARCH: INDIVIDUAL-LEVEL DECISION MAKING OF DISEASE TREATMENT

Identifying optimal treatment strategies for each individual is challenging in clinical practice due to a growing myriad of treatment options, multi-stage nature of clinical trials, and uncertainty in disease progression and treatment effects. In depression, for example, several sequential treatment steps are often needed to obtain remission. The problems of how many treatment steps and which treatments required for each patient remain unsolved. My future research will focus on the development of general methodologies and scalable computational tools to predict treatment effects and design treatment strategies on individuals to facilitate the transition from disease-centered medication to patient-centered medication.

ACKNOWLEDGEMENT

The authors acknowledge funding support from the National Science Foundation under Grant CMMI-1505260 and 1536398.

BIBLIOGRAPHY

[1] Akaike, H. (1974), A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19, 716-723.

[2] Alagoz O, Bryce CL, Shechter S, et al. Incorporating biological natural history in simulation models: Empirical estimates of the progression of end-stage liver disease. Medical Decision Making 2005;25:620-32.

[3] Albert, M. S., Dekosky, S. T., Dickson, D. DICKSON, et al. (2011), Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease, *Alzheimers Dement*, 7, 270–9.

[4] Anderson, T.W. (1951), Estimating linear restrictions on regression coefficients for multivariate normal distributions, *Annals of Mathematical Statistics*, 22, 327-351.

[5] Anstey K.J., Sanden C.V., Sargent-Cox K., et al. "Prevalence and risk factors for depression in a longitudinal, population-based study including individuals in the community and residential care." The American journal of geriatric psychiatry 15.6 (2007): 497-505.

[6] Ashford, J. and Schmitt, F. (2001), Modeling the time-course of Alzheimer dementia, *Current Psychiatry Report*, 3, 20-28.

[7] Ayer, T., Oguzhan, A., and Natasha, K. S., OR Forum-A pomdp approach to personalize mammography screening decisions. *Operations Research*, 60.5 (2012), 1019-1034.

[8] Baayen, R. H., Davidson, D. J., and Bates, M. D. (2008), Mixed-effects modeling with crossed random effects for subjects and items, *Journal of Memory and Language*, 59, 390-412.

[9] Baird, L. C., Reinforcement learning in continuous time: advantage updating. In *Proc. of ICNN* (1994).

[10] Baker F, Kim S (2004) Item Response Theory: Parameter Estimation Techniques. 2nd Edition, CRC Press.

[11] Baker, F.B., and Kim S., eds. Item response theory: Parameter estimation techniques. CRC Press, 2004.

[12] Barrett, E., Howley, E., and Duggan, J., Applying reinforcement learning towards resource allocation and application scalability. *Concurrency and Computation: Practice and Experience*, 25.12 (2013), 1656-1674.

[13] Bartzokis, G., Sultzer, D., Lu, P. H., et al. (2004), Heterogeneous age-related breakdown of white matter structural integrity, *Neurobiol Aging*, 25, 843–851.

[14] Belkin, M. and Niyogi, P. (2001), Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering, *Advances in Neural Information Processing System*, 14, 585-591.

[15] Bellazzi R., Zupan B. (2008) Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics, 77(2):81-97.

[16] Bernd L., et al. "Measuring depression outcome with a brief self-report instrument: sensitivity to change of the Patient Health Questionnaire (PHQ-9)." Journal of affective disorders 81.1 (2004): 61-66.

[17] Biesanz, J. C., Deeb-sossa, N., Aubrecht, A. M., and Curran, P. J. (2004), The role of coding time in estimating and interpreting growth curve models, *Psychological Methods*, 9, 30 – 52.

[18] Bingley PJ (1996) Interactions of age, islet cell antibodies, and first-phase insulin response in predicting risk of progression to IDDM in ICA+ relatives: the ICARUS data set. Diabetes 45: 1720 –1728.

[19] Bonarini A. (2000) An introduction to learning fuzzy classifier systems. In: Learning

Classifier Systems , Springer, Berlin, pp. 83-104.
[20] Bougneres P, Valleron AJ (2009) Causes of early-onset type 1 diabetes: toward data-driven environmental approaches. J Exp Med 205: 2953–2957.
[21] Bradley P. S., Mangasarian O. L. (1998) Feature selection via concave minimization and support vector machines. In: ICML, vol 98, pp. 82-90.
[22] Breiman L (2001) Random forest. Machine Learning 45(1): 5-32.
[23] Bro, R. and Jong, S. De (1997), A fast non-negative-constrained least squares algorithm, *J. Chemometrics*, 11, 393-401.
[24] Bryk, A. S. and Raudenbush, S. (1987), Application of hierarchical linear models to accessing change, *Psychological Bulletin*, 101, 147-158.
[25] Bu, Davis, et al. "Benefits of information technology–enabled diabetes management." *Diabetes Care* 30.5 (2007): 1137-1142.
[26] Bubeck, S., and Cesa-Bianchi, N., Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721* (2012).
[27] Byon, E., Ntaimo, L., and Ding ,Y., Optimal maintenance strategies for wind turbine systems under stochastic weather conditions. *IEEE Transactions on* Reliability, 59.2 (2010), 393-404.
[28] C. E. Rasmussen, "Gaussian processes for machine learning." 2006.
[29] Cai, D., He, X., Han, J., and Huang, T. S. (2011), Graph regularized nonnegative matrix factorization for data representation, *IEEE Transactions*, 33, 1548-1560.
[30] Centers for Disease Control and Prevention web: An Estimated 1 in 10 U.S. Adults Report Depression. Accessed at http://www.cdc.gov/features/dsdepression/. 2012.
[31] Chase HP, Cuthbertson DD, Dolan LM, Kaufman F, Krischer JP, et al. (2001) First phase insulin release during the intravenous glucose tolerance test is a risk factor for type 1 diabetes. J Pediatr 138(2): 244-9.
[32] Clark P., Niblett T. (1989) The cn2 induction algorithm. Machine learning, 3(4):261-283.
[33] CMS.gov (2013) National health expenditures accounts methodology paper, 2013: de_nitions, sources, and methods. http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/downloads/dsm-13.pdf , accessed 12 April 2015.
[34] Cohen W. W. (1995) Fast effective rule induction. In: Proceedings of the twelfth international conference on machine learning , pp. 115-123.
[35] Cole M.G., and Dendukuri N. "Risk factors for depression among elderly community subjects: a systematic review and meta-analysis." American Journal of Psychiatry 160.6 (2003): 1147-1156.
[36] CostHelper (2015) Health & personal care costs and price paid-costhelper.com. http://health.costhelper.com , accessed 17 July 2015.
[37] Cox D. R. (1958) The regression analysis of binary sequences. Journal of the Royal Statistical Society, 20(2):215-242.
[38] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), Maximum Likelihood from Incomplete Data via the Em Algorithm, *J. Royal Statistics Soc. Series B (Methodological),* 39, 1-38.
[39] Deogun J. S., Raghavan V. V., Sarkar A., Sever H. (1997) Data mining: Research trends, challenges, and applications. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.52.337&rep=rep1&type=pdf ,accessed 27 April 2015.

[40] DeSarbo, W. S. and Cron, W. L. (1999), A maximum likelihood methodology for clusterwise linear regression, *J. Classification*, 5(2), 249-282.
[41] Diabetes Prevention Trial–Type 1 Diabetes Study Group (2002) Effects of insulin in relatives of patients with type 1 diabetes mellitus. N Engl J Med 346: 1685–1691.
[42] Ding, C., Zhang, Y., Li, T., and Holbrook, S. R. (2006), Biclustering protein complex interactions with a biclique finding algorithm, *In: Data Mining, ICM Sixth International Conference*, 178–187.
[43] Duan C., Lin Y., Won D., Huang S., You J., and Chaovalitwongse W.A., A cost-sensitive rule-based classification framework for medical diagnosis and decision making, *Annals of Operation Research*, submitted.
[44] Duch W. (2010) Rule-based methods. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.228.9787&rep=rep1&type=pdf , accessed
[45] Duchesne, S., Caroli, A., Geroldi, C., et al. (2009), Relating one-year cognitive change in mild cognitive impairment to baseline MRI features, *Neuroimage*, 47, 1363-1370.
[46] Embretson, Susan E., and Steven P. Reise. Item response theory. Psychology Press, 2013.
[47] Ewout W. Steyerberg, Andrew J. Vickers, Nancy R. Cook, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology 2010;21:128-38.
[48] Fan Y. J., Chaovalitwongse W. A. (2010) Optimizing feature selection to improve medical diagnosis. Annals of Operations Research , 174(1):169-183.
[49] Fox J., Das S. (2000) Safe and sound: arti_cial intelligence in hazardous applications . MIT press, Cambridge, MA.
[50] Frazier, P., Warren, P., and Savas, D., Knowledge-gradient policy for correlated normal beliefs. *INFORMS journal on Computing*, 21.4 (2009), 599-613.
[51] Freed N., Glover F. (1981) Simple but powerful goal programming models for discriminant problems. European Journal of Operational Research, 7(1):44-60.
[52] Friedman J, Popescu BE, Predictive Learning via rule ensemble. Annals of Applied Statistics, 2.3 (2008): 916–954.
[53] Fuurnkranz J. (1999) Separate-and-conquer rule learning. Artificial Intelligence Review , 13(1):3{54.
[54] Gale EA, Bingley PJ, Emmett CL, Collier T, the European Nicotinamide Diabetes Intervention Trial (ENDIT) Group (2004) European Nicotinamide Diabetes Intervention Trial (ENDIT): a randomised controlled trial of intervention before the onset of type 1. Lancet 363: 925–931.
[55] Galecki, A. and Burzykowski, T. (2013), Linear Mixed-Effects Models using R, *Springer Texts in Statistics*.
[56] Gallo, G., Grigoriadis, M. D., and Tarjan, R. E., A fast parametric maximum flow algorithm and applications. *SIAM on Computing*, 18.1 (1989), 30-55.
[57] Gimenez M, Lara ND, Aguilera E, Nicolau J, Castell C, et al. (2007) Relationship between BMI and Age at diagnosis of type 1 diabetes in a mediterranean area in the period of 1990-2004. Diabetes Care 30: 1593-1595.
[58] Goldstein, H. (1987), Multilevel models in education and social research, *New York: Oxford University Process*.
[59] Guan, N., et al., Non-negative patch alignment framework. *IEEE Transactions on Neural Networks*, 22.8 (2011), 1218-1230.

[60] Gunn, J., P. Elliott, K. Densley, A. Middleton, G. Ambresin, C. Dowrick, H. Herrman, K. Hegarty, G. Gilchrist and F. Griffiths, *A trajectory-based approach to understand the factors associated with persistent depressive symptoms in primary care.* J Affect Disord, 2013. **148**(2-3): p. 338-46.

[61] Guyon I., Weston J., Barnhill S., Vapnik V. (2002) Gene selection for cancer classification using support vector machines. Machine Learning , 46(1-3):389-422.

[62] Hammer P. L., Bonates T. O. (2006) Logical analysis of data? An overview: from combinatorial optimization to medical applications. Annals of Operations Research , 148(1):203-225.

[63] Hand D. J. (1981) Discrimination and classification . Wiley Series in Probability and Mathematical Statistics, Chichester: Wiley.

[64] Hardy, J. and Selkoe, D. J. (2002), The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics, *Science,* 297, 353–356.

[65] Hartman M., Martin A. B., Lassman D., Catlin A. (2015) National health spending in 2013: growth slows, remains in step with the overall economy. Health Affairs, 34(1):150-160.

[66] Head, D., Buckner, R. L., Shimony, J. S., et al. (2004), Differential vulnerability of anterior white matter in nondemented aging with minimal acceleration in dementia, *Cerebral Cortex*, 14, 410–423.

[67] Hedeker, D. and Gibbons, R. D. (1997), Application of random-effects pattern-mixture models for missing data in longitudinal studies, *Psychological Methods*, 2, 64-78.

[68] Hensrud D. D. (2000a) Clinical preventive medicine in primary care: background and practice: 1. rationale and current preventive practices. Mayo Clinic Proceedings , 75(2):165-172.

[69] Hertzog, C., Dixon, R. A., Hultsch, D. F., and Macdonald, S. W. (2003), Latent change models of adult cognition, *Psychology and Aging*, 18, 755– 769.

[70] Hoshikawa, T. (2013), Mixture regression for observational data, with application to functional regression models. http://arxiv.org/abs/1307.0170

[71] Huang S.H., LePendu P., Iyer S.V., et al. "Toward personalizing treatment for depression: predicting diagnosis and severity." Journal of the American Medical Informatics Association 21.6 (2014): 1069-1075.

[72] Hunink M.G.M, Weinstein M.C., Wittenberg E., et al. Decision making in health and medicine: integrating evidence and values. Cambridge University Press, 2014.

[73] Ito, K., et al. (2010), Disease progression model for cognitive deterioration from Alzheimer`s Disease Neuroimaging Initiative database, *Alzheimers Dementia,* 6, 39-53.

[74] Islam MA, Chowdhury RI, Huda S. A multistate transition model for analyzing longitudinal depression data. Bulletin of the Malaysian Mathematical Sciences Society 2013;36:.

[75] Izenman, A.J. (1975), Reduced-rank regression for the multivariate linear model, *Journal of Multivatiate Analysis*, 5(2), 248-264.

[76] J. Liu, V. Vitelli, E. Zio, and R. Seraoui, "A Novel Dynamic-Weighted Probabilistic Support Vector Regression-Based Ensemble for Prognostics of Time Series Data". IEEE Transactions on Reliability, 64(4), pp.1203-1213, 2015.

[77] Jack, C. JR., Bernstein, M., Fox, N., et al. (2008), The Alzheimer`s disease neuroimaging initiative (ADNI): MRI methods, *J. Magn. Regon. Imaging*, 27, 685-691.

[78] Jack, C. R. JR., Knopman, D. S., Jagust, W. J., et al. (2010), Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade, *Lancet Neurol*, 9, 119–128.

[79] Jedynak, B. M., et al. (2012), A computational neurodegenerative disease progression scor, *NeuroImage,* 63, 1478–1486.

[80] Jiang, H., et al., Optimal selection of channel sensing order in cognitive radio. *IEEE Transactions on Wireless Communications*, 8.1 (2009), 297-307.
[81] Johnson, K. A., Minoshima, S, Bohnen, N. I., et al. (2013), Appropriate use criteria for amyloid PET: A report of the Amyloid Imaging, *Journal of Nuclear Medicine*, 54, 1-16.
[82] K. A. Doksum, and S. L. T. Normand, "Gaussian models for degradation processes-Part I: Methods for the analysis of biomarker data." Lifetime Data Analysis 1.2: 131-144, 1995.
[83] Kales HC, Kim HM, Austin KL, Valenstein M. Who receives outpatient monitoring during High-Risk depression treatment periods? J. Am. Geriatr. Soc. 2010;58:908-13.
[84] Kendler K.S., Kessler R.C., Neale MC, et al. "The prediction of major depression in women: toward an integrated etiologic model." American Journal of Psychiatry 150 (1993): 1139-1139.
[85] Kessler RC. Gender differences in major depression: Epidemiological findings. In: Frank E, editor. Gender and its effects on psychopathology. Washington, DC: American Psychiatric Press; 2000. pp. 61–84.
[86] King, Michael, et al. "Development and validation of an international risk prediction algorithm for episodes of major depression in general practice attendees: the PredictD study." Archives of General Psychiatry 65.12 (2008): 1368-1376.
[87] Kotsiantis S. B. (2007) Supervised machine learning: a review of classification techniques. Informatica, 33(3):249-268.
[88] Kroenke, K., R.L. Spitzer and J.B. Williams, *The PHQ-9: validity of a brief depression severity measure.* J Gen Intern Med, 2001. **16**(9): p. 606-13.
[89] Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PloS One 2013;8:e66341.
[90] Lawson, C. L. and Hanson, R. J. (1947), Solving least squares problems. *SIAM Classics in Applied Mathematics*.
[91] Leighton, T., and Rao, S., Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *Journal of the ACM (JACM)*, 46.6 (1999), 787-832.
[92] Leisch, F. (2004), FlexMix: A general framework for finite mixture models and latent class regression in R, *Journal of Statistical Software*, 11(8), 1-18.
[93] Lernmark A, Ott J (1998) Sometimes it's hot, sometimes it's not. Nat Genet 19(3): 213–214.
[94] Lin Y., Huang S., Simon G.E., and Liu S., "Analysis of depression trajectory patterns using collaborative learning." Mathematical Bioscience, 828: p.191-203.
[95] Lin Y., Qian X., Krischer J., et al. "A Rule-Based Prognostic Model for Type 1 Diabetes by Identifying and Synthesizing Baseline Profile Patterns." PloS one 9.6 (2014): e91095.
[96] Lin, Y., Liu, K., Byon, E., Qian, X., et al., Domain-knowledge driven cognitive degradation modeling for Alzheimer's disease. In *SIAM International Conference on Data Mining 2015, SDM 2015*, 721-729.
[97] Little, R. J. A. (1995), Modeling the dropout mechanism in repeated measures studies, *Journal of American Statistical Association*, 90, 1112-1121.
[98] Liu, Jie, et al. "Multiple Testing under Dependence via Semiparametric Graphical Models." *ICML*. 2014.
[99] Liu, K., Mei, Y., and Shi, J., An Adaptive Sampling Strategy for Online High-Dimensional Process Monitoring. *Technometrics*, 57.3 (2015), 305-319.

[100] Liu S, Cipriano LE, Holodniy M, Owens DK, Goldhaber-Fiebert JD. New protease inhibitors for the treatment of chronic hepatitis CA cost-effectiveness analysis. Ann. Intern. Med. 2012;156:279-90.
[101] Lopez, O. L., Jagust, W. J., Dekosky, S. T., et al. (2003), Prevalence and classification of mild cognitive impairment in the Cardiovascular Health Study Cognition Study, *Arch Neurol*, 60, 1385-1389.
[102] Lord, F.M. Applications of item response theory to practical testing problems. Routledge, 1980.
[103] Löwe, B., et al., Monitoring depression treatment outcomes with the patient health questionnaire-9. *Medical care*, 42.12 (2004), 1194-1201.
[104] Lunn D, Jackson C, Thomas A, Best N, Spiegelhalter D (2012) The BUGS Book: A Practical Introduction to Bayesian Analysis. Chapman & Hall/CRC.
[105] Matheus C. J., Chan P. K., Piatetsky-Shapiro G. (1993) Systems for knowledge discovery in databases. IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913.
[106] McCullagh P, Nelder J (1989) Generalized Linear Models. Chapman & Hall.
[107] Mcgue, M. and Christensen, K. (2002), Heritability of level of and rate of change in cognitive functioning in Danish twins aged 70 years and older, *Experimental Aging Research*, 28, 435– 451.
[108] McLachlan, C. and Peel, D. (2004), Finite mixture models, *Wiley Series in Probability and Statistics*.
[109] Montgomery D.C., Design and analysis of experiments. John Wiley & Sons, 2008.
[110] Mrena S., Virtanen S. M., Laippala P., Kulmala P., Hannila M. L., Akerblom H. K., Knip M. (2006) Models for predicting type 1 diabetes in siblings of affected children. Diabetes Care, 29(3):662-667.
[111] Mueller, S. G., Weiner, M. W., Thal, L. J., et al. (2005), The Alzheimer's Disease Neuroimaging Initiative, *Neuroimaging Clin North Am*., 15, 869–877.
[112] National Institute of Mental Health website on depression. Accessed at http://www.nimh.nih.gov/health/topics/depression/index.html.
[113] Nelson, B.L., Swann, J., Goldsman, D., et al., Simple procedures for selecting the best simulated system when the number of alternatives is large. Oper. Res., 49 (2001), 950–963.
[114] Oquendo M. A., Turret J., Grunebaum M.F., et al. "Sex differences in clinical predictors of depression: a prospective study." Journal of affective disorders 150.3 (2013): 1179-1183.
[115] Ortiz, Eduardo, and Carolyn M. Clancy. "Use of information technology to improve the quality of health care in the United States." *Health Services Research* 38.2 (2003): xi-xxii.
[116] Oskooyee KS, Rahmani AM, Kashani MMR. Predicting the severity of major depression disorder with the markov chain model.
[117] Pan, S. J. and Yang, Q. (2010), A survey on transfer learning, *IEEE Transactions on Knowledge and Data Engineering*, 22, 1345-1359.
[118] Pearson, R., Kingan, R., and Hochberg, A. (2005), Disease progression modeling from historical databases, *KDD 2005*.
[119] Penny, W. D., Holmes, A. P., and Friston, K. J. (2003), Random effects analysis, *Human brain function II (2nd ed.)*.
[120] Petersen, R. C., Smith, G. E., Waring, S. C., *et al. (1999)*, Mild cognitive impairment: clinical characterization and outcome, *Archives of Neurology*, 56, 303-308.
[121] Pfeiffer P.N., Bohnert K.M., Zivin K., et al. Mobile health monitoring to characterize depression symptom trajectories in primary care. Journal of affective disorders, 174 (2015), 281-

286.
[122] Piccinelli M., and Wilkinson G. "Gender differences in depression." The British Journal of Psychiatry 177.6 (2000): 486-492.
[123] Powell, W. B., and Ryzhov, I. O., Optimal learning. *John Wiley & Sons*, 841 (2012).
[124] Pratt L., Brody D.J., Gu Q. Antidepressant Use in Persons Aged 12 and Over: United States, 2005-2008. NCHS Data Brief, 76 (2011).
[125] Rag, N. (2000), Aging of the brain and its impact on cognitive performance, *In: Handbook of aging and cognition II (Craik FIM, Salthouse TA, eds)*, 1–90.
[126] Rasmusson, D. X., Carson, K. A., Brookmeyer, R., et al. (1996), Predicting rate of cognitive decline in probable Alzheimer's disease, *Brain and Cognition*, 31, 133–147.
[127] REGAN, C., KATONA, C., WALKER, Z., ET AL. (2006), RELATIONSHIP OF VASCULAR RISK TO THE PROGRESSION OF ALZHEIMER DISEASE, NEUROLOGY, 67, 1357-1362.
[128] Reinsel, G.C., and Velu, R.P. (1998), Multivariate reduced rank regression: theory and applications, *New York: Springer*.
[129] Reynolds, C. A., Finkle, D., and Pedersen, N. L. (2002), Sources of influence on rate of cognitive change over time in Swedish twins, *Experimental Aging Research*, 28, 407-433.
[130] Reynolds, C.F.III and E. Frank, US Preventive Services Task Force Recommendation Statement on Screening for Depression in Adults: Not Good Enough. Published online January 26, 2016. http://archpsyc.jamanetwork.com/article.aspx?articleid=2484482. JAMA Psychiatry, 2016.
[131] Richardson LP, McCauley E, Grossman DC, et al. Evaluation of the patient health questionnaire-9 item for detecting major depression among adolescents. Pediatrics 2010;126:1117-23.
[132] Robinson, J. C., Zbigniew Czyzewski, and James W. P., Wireless machine monitoring and communication system. U.S. Patent No. 5,907,491. 25 May 1999.
[133] Royall, D. R., Palmer, R., and Chiodo, L. K. (2005), Normal rates of cognitive change in successful aging, *Journal of Neuropsychological Society*, 11, 899–909.
[134] Sanz J. A., Galar M., Jurio A., Brugos A., Pagola M., Bustince H. (2014) Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system. Applied Soft Computing, 20(7):103-111.
[135] Scholkopf, B. and Smola, A. J. (2002), Learning with Kernels, *MIT Press*.
[136] Simon, G. E., et al. "Cost-effectiveness of a collaborative care program for primary care patients with persistent depression." American Journal of Psychiatry, 158.10 (2001): 1638-1644.
[137] Simon, G.E., Rutter, C.M., Peterson, D., et al., Does response on the PHQ-9 Depression Questionnaire predict subsequent suicide attempt or suicide death? Psychiatr Serv, 64.12 (2013): p. 1195-202.
[138] Sliwinski, M. J., Hofer, S. M., et al. (2003), Modeling memory decline in older adults: The importance of preclinical dementia, *Psychology and Aging*, 18, 658–671.
[139] Smola, A. J., Gretton, A., Song, L., and Scholkopf, B. (2007), A Hilbert Space Embedding for Distributions, *Lecture Notes on Computer Science*, 4757, 13-31.
[140] Solt I., Tikk D., Gal V., Kardkovacs Z. T. (2009) Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier. Journal of the American Medical Informatics Association, 16(4):580-584.
[141] Sosenko J, Krischer J, Palmer J, Mahon J, Cowie C, et al. (2008) A risk score for type 1 diabetes derived from autoantibody-positive participants in the Diabetes Prevention Trial–Type 1. Diabetes Care 31: 528 –533.

[142] Sosenko J, Skyler J, Mahon J, Krischer J, Craig B, et al. (2011) Validation of the Diabetes Prevention Trial-Type 1 risk score in the TrialNet natural history study. Diabetes Care 34: 785-787.
[143] Sosenko J. M., Krischer J. P., Palmer J. P., Mahon J., Cowie C., Greenbaum C. J., Cuthbertson D., Lachin J. M., Skyler J. S. (2008) A risk score for type 1 diabetes derived from autoantibody-positive participants in the diabetes prevention trial-type 1. Diabetes Care, 31(3):528-533.
[144] Sosenko J. M., Skyler J. S., Mahon J., Krischer J. P., Beam C. A., Boulware D. C., Greenbaum C. J., Rafkin L. E., Cowie C., Cuthbertson D. (2011) Validation of the diabetes prevention trial-type 1 risk score in the trialnet natural history study. Diabetes Care, 34(8):1785-1787.
[145] Srikanta S, Ganda OP, Jackson RA, Brink SJ, Flwischnick E, et al. (1984) Pre-type 1 (insulindependent) diabetes: common endocrinological course despite immunological and immunogenetic heterogeneity. Diabetologia 27: 146– 148.
[146] Stern SE, Williams K, Ferrannini E, DeFronzo RA, Bogardus, C, et al. (2005) Identification of individuals with insulin resistance using routine clinical measurements. Diabetes 54: 333-339.
[147] Stettin G.D., et al. "Frequency of follow-up care for adult and pediatric patients during initiation of antidepressant therapy." American Journal of Managed Care 12.8 (2006): 453-463.
[148] Suh, G. H., Ju, Y. S., Yeon, B. K., and Shah, A. (2004), A longitudinal study of Alzheimer's disease: Rates of cognitive decline, *Journal of Geriatric Psychiatry*, 19, 817–824.
[149] Sun, J., Sow, D., Hu, J., and Ebradollahi, S. (2010), Localized supervised metric learning on temporal physiological data, *ICPR* 2010.
[150] Sutin, A.R., A. Terracciano, Y. Milaneschi, Y. An, L. Ferrucci and A.B. Zonderman, *The trajectory of depressive symptoms across the adult life span.* JAMA Psychiatry, 2013. **70**(8): p. 803-11.
[151] Sutton, R. S., and Barto, A. G., Reinforcement learning: An introduction. *MIT press*, 1998.
[152] Swan, Melanie. "Health 2050: The realization of personalized medicine through crowdsourcing, the quantified self, and the participatory biocitizen."*Journal of Personalized Medicine* 2.3 (2012): 93-118.
[153] Tibshirani R., Optimal reinsertion: regression shrinkage and selection via the LASSO. Journal of Royal Statistical Society B, 58.1 (1996): 267–288.
[154] U.S. Food and Drug, Anti-depressant drug use in pediatric population, 2004. http://www.fda.gov/NewsEvents/Testimony/ucm113265.htm
[155] U.S. Food and Drug, FDA proposes new warnings about suicidal thinking, behavior in young adults who take antidepressant medications, 2007. http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2007/ucm108905.htm
[156] Unützer, J., Katon, W., Callahan, C.M., et al., "Collaborative care management of late-life depression in the primary care setting: a randomized controlled trial." Jama, 288.22 (2002): 2836-2845.
[157] Vapnik V. (2000) The nature of statistical learning theory. Springer Science & Business Media, New York.
[158] Vella A, Rizza R (2011) clinical dilemmas in diabetes, Wiley-Blackwell.
[159] Vermunt, J. K. and Magidson, J. (2002), Latent class cluster analysis, *Applied latent class analysis*, 29-106.

[160] Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. Med. Decis. Making 2006;26:565-74.
[161] Wang, F., Sun, J., Hu, J., and Ebadollahi, S. (2011), Imet: Interactive metric learning in healthcare applications, *SDM* 2011.
[162] Weissman M.M., Bland R.C., Canino G.J., et al. Cross-national epidemiology of major depression and bipolar disorder. JAMA, 276.4 (1996):293–299.
[163] Wilkosz, P. A., Seltman, H. J., Devlin, B., et al. (2010), Trajectories of cognitive decline in Alzheimer`s disease, *Int. Psychogeriatric*, 22, 281-290.
[164] Wimo, A., Winblad, B., and Jonsson, L. (2007), An estimate of the total worldwide societal costs of dementia in 2005, *Alzheimer's and Dementia*, 3, 81–91.
[165] Wiser, R., and Bolinger, M., Annual report on US wind power installation, cost, and performance trend: 2007. Technical report, US Department of Energy, Washington, DC, 2008.
[166] Xu P, Wu U, Zhu Y, Dagne G, Johnson G, et al. (2010) Prognostic performance of metabolic indexes in predicting onset of type 1 diabetes. Diabetes Care 33: 2508–1513.
[167] Z. Q. Shen, and F. S. Kong, "Dynamically weighted ensemble neural networks for regression problems". In Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on (Vol. 6, pp. 3492-3496). IEEE.
[168] Zhang, P., et al., Towards personalized medicine: leveraging patient similarity and drug similarity analytics. *AMIA Summits on Translational Science Proceedings* 2014, 132 (2014).
[169] Zhou, J., Chen, J., and Ye, J. (2012), MALSAR: multi-task learning via structural regularization, *Arizona State University*. Available at http://www.public.asu.edu/~jye02/Software/MALSAR.
[170] Zhou, J., Liu, J., Narayan, V. A., and Ye, J. (2013), Modeling disease progression via multi-task learning, *NeuroImage*, 78, 233–248.
[171] Ziegler A. G., Nepom G. T. (2010) Prediction and pathogenesis in type 1 diabetes. Immunity, 32(4):468-478.
[172]

APPENDIX A

Proof of Theorem 4-1:

In what follows we will show the equivalence between the objective function of the MEM model with the objective function of our SCM model, given number of latent classes, K and penalty parameter λ. Firstly, we consider the objective function of MEM. The MEM uses fixed effects, $\boldsymbol{b}_0^{MEM}$, and random effects, $\boldsymbol{b}_{ri}^{MEM}$, to model the degradation of measures on each subject. It also assumes the random effects are correlated which can be formulated as:

$$\boldsymbol{y}_i = \mathbf{X}_i\boldsymbol{b}_0^{MEM} + \mathbf{X}_i\boldsymbol{b}_{ri}^{MEM} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{b}_{ri}^{MEM} \sim N(\mathbf{0}, \mathbf{G}), \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \omega^2\mathbf{I})$$

The conditional distribution of $\boldsymbol{y}_i$ is:

$$\boldsymbol{y}_i|\boldsymbol{b}_{ri}^{MEM} \sim N(\mathbf{X}_i\boldsymbol{b}_0^{MEM} + \mathbf{X}_i\boldsymbol{b}_{ri}^{MEM}, \omega^2\mathbf{I})$$

Based on the conditional distribution of $\boldsymbol{y}_i$, we can derive the log-likelihood function as:

$$ML_{MEM} = \sum_i \left\|\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{b}_0^{MEM} - \mathbf{X}_i\boldsymbol{b}_{ri}^{MEM}\right\|^2 + \omega^2 \sum_i (\boldsymbol{b}_{ri}^{MEM})^{\mathrm{T}}\mathbf{G}^{-1}\boldsymbol{b}_{ri}^{MEM}.$$

Use $\boldsymbol{b}_i^{MEM} = \boldsymbol{b}_0^{MEM} + \boldsymbol{b}_{ri}^{MEM}$ to replace the $\boldsymbol{b}_{ri}^{MEM}$ in the ML_{MEM} we have:

$$ML_{MEM} = \sum_i \left\|\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{b}_i^{MEM}\right\|^2 + \omega^2 \sum_i [(\boldsymbol{b}_i^{MEM} - \boldsymbol{b}_0^{MEM})^{\mathrm{T}}\mathbf{G}^{-1}(\boldsymbol{b}_i^{MEM} - \boldsymbol{b}_0^{MEM})],$$

$$= \sum_i \left\|\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{b}_i^{MEM}\right\|^2 + \omega^2 \sum_i [(\boldsymbol{b}_i^{MEM})^{\mathrm{T}}\mathbf{G}^{-1}\boldsymbol{b}_i^{MEM} - 2(\boldsymbol{b}_0^{MEM})^{\mathrm{T}}\mathbf{G}^{-1}\boldsymbol{b}_i^{MEM} +$$

$$(\boldsymbol{b}_0^{MEM})^{\mathrm{T}}\mathbf{G}^{-1}\boldsymbol{b}_0^{MEM}].$$

Secondly, given K and λ, we consider the objective function of SCM. The objective function of SCM is:

$$ML_{SCM} = \sum_i || \boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i||^2 + \lambda\left[\sum_i (\boldsymbol{c}_i)^{\mathrm{T}}\boldsymbol{c}_i\mathbf{D}_{ii} - \sum_{i,j} (\boldsymbol{c}_i)^{\mathrm{T}}\boldsymbol{c}_j w_{ij}\right].$$

Use $\boldsymbol{b}_i^{SCM} = \mathbf{Q}\boldsymbol{c}_i$ to replace $\boldsymbol{c}_i$ in the objective function ML_{SCM}, we have $\boldsymbol{c}_i = (\mathbf{Q}^{\mathrm{T}}\mathbf{Q})^{-1}\mathbf{Q}^{\mathrm{T}}\boldsymbol{b}_i^{SCM} = \mathbf{Q}^{+}\boldsymbol{b}_i^{SCM}$. $\mathbf{Q}^{+}$ is the pseudo inverse of $\mathbf{Q}$, which means $\mathbf{Q}\mathbf{Q}^{+}\mathbf{Q} = \mathbf{Q}$. Then ML_{SCM} can be rewritten as

$$ML_{SCM} = \sum_i || \boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{b}_i^{SCM}||^2 + \lambda$$

where $\mathbf{\Lambda}^{-1} = (\mathbf{Q}^{+})^{\mathrm{T}}\mathbf{Q}^{+}$, which leads to $\mathbf{\Lambda} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}}$.

When $\mathbf{W}$ is $\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$, e.g., $\mathbf{W}_{ij} = 1, \mathbf{D}_{ii} = N$, we have:

$$ML_{SCM} = \sum_i || \boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{b}_i^{SCM}||^2 + \lambda N \sum_i \Big[\big(\boldsymbol{b}_i^{SCM}\big)^{\mathrm{T}}\mathbf{\Lambda}^{-1}\boldsymbol{b}_i^{SCM} - 2(\boldsymbol{b}_0^{SCM})^{\mathrm{T}}\mathbf{\Lambda}^{-1}\boldsymbol{b}_i^{SCM} + (\boldsymbol{b}_0^{SCM})^{\mathrm{T}}\mathbf{\Lambda}^{-1}\boldsymbol{b}_0^{SCM}\Big],$$

where $\boldsymbol{b}_0^{SCM} = \frac{\sum_i \boldsymbol{b}_i^{SCM}}{N}$.

Comparing the formulation of ML_{SCM} and ML_{MEM}, we observe that, if $\lambda = \frac{\omega}{N}$, the objective function of MEM is equivalent to the objective function of SCM, with a constraint that $rank(\mathbf{G}) = rank(\mathbf{\Lambda}) \leq rank(\mathbf{Q}) = K$.

Proof of Theorem 4-2:

Our proof follows similar ideas used by Ding et al. (2006) and Cai et al. (2011). It's obvious that the objective function in (4.1) is bounded from below by zero, so the Lagrangian L is also bounded from below. The solution will converge if the Lagrangian L is monotonically nonincreasing. To prove Theorem 4-2, firstly, we need to show that the Lagrangian is nonincreasing in each step of the iterative algorithm, then, prove the iterative updates converge to a stationary point. Since in each iteration, the estimated $\mathbf{Q}$ minimizes the objective function**,** we only need to prove that the Lagrangian will be nonincreasing under the updating rule of $\mathbf{C}$ in (4.6).

Since the updating rule (4.6) is essentially element wise, it is sufficient to show that each $L_{ik}(c) = L(c, \mathbf{C}^m_{\backslash i \backslash k})$ is monotonically nonincreasing under the update step of (4.6). Note that $L_{ik}(c)$ only depends on c_{ik} when other values in $\mathbf{C}$ are given. To prove this, we need to use the concept of auxiliary function, which is similar to that used in the expectation-maximization algorithm (Dempster et al. 1997).

A function $G(c', c)$ is an auxiliary function of $L_{ik}(c)$ if

$$G(c', c) \geq L_{ik}(c');\ G(c, c) = L_{ik}(c). \tag{A.1}$$

By constructing $G(c', c)$, we define

$$c^{m+1} = arg\min_{c} G(c, c^m). \tag{A.2}$$

Thus, we have $L_{ik}(c^m) = G(c^m, c^m) \geq G(c^{m+1}, c^m) \geq L_{ik}(c^{m+1})$. This leads to the monotonicity of L_{ik} under the iterative updating rule of (A.2).

To construct an auxiliary function, we write $L_{ik}(c)$ using Taylor expression,

$$L_{ik}(c) = L_{ik}(c^m_{ik}) + L'_{ik}(c^m_{ik})(c - c^m_{ik}) + \frac{1}{2}L''_{ik}(c^m_{ik})(c - c^m_{ik})^2,$$

where

$$L'_{ik} = \frac{\partial L}{\partial c_{ik}} = [-2\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i + 2\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}\boldsymbol{c}^m_i]_k + 2\lambda(\mathbf{L}\mathbf{C}^m)_{ki} + \mu_i,$$

$$L''_{ik} = \frac{\partial L'}{\partial c_{ik}} = 2[\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}]_{kk} + 2\lambda\mathbf{D}_{kk},$$

$$L^{(t)}_{ik} = 0 \quad \text{for } t > 2.$$

Then we have

$$\frac{[\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}\boldsymbol{c}^m_i]_k + (\lambda\mathbf{D}\mathbf{C}^{\mathrm{m}})_{ki}}{c^m_{ik}} \geq [\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}]_{kk} + \lambda\mathbf{D}_{kk} = \frac{1}{2}L''_{ik}(c^m_{ik}),$$

because

$$[\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}\boldsymbol{c}^m_i]_k = \sum_j [\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}]_{kj}c^m_{ij} \geq [\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}]_{kk}c^m_{ik},$$

and

$$(\lambda \mathbf{D}\mathbf{C}^m)_{ki} = \sum_j \lambda \mathbf{D}_{kj} c_{ij}^m \geq \lambda \mathbf{D}_{kk} c_{ik}^m,$$

given that $\mathbf{X}_i\mathbf{Q} \geq \mathbf{0}$, $\boldsymbol{c}_i \geq \mathbf{0}$.

This leads to

$$\frac{[\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i^m]_k + (\lambda \mathbf{D}\mathbf{C}^{\mathrm{m}})_{ki} + [\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]^{\mathrm{T}}\boldsymbol{c}_i^m + \lambda[(\mathbf{W}\mathbf{C}^{\mathrm{m}})_i]^{\mathrm{T}}\boldsymbol{c}_i^m}{c_{ik}^m}$$

$$\geq [\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}]_{kk} + \lambda \mathbf{D}_{kk}\ .$$

Therefore, we can show the following function is an auxiliary function of $L_{ik}(c)$:

$$G(c, c_{ik}^m) = L_{ik}(c_{ik}^m) + L'_{ik}(c_{ik}^m)(c - c_{ik}^m)$$

$$+ \frac{[\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i^m]_k + (\lambda \mathbf{D}\mathbf{C}^{\mathrm{m}})_{ki} + [\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]^{\mathrm{T}}\boldsymbol{c}_i^m + \lambda[(\mathbf{W}\mathbf{C}^{\mathrm{m}})_i]^{\mathrm{T}}\boldsymbol{c}_i^m}{c_{ik}^m}(c - c_{ik}^m)^2.$$

This is because $G(\cdot,\cdot)$ satisfies the conditions in (A.1): the last term in $G(c, c_{ik}^m) \geq L_{ik}(c)$; and the equality holds when $c = c_{ik}^m$.

The minimum for (A.2) can be obtained by setting the gradient to zero, i.e.,

$$\frac{\partial G(c, c_{ik}^m)}{\partial c} = L'_{ik}(c_{ik}^m) + 2\frac{[\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i^m]_k + (\lambda \mathbf{D}\mathbf{C}^{\mathrm{m}})_{ki} + [\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]^{\mathrm{T}}\boldsymbol{c}_i^m + \lambda[(\mathbf{W}\mathbf{C}^{\mathrm{m}})_i]^{\mathrm{T}}\boldsymbol{c}_i^m}{c_{ik}^m}(c - c_{ik}^m),$$

$$= 2\left[[-\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]_k - \lambda(\mathbf{W}\mathbf{C}^m)_{ki} + \frac{1}{2}\mu_i - [\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]^{\mathrm{T}}\boldsymbol{c}_i^m - \lambda[(\mathbf{W}\mathbf{C}^{\mathrm{m}})_i]^{\mathrm{T}}\boldsymbol{c}_i^m + \right.$$

$$\left. \frac{[\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i^m]_k + (\lambda \mathbf{D}\mathbf{C}^{\mathrm{m}})_{ki} + [\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]^{\mathrm{T}}\boldsymbol{c}_i^m + \lambda[(\mathbf{W}\mathbf{C}^{\mathrm{m}})_i]^{\mathrm{T}}\boldsymbol{c}_i^m}{c_{ik}^m} c \right] = 0.$$

This leads to the following solution:

$$c = c_{ik}^m \frac{[\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]_k + \lambda(\mathbf{W}\mathbf{C}^m)_{ki} + [\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]^{\mathrm{T}}\boldsymbol{c}_i^m + \lambda[(\mathbf{W}\mathbf{C}^{\mathrm{m}})_i]^{\mathrm{T}}\boldsymbol{c}_i^m - \frac{1}{2}\mu_i}{[\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i^m]_k + (\lambda \mathbf{D}\mathbf{C}^m)_{ki} + [\mathbf{Q}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]^{\mathrm{T}}\boldsymbol{c}_i^m + \lambda[(\mathbf{W}\mathbf{C}^{\mathrm{m}})_i]^{\mathrm{T}}\boldsymbol{c}_i^m}.$$

By substituting the Lagrange multiplier μ_i as shown in (3.5) in the equation above, we recover the updating rule (4.6).

Next, we prove the iterative updates converge to a stationary point that satisfies the Karush Kuhn Tucker conditions.

Lemma 1. *Starting from an arbitrary feasible nonzero point* $\mathbf{C}^0$ *and* $\mathbf{Q}^0$*, the iterative procedure based on updating rule in (4.6) converge to a point that satisfies the KKT conditions for the optimization problem:*

$$min_{\boldsymbol{c}_i, i=1,\dots,k} \sum_i \left|\left|\boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}^*\boldsymbol{c}_i\right|\right|_{\mathrm{F}}^2 + \lambda \mathrm{Tr}(\mathbf{C}^{\mathrm{T}}\mathbf{L}\mathbf{C}), \tag{A.3}$$

$$subject\ to: \boldsymbol{c}_i \geq 0, \boldsymbol{c}_i{}^{\mathrm{T}}\mathbf{1} = 1, \forall\, i = 1, \dots, N.$$

Proof of Lemma 1:

We first write the KKT conditions for the optimization problem in (A.3):

- Stationarity:

$$-2(\mathbf{Q}^*)^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i + 2(\mathbf{Q}^*)^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^*\boldsymbol{c}_i + 2(\lambda\mathbf{L}\mathbf{C})_i + \boldsymbol{\varphi}_i + \mu_i\mathbf{1} = 0, \forall\, i = 1, \dots, N,$$

- Primal feasibility:

$$\boldsymbol{c}_i \geq \mathbf{0}, \forall\, i = 1, \dots, N,$$

$$\boldsymbol{c}_i{}^{\mathrm{T}}\mathbf{1} = 1, \forall\, i = 1, \dots, N,$$

- Dual feasibility:

$$\boldsymbol{\varphi}_i \geq \mathbf{0}, \forall\, i = 1, \dots, N,$$

- Complementary slackness:

$$\varphi_{ik}c_{ik} = 0, \forall\, i = 1, \dots, N, k = 1, \dots, K.$$

It is straightforward to observe that, starting with non-negative nonzero $\mathbf{C}^0$, the updating rule in (4.6) always keeps $\boldsymbol{c}_i^m$ non-negative nonzero since the nonnegative assumption of the cognitive measures, $\boldsymbol{y}_i$ and $\mathbf{X}_i\mathbf{Q}^*$.

Assuming $\boldsymbol{c}_i^m$ converge to $\boldsymbol{c}_i^*$, we have

$$c_{ik}^{*} = c_{ik}^{*} \frac{[(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i + (\lambda \mathbf{W}\mathbf{C}^{*})_i]_k + [(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^{*}\boldsymbol{c}_i^{*}]^{\mathbf{T}}\boldsymbol{c}_i^{*} + \lambda[(\mathbf{D}\mathbf{C}^{*})_i]^{\mathrm{T}}\boldsymbol{c}_i^{*}}{[(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^{*}\boldsymbol{c}_i^{*} + (\lambda \mathbf{D}\mathbf{C}^{*})_i]_k + [(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]^{\mathrm{T}}\boldsymbol{c}_i^{*} + \lambda[(\mathbf{W}\mathbf{C}^{*})_i]^{\mathrm{T}}\boldsymbol{c}_i^{*}}. \quad (A.4)$$

The equation (A.4) implies:

$$c_{ik}^{*}\left\{\frac{[(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i + (\lambda \mathbf{W}\mathbf{C}^{*})_i]_k + [(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^{*}\boldsymbol{c}_i^{*}]^{\mathbf{T}}\boldsymbol{c}_i^{*} + \lambda[(\mathbf{D}\mathbf{C}^{*})_i]^{\mathrm{T}}\boldsymbol{c}_i^{*}}{[(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^{*}\boldsymbol{c}_i^{*} + (\lambda \mathbf{D}\mathbf{C}^{*})_i]_k + [(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]^{\mathrm{T}}\boldsymbol{c}_i^{*} + \lambda[(\mathbf{W}\mathbf{C}^{*})_i]^{\mathrm{T}}\boldsymbol{c}_i^{*}} - 1\right\} = 0. \quad (A.5)$$

Therefore, we obtain

$$\sum_k \left\{c_{ik}^{*}\left\{\frac{[(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i + (\lambda \mathbf{W}\mathbf{C}^{*})_i]_k + [(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^{*}\boldsymbol{c}_i^{*}]^{\mathbf{T}}\boldsymbol{c}_i^{*} + \lambda[(\mathbf{D}\mathbf{C}^{*})_i]^{\mathrm{T}}\boldsymbol{c}_i^{*}}{[(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^{*}\boldsymbol{c}_i^{*} + (\lambda \mathbf{D}\mathbf{C}^{*})_i]_k + [(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]^{\mathrm{T}}\boldsymbol{c}_i^{*} + \lambda[(\mathbf{W}\mathbf{C}^{*})_i]^{\mathrm{T}}\boldsymbol{c}_i^{*}} - 1\right\}\right\} = 0,$$

and

$$\{[(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^{*}\boldsymbol{c}_i^{*}]^{\mathbf{T}}\boldsymbol{c}_i^{*} + \lambda[(\mathbf{D}\mathbf{C}^{*})_i]^{\mathrm{T}}\boldsymbol{c}_i^{*} + [(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]^{\mathrm{T}}\boldsymbol{c}_i^{*} + \lambda[(\mathbf{W}\mathbf{C}^{*})_i]^{\mathrm{T}}\boldsymbol{c}_i^{*}\}(\sum_k c_{ik}^{*} - 1) = 0.$$

This leads to $\sum_k c_{ik}^{*} - 1 = 0$, because $\boldsymbol{c}_i^{*}$ is nonnegative nonzero under the updating rule. Consequently, the first term in the equation is nonzero. The updates $\boldsymbol{c}_i^{m}$ converge to a feasible solution.

By (3.5) it is known that

$$\frac{1}{2}\mu_i = [(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i]^{\mathrm{T}}\boldsymbol{c}_i^{*} + \lambda[(\mathbf{W}\mathbf{C}^{*})_i]^{\mathrm{T}}\boldsymbol{c}_i^{*} - [(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^{*}\boldsymbol{c}_i^{*}]^{\mathbf{T}}\boldsymbol{c}_i^{*} - \lambda[(\mathbf{D}\mathbf{C}^{*})_i]^{\mathrm{T}}\boldsymbol{c}_i^{*}.$$

With the equation (A.5), we have:

$$c_{ik}^{*}\left[[(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i + (\lambda \mathbf{W}\mathbf{C}^{*})_i]_k - [(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^{*}\boldsymbol{c}_i^{*} + (\lambda \mathbf{D}\mathbf{C}^{*})_i]_k - \frac{1}{2}\mu_i\right] = 0.$$

with

$$\varphi_{ik}^{*} = 2[(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{y}_i + (\lambda \mathbf{W}\mathbf{C}^{*})_i]_k - 2[(\mathbf{Q}^{*})^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^{*}\boldsymbol{c}_i^{*} + (\lambda \mathbf{D}\mathbf{C}^{*})_i]_k - \mu_i,$$

Showing that both the complementary slackness and stationarity conditions are satisfied. Because c_{ik}^{*} is nonnegative nonzero, φ_{ik}^{*} is necessary to be zero. Thus, all the KKT conditions are satisfied in $\boldsymbol{c}_i^{*}$ based on our updating rules.

Proof of Theorem 5-1:

It's obvious that the log-likelihood function in (5.4) is bounded from above by zero, so the Lagrangian L is also bounded from above. The solution will converge if the Lagrangian L is monotonically nondecreasing. To prove Theorem 5-1 we need to show that the Lagrangian is nondecreasing under the updating rules of canonical models and $\mathbf{C}$ in (5.9) and (5.13) in each step of the iterative algorithm.

Since the updating rules are essentially element wise, it is sufficient to show that $L_{ks}(\theta) = L\left(\theta, \boldsymbol{\theta}^{m}_{\backslash k\backslash s}, \boldsymbol{\Pi}^{m}, \mathbf{C}^{m}\right)$, $L_{ks_1s_2}(\pi) = L\left(\pi, \boldsymbol{\Pi}^{m}_{\backslash k\backslash s_1\backslash s_2}, \boldsymbol{\theta}^{m}, \mathbf{C}^{m}\right)$, and $L_{ik}(c) = L(c, \mathbf{C}^{m}_{\backslash i\backslash k}, \boldsymbol{\theta}^{m}, \boldsymbol{\Pi}^{m})$ are monotonically nondecreasing under the update steps in (5.9) and (5.13). To prove this, we first decompose the Lagrangian function into three subfunctions:

$$F_1(\boldsymbol{\theta}, \mathbf{C}) = \sum_{i=1}^{N} \sum_{s} e_{is} \log[\sum_{k} c_{ik}\theta_{ks}] - \sum_{k} a_k(\sum_{\boldsymbol{s}} \theta_{ks} - 1),$$

$$F_2(\boldsymbol{\Pi}, \mathbf{C}) = \sum_{i=1}^{N} \sum_{s_1,s_2} N_i(s_1, s_2)\log[\sum_{k} c_{ik}\boldsymbol{\Pi}_k(s_1, s_2)] - \sum_{k,s} b_{ks}\left(\sum_{s_2} \boldsymbol{\Pi}_k(s, s_2) - 1\right),$$

$$F_3(\boldsymbol{\theta}, \boldsymbol{\Pi}, \mathbf{C}) = \sum_{i=1}^{N}\left\{\sum_{s} e_{is} \log[\sum_{k} c_{ik}\theta_{ks}] + \sum_{s_1,s_2} N_i(s_1, s_2)\log[\sum_{k} c_{ik}\boldsymbol{\Pi}_k(s_1, s_2)]\right\} -$$

$$\sum_{i=1}^{N} \mu_i(\boldsymbol{c_i}\mathbf{1} - 1),$$

The problem is simplified to prove $F_1\left(\theta, \boldsymbol{\theta}^{m}_{\backslash k\backslash s}, \mathbf{C}^{m}\right)$, $F_2\left(\pi, \boldsymbol{\Pi}^{m}_{\backslash k\backslash s_1\backslash s_2}, \mathbf{C}^{m}\right)$, and $F_3\left(c, \mathbf{C}^{m}_{\backslash i\backslash k}, \boldsymbol{\theta}^{m}, \boldsymbol{\Pi}^{m}\right)$ monotonically nondecreasing under the update steps in (5.9) and (5.10). We use the concept of auxiliary function, which is similar to that used in the expectation-maximization algorithm (Dempster et al. 1997) and two inequalities in Lemma 1 and Lemma 2 to prove it.

A function $G(x', x)$ is an auxiliary function of $L(x)$ if

$$G(x', x) \leq L(x);\ G(x, x) = L(x)$$

By constructing $G(x', x)$, we define

$$x^{m+1} = arg\max_{x} G(x, x^m)$$

Thus, we have $L(x^m) = G(x^m, x^m) \leq G(x^{m+1}, x^m) \leq L(x^{m+1})$. This leads to the monotonicity of L under the iterative updating rule of (A.2).

To construct the auxiliary functions for $L_{ks}(\theta)$ and $L_{ks_1s_2}(\pi)$, we need to use the inequality in Lemma 1.

*Lemma 1:*For any positive variables x_k, $\log(\sum_k x_k) \geq \sum_k q_k \log\left(\frac{x_k}{q_k}\right)$, with $\sum_k q_k = 1, q_k \geq 0$

To construct the auxiliary functions for $L_{ik}(c)$, we need to use the inequality in Lemma 2.

*Lemma 2:*For any positive variable x, $\log(x) + 1 \leq x$.

Using Lemma 1, we have the following observations:

$$\log[\sum_k c_{ik}\theta_{ks}] \geq \sum_k \frac{c_{ik}^m \theta_{ks}^m}{\sum_k c_{ik}^m \theta_{ks}^m}\left[\log(c_{ik}\theta_{ks}) - \log\left(\frac{c_{ik}^m \theta_{ks}^m}{\sum_k c_{ik}^m \theta_{ks}^m}\right)\right],$$

$$\log[\sum_k c_{ik}\mathbf{\Pi}_k(s_1, s_2)] \geq \sum_k \frac{c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m}{\sum_k c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m}\left[\log\left(c_{ik}\mathbf{\Pi}_k(s_1, s_2)\right) - \log\left(\frac{c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m}{\sum_k c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m}\right)\right],$$

the equalities are achieved if and only if $\theta_{ks} = \theta_{ks}^m$, and $\mathbf{\Pi}_k(s_1, s_2) = \mathbf{\Pi}_k(s_1, s_2)^m$.

Thus we get the auxiliary functions for F_1 and F_2 as:

$$G_1\left(\theta, \boldsymbol{\theta}_{\backslash k \backslash s}^m, \mathbf{C}^m\right) = \sum_i e_{is} \frac{c_{ik}^m \theta_{ks}^m}{\sum_k c_{ik}^m \theta_{ks}^m}\left[\log(c_{ik}^m \theta) - \log\left(\frac{c_{ik}^m \theta_{ks}^m}{\sum_k c_{ik}^m \theta_{ks}^m}\right)\right] - a_k\theta + M_1,$$

$$G_2\left(\pi, \mathbf{\Pi}_{\backslash k \backslash s_1 \backslash s_2}^m, \mathbf{C}^m\right) = \sum_i \frac{c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m}{\sum_k c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m}\left[\log(c_{ik}^m \pi) - \log\left(\frac{c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m}{\sum_k c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m}\right)\right] - b_{ks}\pi + M_2,$$

where M_1 and M_2 represent the parts unrelated to θ_{ks} and $\mathbf{\Pi}_k(s_1, s_2)$ in F_1 and F_2 respectively. Letting the partial deviations of G_1 and G_2 with respect to $\theta_{ks}, \mathbf{\Pi}_k(s_1, s_2)$ to be zeros, we have:

$$\frac{\partial G_1}{\partial \theta} = \frac{1}{\theta}\sum_i \frac{c_{ik}^m \theta_{ks}^m}{\sum_k c_{ik}^m \theta_{ks}^m} - a_k = 0,$$

$$\frac{\partial G_2}{\partial \pi} = \frac{1}{\pi}\sum_i \frac{c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m}{\sum_k c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m} - b_{ks} = 0,$$

which lead to the solutions of θ_{ks}^{m+1} and $\mathbf{\Pi}_k(s_1, s_2)^{m+1}$ that maximize F_1 and F_2.

$$\theta_{ks}^{m+1} = \frac{\sum_i \frac{c_{ik}^m \theta_{ks}^m}{\sum_k c_{ik}^m \theta_{ks}^m}}{a_k},$$

$$\mathbf{\Pi}_k(s_1, s_2)^{m+1} = \frac{\sum_i \frac{c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m}{\sum_k c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m}}{b_{ks}},$$

Using the expressions of a_k and b_{ks} in (5.8), we can obtain the updating rules in (5.9).

To find the auxiliary functions for F_3 we approximate the $\frac{\lambda}{2}\mathrm{Tr}(\mathbf{C^T LC})$ term using Tylor expression at c_{ik}^m:

$$\frac{\lambda}{2}\mathrm{Tr}(\mathbf{C^T LC}) = \frac{\lambda}{2}\mathrm{Tr}\left(\mathbf{C}^{m\mathbf{T}}\mathbf{L}\mathbf{C}^m\right) + \lambda(\mathbf{L}\mathbf{C}^m)_{ik}(c - c_{ik}^m) + \frac{\lambda}{2}\mathbf{D}_{ii}(c - c_{ik}^m)^2 = \frac{\lambda}{2}\mathrm{Tr}\left(\mathbf{C}^{m\mathbf{T}}\mathbf{L}\mathbf{C}^m\right)$$

$$+\lambda\left[(\mathbf{D}\mathbf{C}^m)_{ik} + (\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right]\left[(c - c_{ik}^m) + \frac{\mathbf{D}_{ii}}{2(\mathbf{D}\mathbf{C}^m)_{ik} + 2(\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathbf{T}}}(c - c_{ik}^m)^2\right] - \lambda\left[(\mathbf{W}\mathbf{C}^m)_{ik} +\right.$$

$$\left.(\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right] c_{ik}^m \left(\frac{c}{c_{ik}^m} - 1\right),$$

By using Lemma 2, we have:

$$\frac{\lambda}{2}\mathrm{Tr}(\mathbf{C^T LC}) \le \frac{\lambda}{2}\mathrm{Tr}\left(\mathbf{C}^{m\mathbf{T}}\mathbf{L}\mathbf{C}^m\right) + \lambda\left[(\mathbf{D}\mathbf{C}^m)_{ik} + (\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right]\left[(c - c_{ik}^m) +\right.$$

$$\left.\frac{\mathbf{D}_{ii}}{2(\mathbf{D}\mathbf{C}^m)_{ik} + 2(\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathbf{T}}}(c - c_{ik}^m)^2\right] - \lambda\left[(\mathbf{W}\mathbf{C}^m)_{ik} + (\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right] c_{ik}^m \log\left(\frac{c}{c_{ik}^m}\right),$$

the equality is achieved if and only if $c = c_{ik}^m$. Thus we get the auxiliary function for F_3 as:

$$G_3\left(c, \mathbf{C}_{\backslash i\backslash k}^m, \boldsymbol{\theta}^m, \mathbf{\Pi}^m\right) = G_1\left(c, \mathbf{C}_{\backslash i\backslash k}^m, \boldsymbol{\theta}^m\right) + G_2\left(c, \mathbf{C}_{\backslash i\backslash k}^m, \mathbf{\Pi}^m\right) - \frac{\lambda}{2}\mathrm{Tr}\left(\mathbf{C}^{m\mathbf{T}}\mathbf{L}\mathbf{C}^m\right) -$$

$$\lambda\left[(\mathbf{D}\mathbf{C}^m)_{ik} + (\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right]\left[(c - c_{ik}^m) + \frac{\mathbf{D}_{ii}}{2(\mathbf{D}\mathbf{C}^m)_{ik} + 2(\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathbf{T}}}(c - c_{ik}^m)^2\right] + \lambda\left[(\mathbf{W}\mathbf{C}^m)_{ik} +\right.$$

$$\left.(\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right] \log\left(\frac{c}{c_{ik}^m}\right) - \mu_i c + M_3,$$

Letting the partial deviation of G_3 with respect to c_{ik} to be zero, we have:

$$\frac{\partial G_3}{\partial c} = \frac{1}{c}\sum_i \frac{c_{ik}^m \theta_{ks}^m}{\sum_k c_{ik}^m \theta_{ks}^m} + \frac{1}{c}\sum_i \frac{c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m}{\sum_k c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m} - \lambda\left[(\mathbf{D}\mathbf{C}^m)_{ik} + (\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right] + \lambda\left[(\mathbf{W}\mathbf{C}^m)_{ik} +\right.$$

$$\left.(\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right]\frac{1}{c} - \mu_i = 0,$$

We can obtain the solution of c_{ik}^{m+1} that maximize the auxiliary function G_3 as:

$$c_{ik}^{m+1} = \frac{\sum_i \frac{c_{ik}^m \theta_{ks}^m}{\sum_k c_{ik}^m \theta_{ks}^m} + \sum_i \frac{c_{ik}^m \boldsymbol{\Pi}_k(s_1,s_2)^m}{\sum_k c_{ik}^m \boldsymbol{\Pi}_k(s_1,s_2)^m} + \lambda\left[(\mathbf{WC}^m)_{ik} + (\mathbf{DC}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right]}{\mu_i + \lambda\left[(\mathbf{DC}^m)_{ik} + (\mathbf{DC}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right]},$$

By substituting the Lagrange multiplier μ_i as shown in (5.11) in the equation above, we recover the updating rule (5.13).

Parameters of canonical models in simulation study:

$$\boldsymbol{\Pi}_1 = \begin{bmatrix} 0.20 & 0.50 & 0.10 & 0.10 & 0.10 \\ 0.00 & 0.20 & 0.50 & 0.15 & 0.15 \\ 0.00 & 0.00 & 0.20 & 0.50 & 0.30 \\ 0.00 & 0.00 & 0.00 & 0.20 & 0.80 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}$$

$$\boldsymbol{\Pi}_2 = \begin{bmatrix} 0.56 & 0.44 & 0.00 & 0.00 & 0.00 \\ 0.20 & 0.65 & 0.15 & 0.00 & 0.00 \\ 0.17 & 0.50 & 0.33 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.25 & 0.25 & 0.50 \\ 0.00 & 0.00 & 0.00 & 0.27 & 0.73 \end{bmatrix}$$

$$\boldsymbol{\Pi}_3 = \begin{bmatrix} 0.98 & 0.02 & 0.00 & 0.00 & 0.00 \\ 0.05 & 0.94 & 0.01 & 0.00 & 0.00 \\ 0.00 & 0.19 & 0.80 & 0.01 & 0.00 \\ 0.00 & 0.00 & 0.38 & 0.61 & 0.01 \\ 0.00 & 0.00 & 0.00 & 0.43 & 0.57 \end{bmatrix}$$

$$\boldsymbol{\theta}_1 = [0.80 \quad 0.20 \quad 0.00 \quad 0.00 \quad 0.00]$$

$$\boldsymbol{\theta}_2 = [0.00 \quad 0.00 \quad 0.20 \quad 0.50 \quad 0.30]$$

$$\boldsymbol{\theta}_3 = [0.20 \quad 0.20 \quad 0.20 \quad 0.20 \quad 0.20]$$

APPENDIX B

Proof of Theorem 1:

It's obvious that the log-likelihood function in (2.4) is bounded from above by zero, so the Lagrangian L is also bounded from above. The solution will converge if the Lagrangian L is monotonically nondecreasing. To prove Theorem 1 we need to show that the Lagrangian is nondecreasing under the updating rules of canonical models and $\mathbf{C}$ in (2.9) and (2.13) in each step of the iterative algorithm.

Since the updating rules are essentially element wise, it is sufficient to show that $L_{ks}(\theta) = L\left(\theta, \boldsymbol{\theta}^{m}_{\backslash k\backslash s}, \boldsymbol{\Pi}^{m}, \mathbf{C}^{m}\right)$, $L_{ks_1s_2}(\pi) = L\left(\pi, \boldsymbol{\Pi}^{m}_{\backslash k\backslash s_1\backslash s_2}, \boldsymbol{\theta}^{m}, \mathbf{C}^{m}\right)$, and $L_{ik}(c) = L(c, \mathbf{C}^{m}_{\backslash i\backslash k}, \boldsymbol{\theta}^{m}, \boldsymbol{\Pi}^{m})$ are monotonically nondecreasing under the update steps in (2.9) and (2.13). To prove this, we first decompose the Lagrangian function into three subfunctions:

$$F_1(\boldsymbol{\theta}, \mathbf{C}) = \sum_{i=1}^{N}\sum_{s} e_{is}\log[\sum_{k} c_{ik}\theta_{ks}] - \sum_{k} a_k(\sum_{\boldsymbol{s}} \theta_{ks} - 1),$$

$$F_2(\boldsymbol{\Pi}, \mathbf{C}) = \sum_{i=1}^{N}\sum_{s_1,s_2} N_i(s_1, s_2)\log[\sum_{k} c_{ik}\boldsymbol{\Pi}_k(s_1, s_2)] - \sum_{k,s} b_{ks}\left(\sum_{s_2}\boldsymbol{\Pi}_k(s, s_2) - 1\right),$$

$$F_3(\boldsymbol{\theta}, \boldsymbol{\Pi}, \mathbf{C}) = \sum_{i=1}^{N}\left\{\sum_{s} e_{is}\log[\sum_{k} c_{ik}\theta_{ks}] + \sum_{s_1,s_2} N_i(s_1, s_2)\log[\sum_{k} c_{ik}\boldsymbol{\Pi}_k(s_1, s_2)]\right\} -$$

$$\sum_{i=1}^{N}\mu_i(\boldsymbol{c_i}\mathbf{1} - 1),$$

The problem is simplified to prove $F_1\left(\theta, \boldsymbol{\theta}^{m}_{\backslash k\backslash s}, \mathbf{C}^{m}\right)$, $F_2\left(\pi, \boldsymbol{\Pi}^{m}_{\backslash k\backslash s_1\backslash s_2}, \mathbf{C}^{m}\right)$, and $F_3\left(c, \mathbf{C}^{m}_{\backslash i\backslash k}, \boldsymbol{\theta}^{m}, \boldsymbol{\Pi}^{m}\right)$ monotonically nondecreasing under the update steps in (2.9) and (2.10). We use the concept of auxiliary function, which is similar to that used in the expectation-maximization algorithm (Dempster et al. 1997) and two inequalities in Lemma 1 and Lemma 2 to prove it.

A function $G(x', x)$ is an auxiliary function of $L(x)$ if

$$G(x', x) \le L(x);\ G(x, x) = L(x)$$

By constructing $G(x', x)$, we define

$$x^{m+1} = arg\max_{x} G(x, x^m)$$

Thus, we have $L(x^m) = G(x^m, x^m) \leq G(x^{m+1}, x^m) \leq L(x^{m+1})$. This leads to the monotonicity of L under the iterative updating rule of (A.2).

To construct the auxiliary functions for $L_{ks}(\theta)$ and $L_{ks_1s_2}(\pi)$, we need to use the inequality in Lemma 1.

*Lemma 1:*For any positive variables x_k, $\log(\sum_k x_k) \geq \sum_k q_k \log\left(\frac{x_k}{q_k}\right)$, with $\sum_k q_k = 1, q_k \geq 0$

To construct the auxiliary functions for $L_{ik}(c)$, we need to use the inequality in Lemma 2.

*Lemma 2:*For any positive variable x, $\log(x) + 1 \leq x$.

Using Lemma 1, we have the following observations:

$$\log[\sum_k c_{ik}\theta_{ks}] \geq \sum_k \frac{c_{ik}^m \theta_{ks}^m}{\sum_k c_{ik}^m \theta_{ks}^m}\left[\log(c_{ik}\theta_{ks}) - \log\left(\frac{c_{ik}^m \theta_{ks}^m}{\sum_k c_{ik}^m \theta_{ks}^m}\right)\right],$$

$$\log[\sum_k c_{ik}\boldsymbol{\Pi}_k(s_1, s_2)] \geq \sum_k \frac{c_{ik}^m \boldsymbol{\Pi}_k(s_1,s_2)^m}{\sum_k c_{ik}^m \boldsymbol{\Pi}_k(s_1,s_2)^m}\left[\log\left(c_{ik}\boldsymbol{\Pi}_k(s_1, s_2)\right) - \log\left(\frac{c_{ik}^m \boldsymbol{\Pi}_k(s_1,s_2)^m}{\sum_k c_{ik}^m \boldsymbol{\Pi}_k(s_1,s_2)^m}\right)\right],$$

the equalities are achieved if and only if $\theta_{ks} = \theta_{ks}^m$, and $\boldsymbol{\Pi}_k(s_1, s_2) = \boldsymbol{\Pi}_k(s_1, s_2)^m$.

Thus we get the auxiliary functions for F_1 and F_2 as:

$$G_1\left(\theta, \boldsymbol{\theta}_{\backslash k \backslash s}^m, \mathbf{C}^m\right) = \sum_i e_{is} \frac{c_{ik}^m \theta_{ks}^m}{\sum_k c_{ik}^m \theta_{ks}^m}\left[\log(c_{ik}^m \theta) - \log\left(\frac{c_{ik}^m \theta_{ks}^m}{\sum_k c_{ik}^m \theta_{ks}^m}\right)\right] - a_k\theta + H_1,$$

$$G_2\left(\pi, \boldsymbol{\Pi}_{\backslash k \backslash s_1 \backslash s_2}^m, \mathbf{C}^m\right) = \sum_i \frac{c_{ik}^m \boldsymbol{\Pi}_k(s_1,s_2)^m}{\sum_k c_{ik}^m \boldsymbol{\Pi}_k(s_1,s_2)^m}\left[\log(c_{ik}^m \pi) - \log\left(\frac{c_{ik}^m \boldsymbol{\Pi}_k(s_1,s_2)^m}{\sum_k c_{ik}^m \boldsymbol{\Pi}_k(s_1,s_2)^m}\right)\right] - b_{ks}\pi + H_2,$$

where H_1 and H_2 represent the parts unrelated to θ_{ks} and $\boldsymbol{\Pi}_k(s_1, s_2)$ in F_1 and F_2 respectively. Letting the partial deviations of G_1 and G_2 with respect to $\theta_{ks}, \boldsymbol{\Pi}_k(s_1, s_2)$ to be zeros, we have:

$$\frac{\partial G_1}{\partial \theta} = \frac{1}{\theta}\sum_i \frac{c_{ik}^m \theta_{ks}^m}{\sum_k c_{ik}^m \theta_{ks}^m} - a_k = 0,$$

$$\frac{\partial G_2}{\partial \pi} = \frac{1}{\pi}\sum_i \frac{c_{ik}^m \boldsymbol{\Pi}_k(s_1,s_2)^m}{\sum_k c_{ik}^m \boldsymbol{\Pi}_k(s_1,s_2)^m} - b_{ks} = 0,$$

which lead to the solutions of θ_{ks}^{m+1} and $\mathbf{\Pi}_k(s_1,s_2)^{m+1}$ that maximize F_1 and F_2.

$$\theta_{ks}^{m+1} = \frac{\sum_i \frac{c_{ik}^m \theta_{ks}^m}{\sum_k c_{ik}^m \theta_{ks}^m}}{a_k},$$

$$\mathbf{\Pi}_k(s_1,s_2)^{m+1} = \frac{\sum_i \frac{c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m}{\sum_k c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m}}{b_{ks}},$$

Using the expressions of a_k and b_{ks} in (2.8), we can obtain the updating rules in (2.9).

To find the auxiliary functions for F_3 we approximate the $\frac{\lambda}{2}\mathrm{Tr}(\mathbf{C^T LC})$ term using Tylor expression at c_{ik}^m:

$$\frac{\lambda}{2}\mathrm{Tr}(\mathbf{C^T LC}) = \frac{\lambda}{2}\mathrm{Tr}\left(\mathbf{C}^{m\mathbf{T}}\mathbf{L}\mathbf{C}^m\right) + \lambda(\mathbf{L}\mathbf{C}^m)_{ik}(c - c_{ik}^m) + \frac{\lambda}{2}\mathbf{D}_{ii}(c - c_{ik}^m)^2 = \frac{\lambda}{2}\mathrm{Tr}\left(\mathbf{C}^{m\mathbf{T}}\mathbf{L}\mathbf{C}^m\right)$$

$$+\lambda\left[(\mathbf{D}\mathbf{C}^m)_{ik} + (\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right]\left[(c - c_{ik}^m) + \frac{\mathbf{D}_{ii}}{2(\mathbf{D}\mathbf{C}^m)_{ik} + 2(\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}}(c - c_{ik}^m)^2\right] - \lambda\left[(\mathbf{W}\mathbf{C}^m)_{ik} +\right.$$

$$\left.(\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right] c_{ik}^m \left(\frac{c}{c_{ik}^m} - 1\right),$$

By using Lemma 2, we have:

$$\frac{\lambda}{2}\mathrm{Tr}(\mathbf{C^T LC}) \le \frac{\lambda}{2}\mathrm{Tr}\left(\mathbf{C}^{m\mathbf{T}}\mathbf{L}\mathbf{C}^m\right) + \lambda\left[(\mathbf{D}\mathbf{C}^m)_{ik} + (\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right]\left[(c - c_{ik}^m) +\right.$$

$$\left.\frac{\mathbf{D}_{ii}}{2(\mathbf{D}\mathbf{C}^m)_{ik} + 2(\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}}(c - c_{ik}^m)^2\right] - \lambda\left[(\mathbf{W}\mathbf{C}^m)_{ik} + (\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right] c_{ik}^m \log\left(\frac{c}{c_{ik}^m}\right),$$

the equality is achieved if and only if $c = c_{ik}^m$. Thus we get the auxiliary function for F_3 as:

$$G_3\left(c, \mathbf{C}_{\backslash i\backslash k}^m, \boldsymbol{\theta}^m, \mathbf{\Pi}^m\right) = G_1\left(c, \mathbf{C}_{\backslash i\backslash k}^m, \boldsymbol{\theta}^m\right) + G_2\left(c, \mathbf{C}_{\backslash i\backslash k}^m, \mathbf{\Pi}^m\right) - \frac{\lambda}{2}\mathrm{Tr}\left(\mathbf{C}^{m\mathbf{T}}\mathbf{L}\mathbf{C}^m\right) -$$

$$\lambda\left[(\mathbf{D}\mathbf{C}^m)_{ik} + (\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right]\left[(c - c_{ik}^m) + \frac{\mathbf{D}_{ii}}{2(\mathbf{D}\mathbf{C}^m)_{ik} + 2(\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}}(c - c_{ik}^m)^2\right] + \lambda\left[(\mathbf{W}\mathbf{C}^m)_{ik} +\right.$$

$$\left.(\mathbf{D}\mathbf{C}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right] \log\left(\frac{c}{c_{ik}^m}\right) - \mu_i c + H_3,$$

Letting the partial deviation of G_3 with respect to c_{ik} to be zero, we have:

$$\frac{\partial G_3}{\partial c} = \frac{1}{c}\sum_i \frac{c_{ik}^m \theta_{ks}^m}{\sum_k c_{ik}^m \theta_{ks}^m} + \frac{1}{c}\sum_i \frac{c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m}{\sum_k c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m} - \lambda\left[(\mathbf{DC}^m)_{ik} + (\mathbf{DC}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right] + \lambda\left[(\mathbf{WC}^m)_{ik} + (\mathbf{DC}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right]\frac{1}{c} - \mu_i = 0,$$

We can obtain the solution of c_{ik}^{m+1} that maximize the auxiliary function G_3 as:

$$c_{ik}^{m+1} = \frac{\sum_i \frac{c_{ik}^m \theta_{ks}^m}{\sum_k c_{ik}^m \theta_{ks}^m} + \sum_i \frac{c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m}{\sum_k c_{ik}^m \mathbf{\Pi}_k(s_1,s_2)^m} + \lambda\left[(\mathbf{WC}^m)_{ik} + (\mathbf{DC}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right]}{\mu_i + \lambda\left[(\mathbf{DC}^m)_{ik} + (\mathbf{DC}^m)_i \boldsymbol{c_i}^{m\mathrm{T}}\right]},$$

By substituting the Lagrange multiplier μ_i as shown in (2.11) in the equation above, we recover the updating rule (2.13).

Parameters selection in simulation study:

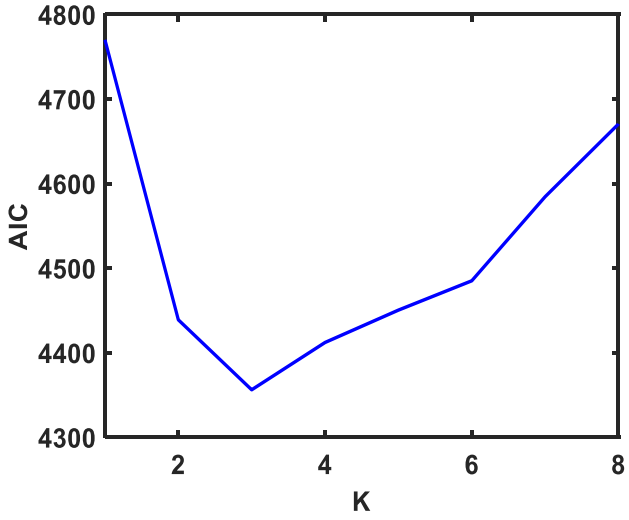


Figure B-1. The AIC value versus K for CM ($\zeta = 5$, $\varepsilon = 0.05$).

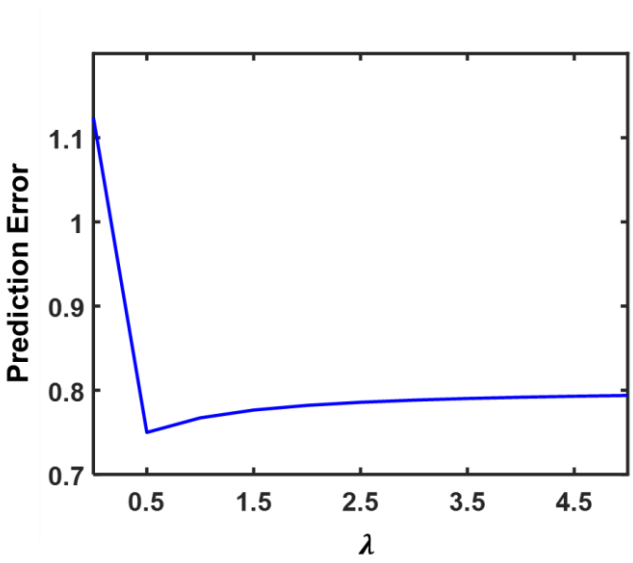


Figure B-2. The prediction error in LOOCV under different values of λ for CM ($\zeta = 5$, $\varepsilon = 0.05$).

Parameters of canonical models in simulation study:

$$\boldsymbol{\Pi}_1^S = \begin{bmatrix} 0.20 & 0.50 & 0.10 & 0.10 & 0.10 \\ 0.00 & 0.20 & 0.50 & 0.15 & 0.15 \\ 0.00 & 0.00 & 0.20 & 0.50 & 0.30 \\ 0.00 & 0.00 & 0.00 & 0.20 & 0.80 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}$$

$$\boldsymbol{\Pi}_2^S = \begin{bmatrix} 0.56 & 0.44 & 0.00 & 0.00 & 0.00 \\ 0.20 & 0.65 & 0.15 & 0.00 & 0.00 \\ 0.17 & 0.50 & 0.33 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.25 & 0.25 & 0.50 \\ 0.00 & 0.00 & 0.00 & 0.27 & 0.73 \end{bmatrix}$$

$$\boldsymbol{\Pi}_3^S = \begin{bmatrix} 0.98 & 0.02 & 0.00 & 0.00 & 0.00 \\ 0.05 & 0.94 & 0.01 & 0.00 & 0.00 \\ 0.00 & 0.19 & 0.80 & 0.01 & 0.00 \\ 0.00 & 0.00 & 0.38 & 0.61 & 0.01 \\ 0.00 & 0.00 & 0.00 & 0.43 & 0.57 \end{bmatrix}$$

$$\boldsymbol{\theta}_1^S = [0.80 \quad 0.20 \quad 0.00 \quad 0.00 \quad 0.00]$$

$$\boldsymbol{\theta}_2^S = [0.00 \quad 0.00 \quad 0.20 \quad 0.50 \quad 0.30]$$

$$\boldsymbol{\theta}_3^S = [0.20 \quad 0.20 \quad 0.20 \quad 0.20 \quad 0.20]$$

Sensitivity analysis in depression application:

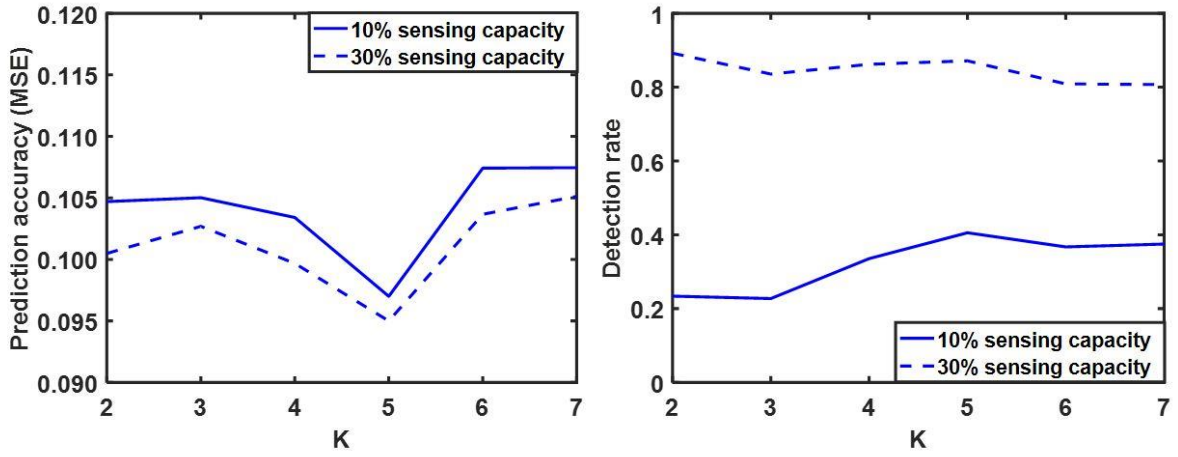


Figure B-3. Prediction and detection accuracy of the proposed method under different K in depression application.

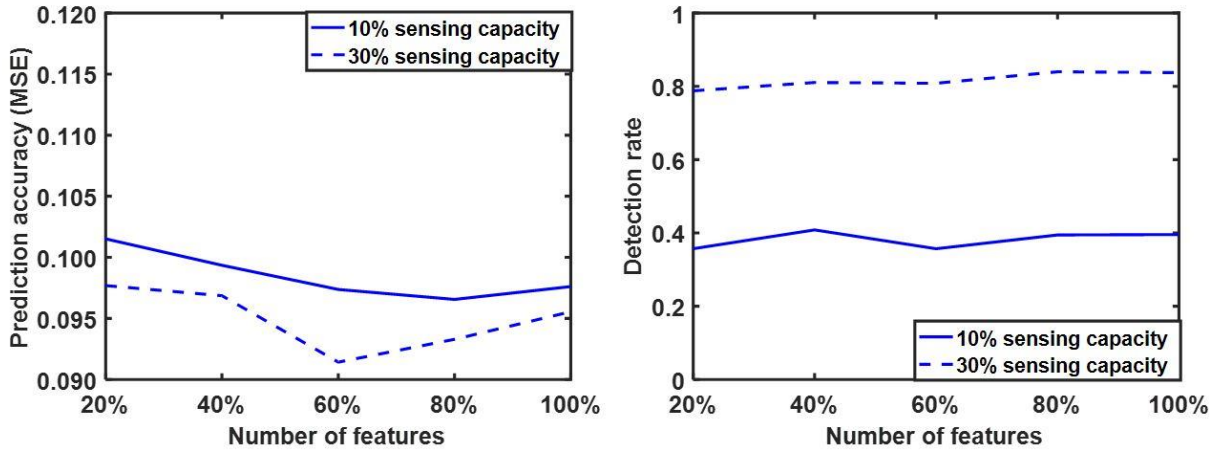


Figure B-4. Prediction and detection accuracy of the proposed method under different similarity structure in depression application.

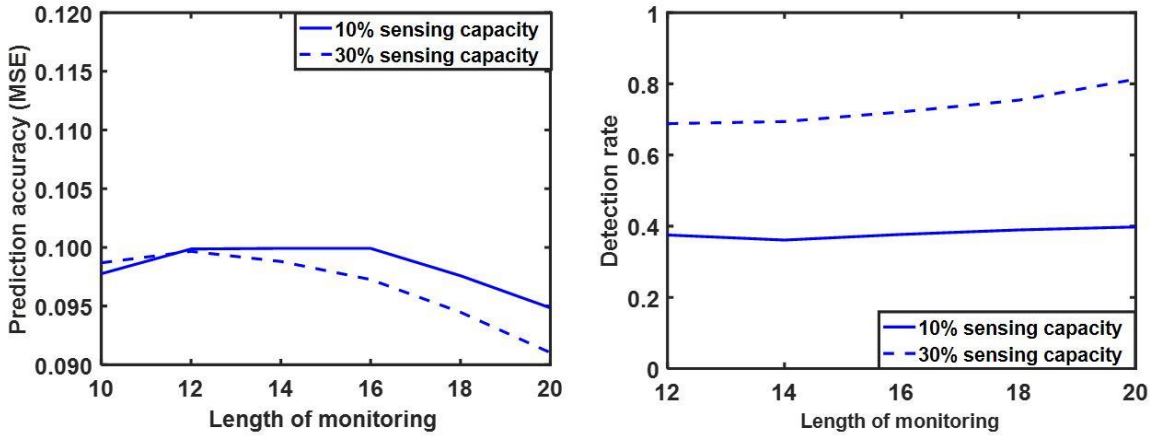


Figure B-5. Prediction and detection accuracy of the proposed method under different lengths of monitoring in depression application.

APPENDIX C

Table C-1: Significant predictors identified by logistic regression model ($\alpha = 0.05$).

Risk factors	Coefficient	SE	P-value
Intercept	4.49	1.88	0.02*****
Age between 30 to 44 years old	0.19	0.24	0.43
Age between 45 to 64 years old	-0.04	0.23	0.86
Age greater or equal to 65 years old	0.53	0.27	0.05*
Sex (1 = female)	-0.40	0.16	0.01*
Number of observations	0.04	0.06	0.49
Time stamp of first observation	0.00	0.00	0.81
Observing density	0.21	1.61	0.90
First Charlson comorbidity score	0.01	0.33	0.98
Maximal Charlson comorbidity score	0.18	0.70	0.80
Minimal Charlson comorbidity score	0.18	0.71	0.80
Range of Charlson comorbidity score	0.07	1.11	0.95
Average Charlson comorbidity score	0.08	0.59	0.89
Median of Charlson comorbidity score	-0.07	0.45	0.88
25% percentile of Charlson comorbidity score	-0.48	0.66	0.47
75% percentile of Charlson comorbidity score	0.01	0.59	0.99
Volatility of Charlson comorbidity score	-0.19	2.08	0.93
First item 9's score	0.02	0.14	0.92
Maximal item 9's score	0.17	0.68	0.81
Minimal item 9's score	-0.11	0.80	0.89
Range of item 9's score	0.23	0.69	0.75
Average item 9's score	0.16	1.31	0.91
Median of item 9's score	0.03	0.41	0.94
25% percentile of item 9's score	0.19	0.57	0.74
75% percentile of item 9's score	-0.20	0.45	0.65
Volatility of item 9's score	-0.45	0.96	0.64
First PHQ-9 score	0.03	0.03	0.30
Latest PHQ-9 score	-0.06	0.03	0.02*
Maximal PHQ-9 score	-0.06	0.11	0.59
Minimal PHQ-9 score	-0.08	0.11	0.51
Range of PHQ-9 score	0.00	0.11	0.99
Average PHQ-9 score	-0.10	0.18	0.58
Median of PHQ-9 score	0.04	0.06	0.54
25% percentile of PHQ-9 score	-0.04	0.09	0.66
75% percentile of PHQ-9 score	-0.09	0.08	0.26
Volatility of PHQ-9 score	0.04	0.18	0.84
Deepest increase between consecutive PHQ9 scores	-0.01	0.04	0.85
Deepest decrease between consecutive PHQ9 scores	-0.01	0.04	0.85
Volatility of difference between nearby PHQ9 scores	-0.02	0.05	0.75
Percentage of depression free	-0.95	1.77	0.59
Percentage of mild depression	-0.56	1.56	0.72
Percentage of moderate depression	-0.61	1.45	0.68

Percentage of moderately severe depression	0.06	1.44	0.97
Percentage of severe depression	0.68	1.53	0.66

Table C-2: Means and standard deviations (in brackets) of 45 risk factors (1,762 patients).

Risk factors	**Low-risk group (Y_i=1), 876(49.72%)**	**High-risk group (Y_i=0), 886(50.28%)**
Age	3.71(0.93)	3.57(0.87)
Sex n(%)		
Female	578(65.98%)	643(72.57%)
Male	298(34.02%)	243(27.43%)
Statistical Summarization		
Charlson comorbidity score		
First observation	0.66(1.21)	0.65(1.24)
Median observation	0.67(1.24)	0.68(1.27)
Maximal observation	0.71(1.28)	0.73(1.30)
Minimal observation	0.60(1.18)	0.62(1.21)
Range of observations	0.10(0.46)	0.11(0.47)
Mean of observations	0.67(1.23)	0.68(1.25)
Volatility of observations	0.06(0.25)	0.06(0.26)
25% percentile of observations	0.62(1.18)	0.64(1.22)
75% percentile of observations	0.69(1.26)	0.71(1.28)
9^{th} question score		
First observation	0.40(0.79)	0.73(1.00)
Median observation	0.24(0.54)	0.62(0.86)
Maximal observation	0.70(1.10)	1.27(1.17)
Minimal observation	0.06(0.27)	0.24(0.56)
Range of observations	0.64(0.93)	1.04(1.03)
Mean of observations	0.31(0.51)	0.69(0.77)
Volatility of observations	0.31(0.43)	0.50(0.49)
25% percentile of observations	0.11(0.34)	0.36(0.64)
75% percentile of observations	0.49(0.77)	1.02(1.02)
PHQ-9 score		
First observation	11.69(6.57)	16.31(6.05)
Median observation	9.45(5.31)	15.47(5.18)
Maximal observation	11.42(6.23)	19.88(4.76)
Minimal observation	5.11(4.06)	10.71(5.60)
Range of observations	9.31(5.60)	9.17(5.26)
Mean of observations	9.60(4.66)	15.38(4.63)
Volatility of observations	4.20(2.49)	4.18(2.36)
25% percentile of observations	6.59(4.37)	12.39(5.29)
75% percentile of observations	12.57(5.71)	18.37(4.65)
Percentage of healthy states	0.23(0.29)	0.05(0.13)
Percentage of mildly depressive states	0.32(0.28)	0.14(0.21)
Percentage of moderately depressive states	0.23(0.24)	0.24(0.26)

Percentage of moderately severe states	0.14(0.20)	0.27(0.26)
Percentage of severely depressive states	0.09(0.18)	0.29(0.32)
Progression Trajectories		
Number of observations	4.47(2.11)	4.34(2.10)
Time of first observation since initial treatment (days)	217.64(413.16)	239.17(418.29)
Observing density	0.04(0.03)	0.04(0.05)
Lastest PHQ-9 score	7.53(5.41)	14.14(6.35)
Deepest increasing between consecutive PHQ-9 scores	4.07(4.25)	4.79(4.55)
Deepest decreasing between consecutive PHQ-9 scores	6.72(4.88)	6.16(4.59)
Volatility of difference between nearby PHQ-9 score	5.52(4.26)	5.63(4.01)
Non-random Longitudinal Pattern		
WE1 n(%)		
Support	253(28.88%)	191(10.84%)
Not support	623(71.12%)	695(89.16%)
WE2 n(%)		
Support	112(12.79%)	88(4.99%)
Not support	764(87.21%)	798(95.01%)
WE5 n(%)		
Support	209(23.86%)	219(12.43%)
Not support	667(76.14%)	667(87.57%)
WE6 n(%)		
Support	131(14.95%)	97(5.51%)
Not support	745(85.05%)	789(94.49%)

APPENDIX D

Table D-1: Statistical summarization of 38 features on training data.

Risk factors	Low-risk group (Y_i=0), 633(65.60%)	High-risk group (Y_i=1), 332(34.40%)
Age	3.73(0.85)	3.64(0.81)
Sex n(%)		
Female	418(66.03%)	235(70.78%)
Male	215(33.97%)	91(29.22%)
Statistical Summarization		
Charlson comorbidity score		
First observation	0.77(1.21)	0.69(1.24)
Median observation	1.47(1.06)	1.50(1.06)
Maximal observation	0.56(0.90)	0.45(0.76)
Minimal observation	0.91(0.82)	1.05(0.85)
Range of observations	0.98(0.97)	0.91(0.88)
Mean of observations	0.94(1.08)	0.84(0.99)
Volatility of observations	0.67(0.97)	0.55(0.82)
25% percentile of observations	1.29(1.06)	1.26(1.04)
75% percentile of observations	0.46(0.42)	0.54(0.44)
9^{th} question score		
First observation	0.43(0.76)	0.88(0.98)
Median observation	0.93(0.76)	1.54(0.98)
Maximal observation	0.10(0.28)	0.28(0.38)
Minimal observation	0.83(0.68)	1.27(0.82)
Range of observations	0.45(0.43)	0.86(0.63)
Mean of observations	0.39(0.46)	0.82(0.70)
Volatility of observations	0.18(0.33)	0.44(0.47)
25% percentile of observations	0.73(0.61)	1.29(0.86)
75% percentile of observations	0.40(0.31)	0.60(0.38)
PHQ-9 score		
First observation	11.88(6.31)	17.00(5.84)
Median observation	9.95(5.07)	17.38(4.33)
Maximal observation	14.13(5.86)	20.91(4.12)
Minimal observation	6.68(4.58)	13.66(5.26)
Range of observations	7.46(4.31)	7.25(4.14)
Mean of observations	10.17(4.85)	17.33(4.33)
Volatility of observations	3.35(1.92)	3.27(1.82)
25% percentile of observations	7.78(4.67)	14.98(4.87)
75% percentile of observations	12.57(5.39)	19.68(4.15)
Percentage of healthy states	0.20(0.30)	0.01(0.07)
Percentage of mildly depressive states	0.30(0.28)	0.08(0.18)
Percentage of moderately depressive states	0.27(0.27)	0.22(0.26)
Percentage of moderately severe states	0.17(0.23)	0.33(0.27)
Percentage of severely depressive states	0.07(0.16)	0.36(0.35)

Progression Trajectories		
Lastest PHQ-9 score	8.96(5.17)	18.21(4.94)
Deepest increasing between consecutive PHQ-9 scores	3.30(3.08)	4.80(3.40)
Deepest decreasing between consecutive PHQ-9 scores	5.27(3.95)	4.26(3.74)
Volatility of difference between nearby PHQ-9 score	4.43(2.97)	4.69(3.05)

Table D-2: 12 identified rules

Rule 1	Latest PHQ-9 < 9.29	Rule 7	Age >= 2.5 & minimal PHQ-9 < 11.43
Rule 2	Latest PHQ-9 < 13.79 & percentage of moderate < 87.5%	Rule 8	Deepest increasing between nearby PHQ-9 > 0.47 & median of PHQ-9 >18.01
Rule 3	Average PHQ-9 > 15.09 & deepest decreasing between nearby PHQ-9 scores < 5.197	Rule 9	Latest PHQ-9 > 17.9 & Average PHQ-9 > 12.34
Rule 4	Latest PHQ-9 > 17.96	Rule 10	Latest PHQ-9 < 12.53
Rule 5	Latest PHQ-9 >16.03 & maximal PHQ-9 > 17.18	Rule 11	Minimal PHQ-9 < 14.05 & deepest increasing between nearby PHQ-9 < 2.09
Rule 6	Latest PHQ-9 >13.70 & 75% quantile of PHQ-9 > 14.36	Rule 12	Latest PHQ-9 >16

Figure D-1: Proportion of high-risk patients (PHQ-9 score in 5th month ≥ 15) in rule endorsing group

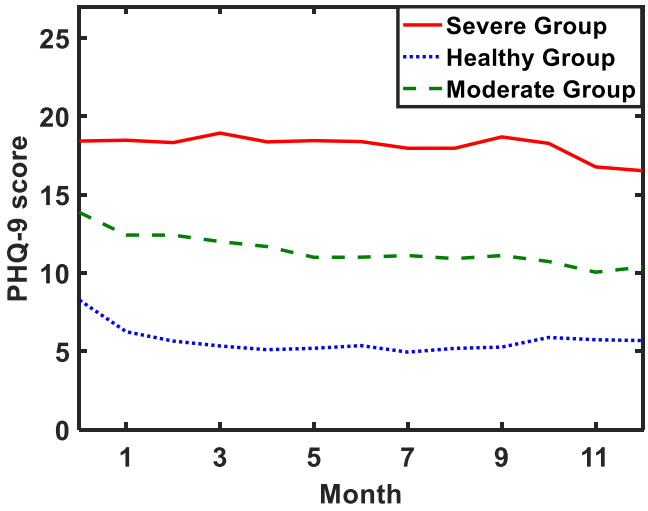
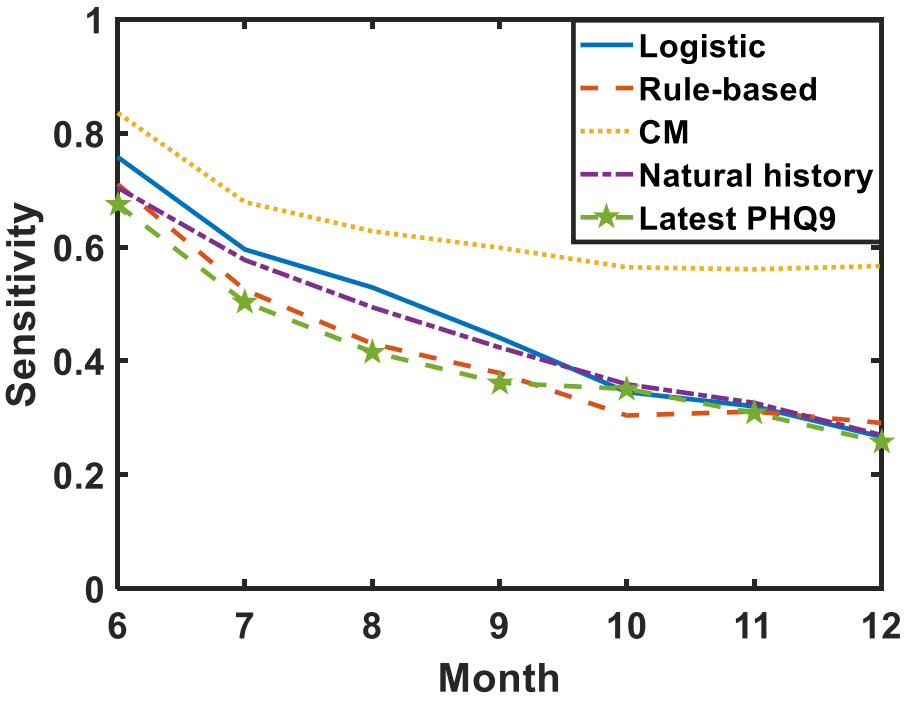
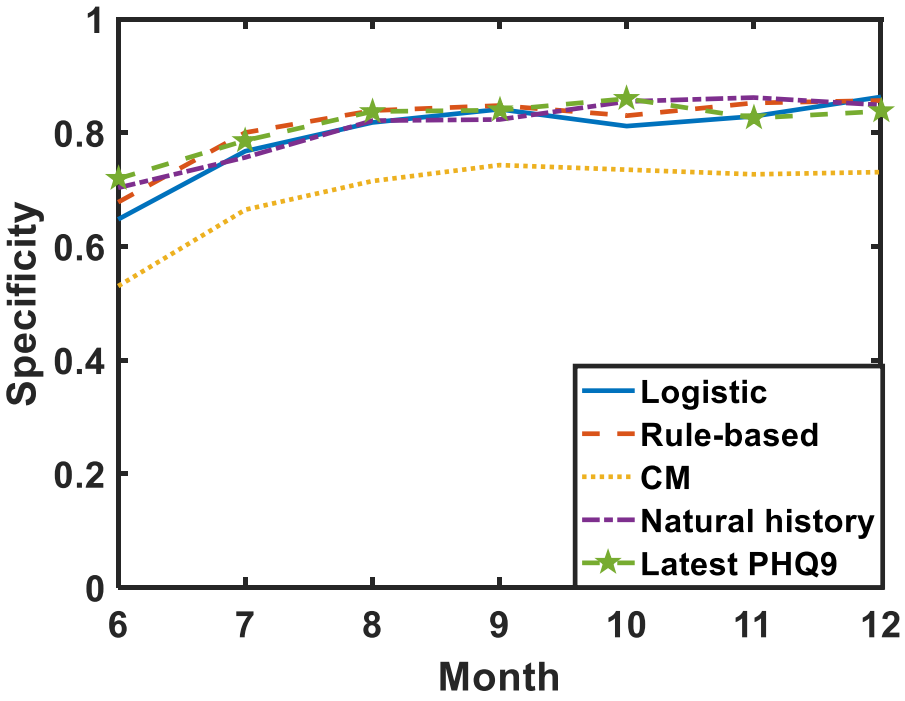


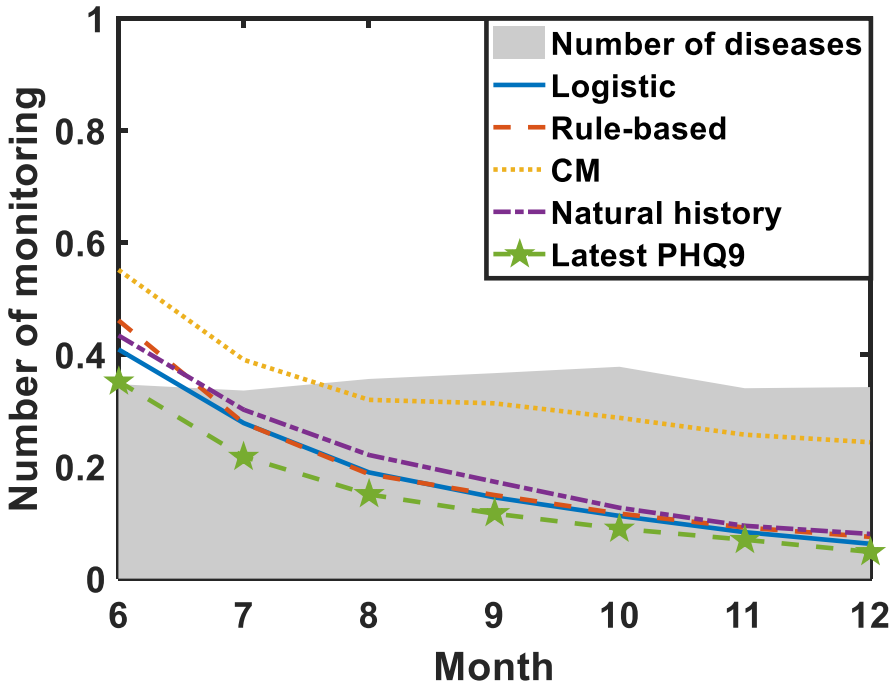
Figure D-2: Three depression trajectory patterns in Markov-based collaborative learning.


Sensitivity
Month
Logistic
Rule-based
CM
Natural history
Latest PHQ9
1
0.8
0.6
0.4
0.2
0
6
7
8
9
10
11
12

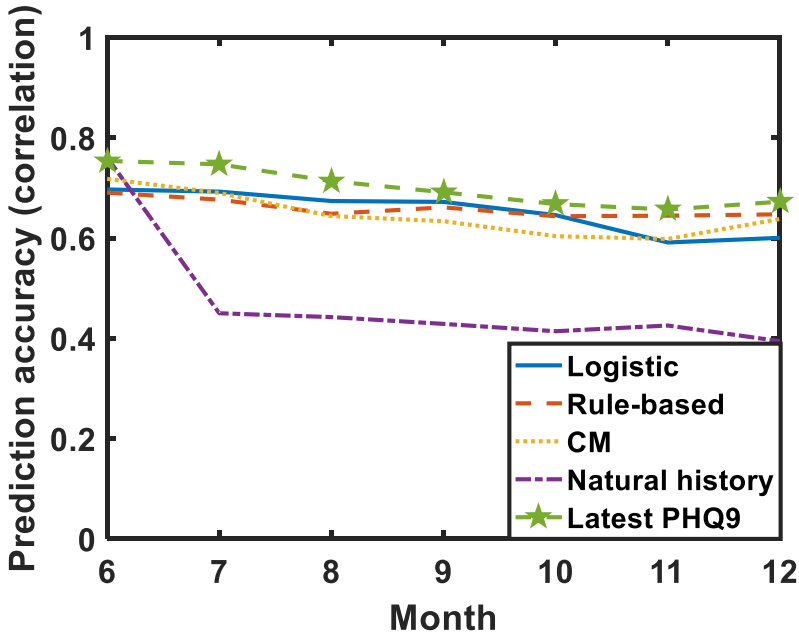
(a)


Specificity
Month
Logistic
Rule-based
CM
Natural history
Latest PHQ9
1
0.8
0.6
0.4
0.2
0
6
7
8
9
10
11
12

(b)



(c)



(d)

Figure D-3: Comparisons of (a) sensitivity, (b) specificity, (c) number of monitoring and (d) correlation between predicted risks and real risks in each month.