Cost-Effectiveness Analysis of Prognostic-Based Depression Monitoring

Ying Lin, PhD[1], Shuai Huang, PhD[2], Gregory E. Simon, MD, MPH[3,4], Shan Liu, PhD*[2]

1. Department of Industrial Engineering, University of Houston, 4722 Calhoun Road, Houston, TX 77204, United States
2. Department of Industrial and Systems Engineering, University of Washington, Box 352650, Seattle, WA 98195, United States
3. Kaiser Permanente Washington Health Research Institute, 1730 Minor Ave, Suite 1600, Seattle, WA 98101, United States
4. Psychiatry and Behavioral Sciences, University of Washington, Box 356560, Seattle, WA 98195, United States

* Corresponding author: liushan@uw.edu

Abstract

Chronic depression monitoring often relies on one-size-fits-all routine monitoring guideline. Without considering heterogeneity in patients' disease progression, routine monitoring guideline may lead to inadequate monitoring on sick individuals and unnecessary monitoring on healthy individuals. Prognostic-based monitoring that stratifies the individual's disease progression risk into different levels and adaptively allocates monitoring resource to high-risk individuals has the potential to improve patient health outcome and cost-effectiveness of the monitoring service. However, challenges include how to best apply prognostic models to inform the design of monitoring strategies and identify the cost-effective strategies. To address these challenges, we develop a decision support framework that integrates individual prognostics, monitoring strategy design and cost-effectiveness analysis. We apply the proposed framework to simulate the adaptive monitoring of a depression treatment population from electronic health record data. Several prediction algorithms with increasing complexity, including natural history matching, logistic regression, rule-based method and Markov-based collaborative model, are simulated to monitor the high-risk individuals for severe depression over time. We find six cost-effective monitoring strategies and demonstrate that two routine monitoring strategies are dominated by the prognostic-based monitoring strategies. Methods from this research show promise to implement prognostic-based monitoring of chronic conditions in clinical practice.

Keywords: Prognostic-based monitoring, cost-effectiveness analysis, depression monitoring, electronic health record, mental health, machine learning, statistical learning, Markov model, chronic conditions, health services

Cost-effectiveness analysis, Markov chains, monitoring, ROC curve, clinical decision model

1. Introduction

Monitoring of chronic conditions is an essential healthcare service to assess patients' disease progressions, treatment outcomes and development of complications. This study focuses on major depressive disorder, which is one of the most common mental disorders with a prevalence of 7.6% among people who are 12 years and over in the United States (CDCMH). Chronic depression can lead to reduced quality of life and productivity, and increased morbidity and mortality due to comorbidities and suicide (CDCMH). Finding appropriate monitoring strategies for major depression is critical to improve the well-beings of people living with the disease (Simon et al., 2000). The Food and Drug Administration (FDA) recommends monitoring of depressed patients on antidepressant medications every six months to one year (NIMH). However, current recommendations for follow-up care are based almost entirely on expert opinions (Reynolds et al., 2016), which do not account for significant heterogeneity in the course of depression between individuals and within individuals over time. Given that as many as 30 million Americans use antidepressants, even minor changes in recommendations for follow-up frequency have major implications for health care utilization. Depression diagnoses and outcomes are regularly entered in the electronic health record (EHR) at outpatient psychiatry and primary care visits in large healthcare systems. However, numbers of follow-up visits differ across patients and systems, and they are often constrained by providers' capacity. For example, demand for individual psychotherapy visit is much greater than supply at many healthcare systems. The problem on how to move the right patients into effective care at the right time remains a major challenge. The objective of this research is to create a data-driven decision-support framework to identify individuals at high risk of major depression, recommend monitoring schedules tailored to each patient, and identify cost-effective monitoring strategies.

Advances in medicine and information technology have provided better understanding of the natural history of many chronic conditions. However, monitoring and treatment are typically reactive and rely on the routine visit of at-risk individuals (Aronson et al., 2015; Boult and Wieland, 2010). Due to inadequate understanding of significant heterogeneity in disease progression and treatment outcome, routine monitoring strategies may lead to inadequate follow-up of high-risk or severely sick individuals and unnecessary monitoring of low-risk or healthy individuals (Kales et al., 2010). On the other hand, personalized prognostic-based monitoring of chronic conditions has the potential to deliver appropriate care to the right people at the right time, and lead to cost-effective resource utilization in clinical practice. In recent years, prognostic-based monitoring is enabled by the growing availability of sensing and information technology such as the EHR, and recent advances in using big-scale data to train prognostic models. For example, feature-based prognostic models summarize the longitudinal sensing information as a set of risk predictive features and stratify the individual's risk of disease onset based on his/her feature profiles (Huang et al., 2014; Lasko et al., 2013). Trajectory-based prognostic models, on the other hand, assess the individual's risk by modeling the trajectory of disease progression over time (Lin et al., 2016; Oskooyee et al., 2011; Sutin et al., 2013). Implementing these prognostic models for personalized monitoring still needs a seamless combination of data analysis and decision-making. These models can provide assessment for an individual's disease progression risk, but optimally allocating monitoring resources to different risk groups remains a challenge. Furthermore, prognostic models at various level of accuracy and complexity may lead to different monitoring strategies on the same individual. Understanding the cost-effectiveness of prognostic-based monitoring strategies compared to routine monitoring is important to operationalize adaptive monitoring in practice. Existing evaluations of prognostic models mainly focus on the prediction accuracy of health outcomes or net benefits resulted from detecting critical events (Vickers and Elkin, 2006). Without considering the monitoring capability associated with each prognostic model and their cost-effectiveness, resulting models are often inadequate or unrealistic to be implemented into the clinical flow.

To overcome these challenges, we develop a prognostic-based monitoring framework for chronic depression that can automatically use sensing data in disease risk prediction and identify the design of cost-effective monitoring strategy. First, we apply four prognostic models at various level of computational complexity to identify individuals at high risk of progressing to severe depression state over time. These models include both feature-based models (i.e. logistic regression and rule-based method) and trajectory-based models (i.e. Markov-based collaborative model and natural history matching). Under feature-based prognostic models, we first summarize the longitudinal depression severity measurements within a certain period as a set of risk predictive features. We then predict severity in the next monitoring period using a *logistic regression* model. To better interpret the heterogeneous depression progression patterns, we further apply a *rule-based method* to characterize the progression patterns as a set of humanly interpretable rules discovered from the risk predictive features; each rule segments the population into subgroups with different risk indications (Lin et al., 2014; 2018a). Under the trajectory-based models, we assume the trajectory of depression progression as a Markov process (Bhattacharya, 2014; Islam et al., 2013; Lin et al. 2018b). Heterogeneity in disease progression can be modeled by learning the latent structure in the population and similarity between individuals using a *Markov-based collaborative model* (Lin et al., 2016; 2017; 2018b). Furthermore, to better capture short-term stochastic changes in the depression progression, we build on a *natural history matching* idea from Alagoz et al., 2005. The matching model predicts progression on an index/new patient by searching for a number of most similar patients in an existing database, and then using their next-period disease states to assign a weighted risk to the index patient. This simple matching process ensures that the natural history matching is more sensitive to small variations in the observations. In the monitoring strategy design phase, we monitor individuals with predicted progression risks higher than a pre-defined threshold at each period. We update individuals with predicted low risks by monitoring a certain percentage of the low-risk group. To inform which monitoring strategies are cost-effective, we further compare these prognostic-based monitoring strategies with routine monitoring in a cost-effectiveness analysis.

Methodology contribution of this research includes comparing four prognostic models at various level of complexity to enable adaptive depression monitoring, and evaluating their operational value using a cost-effectiveness analysis. The proposed framework is potentially generalizable to developing empirically supported monitoring recommendations for other chronic conditions. The U.S. Department of Health & Human Services defines chronic conditions as "conditions that last a year or more and require ongoing medical attention and/or limit activities of daily living" (USHHS, 2010). These conditions include both physical illness such as diabetes, cancer, and HIV infection, as well as mental and cognitive disorders, such as depression, substance addiction, and dementia (Ward et al., 2012). Two-thirds of American older adults live with two or more chronic conditions, accounting for 66% of the total health care spending in the United States (Venkatesh et al., 2014). The long-term impact of this research includes demonstrating the projected value of computerized decision-support tools in improving health outcomes of patients. The value is gained through smart monitoring and facilitating efficient allocation of health providers' limited resources.

This paper is organized as follows. Section 2 reviews related work on modeling chronic conditions' progression, adaptive monitoring, and cost-effectiveness analysis. Section 3 illustrates technical details of the proposed monitoring framework. Section 4 presents results of simulating a depression treatment population from EHR data. Section 5 draws the conclusion and discusses limitations.

2. Related work

This work is relevant to two research areas in the literature. The first area involves building disease progression models using clinical data, and using these models to optimize healthcare interventions through simulated experiments. Modeling disease progression involves estimating a mathematical model to describe and predict the time course of the disease. Common models include regression, Bayesian updating, Markov models, optimal control, neural networks, and reinforcement learning. Applications of these methods can be seen in modeling CD4 count decline in HIV patients (Shechter et al., 2008), liver deterioration on the transplant waiting list (Alagoz et al., 2004; 2007; Sandikci et al., 2008), hepatitis (Salomon et al., 2002; Hutton et al., 2007; Liu et al., 2012), liver cancer (Lee et al., 2015), depression (Sutin et al., 2013; Gunn et al., 2013; Lin et al., 2016), glaucoma (Helm et al., 2015; Kazemian et al., 2015), diabetes (Mason et al., 2012), breast cancer (Ayer et al., 2012; Chen et al., 2017), and chronic obstructive pulmonary disease progression (Wang et al., 2014). The majority of these studies learn a single disease model from population-level data without explicitly considering individual patient's heterogeneity. Furthermore, the literature on using stochastic and dynamic models to optimize disease monitoring and treatment decisions over time is growing. For example, Markov Decision Processes (MDPs) and dynamic programming algorithms are common methods used to optimize monitoring and control of disease progression through simulation (Brandeau et al., 2004). Despite the successful application of these methods in a number of health applications (Shechter et al., 2008; Alagoz et al., 2004; 2007; Sandikci et al., 2008; 2013; Liu et al., 2017), they often require extensive data to estimate the transition probabilities and rewards for each possible action, and often do not incorporate real-time updating of the transition probabilities. Furthermore, it is unclear how well some of these models would perform when the disease dynamic is widely fluctuating. For instance, a disease model for depression must be able to accommodate complex fluctuations in the disease trajectories. Such a model may output very different monitoring schedules compared to diseases with well-defined natural history. In summary, these studies focus on developing new optimization methods and output policy recommendations. They rarely consider the operational value of online decision-support tools. Therefore, these models generally have strong assumptions and limited usability in clinical practice, which is a research challenge addressed in this paper.

The second area is on the evaluation of prognostic-based disease monitoring strategies. Current assessment of prognostic models usually focus on their discriminative ability of binary outcome measured by area under the receiver operating characteristic (ROC) curve, and overall statistics of prediction accuracy such as R^2. Very few studies have compared the operational outcomes of prognostic models in the adaptive monitoring process. Recent developments in the decision curve analysis consider the clinical usefulness of prognostic models by estimating net benefits (Vickers et al., 2006). However, the decision curve analysis focuses on the clinical consequence of a critical event, such as cost of recurrence after prostate cancer surgery, which is inadequate to evaluate a sequential monitoring process. Cost-effectiveness analysis (CEA) can be a crucial methodological component when evaluating the design of adaptive care strategies, as well as an enabler for the adoption of these strategies into routine clinical practice. CEA is a formal health economic evaluation method that compares the downstream health benefit and cost of alternative interventions to determine whether the intervention is worth doing (Weinstein et al., 1996; 2003). The outcome measure is called the incremental cost-effectiveness ratio (ICER). ICER is used to assess the value of an intervention by providing the ratio of additional cost required to achieve a defined improvement in benefit, compared with the next best intervention. Benefits are often measured by the natural units of the intervention (e.g. reduction in hospital days, infections averted, etc.), or a common metric such as quality-adjusted life years that takes account of both survival and quality of life. Simulation-based and clinical trial based CEA of depression care interventions are limited in the literature (Hay et al., 2018; Valenstein et al., 2001; Simon et al., 2001). Two important questions regarding chronic depression care include whether routine monitoring is justified and how to coordinate long-term follow-up. These questions are challenging due to heterogeneity in treatment response within the population. A CEA can link the data-driven designs of prognostic-based monitoring with evaluation of their short-term and long-term patient outcome and cost. Based on the ICERs, a healthcare provider can make a value judgement on whether a prognostic-based monitoring strategy enabled by a decision-support system is worth implementing by their own willingness-to-pay threshold.

3. Method

The proposed framework integrates individual prognostics, monitoring strategy and cost-effectiveness analysis. An overview of the method is presented in Figure 1. It consists of a model training phase (offline) and an adaptive monitoring/cost-effectiveness analysis phase (online). The model training phase learns the prognostic models from a set of existing EHR data. It includes missing value imputation, feature extraction and model learning steps. Next we simulate a real-world adaptive monitoring process in the online phase. Based on an individual patient i's predicted risk score at time t, denoted as r_{it}, we select a monitoring decision based on a threshold, θ. If $r_{it} \geq \theta$, then patient i is categorized as high-risk and selected for monitoring in the next period. Otherwise, we use an exploration approach to select a portion of the low-risk patients for next-period assessment. The cost-effectiveness analysis evaluates each prognostic model by iteratively stratifying the individual risks, making a monitoring decision on each individual for the next period, and incorporating new measurement for risk updates. We use missing value imputation methods if some patients do not show up for their appointments. By comparing the cost and effect of each monitoring strategy in the online phase, we can identify a set of cost-effective monitoring strategies. We conduct a simulation of chronic depression monitoring to illustrate the framework in each step.

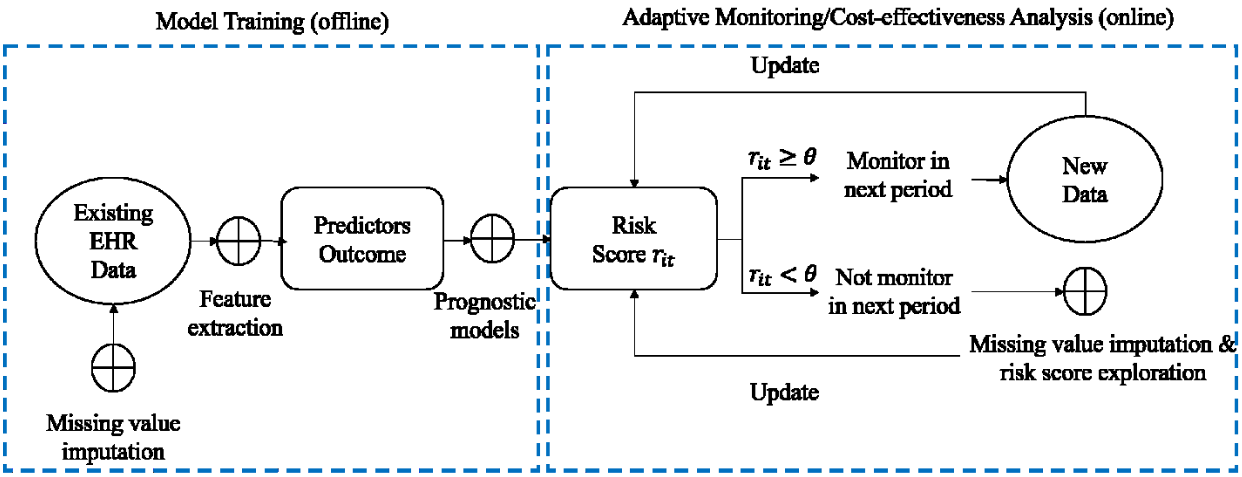


Figure 1. The framework of prognostic-based monitoring. r_{it} denotes the risk score of individual i in t^{th} monitoring period and θ represents the threshold for monitoring.

3.1. Data description

We study a depression treatment population of 965 individuals from the EHR of four health systems participating in the Mental Health Research Network (Kaiser Permanente Washington, HealthPartners, and the Colorado and Southern California regions of Kaiser Permanente) (Simon et al., 2013). Depression severity is assessed by the Patient Health Questionnaire (PHQ-9), a self-administrated questionnaire. All four health systems recommend routine use of the PHQ-9 questionnaire at all specialty mental health visits and at all primary care visits including diagnosis or treatment of depression. The dataset includes individuals' longitudinal PHQ-9 scores in EHR between year 2007 and 2012, and are linked to relative time between measurements, type of providers (primary care, specialist, mental health) where the questionnaire was conducted, individuals' age, sex, diagnosis and treatment status, and the Charlson Comorbidity Score (a standard indicator of medical disease burden).

The 965 individuals with on-going treatment were closely monitored for one year and had at least 6 measurements in this time window. The PHQ-9 score ranges from 0 to 27 and indicates severe depression when it is greater or equal to 15. The PHQ-9 score can further stratify the depression severity into five levels including no depression (0-4), mild depression (5-9), moderate depression (10-14), moderate severe depression (15-19) and severe depression (20-27). Each PHQ-9 score also records the 9th question score concerning to suicidal ideation. We conduct analysis on this group by using a monthly monitoring time window and regarding the first 6 months as model training phase and the remaining 7 months as cost-effectiveness analysis phase. In the model training phase, we further split measurements in the first five months as training data and leave the measurements in the 6th month as validation data.

Since the EHR data has irregular and sparse measurements on each individual, we impute the missing values using a two-step approach. Specifically, we first impute the missing values between the initial and last measurements by fitting a smoothed B-spline model for each individual and follow a similar process described in Lin et al., 2016. As shown in Figure 2, each individual may have different length of observation. To impute missing values outside of the observed window, we use measurements from other individuals. We find the 10 nearest neighbors of each individual that have the most similar baseline features. Then we impute the missing values after the last observation using the average value of measurements from the 10 neighbors at corresponding time points. The baseline features include patient's demographic features and the coefficients fitted from the B-spline models. Next, random error from a standard normal distribution is added to simulate a random noise. The missing value imputation of four randomly selected individuals shown in Figure 2 demonstrate that, 1) the trajectory fitted from smoothed B-spline is able to capture the individual depression trajectory; and 2) the imputed PHQ-9 scores follow these progression trajectories.

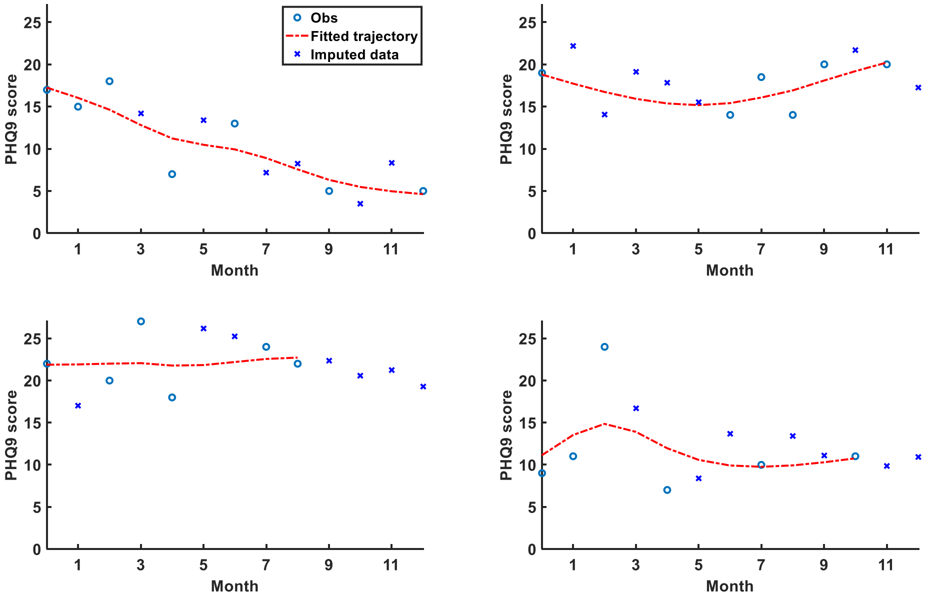


Figure 2. Missing value imputation on four randomly selected individuals.

3.2. Prognostic models

Both feature-based and trajectory-based models, including logistic regression, rule-based method, Markov-based collaborative model and natural history matching, have been designed to predict depression severities. We consider these models at various level of complexity to identify the ones that can lead to more cost-effective monitoring strategies. The prognostic models are trained on the first five months' measurements and validated in the 6th month. Detailed description of each model is provided in section 3.2.1 - 3.2.3.

3.2.1. Feature-based prognostic model

The feature-based prognostic model uses a set of risk-predictive features to describe the disease progression in a certain time window and predicts the health outcome in the following time point. In depression monitoring, we transform the measurements in every four months to 38 features and predicts the depression severity in the 5th month. These features characterize demographic factors of the population, statistical summarization, progression trajectories as well as abnormal patterns in the longitudinal measurements. These features are summarized in Table A-1 in the Appendix. The depression severity in the 5th month is measured by the PHQ-9 score, with no less than 15 indicating a severe depression. First, logistic regression model is used to predict the risk of progressing to severe depression in the 5th month from the 38 extracted features. To capture the complex interactions between risk-predictive features and the heterogeneity in depression progression, the rule-based method (Lin et al., 2014; 2018a) is further applied to identify a set of longitudinal patterns from the 38 features that segment the population into subgroups. The rule-based method is a machine learning technique built over the random forest model. It generates a set of rules by decomposing the classification trees in random forest into rules and selecting the most predictive rules by a rule pruning process. Each rule is interpreted as a longitudinal pattern that consists of several interacting risk-predictive features and their ranges. Individuals endorsing the same patterns have more homogeneous risk indications. 12 longitudinal patterns are discovered from the first four months measurements and summarized in Table A-2 and Figure A-1 in the Appendix. Denote an individual's endorsements of 12 rules as $[R_1, \dots, R_{12}]$, with $R_j = 1$ if the jth rule is endorsed and $R_j = 0$ otherwise. The identified rules span a feature space for assessing the individual's risk. We use a logit function to predict the individual's risk of disease onset from his/her rule endorsement profile.

3.2.2. Markov-based collaborative model

A Markov model is used to characterize depression progression by modeling the probability of transition between five depression severity levels defined in section 3.1. Traditional Markov model is estimated using maximal likelihood estimator, which is inadequate to deal with the sparse and irregular individual measurements under adaptive monitoring. To accurately estimate the Markov model for each individual, we use the similarity-based collaborative model developed in Lin et al., 2016, 2017, 2018b. Collaborative model captures the heterogeneous disease progression using K numbers of canonical Markov models; each with distinct initial distribution and transition matrix. Each canonical model represents a main pattern of disease progression in the population. Each individual-level Markov model is further assembled as a weighted combination of these canonical models. Specifically, we denote the transition matrix and initial distribution of the kth canonical model as $\mathbf{\Pi}_k$ and $\boldsymbol{\theta}_k$, respectively. We then assign each individual a distinct weight vector on the canonical models, $\boldsymbol{c}_i$, to capture the individual to individual variations. Then the individual Markov model is expressed as:

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

where $\mathbf{P}_i$ and $\boldsymbol{\nu}_i$ represent the transition matrix and initial distribution of the ith individual.

The collaborative model further incorporates the similarity between individuals as a regularization on the weight vectors to enhance learning of the Markov models. We assume similar individuals are more likely to have similar weight vectors on the canonical models. The risk predictive features in section 3.2.1 are used to measure the similarity between individuals. To initialize the canonical models and weight vectors, we cluster the individuals into groups based on their risk predictive features. The first five months' measurements in the same group are used to initialize the canonical models, and the cluster indices are used to initialize the weight vector of each individual. We then run the collaborative model algorithm to estimate the canonical models and weight vectors for all individual-level Markov models (Lin et al., 2018b). We find that three canonical models give the best model fitting in this population. These patterns shown in Figure A-2 in the Appendix represent the stable high, stable low, and moderate depression trajectories. The risk of each individual is predicted as the probability she/he transitions from the latest observed state to the severe depression states (PHQ-9 score no less than 15).

3.2.3. Natural history matching

Natural history matching searches for the most similar disease progression patterns in an existing database to predict an individual's disease severity in next monitoring period (Alagoz et al., 2005). To capture the depression trajectories, we regard every three sequential PHQ-9 scores on an individual as a triplet and build a database consisting of all triplets segmented from the training data. Each individual is also associated with the 38 features shown in Table A-1. For an index individual i, we consider his/her PHQ-9 scores in the previous period $t-1$ and current period t to determine the next period $t+1$'s depression severity, Y_{it+1}. We search for 10 most similar individuals in the database and identify a set of triplets that may have the closest depression trajectories among the similar individuals. The similarity between individuals are measured from the Euclidean distance on their 38 features. When these triplets are 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 third measurements in the 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. For example, the closeness is measured by the differences between their previous and current period PHQ-9 scores, i.e. $w_{ji} = \frac{1}{C_i} \times \frac{1}{(Y_{it-1}-Z_{j1})^2+(Y_{it}-Z_{j2})^2}$, where C_i is a normalization term to guarantee the weights sum to one. The risk of each individual can be obtained by rescaling the predicted outcome $\widehat{Y_{it+1}}$ to a value between 0 and 1.

3.3. Prognostic-based monitoring

3.3.1. Monitoring strategies

The prognostic models presented in section 3.2 stratify the individuals' risks of progressing to severe depression in the next monitoring period to different levels. To decide on who should be monitored in the next period, we segment the population into high-risk and low-risk groups by comparing the predicted risks, r_{it}'s, with a predefined threshold θ. All individuals in the high-risk group ($r_{it} \geq \theta$) are monitored. Due to uncertainty in the risk prediction, we randomly select 10% individuals from the low-risk group ($r_{it} < \theta$) to improve model updating. We select 10% of low-risk individuals for exploration base on the fact that monitoring resource in a healthcare system is very constrained and achieving the objective of comparing the prognostic-based monitoring strategies.

The monitoring accuracy of each strategy can be measured by the percentage of severely depressive patients being monitored (sensitivity) and the percentage of healthy to moderately depressive patients not monitored (specificity). Different choice of threshold θ may result in different levels of sensitivity and specificity. For instance, increasing the threshold leads to an increase in specificity and a decrease in sensitivity. To ensure the best tradeoff between sensitivity and specificity, we find the optimal threshold $\hat{\theta}$ for each prognostic model using the validation data at 6th month, that minimizes the distance between monitoring accuracy and perfect monitoring (sensitivity = 1 and specificity = 1), i.e.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}((1 - \text{sensitivity})^2 + (1 - \text{specificity})^2) \quad (3.1)$$

where sensitivity= $\sum_{i=1}^{n} 1(r_{i6} \geq \theta)$/number of severly depressive patients at 6th month, specificity= $\sum_{i=1}^{n} 1(r_{i6} < \theta)$/number of healthy to moderate patients at 6th month.

3.3.2. Missing value imputation and risk update

Adaptively monitoring the high-risk individuals will lead to increasing number of missing values on the low-risk individuals. To address this issue during the online phase, we impute the missing values between initial and last measurements for each individual by fitting a smoothed B-spline model on the observations (Lin et al., 2016). When new measurements are collected, missing values are further updated by refitting the B-spline model.

To update all individuals' risks over time under the feature-based prognostic models, we update the features of monitored individuals by including the collected measurements at each monitoring period. For the Markov-based model, we update the weight vectors in the collaborative model to re-estimate the individual transition matrix. The probability of transitioning from the current state to severe depression is further updated. For the natural history matching, we incorporate the new measurements as triplets in the database.

3.4. Cost-effectiveness analysis

To evaluate prognostic-based monitoring strategies, we compare them using a cost-effectiveness analysis. In the context of depression monitoring, routine monitoring (status quo) and a latest PHQ-9 based strategies are commonly used in clinical practice (Kroenke and Spitzer, 2002; Untzer et al., 2002). The status quo strategy monitors all patients under a fixed frequency. This may lead to unnecessary monitoring of low-risk patients (false positives), incurring higher cost to the healthcare system. The latest PHQ-9 based strategy, on the other hand, predicts the depression severity of each patient using his/her most recent score, and adaptively monitors the ones with high scores. Without considering the trajectories of depression progression, the latest PHQ-9 strategy may not be able to capture patients with widely fluctuating depression levels. We compare the four prognostic-model enabled monitoring strategies with these existing monitoring strategies. Furthermore, we consider different frequencies in the status quo, including monitoring monthly (SQ_I), every two months (SQ_II), every three months (SQ_III), and monitoring at the following 1st, 3rd and 6th months (SQ_IV) in the online phase. In the latest PHQ-9 based monitoring, patients with their latest scores greater than or equal to 15 will be monitored in the next month. We also randomly select 10% of low-risk individuals to monitor their health conditions.

To investigate which monitoring strategies can lead to cost-effective usage of monitoring resources, we quantify the monitoring cost and effect of each strategy. We estimate the cost of one-time monitoring to be $107 from the current procedural terminology (CPT) code. The total cost is calculated as $107 multiplied by the total number of individuals monitored under each strategy. The SQ_I strategy always captures all severe patients at the highest cost by monitoring all individuals every month. The monitoring effect is measured by the number of severely depressive 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 cost and calculate the incremental cost-effectiveness ratio 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 $C_1 > C_0$. The ICER between two strategies is calculated as:

$$ICER = \frac{C_1 - C_0}{E_1 - E_0}, \tag{3.2}$$

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

In addition to the incremental cost-effectiveness ratio, we further consider how much cost the adaptive monitoring strategies can save compared to routine monthly monitoring by calculating the number of healthy to moderate patients not monitored, which is denoted as true negatives (TN). The cost savings of each monitoring strategy is calculated as the number of TN multiplied by $107.

4. Result

4.1. Prediction accuracy

We first compare the prediction accuracy of four prognostic models in the training phase. Specifically, we train the models in the first 5 months and validate the models in the 6th month. The prediction accuracy in the 6th month is summarized in Table 1. It is measured by the area under the ROC curves (AUC), the correlation coefficient between predicted risks and real risks, and the root of mean square error (rMSE) between predicted risks and real risks. The real risks are estimated from the time to severe depression onset using a survival function. It can be observed that the natural history matching has the highest AUC and correlation, while the Markov-based collaborative model (CM) and the rule-based model have lower rMSE. This result indicates that the natural history matching can better distinguish the high-risk and low-risk individuals by capturing the stochastic changes in depression progression but is inadequate to predict individual risk well due to its sensitivity to the noise in PHQ-9 observations. Markov-based CM and rule-based method can achieve more accurate prediction of the individual risk by explicitly exploiting and smoothing the heterogeneous depression progression patterns.

Table 1: Prediction accuracy of four prognostic models in the 6th month.

	Logistic	**Rule-based**	**CM**	**Natural history**
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

4.2. Monitoring accuracy

We further compare the four prognostic-based models with existing strategies during the adaptive monitoring phase. The accuracy of each monitoring strategy is measured by the average sensitivity and specificity over seven monitoring periods, summarized in Table 2 and Figure 3. As shown in the result, the latest PHQ-9 based monitoring strategy outperforms the routine monitoring strategies, including monitoring every two months (SQ_II), every three months (SQ_III), and at the following 1st, 3rd and 6th months (SQ_IV), on both sensitivity and specificity. The prognostic-based monitoring strategies have higher sensitivity than the latest PHQ-9 based strategy, which indicates the prognostic-based monitoring strategies have the potential to accurately monitor the high-risk individuals. The feature-based monitoring strategies have similar monitoring accuracy; the rule-based method has slightly higher specificity and lower sensitivity than the logistic regression model. Markov-based CM has higher sensitivity and lower specificity compared to the feature-based method. The natural history matching method has similar sensitivity with monthly monitoring strategy (SQ_I) but greatly improves the specificity of routine monitoring, which indicates that the natural history matching has the potential to save monitoring resources on healthy patients compared to monthly monitoring. The overall accuracy measured by the distance between monitoring accuracy and the perfect monitoring indicates that Markov-based CM has the best monitoring accuracy overall. We further show the monitoring accuracy of different methods in each monitoring period in Figure A-3 (a) - (d) in the Appendix. We observed that the natural history matching based monitoring strategy tends to use more resources for monitoring, leading to higher sensitivity. Other monitoring strategies have better performance on saving resources on the healthier individuals. Due to the increase in missing values during adaptive monitoring, natural history matching is inadequate to find similar progression patterns from existing triplets, which is demonstrated by its lower prediction accuracy after the first month in the online phase. On the other hand, the latest PHQ-9 based method is effective in predicting the individuals with stable high and stable low depression severities, leading to its high prediction accuracy. The rule-based method has higher prediction accuracy than the logistic regression method by exploiting the complex interactions between risk-predictive features to capture the heterogeneous disease progression process.

Table 2: Average sensitivity and specificity of different strategies in the adaptive monitoring period.

Method	**Sensitivity**	**Specificity**	$(\mathbf{1}-\textbf{ Sensitivity})^{\mathbf{2}}+(\mathbf{1}-\textbf{ Specificity})^{\mathbf{2}}$
SQ_III	0.29	0.71	0.59
SQ_II	0.43	0.57	0.51
SQ_IV	0.43	0.57	0.51
Latest-PHQ9	0.56	0.85	0.22
Rule-based	0.64	0.82	0.16
Logistic	0.66	0.79	0.16
CM	0.76	0.74	0.13
Natural history	0.95	0.43	0.33
SQ_I	1.00	0.00	1

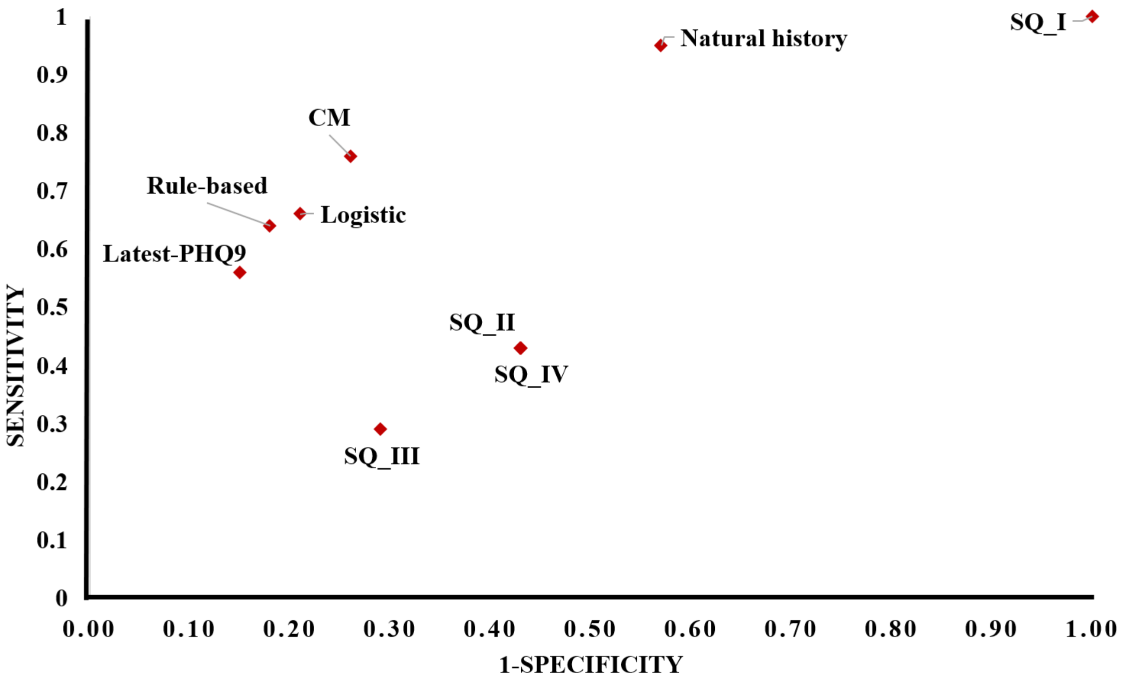


Figure 3. Average sensitivity and specificity of different strategies in the adaptive monitoring period.

4.3. Cost-effectiveness analysis

We evaluate the cost-effectiveness of all monitoring strategies in the online phase. Results are shown in Table 3. Figure 4 displays the cost-effectiveness frontier. We observed that logistic regression, status quo of monitoring every two months (SQ_II) and monitoring at 1st, 3rd and 6th months (SQ_IV) are dominated. The latest PHQ-9 based strategy costs $8 to monitor an additional high-risk patient (i.e. true positive) compared to the status quo of monitoring every 3 months (SQ_III). The status quo strategy of monitoring all patients every month (SQ_I) costs $2,133 to monitor an additional high-risk patient compared to the next-best prognostic-based strategy, which is natural history matching. Implementing the adaptive monitoring strategies (except natural history matching) can save $2,354 to $59,599 compared to the status quo strategies in this population. Therefore, the adaptive monitoring strategies may be preferred when monitoring resource is constrained, while routine monitoring patients monthly may be preferred if healthcare providers are willing to pay a substantial amount to identify additional true positive patients. Among the adaptive monitoring strategies, latest PHQ-9 based monitoring, rule-based monitoring, Markov-based CM monitoring, and natural history matching based monitoring are on the cost-effective frontier.

Table 3: Cost-effectiveness analysis results

Method	Effect (TP)	Cost ($)	ICER ($/TP)	TN	Saving ($)
SQ_III	86	29,501		469	50,183
Latest-PHQ9	181	30,281	8	557	59,599
Rule-based	196	33,705	228	539	57,673
Logistic	204	36,380	Dominated	522	55,854
CM	233	43,442	263	491	52,537
SQ_II	128	44,252	Dominated	373	39,911
SQ_IV	131	44,252	Dominated	376	40,232
Natural history	292	71,262	472	284	30,388
SQ_I	307	103,255	2,133	0	0

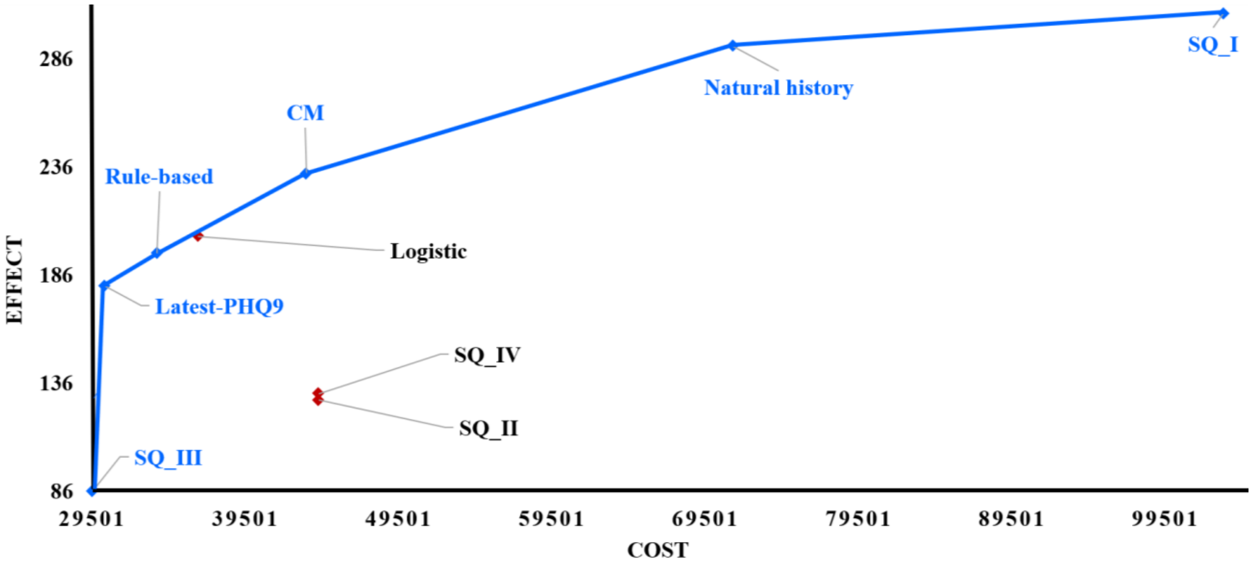
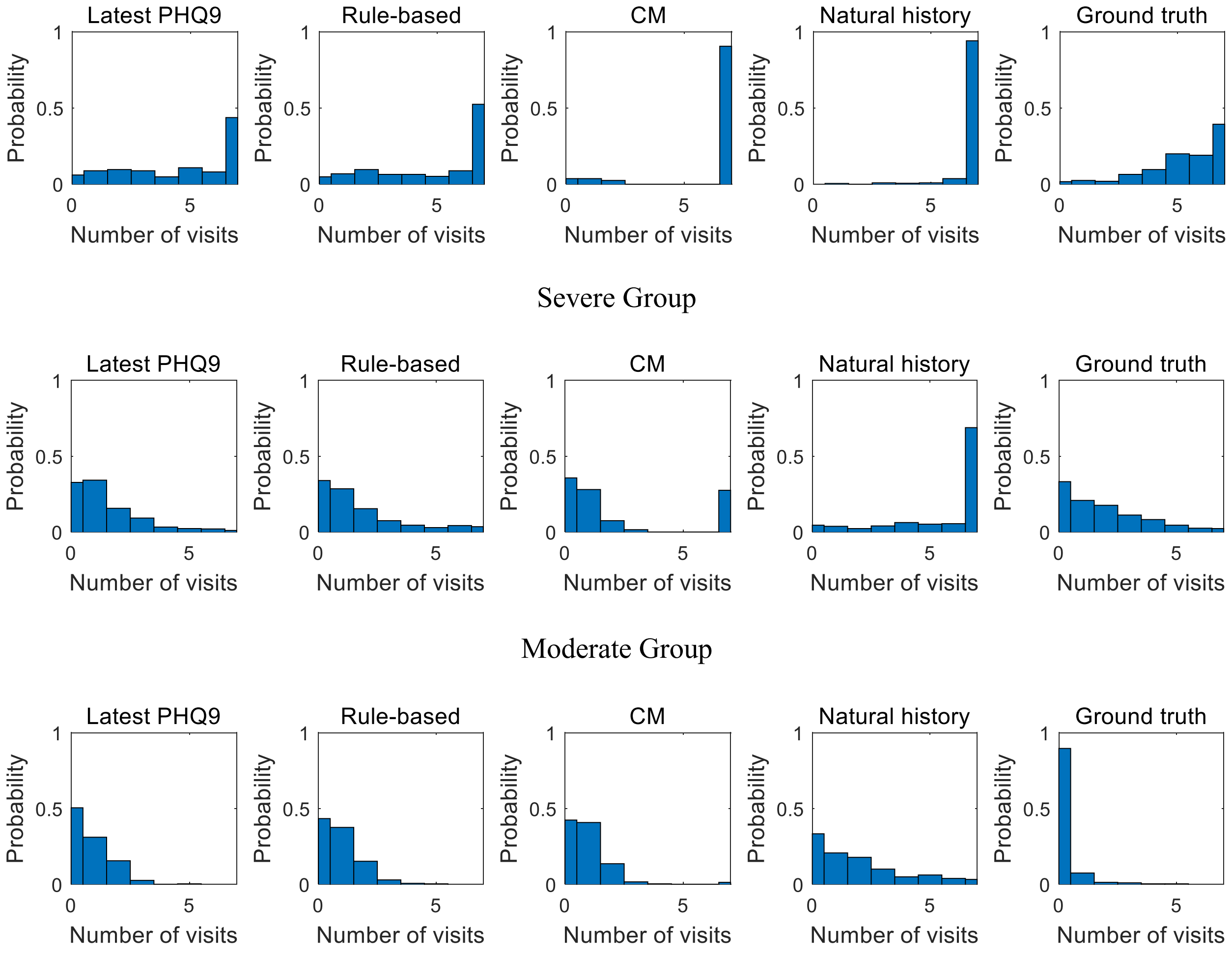


Figure 4. Cost-effectiveness frontier. The cost-effective strategies are represented by blue dots and the dominated strategies are denoted by red dots.

4.4. Analysis of cost-effective monitoring strategies

We further compare the cost-effective adaptive monitoring strategies 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 a severe group with stable-high PHQ-9 scores, a healthy group with stable-low PHQ-9 scores, and a moderate group that has PHQ-9 scores fluctuating between low and high values. The distributions of monitoring frequency under four monitoring strategies are compared to ground truth in Figure 5. It can be observed that all adaptive monitoring strategies tend to allocate more monitoring resources in the severe group and save monitoring resources in the healthy group. Specifically, Markov-based CM and natural history based strategies assign frequent monitoring to the whole severe group while the latest PHQ-9 and rule-based strategies have similar distributions with the ground truth in the moderate group.



Severe Group

Moderate Group

Healthy Group

Figure 5. Comparison of monitoring frequencies in severe, moderate and healthy groups.

Results may also indicate that the performance of these prognostic models may differ by disease trajectory patterns; thus the cost-effectiveness of monitoring strategies may vary by patient population and dataset. Furthermore, we would like to know if the distributions of the cost-effective monitoring strategies are significantly different. In a pure computer simulated CEA, we can easily conduct sensitivity analyses on model parameters (e.g. monitoring cost, prediction accuracy, etc.) through deterministic scenario analysis and probabilistic sensitivity analysis (i.e. a Monte Carlo simulation method that characterizes the distributions of outcomes from the uncertainty around all input parameters) (Weinstein et al., 1996; 2003). However, our CEA study is based on analyzing an existing EHR dataset. One way to test uncertainties in the ICER outcome is to apply the offline and online phases multiple times using different datasets, or for different time windows in the same dataset. Due to the limited sample size of our data, we do not have the capacity to perform these types of sensitivity analyses, but we highlight the importance of understanding uncertainty in the CEA outcome using a large dataset or multiple datasets.

5. Conclusion

We establish a prognostic-based monitoring framework to translate the existing EHR data into evidence to support cost-effective monitoring strategy design. The proposed framework can act as a decision-support tool for healthcare professionals in conducting adaptive monitoring by integrating the individual prognostics, monitoring strategy design and cost-effectiveness analysis. The proposed method has the potential to enable better use of EHR data for chronic depression monitoring, leading to quality improvement in healthcare resource delivery.

We simulated the adaptive monitoring of a depression treatment population and compared four types of prognostic models (logistic regression, rule-based method, Markov-based collaborative model and natural history model). We further compared the prognostic-based monitoring strategies with current monitoring strategies used in clinical practice. We identified that the latest PHQ-9 based method, rule-based method, Markov-based collaborative model and natural history matching have the potential to be cost-effective strategies for depression monitoring. Specifically, the latest PHQ-9 score is a simple approach for prognostics, but it can provide comparable performance with more advanced prognostic models in patients with stable patterns. 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 9th question, is predictive to subsequent suicide attempts (Richardson et al., 2010). However, we discovered that the latest PHQ-9 based monitoring strategy has advantage in saving resources on healthy individuals but may be inadequate for high-risk individuals. This is due to the fact that the latest PHQ-9 based method is not able to capture individuals with increasing and fluctuating risks. The rule-based method which exploits the complex interactions between risk-predictive features can enable more accurate risk assessment than the logistic regression model, which only reflects the average effect of risk-predictive features over the population. Markov-based CM monitoring strategy and natural history matching assign more frequent monitoring to the severe and moderate individuals. However, natural history matching has the lowest prediction accuracy among the adaptive models in the online phase (Figure A-3 d), which indicates that the short-term stochastic changes of PHQ-9 score captured in the triplets are not very predictive of future disease progression.

In this study, we initialize the monitoring threshold by maximizing the sensitivity and specificity on the 6th month validation data. It is notable that in other applications, such as the monitoring of seminal vesicle invasion prior to or during surgery (Vickers and Elkin, 2006), monitoring a low risk patient and missing a severely diseased patient may result in different costs, the objective function in (3.1) could be further extended to minimize the total cost by incorporating the cost per false positive and false negative.

One major limitation of our research is that the depression EHR dataset has limited information on the treatment types, and lack frequent assessments that cover an extended period (i.e. >2 years) of depression progression. Therefore, we are unable to conduct a cost-effectiveness analysis that projects the long-term health outcomes of the adaptive monitoring strategies. For example, an outcome measure such as the quality-adjusted life years (QALYs) gained is affected by the drop-off rate and mortality rate during depression treatment follow-up. In the future, we plan to conduct a more sophisticated cost-effectiveness analysis with consideration of downstream treatment scenarios to accurately estimate long-term health outcomes (e.g. QALYs gained). A possible approach is to build a decision-analytic Markov model (Liu et al., 2016) to simulate the long-term monitoring outcomes and costs under different monitoring strategies and treatment scenarios. Furthermore, conclusions obtained from the ICER outcomes may depend on the dataset used. However, the insights on depression progression patterns, such as the latest PHQ-9 based method performs well under stable depression trajectory, but lacks accuracy under fluctuating and increasing depression patterns, are likely to be generalizable to other populations. To obtain insights on ICERs accounting for uncertainties, we plan to apply the proposed method to different datasets, or for different time windows in the same dataset in the future.

In summary, we demonstrate a decision support framework to adaptively and cost-effectively monitor a heterogeneous depression treatment population. The proposed method may be adaptable to other chronic conditions by integrating the individual prognostic, monitoring strategy design and cost-effectiveness analysis using large-scale EHR data. By applying the proposed framework to chronic depression, we discover four adaptive monitoring strategies that have the potential to improve the current recommendation, and contribute to evidence-based strategies in clinical practice.

Reference

Alagoz, O., Bryce, C. L., Shechter, S., Schaefer, A., Chang, C. C. H., Angus, D. C., and Roberts, M. S. (2005) Incorporating biological natural history in simulation models: Empirical estimates of the progression of end-stage liver disease. *Medical Decision Making*, 25, 620-32.

Alagoz, O., Maillart, L. M., Schaefer, A. J. and Roberts, M. S. (2007) Determining the acceptance of cadaveric livers using an implicit model of the waiting list. *Operations Research*, 55(1), 24-36.

Alagoz, O., Maillart, L. M., Schaefer, A. J. and Roberts, M. S. (2004) The optimal timing of living-donor liver transplantation. *Management Science*, 50(10), 1420-1430.

Aronson, L., Bautista, C. A. and Covinsky, K. (2015) Medicare and care coordination: expanding the clinician's toolbox. *JAMA*, 313(8), 797-8.

Ayer, T., Alagoz, O. and Stout, N. K. (2012) OR forum—a POMDP approach to personalize mammography screening decisions. *Operations Ressearch*, 60, 1019-34.

Bhattacharya, S. (2014). Markov chain model to explain the dynamics of human depression. *Journal of Nonlinear Dynamics*, 2014.

Boult, C. and Wieland, G. D. (2010) Comprehensive primary care for older patients with multiple chronic conditions: "Nobody rushes you through". *JAMA*, 304(17), 1936-43.

Brandeau, M. L., Sainfort, F. and Pierskalla, W. P. (2004) Operations Research and Health Care: A Handbook of Methods and Applications. *International Series in Operations Research & Management Science 70*. Boston, Mass.: Kluwer Academic. viii, 872 p.

Centers for Disease Control and Prevention Mental Health website on depression. Accessed at http://www.cdc.gov/mentalhealth/basics/mental-illness/depression.htm.

Chen, X., Shachter, R.D., Kurian, A.W., and Rubin, D.L. (2017) Dynamic strategy for personalized medicine: An application to metastatic breast cancer. *J Biomed Inform*, 68:50-57.

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

Hay, J.W., Lee, P.L., Jin, H., Guterman, J.J., Gross-Schulman, S., Ell, K., Wu, S. (2018) Cost-effectiveness of a technology-facilitated depression care management adoption model in safety-net primary care patients with type 2 diabetes. *Value in Health*, 21(5), 561-568.

Helm, J. E., Lavieri, M. S., Van Oyen, M. P., Stein, J. D. and Musch, D. C. (2015) Dynamic Forecasting and Control Algorithms of Glaucoma Progression for Clinician Decision Support. *Operations Research*, 63(5), 979-999.

Huang, S. H., LePendu, P., Iyer, S. V., Tai-Seale, M., Carrell, D. and Shah, N. H. (2014) Toward personalizing treatment for depression: Predicting diagnosis and severity. *Journal of the American Medical Informatics Association*, 21, 1069-75.

Hutton, D. W., Tan, D., So, S. K. and Brandeau, M. L. (2007) Cost-effectiveness of screening and vaccinating Asian and pacific islander adults for hepatitis B. *Ann Intern Med*, 147(7), 460-469.

Islam, M. A., Chowdhury, R. I., and Huda, S. (2013). A multistate transition model for analyzing longitudinal depression data. *Bulletin of the Malaysian Mathematical Sciences Society*, 36(3).

Kales, H. C., Kim, H. M., Austin, K. L. and Valenstein, M. (2010) Who receives outpatient monitoring during High-Risk depression treatment periods? *J. Am. Geriatr. Soc.*, 58, 908-13.

Kazemian, P., Helm, J. E., Lavieri, M. S., Stein, J. D. and Van Oyen, M. P. (2015) Dynamic Personalized Monitoring and Treatment Control of Glaucoma. Accessed at https://www.researchgate.net/publication/314419495_Dynamic_Personalized_Monitoring_and_Treatment_Control_of_Glaucoma. Under review.

Kroenke, K. and Spitzer, R. L. (2002) The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals*, 32, 509-15.

Lasko, T. A., Denny, J. C. and Levy, M. A. (2013) Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS One*, 8, e66341.

Lee, E., Lavieri, M. S., Volk M. L. and Xu, Y. (2015) Applying reinforcement learning techniques to detect hepatocellular carcinoma under limited screening capacity. *Health Care Manag Sci*, 2015. 18(3), 363-75.

Lin, Y., Huang, S., Simon, G. E. and Liu, S. (2016) Analysis of depression trajectory patterns using collaborative learning. *Mathematical Biosciences*, 282, 191-203.

Lin, Y., Huang, S., Simon, G.E., and Liu, S. (2018a) Data-based decision rules to personalize depression follow-up. *Scientific Reports* 8(1), 5064.

Lin, Y., Liu, K., Byon, E., Qian, X., Liu, S. and Huang, S. (2017) A collaborative learning framework for estimating many individualized regression models in a heterogeneous population. *IEEE Transactions on Reliability*, 99, 1-14.

Lin, Y., Liu, S. and Huang, S. (2018b) Selective sensing of a heterogeneous population of units with dynamic health conditions. *IIE Transactions*, accepted.

Lin, Y., Qian, X., Krischer, J., Vehik, K., Lee, H. and Huang, S. (2014) A rule-based prognostic model for type 1 diabetes by identifying and synthesizing baseline profile patterns. *PloS One*, 9, e91095.

Liu S, Huang S, Lin Y, Yang X, Huang J, Shang W. (2016) Depression care management: personalized assessment to cost-effective population interventions. *INFORMS Annual Conference*, Nashville, Tennessee. November 2016.

Liu, S., Brandeau, M. and Goldhaber-Fiebert, J. D. (2017) Optimizing patient treatment decisions in an era of rapid technological advances: the case of hepatitis C treatment. *Health care management science*, 20(1), 16-32.

Mason, J. E., England, D. A., Denton, B. T., Smith, S. A., Kurt, M. and Shah, N. D. (2012) Optimizing statin treatment decisions for diabetes patients in the presence of uncertain future adherence. *Med Decis Making*, 32(1), 154-66.

National institute of mental health website on depression. Accessed at http://www.nimh.nih.gov/health/topics/depression/index.html.

Oskooyee, K. S., Rahmani, A. M. and Kashani, M. M. R. (2011) Predicting the severity of major depression disorder with the markov chain model. *In Proceedings of the International Conference on Bioscience, Biochemistry and Bioinformatics*, 5.

Reynolds, C. F. R. and Frank, E. (2016) US Preventive Services Task Force Recommendation Statement on Screening for Depression in Adults: Not Good Enough. *JAMA Psychiatry*, Published online January 26, 2016. http://archpsyc.jamanetwork.com/article.aspx?articleid=2484482.

Richardson, L. P., McCauley, E., Grossman, D. C., McCarty, C. A., Richards, J., Russo, J. E., Rockhill, C. and Katon, W. (2010) Evaluation of the patient health questionnaire-9 item for detecting major depression among adolescents. *Pediatrics*, 126, 1117-23.

Salomon, J. A., Weinstein, M. C. Hammitt, J. K. and Goldie, S. J. (2002) Empirically calibrated model of hepatitis C virus infection in the United States. *American Journal of Epidemiology*, 156(8), 761-773.

Sandikci, B., Maillart, L. M., Schaefer, A. J. and Roberts, M. S. (2013) Alleviating the Patient's Price of Privacy Through a Partially Observable Waiting List. *Manage Sci*, 59(8), 1836-1854.

Sandikci, B., Maillart, L. M., Shaefer, A. J., and Alagoz, O. (2008) Estimating the Patient's Price of Privacy in Liver Transplantation. *Operations Research*, 56(6), 1393-1410.

Shechter, S. M., Bailey, M. D., Schaefer, A. J. and Roberts, M. S. (2008) The Optimal Time to Initiate HIV Therapy Under Ordered Health States. *Operations Research*, 56(1), 20-33.

Simon, G.E., Manning, W.G., Katzelnick, D.J., Pearson, S.D., Henk, H.J., Helstad, C.S. (2001) Cost-effectiveness of systematic depression treatment for high utilizers of general medical care. *Archives of general psychiatry,* 58.2: 181-187.

Simon, G. E., Rutter, C. M., Peterson, D., Oliver, M., Whiteside, U., Operskalski, B. and Ludman, E. J. (2013) Does response on the PHQ-9 Depression Questionnaire predict subsequent suicide attempt or suicide death? *Psychiatr Serv*, 64(12), 1195-202.

Simon, G. E., VonKorff, M., Rutter, C. and Wagner, E. (2000) Randomised trial of monitoring, feedback, and management of care by telephone to improve treatment of depression in primary care. *BMJ*, 320(7234), 550-4.

Sutin, A. R., Terracciano, A., Milaneschi, Y., An, Y., Ferrucci, L. and Zonderman, A. B. (2013) The trajectory of depressive symptoms across the adult life span. *JAMA Psychiatry*, 70, 803-11.

U.S. Department of Health & Human Services. Multiple Chronic Conditions: A Strategic Framework. Optimum Health and Quality of Life for Individuals with Multiple Chronic Conditions. Accessed at http://www.hhs.gov/sites/default/files/ash/initiatives/mcc/mcc_framework.pdf. 2010: Washington, DC.

Untzer, J., Katon, W., Callahan, C. M., Williams Jr, J. W., Hunkeler, E., Harpole, L., Hoffing, M., Della Penna, R. D., Noël, P. H., Lin, E. H. and Areán, P.A. (2002) Collaborative care management of late-life depression in the primary care setting: A randomized controlled trial. *Jama*, 288, 2836-45.

Valenstein, M., Vijan, S., Zeber, J.E., Boehm, K., Buttar, A. (2001) The cost-utility of screening for depression in primary care. *Ann Intern Med*, 134(5): 345-60.

Venkatesh, A., Goodrich, K. and Conway, P.H. (2014) Opportunities for quality measurement to improve the value of care for patients with multiple chronic conditions. *Ann Intern Med*, 161(10), 76-80.

Vickers, A. J. and Elkin, E. B. (2006) Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26, 565-74.

Wang, X., Sontag, D. and Wang, F. (2014) Unsupervised Learning of Disease Progression Models. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Ward, B. W., Schiller, J. S. and Goodman, R. A. (2012) Multiple chronic conditions among US adults: a 2012 update. *Prev Chronic Dis*, 11, E62.

Weinstein, M. C., O'Brien, B., Hornberger, J., Jackson, J., Johannesson, M., McCabe, C. and Luce, B. R. (2003) Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices—Modeling Studies. *Value Health*, 6(1), 9-17.

Weinstein, M. C., Siegel, J. E., Gold, M. R., Kamlet, M. S. and Russell, L. B. (1996) Recommendations of the Panel on Cost-effectiveness in Health and Medicine. *JAMA*, 276(15), 1253-8.

Appendix

Table A-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, n(%)		
18-29	56(8.85%)	35(10.54%)
30-44	168(26.54%)	84(25.30%)
45-64	297(46.92%)	177(53.31%)
≥65	112(17.69%)	36(10.84%)
Sex, n(%)		
Female	418(66.03%)	235(70.78%)
Male	215(33.97%)	91(29.22%)
Statistical Summarization, *mean (standard deviation)*		
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)
9th 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 A-2: 12 identified rules in the rule-based model.

Rule 1	Average PHQ-9 < 11.32 & standard deviation of Charlson comorbidity score <0.84	Rule 7	29< Age <65 & average PHQ-9 > 16.95
Rule 2	Latest PHQ-9 < 17.80 & percentage of moderate < 37.5%	Rule 8	Age > 29 & Latest PHQ-9 >18.93 & 75% quantile of PHQ-9 > 18.37
Rule 3	Median of 9th question score < 0.81 & Latest PHQ-9 < 9.92	Rule 9	Age > 29 & 25% quantile of PHQ-9 > 15.17
Rule 4	Median of 9th question score < 1.63 & minimal PHQ-9 < 12.55	Rule 10	Latest PHQ-9 > 15.10
Rule 5	Average 9th question score < 0.92 & first PHQ-9 < 20.58	Rule 11	Minimal PHQ-9 > 12.46 & average PHQ-9 > 15.47
Rule 6	Latest PHQ-9 < 12.95 & 25% quantile of 9th question score < 0.69	Rule 12	75% quantile of PHQ-9 > 21.25

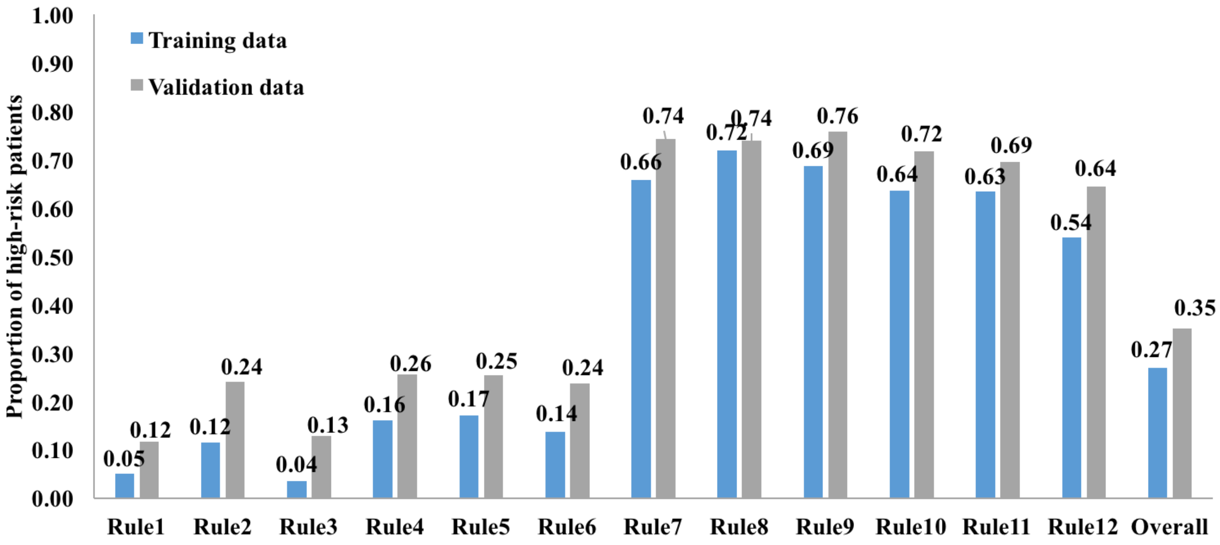


Figure A-1: Proportion of high-risk patients (PHQ-9 score ≥ 15) in rule endorsing groups on training (5th month) and validation (6th month) data.

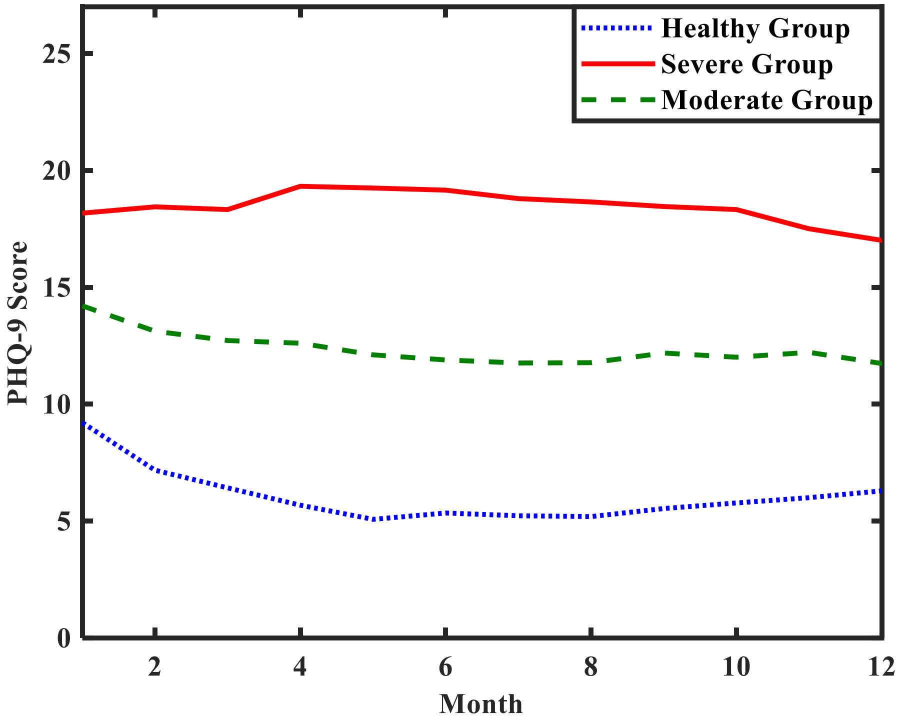
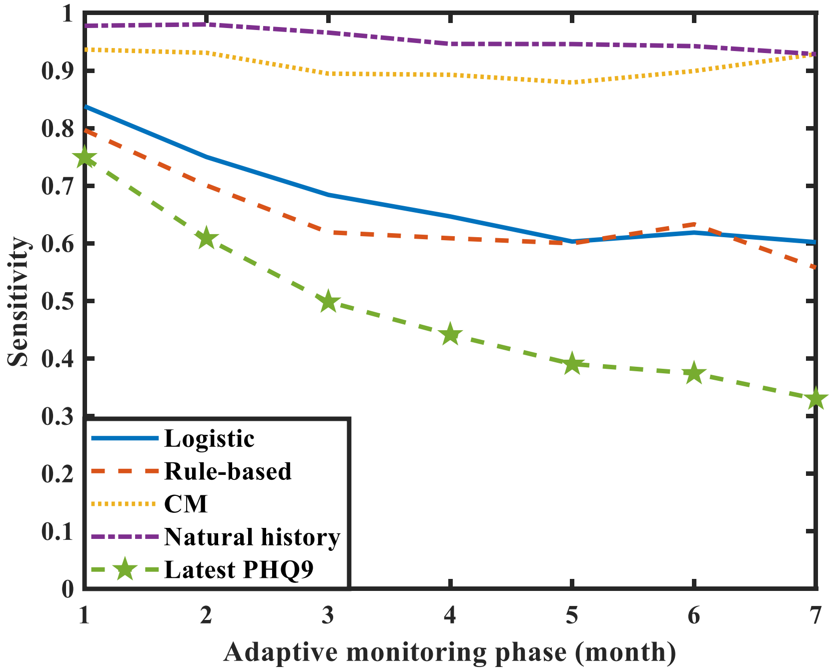
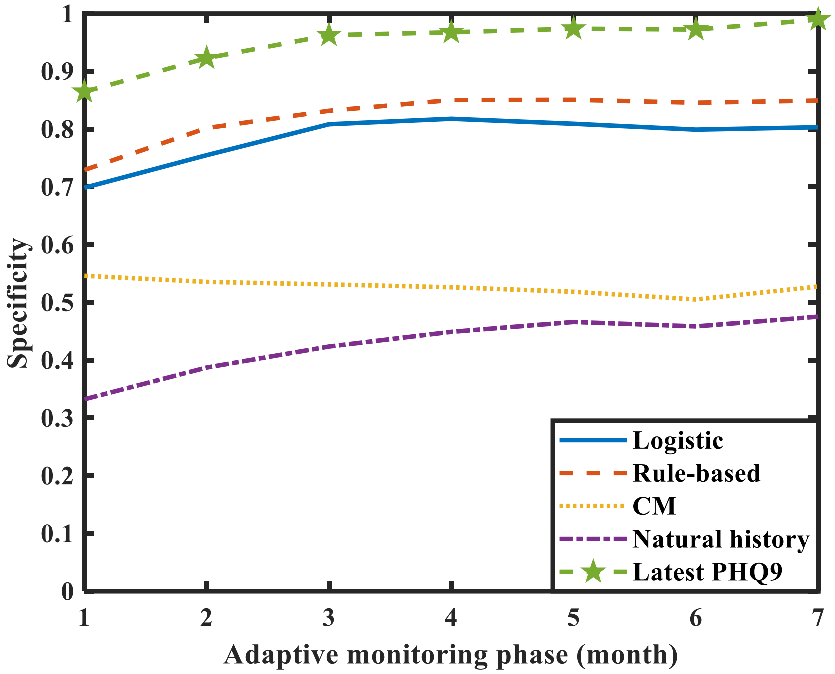


Figure A-2: Three depression trajectory patterns in the Markov-based collaborative model.

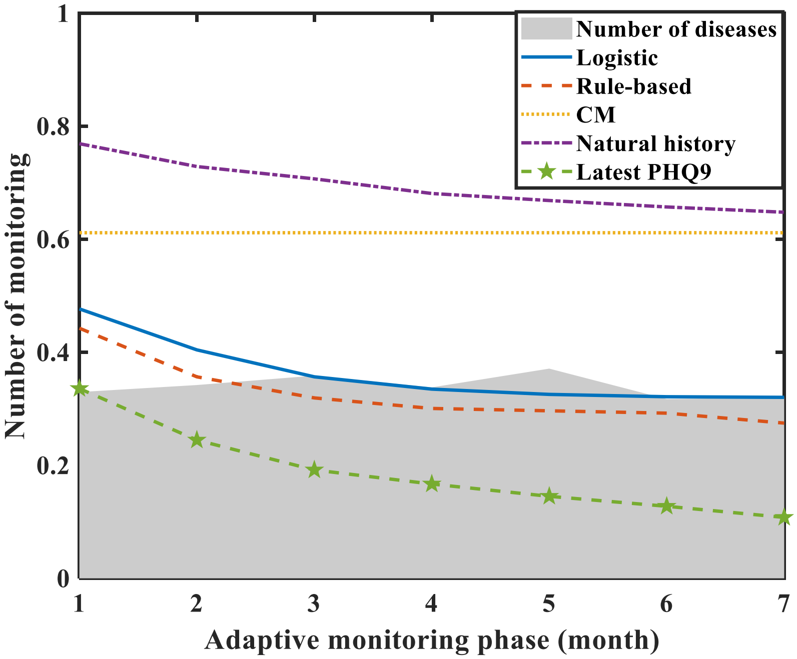


(a)



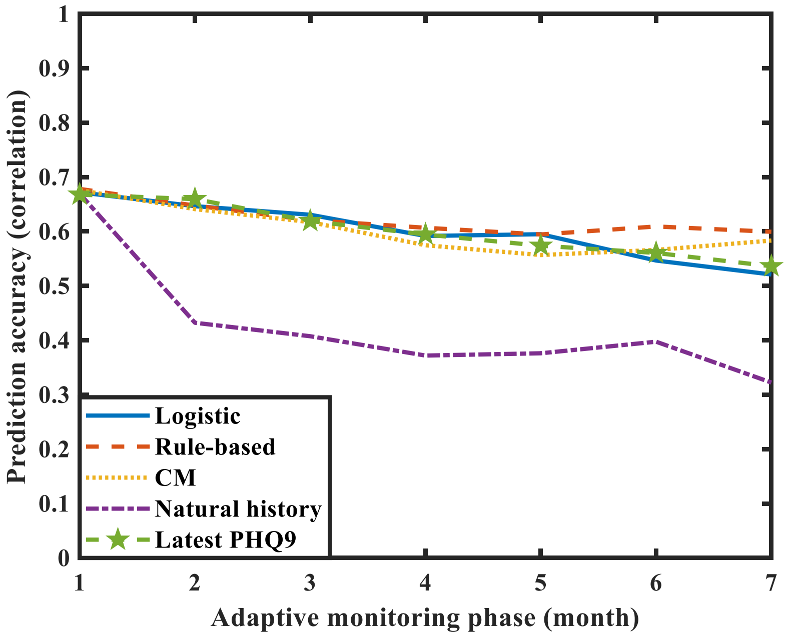
Specificity
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1
2
3
4
5
6
7
Adaptive monitoring phase (month)
Logistic
Rule-based
CM
Natural history
Latest PHQ9

(b)



Number of monitoring
0
0.2
0.4
0.6
0.8
1
1
2
3
4
5
6
7
Adaptive monitoring phase (month)
Number of diseases
Logistic
Rule-based
CM
Natural history
Latest PHQ9

(c)



(d)

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