Journal of Biomedical Informatics 73 (2017) 115–124

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

CHI: A contemporaneous health index for degenerative disease monitoring using longitudinal measurements

Yijun Huang [a], Qiang Meng [b], Heather Evans [c], William Lober [d], Yu Cheng [e], Xiaoning Qian [f], Ji Liu [a], Shuai Huang [b,*]

[a] Department of Computer Science, University of Rochester, United States
[b] Department of Industrial & Systems Engineering, University of Washington, United States
[c] Department of Surgery, University of Washington, United States
[d] Department of Biomedical Informatics and Medical Education, University of Washington, United States
[e] Healthcare Analytic Research, IBM T.J. Watson Research Center, United States
[f] Department of Electrical & Computer Engineering, Texas A&M University, United States

ARTICLE INFO

Article history:
Available online 18 July 2017

Keywords:
Risk monitoring
Longitudinal measurements
Degenerative disease
Machine learning
Convex optimization

ABSTRACT

In this paper, we develop a novel formulation for contemporaneous patient risk monitoring by exploiting the emerging data-rich environment in many healthcare applications, where an abundance of longitudinal data that reflect the degeneration of the health condition can be continuously collected. Our objective, and the developed formulation, is fundamentally different from many existing risk score models for different healthcare applications, which mostly focus on predicting the likelihood of a certain outcome at a pre-specified time. Rather, our formulation translates multivariate longitudinal measurements into a contemporaneous health index (CHI) that captures patient condition changes over the course of progression. Another significant feature of our formulation is that, CHI can be estimated with or without label information, different from other risk score models strictly based on supervised learning. To develop this formulation, we focus on the degenerative disease conditions, for which we could utilize the monotonic progression characteristic (either towards disease or recovery) to learn CHI. Such a domain knowledge leads us to a novel learning formulation, and on top of that, we further generalize this formulation with a capacity to incorporate label information if available. We further develop algorithms to mitigate the challenges associated with the nonsmooth convex optimization problem by first identifying its dual reformulation as a constrained smooth optimization problem, and then, using the block coordinate descent algorithm to iteratively solve the optimization with a derived efficient projection at each iteration. Extensive numerical studies are performed on both synthetic datasets and real-world applications on Alzheimer's disease and Surgical Site Infection, which demonstrate the utility and efficacy of the proposed method on degenerative conditions that include a wide range of applications.

1. Introduction

Although there is no universal definition of the concept "patient condition", it has been a crucial concept in the communications between clinicians and frequently referenced by healthcare providers. Developing a precise contemporaneous longitudinal index that can faithfully reflect the underlying patient condition across the course of the condition's progression holds great value for facilitating a range of clinical decision-makings. For instance, it will help early detection of patient deterioration to help reduce the number of serious incidents, as the National Patient Safety Agency has recently reported that 11% of serious incidents are a function of deterioration not acted upon mainly due to the failure to recognize the sign of deterioration [2,11]. It will also help enhance the continuity of care since a longitudinal perspective of the patient condition can be provided for clinicians and healthcare providers during the entire care process. Towards this goal, technological innovations are emerging in many healthcare applications, which have given rise to a data-rich environment where an abundance of longitudinal clinical measurements that reflect the degeneration of the health condition can be continuously collected. In what follows, we present several examples:

- Alzheimer's disease (AD): A number of biomarkers have been developed to measure the degeneration of the neural systems,

[*] Corresponding author.
E-mail address: shuai.huang.ie@gmail.com (S. Huang).

http://dx.doi.org/10.1016/j.jbi.2017.07.003
1532-0464/© 2017 Elsevier Inc. All rights reserved.

including the growing array of neuroimaging modalities [27,26]. Fig. 1 shows examples of PET images that represent the typical patterns of the AD progression process from normal aging to mild cognitive impairment, to AD onset. As shown in Fig. 1, the PET scan imaging technique provides a objective measure to capture the progressive loss of neural activities. While AD is essentially a multi-factorial complex disease, the biological condition that the PET scan can capture only represents one part of the degeneration process; therefore, there are many other biomarker measurement techniques have been developed, evidenced by the newly developed imaging modalities such as PiB-PET [29] and proteomics-based measurements [16].

- Surgical site infection (SSI): Another example is the surgical site infections to assess the risk of infection after surgeries that may greatly help reduce the healthcare cost due to readmission [8,18,20,31,14,35,4]. While AD and SSI are two health conditions of different nature, analytically the challenges in these two bear a great similarity. Just like in AD, for better preventing the SSI, many data collection methods have been developed to closely monitor the patients who are subject to the risk of developing SSI. Daily wound measurements, such as the temperature, granularity, distance of the wound, could be acquired to assess the condition of the wound, together with other non-wound related but important clinical signals such as heart rate, morning body temperature, and NG tube presence. Fig. 2 shows a real-world longitudinal data of a wound-related variable, which clearly shows the monotonic degradation process of the SSI group. While one variable shows a large uncertainty on the patient condition, how to combine all these longitudinal data to create the envisioned CHI that captures the underlying disease process is a very interesting question.

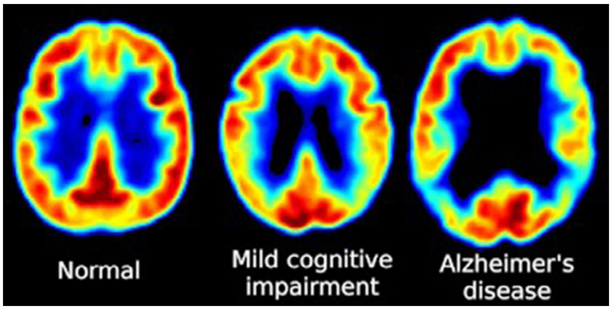

Fig. 1. The PET (Positron emission tomography) scan imaging provides a objective measure to capture the progressive loss of neural activities. *Source:* http://healthncare.info/.

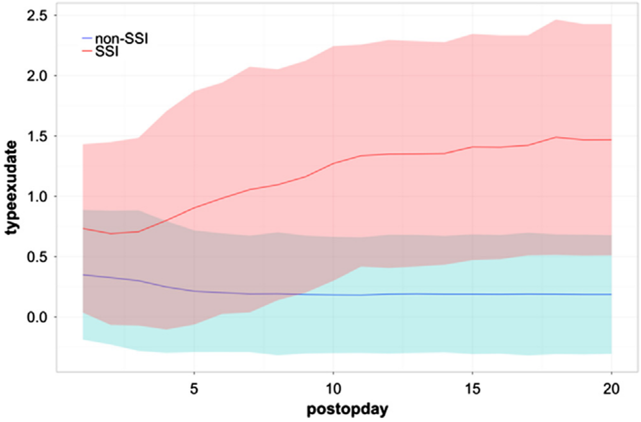

Fig. 2. Examples of the longitudinal data of wound assessment that could gradually separate the SSI group with the non-SSI group as the condition progresses over time.

1.1. Distinct features of CHI and challenges

For the envisioned contemporaneous health index (CHI), we'd like to highlight that our objective, and the developed formulation, are fundamentally different from many existing risk score models for different healthcare applications, which mostly focus on predicting the likelihood of a certain outcome at a pre-specified time (i.e., 10 years risk or within-30 days readmission risk) by utilizing the associations between some clinical measurements with the health status of interest. First, our formulation can be used **even without any label information**. Rather, our strategy is to utilize the monotonic progression characteristic (either towards disease or recovery) underlying the longitudinal measurements to learn CHI. Thus, although we further generalize our formulation to be able to incorporate possible label information to enhance the learning, our problem is not strictly a supervised learning problem. In both simulation studies and real-world applications, we demonstrate that, without label information, our CHI method can still be trained and used to predict, which could generate comparable performance as the supervised learning methods that use the label information in some applications. Second, our CHI method actually aims to visualize the progression process over time, rather than predicting a particular location in the progression process. In other words, to enable the application of supervised learning methods on longitudinal measurements, the number of time points needed as input actually results in an implicit limitation that we could only detect the disease after the observation time. As a comparison, our method has no such a limitation and can be flexibly applied.

That said, there are challenges to develop the learning formulation of CHI. **Model-wise**, the learned health index should be *monotonic* such that the degradation of the health condition can be detected. Also, for robust clinical decision-making, it is better that the health index is *homogeneous* within the same group of subjects, such that normal subjects tend to show a similar health index within a certain range, while another diseased group tends to concentrate in another range of CHI. **Data-wise**, first, we will deal with *irregular* time series data, collected from different individuals at different time points within different time duration. These characteristics have violated the basic assumptions of many statistical models for multivariate time series data analysis such as the ARMA model and its extensions [39,5], spectral analysis methods [28], and state space models [9]. Second, for some applications, there are *label information* for each individual, e.g., we could track many individuals longitudinally and then label each individual as either diseased or not diseased. How to combine the label information with the time series data is rarely investigated in the literature to the best of our knowledge. Third, while many clinical variables can be measured longitudinally, there might be many variables that are not useful or redundant with others. Existing sparse learning methods provide inspiration to solve this problem, but are not readily applicable here for irregular time series data with label information. All those challenges call for a new formulation.

1.2. Our approach and contributions

Thus, in this paper, we aim to fill in the gap and investigate proper computational formulations for optimal combination of multiple longitudinal clinical signals to create a contemporaneous longitudinal index. Sparse learning is also incorporated into this formulation. The learning formulation is found to be a nonsmooth convex optimization, which presents a computational challenge. We further develop algorithms to mitigate the challenges associated with the nonsmooth convex optimization by first identifying its dual reformulation as a constrained smooth optimization problem, and then, using the block coordinate descent algorithm to iterate the optimization process with each iteration solved by an efficient projection operator. Extensive numerical studies are performed on both synthetic datasets and real-world applications on AD and SSI, which demonstrate the utility and efficacy of the proposed method.

The paper is organized as follows. In Section 2, related work in the literature will be reviewed and discussed. In Section 3, the proposed analytic framework will be presented and the corresponding computational algorithm will be derived. In Section 4, the proposed method will be implemented and validated using two real-world applications; one is for monitoring of brain health in AD and the other is monitoring of SSI. We will conclude the study in Section 5.

2. Related work

Our proposed formulation is related to several topics in the literature of medical risk score models, time series models, and supervised learning. Comparing with the former two, our problem has a different objective by studying irregular multivariate longitudinal data to develop a contemporaneous index that is required to fit the underlying degradation process. Comparing with supervised learning, our method is also different since our method can be trained without label information, which could still generate comparable performance with the supervised learning methods as we will show in both simulation studies and real-world applications later.

2.1. Existing risk score models

Risk score models have been widely developed for different healthcare applications. Despite the diversity of the application background, most of them focus on predicting the likelihood of a certain outcome at a pre-specified time (i.e., 10 years risk or within-30 days readmission risk) based on some static measurements or static features extracted from time series data with given length. A typical formulation of these risk score models is to use a regression model or any other weighted sum model to combine multiple static clinical measurements for predicting the care outcome such as disease onset or re-admission. For example, many risk score models for AD have been using the genetic variables, neuropsychological variables, baseline imaging markers, other kinds of biomarkers, to predict the onset of AD [38,37]. SSI Risk Score (SSIRS) models [8,18,20,31,14,35,4] range from simple (e.g. NNIS which includes only 3 predictors) to complex (e.g. SSIRS with 12 covariates and 4 interactions). Similar observations can be made on many other risk score models for a diverse range of health conditions. Despite some exceptions such as [45,44] that investigated how to predict the decline of some AD-related score over time as a multi-task learning model, these existing efforts have been limited on combining static data rather than longitudinal data. For example, multi-modality data integration methods have been developed in [40–42] to combine neuroimaging data, genomic data, clinical data, etc.; however, these data are all static measurements. Similar models have also been developed in engineering applications [23]. Our problem has a fundamentally different objective from the existing risk score models by studying irregular multivariate longitudinal data to develop a contemporaneous index that is required to fit the underlying degradation process.

2.2. Existing time series models

Most of the existing time series models in the statistics literature rely on the assumption that the time series data are measured on regular time points. Most time series models, such as the ARMA model and its extensions [39,5], spectral analysis methods [28], and state space models [9], assume fundamentally different data-generating mechanisms than ours. For example, the ARMA models investigate patterns on the autocorrelation and assume no latent construct that is underlying the multivariate time series data. Another array of models, developed in the biostatistics community, show more relevance with ours. These methods, including the growth mixture modeling [17], latent class analysis [17,21], and latent class growth analysis [34], assume that there is a latent structure such as a growth pattern beyond the time series data of some biological traits such as BMI index. However, those models are only suited to analyze longitudinal cohort data, in which a cohort of subjects are followed and assessed at regular time intervals. Although few missing data points are allowed, but still the assumption of the regular time points has been crucial for valid and effective applications of these methods, so time series of individuals can be conceptualized as random samples from multivariate distributions, paving the way for bringing the distribution theory (together with the extended analysis methods such as ANOVA and mixed effect models) into the regime of trajectory data analysis. Recently, there are some developments of probabilistic graphical models for trajectory modeling [10,30,36], exhibiting powerful capabilities on characterizing the nonlinearity and non-Gaussian patterns in the time series data. However, many of these methods still assume that the time series data are measured on regular time points; and further, they could not incorporate the label information, nor to support the required decision-making capabilities such as the monotonicity of the health index and the homogeneity of the health index.

2.3. Related supervised learning methods

Besides the statistical models for timer series data analysis, there have been efforts on the extension of many supervised learning methods for time series data analysis. A common approach is to first extract features from the time series data and then treat it as a regression or classification problem by using SVM [32], logistic regression [15], or linear regression [12]. It is case-dependent regarding which features should be extracted from the time series data. Apparently, our problem and method are different from these models, since they aim to discriminate classes or predict outcomes based on a time series input, while our problem is to learn the underlying health index as a quantification of the disease progression. Another category of approaches is based on HMMs [7], which model the time series data by a stochastic process, including the Maximum Entropy Markov model [25] and Input-Output HMMs [3], to name a few. However, these models didn't specifically address irregular time series data with label information; nor to incorporate the decision-making requirements of a health index that should be monotonic and homogeneous within groups. Recently there are some other developments to advance the data mining methodologies for longitudinal data from healthcare applications, such as [46,43,13], which have different objectives and targeted data types from ours. Also, one important aspect of our method is that our method can be trained without label information, which still achieves satisfactory performance that could be useful in clinical applications.

3. Methodology

Denote $h_{n,t}$ as the CHI value for subject n at time t. One conceptual challenge of developing this envisioned CHI is the lack of generally accepted definition of patient condition. We limit our scope in this study to degenerative conditions such as the AD or SSI, where the condition of the patient will deteriorate over time once the disease progression is triggered and no intervention is applied to delay or stop the progression. The degenerative nature of these degenerative conditions lead to the monotonic assumption that CHI should be monotonic, i.e., $h_{n,t_1} \geqslant h_{n,t_2}$ if $t_1 \geqslant t_2$, if we assume that higher index represents more severe condition. Since CHI is a latent construct that is not directly measurable, clinical variables associated with it can be measured over time, which provide us data to learn it.

3.1. The formulation of the proposed method

Denote the training data by

$$\mathbf{x}_{n,t} = [x_{n,1,t}, x_{n,2,t}, \ldots, x_{n,d,t}]^\top \in \mathbb{R}^d,$$

where $x_{n,k,t}$ is the value of the k-th variable for individual n at time t. $n \in \{1, \ldots, N\}$ is the sample index and $t \in \{1, \ldots, T_n\}$ is the time index. Note that different individuals could be measured with different length of time and at different time locations. Let $y_n \in \{1, -1\}$ be the label of the nth sample. Converting the measurements $\mathbf{x}_{n,t}$ into $h_{n,t}$ needs a mathematical model for $h_{n,t} = f(\mathbf{x}_{n,t})$. Here, for simplicity and interpretability, we start with the linear models, i.e.,$h_{n,t} = \mathbf{x}_{n,t} \cdot \mathbf{w}$, where $\mathbf{w} \in \mathbb{R}^{d\times 1}$ is a vector of weight coefficients to combine the d variables.

Then, we propose a learning framework to learn the health index from the irregular time series data with label information, which also simultaneously encourages the quality of the health index for decision-making requirements such as the monotonicity and homogeneity of the health index.

Let $\mathbf{z}_{n,t}$ be the difference of two successive data vectors

$$\mathbf{z}_{n,t} := \mathbf{x}_{n,t+1} - \mathbf{x}_{n,t}.$$

N^+ and N^- denote the total number of positive and negative samples, respectively,

$$N^+ := |\{n|y_n = 1\}| \quad \text{and} \quad N^- := |\{n|y_n = -1\}|.$$

$\bar{\mathbf{x}}^+_{T_n}$ and $\bar{\mathbf{x}}^-_{T_n}$ denote the center of data vectors at time T_n for all positive and negative samples, respectively, that is,

$$\bar{\mathbf{x}}^+_{T_n} := \frac{1}{N^+} \sum_{n\in\{n|y_n=1\}} \mathbf{x}_{n,T_n}$$

$$\bar{\mathbf{x}}^-_{T_n} := \frac{1}{N^-} \sum_{n\in\{n|y_n=-1\}} \mathbf{x}_{n,T_n}.$$

The formulation of the generic learning framework is shown in below:

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 + \tag{1a}$$

$$\beta \sum_{n\in\{1,\cdots,N\}} \max(0, 1 - y_n(\mathbf{x}^\top_{n,T_n}\mathbf{w} + b)) + \tag{1b}$$

$$\alpha \sum_{\substack{n\in\{1,\cdots,N\}\\ t\in\{1,\cdots,T_n-1\}}} \max(0, 1 - \mathbf{z}^\top_{n,t}\mathbf{w}) + \tag{1c}$$

$$\frac{\lambda}{2}\left(\frac{1}{N^+} \sum_{n\in\{N^+|y_n=1\}} ((\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}^+_{T_n})^\top \mathbf{w})^2\right) + \tag{1d}$$

$$\frac{\lambda}{2}\left(\frac{1}{N^-} \sum_{n\in\{N^-|y_n=-1\}} ((\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}^-_{T_n})^\top \mathbf{w})^2\right) + \tag{1e}$$

$$\gamma\|\mathbf{w}\|_1. \tag{1f}$$

Items in (1f) can be explained as follows:

- The first term (1a) and second term (1b) is the SVM formulation that aims to utilize the label information to enhance the discriminatory power of CHI.
- The term (1c) is invented to enforce the monotonicity of the learned health index.
- Items (1d) and (1f) are invented to encourage the homogeneity of CHI within the group that has the same health status.
- The last term, (1f), is the L_1-norm penalty that is used to encourage sparsity of the features.

Note that the proposed formulation is inspired by some data fusion models in engineering sensor fusion such as [23] and the linear discriminant analysis that encourage monotonicity and homogeneity of the combined index. Building on that, the proposed formulation generalized many existing models, such as SVM, sparse SVM, and particularly the spirit of LASSO. We will derive an efficient algorithm, ensuring the optimal computational complexity, to solve this generic model.

3.2. Optimization

The proposed formulation (1) is convex but contains multiple nonsmooth terms such as (1b), (1c) and (1f). Since this is nonsmooth convex optimization problem, existing solver might not solve it directly and efficiently. The general nonsmooth optimization algorithms, for example, the subgradient descent methods, need complexity $O(1/\epsilon^2)$ to obtain an approximate solution with precision ϵ. In this section, we propose an efficient algorithm that can exploit the optimization structure of the proposed formulation, and come up with a more efficient optimization algorithm to solve the proposed formulation. The proposed algorithm will reduce the complexity to $O(1/\epsilon^{0.5})$. The basic idea is, we first simplify the formulation by merging (1a), (1d), and (1e); then we derive its dual optimization problem; and finally we propose an efficient algorithm to solve the dual formulation, whose solution can be used to retrieve the primal solution.

3.2.1. Simplification of the proposed formulation (1)

We can simplify Eq. (1) by merging (1a), (1d), and (1e), as all of them are in quadratic forms. Define

$$\|\mathbf{w}\|^2_Q := \mathbf{w}^\top Q\mathbf{w} = \|\mathbf{w}\|^2 + \lambda\left(\frac{1}{N^+}\sum_{n\in\{n|y_n=1\}} ((\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}^+_{T_n})^\top \mathbf{w})^2\right) + \lambda\left(\frac{1}{N^-}\sum_{n\in\{n|y_n=-1\}} ((\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}^-_{T_n})^\top \mathbf{w})^2\right),$$

where Q is defined as

$$Q := I + \lambda\left(\frac{1}{N^+}\sum_{n\in\{n|y_n=1\}} (\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}^+_{T_n})(\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}^+_{T_n})^\top + \frac{1}{N^-}\sum_{n\in\{n|y_n=-1\}} (\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}^-_{T_n})(\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}^-_{T_n})^\top\right).$$

With that, (1) can be simplified to

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2_Q + \gamma\|\mathbf{w}\|_1 + \alpha \sum_{\substack{n\in\{1,\cdots,N\}\\ t\in\{1,\cdots,T_n-1\}}} \max(0, 1 - \mathbf{z}^\top_{n,t}\mathbf{w}) + \beta \sum_{n\in\{1,\cdots,N\}} \max(0, 1 - y_n(\mathbf{x}^\top_{n,T_n}\mathbf{w} + b)). \tag{2}$$

As the two hinge loss terms in (2) are nonsmooth, we reformulate them as the constraints by introducing two relaxation variables ϵ and ξ. Then, (1) is equivalent to

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_Q^2 + \alpha \mathbf{1}^\top \xi + \beta \mathbf{1}^\top \epsilon + \gamma \|\mathbf{w}\|_1 \\ \text{s.t.} \quad & \mathbf{1} - Z^\top \mathbf{w} - \xi \leqslant \mathbf{0} \\ & \mathbf{1} - \widehat{X}^\top \mathbf{w} - b\mathbf{y} - \epsilon \leqslant \mathbf{0} \end{aligned} \tag{3}$$

where

$$\begin{aligned} & \xi = (\xi_{1,1}, \cdots, \xi_{1,T_1-1}, \cdots, \xi_{N,1}, \cdots, \xi_{N,T_N-1})^\top, \\ & Z = (Z_{1,1}, \cdots, Z_{1,T_1-1}, \cdots, Z_{N,1}, \cdots, Z_{N,T_N-1}), \\ & \epsilon = (\epsilon_1, \cdots, \epsilon_N)^\top, \\ & \mathbf{y} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)^\top, \\ & \widehat{X} = (\mathbf{y}_1 X_{1,T_1}, \cdots, \mathbf{y}_N X_{N,T_N}). \end{aligned}$$

3.2.2. Dual reformulation

We then derive the dual formulation of (3) to achieve more efficient optimization solution. First, we substitute the ℓ_1-norm penalty in (3) by its conjugate norm $\|\mathbf{w}\|_1 = \max_{\|\mathbf{s}\|_\infty \leqslant 1} \langle \mathbf{s}, \mathbf{w} \rangle = \max_{\|\mathbf{s}\|_\infty \leqslant 1} - \langle \mathbf{s}, \mathbf{w} \rangle$, which leads to the following formulation:

$$\begin{aligned} \min_{\substack{\mathbf{w},b, \\ \epsilon \geqslant \mathbf{0} \xi \geqslant \mathbf{0}}} \max_{\|\mathbf{s}\|_\infty \leqslant 1} \quad & \frac{1}{2}\|\mathbf{w}\|_Q^2 + \alpha \mathbf{1}^\top \xi + \beta \mathbf{1}^\top \epsilon - \gamma \langle \mathbf{w}, \mathbf{s} \rangle \\ \text{s.t.} \quad & \mathbf{1} - Z^\top \mathbf{w} - \xi \leqslant \mathbf{0} \\ & \mathbf{1} - \widehat{X}^\top \mathbf{w} - b\mathbf{y} - \epsilon \leqslant \mathbf{0}. \end{aligned}$$

To remove the constraints, we further introduce two new dual variables **u** and **v**:

$$\begin{aligned} \min_{\substack{\mathbf{w},b \\ \epsilon \geqslant \mathbf{0} \xi \geqslant \mathbf{0}}} \max_{\substack{\mathbf{u} \geqslant \mathbf{0} \\ \mathbf{v} \geqslant \mathbf{0} \|\mathbf{s}\|_\infty \leqslant \gamma}} \quad & \frac{1}{2}\|\mathbf{w}\|_Q^2 + \alpha \mathbf{1}^\top \xi + \beta \mathbf{1}^\top \epsilon - \langle \mathbf{w}, \mathbf{s} \rangle + \\ & \langle \mathbf{u}, \mathbf{1} - Z^\top \mathbf{w} - \xi \rangle + \langle \mathbf{v}, \mathbf{1} - \widehat{X}^\top \mathbf{w} - b\mathbf{y} - \epsilon \rangle. \end{aligned}$$

By strong duality due to convexity, we can safely swap min and max and obtain the optimal form for $\mathbf{w} : \mathbf{w}^* = Q^{-1}(\mathbf{s} + Z\mathbf{u} + \widehat{X}\mathbf{v})$, as $Q\mathbf{w} - \mathbf{s} - Z\mathbf{u} - \widehat{X}\mathbf{v} = \mathbf{0}$. Plugging the optimal form for **w**, we obtain an equivalent form as:

$$\begin{aligned} & \max_{\substack{\mathbf{u} \geqslant \mathbf{0} \\ \mathbf{v} \geqslant \mathbf{0} \|\mathbf{s}\|_\infty \leqslant \gamma}} \min_{\substack{b \\ \epsilon \geqslant \mathbf{0} \xi \geqslant \mathbf{0}}} -\frac{1}{2}\|\mathbf{s} + Z\mathbf{u} + \widehat{X}\mathbf{v}\|_{Q^{-1}}^2 + \langle \mathbf{1}, \mathbf{u} \rangle + \\ & \quad \langle \mathbf{1}, \mathbf{v} \rangle + \langle \alpha \mathbf{1} - \mathbf{u}, \xi \rangle + \langle \beta \mathbf{1} - \mathbf{v}, \epsilon \rangle - b \langle \mathbf{v}, \mathbf{y} \rangle \\ & = max_{\substack{\mathbf{u} \geqslant \mathbf{0} \\ \mathbf{v} \geqslant \mathbf{0} \|\mathbf{s}\|_\infty \leqslant \gamma}} \begin{cases} -\frac{1}{2}\|\mathbf{s} + Z\mathbf{u} + \widehat{X}\mathbf{v}\|_{Q^{-1}}^2 + \langle \mathbf{1}, \mathbf{u} \rangle + \langle \mathbf{1}, \mathbf{v} \rangle & \text{if } \mathbf{u} \leqslant \alpha \mathbf{1}, \mathbf{v} \leqslant \beta \mathbf{1}, \langle \mathbf{v}, \mathbf{y} \rangle = 0 \\ +\infty & \text{otherwise.} \end{cases} \end{aligned}$$

This can be rewritten as the following constrained smooth convex optimization problem, which can be solved efficiently:

$$\begin{aligned} \min_{\mathbf{s},\mathbf{u},\mathbf{v}} \quad & F(\mathbf{s}, \mathbf{u}, \mathbf{v}) := \frac{1}{2}\|\mathbf{s} + Z\mathbf{u} + \widehat{X}\mathbf{v}\|_{Q^{-1}}^2 - \langle \mathbf{1}, \mathbf{u} \rangle - \langle \mathbf{1}, \mathbf{v} \rangle \\ \text{s.t.} \quad & \mathbf{0} \leqslant \mathbf{u} \leqslant \alpha \mathbf{1} \\ & \mathbf{0} \leqslant \mathbf{v} \leqslant \beta \mathbf{1} \\ & \langle \mathbf{v}, \mathbf{y} \rangle = 0 \\ & \|\mathbf{s}\|_\infty \leqslant \gamma. \end{aligned} \tag{4}$$

3.2.3. Algorithm

To solve (4), we can apply the block coordinate descent (BCD) algorithm [33], which iteratively updates a variable or a block of variables, while fixing other variables at their latest values. Since every update decreases the objective function value, convergence is automatically guaranteed. Specifically, the following iterative procedure is proposed for solving (1):

*Update **s** while **u** and **v** are fixed:*

$$\begin{aligned} \mathbf{s} & = \text{Proj}_{-\gamma \mathbf{1} \leqslant \mathbf{s} \leqslant \gamma \mathbf{1}}(\mathbf{s} - \eta_\mathbf{s} \nabla_\mathbf{s} F(\mathbf{s}, \mathbf{u}, \mathbf{v})) \\ & = \max(-\gamma \mathbf{1}, \min(\gamma \mathbf{1}, \mathbf{s} - \eta_\mathbf{s} \nabla_\mathbf{s} F(\mathbf{s}, \mathbf{u}, \mathbf{v}))), \end{aligned}$$

where the projection operator $\text{Proj}_{\mathbf{0} \leqslant \mathbf{s} \leqslant \gamma \mathbf{1}}(\mathbf{y})$ means to project the point **y** onto the set $\{\mathbf{s} | \mathbf{0} \leqslant \mathbf{s} \leqslant \gamma \mathbf{1}\}$. $\eta_\mathbf{s}$ is the step size for variable **s**. One can either set it to a safe constant $\eta_\mathbf{s} = \frac{1}{\|Q^{-1}\|}$, or adaptively tune it using the line search strategy [6, Algorithm 9.2] such that the objective decreases sufficiently.

*Update **u** while **s** and **v** are fixed:*

$$\begin{aligned} \mathbf{u} & = \text{Proj}_{\mathbf{0} \leqslant \mathbf{u} \leqslant \alpha \mathbf{1}}(\mathbf{u} - \eta_\mathbf{u} \nabla_\mathbf{u} F(\mathbf{s}, \mathbf{u}, \mathbf{v})) \\ & = \max(\mathbf{0}, \min(\alpha \mathbf{1}, \mathbf{u} - \eta_\mathbf{u} \nabla_\mathbf{u} F(\mathbf{s}, \mathbf{u}, \mathbf{v}))), \end{aligned}$$

where the safe step size is $\eta_\mathbf{u} = \frac{1}{\|Z^\top Q^{-1} Z\|}$.

*Update **v** while **u** and **s** are fixed:*

$$\mathbf{v} = \text{Proj}_{\substack{\mathbf{0} \leqslant \mathbf{v} \leqslant \beta \mathbf{1} \\ \langle \mathbf{v}, \mathbf{y} \rangle = 0}}(\mathbf{v} - \eta_\mathbf{v} \nabla_\mathbf{v} F(\mathbf{s}, \mathbf{u}, \mathbf{v})), \tag{5}$$

where the safe step size is $\eta_\mathbf{v} = \frac{1}{\|\widehat{X}^\top Q^{-1} \widehat{X}\|}$.

Algorithm 1. Block Coordinate Descent for Solving (1)

```
Require: Problem parameters $\{Z, Q, \widehat{X}, \mathbf{y}, \alpha, \beta, \gamma\})$ and
    Optimization parameters $\{\eta_\mathbf{s}, \eta_\mathbf{u}, \eta_\mathbf{v}, \rho \in (0, 1)\}$ (step sizes
    $\eta_\mathbf{s}, \eta_\mathbf{u}$, and $\eta_\mathbf{v}$ can be adaptively decided using linear search
    alternatively)
Ensure: $\mathbf{w}^*, b^*$
Initialize $k = 0$
while Not converge do
$\mathbf{s}_{k+1} = \max(-\gamma \mathbf{1}, \min(\gamma \mathbf{1}, \mathbf{s}_k - \eta_\mathbf{s} \nabla_\mathbf{s} F(\mathbf{s}_k, \mathbf{u}_k, \mathbf{v}_k)))$
$\mathbf{u}_{k+1} = \max(\mathbf{0}, \min(\alpha \mathbf{1}, \mathbf{u}_k - \eta_\mathbf{u} \nabla_\mathbf{s} F(\mathbf{s}_{k+1}, \mathbf{u}_k, \mathbf{v}_k)))$
$\mathbf{v}_{k+1} = \text{Proj}_{\substack{\mathbf{0} \leqslant \mathbf{v} \leqslant \beta \mathbf{1} \\ \langle \mathbf{v}, \mathbf{y} \rangle = 0}}(\mathbf{v}_k - \eta_\mathbf{v} \nabla_\mathbf{v} F(\mathbf{s}_{k+1}, \mathbf{u}_{k+1}, \mathbf{v}_k))$
$k \leftarrow k + 1$
Recover the primal variables

$\mathbf{w}^* = Q^{-1}(\mathbf{s} + Z\mathbf{u} + \widehat{X}\mathbf{v})$
$b^* = \sum_{\{i | \mathbf{v}_i \in (0, \beta)\}} \mathbf{y}_i - \mathbf{w}^{*\top} \mathbf{x}_i$
```

Algorithm 1 summarizes the BCD algorithm, in which updating **s** and **u** is efficient since they admit closed-form solutions. However, for the update of **v**, it does not have a closed form. An efficient algorithm is needed to find an exact solution to (5). In what follows, we consider a more general problem by considering an arbitrarily positive vector **c**, for which $\mathbf{1}\beta$ in (5) is the special case:

$$\begin{aligned} \text{Proj}_{\substack{\mathbf{0} \leqslant \mathbf{v} \leqslant \mathbf{c} \\ \langle \mathbf{v}, \mathbf{a} \rangle = 0}}(\bar{\mathbf{v}}) := \arg\min_{\mathbf{v} \in \mathbb{R}^p} \quad & \frac{1}{2}\|\mathbf{v} - \bar{\mathbf{v}}\|^2 \\ \text{s.t.} \quad & \mathbf{0} \leqslant \mathbf{v} \leqslant \mathbf{c} \\ & \langle \mathbf{a}, \mathbf{v} \rangle = 0 \end{aligned} \tag{6}$$

This is essentially to find the projection of $\bar{\mathbf{v}}$ onto the intersection of two sets: a cubic $\mathbf{0} \leqslant \mathbf{v} \leqslant \mathbf{c}$ and a hyperplane $\langle \mathbf{a}, \mathbf{v} \rangle = 0$. The projection onto either of them can be solved exactly in linear time $O(p)$ with p being the dimension of **v**; but projection onto their intersection is nontrivial. We first provide the following theorem to guarantee that this projection can be solved in approximately linear time $O(p \log p)$:

Theorem 1. Eq. (6) *can be solved exactly with complexity* $O(p\log p)$.

The proof can be found in Appendix. Correspondingly, we can derive an efficient implementation, as illustrated in Algorithm 2, to find the solution to (6) in such complexity.

It is worth to mention that the subproblem (6) was also considered in [22]. The key difference lies on that their algorithm can only guarantee to find an approximate solution to (6) using an iterative algorithm; but our method here can exactly solve it with complexity $O(p\log p)$. Therefore, our method is more efficient and accurate.

Algorithm 2. Hyperplane Cue Projection Algorithm for solving (6).

```
Require: $\mathbf{a} \in \mathbb{R}^p, \bar{\mathbf{v}} \in \mathbb{R}^p, \mathbf{c} \in \mathbb{R}^p_+$
Ensure: $\mathbf{v}^*$
Construct two sequences $\Delta$ and $\delta$ with the length of $2n$:
for $i = 1 : p$ do
    $\delta_{2i-1} = -i$
    $\delta_{2i} = i$
    if $a_i > 0$ then
        $\Delta_{2i-1} = -\frac{\bar{v}_i}{a_i}$
        $\Delta_{2i} = \frac{c_i}{a_i} - \frac{\bar{v}_i}{a_i}$
    else
        $\Delta_{2i-1} = \frac{c_i}{a_i} - \frac{\bar{v}_i}{a_i}$
        $\Delta_{2i} = -\frac{\bar{v}_i}{a_i}$
Sort the sequence $\Delta$ in the increasing order and keep $\delta$
  with the same order;
Initialize $g_1 = 0$;
for $i = 1 : p$ do
    if $a_i < 0$ then
        $g_1 = g_1 + a_i \times c_i$;
Initialize $sp = 0$;
for $k = 2 : 2p$ do
    $sp = sp - \mathrm{sgn}(\delta_{k-1}) \times a^2_{-\mathrm{sgn}(\delta_{k-1})\times\delta_{k-1}}$;
    $g_k = g_{k-1} + (\Delta_k - \Delta_{k-1}) \times sp$
    if $g_k \geqslant 0$ then
        $\lambda^* = \Delta_{k-1} - \frac{g_{k-1}}{sp}$;
        $\mathbf{p}^* = \min(0, -(\lambda^*\mathbf{a} + \bar{\mathbf{v}}))$;
        $\mathbf{q}^* = \max(0, \max(\lambda^*\mathbf{a} + \bar{\mathbf{v}} - \mathbf{c}, 0))$;
        $\mathbf{v}^* = \lambda^*\mathbf{a} + \mathbf{p}^* - \mathbf{q}^* + \bar{\mathbf{v}}$;
```

Putting all things together, in Algorithm 1, the complexity of BCD is proven to be $O(1/\epsilon)$ where ϵ is the precision of the solution [1]. Note that the algorithm can be easily extended to the accelerated BCD algorithm [1], which can improve the complexity to $O(1/\sqrt{\epsilon})$.

Note that the proposed approach can incorporate nonlinear models by using nonlinear basis functions of predictors, e.g., polynomial basis functions are typical examples. 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 into the reproducing kernel Hilbert space defined by a certain kernel function. Here, we use the Gaussian process as an example for explanation. Assuming the degradation model $f(\mathbf{x})$ takes the form as a Gaussian process which can be defined as

$$f(\mathbf{x}) = \mathbf{x}\beta + \sum_{j=1}^{n_i} \alpha_j k(\mathbf{x}, \mathbf{x}_j)$$

where $k(\mathbf{x}, \mathbf{x}_j))$ is the covariance function of the two vectors $\mathbf{x}$ and $\mathbf{x}_j$, $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_n]^\top = (K_{y,y} + \sigma^2 I)^{-1}(\mathbf{y} - X\beta)$, σ^2 is the variance parameter, I is the identity matrix, and $K_{y,y} = [k(\mathbf{x}_i, \mathbf{x}_j)]_{n\times n}$. The covariance function could take highly nonlinear forms such as the Gaussian covariance function or the polynomial covariance function. 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.

3.3. Online implementation of CHI

The model can be trained offline via implementing Algorithm 1, using training samples from a cohort of subjects whose conditions deteriorate over time. The online implementation can be applied to any subject whose condition is actually unknown. Our method can be applied on an individual's data from day 1 since we don't have to input a segment of time series data over a few days for prediction. Rather, our method can gradually generate a longitudinal trajectory representing the condition's progression. There could be more than one way to interpret this trajectory and generate prediction. For example, three common rules are shown in below: (1) Rule 1: Health index at any time ⩾a certain cutoff value (2) Rule 2: Health index at two consecutive time ⩾a certain cutoff value (3) Rule 3: The change of health index at any time ⩾a certain cutoff value Here, Rule 1 aims to identify a critical point, larger than which the patient has a high likelihood of developing the disease. Rule 2, which aims to be more cautious than Rule 1, will identify the cutoff value, i.e., only if the patient's CHI value is larger than this cutoff value for two consecutive time points, the patient is predicted as diseased. Different from Rules 1 and 2 that focus on the absolute value of the health index, Rule 3 aims to capture meaningful change, motivated by the hypothesis that if the health index changes rapidly, even the absolute value of the index is within normality, it still indicates increased risk. The cutoff values can be optimized by learning from training data.

4. Numerical studies

We implement and test CHI using both simulated datasets and real-world applications. We demonstrate that the CHI method can effectively recover the underlying disease progression process that has been buried in the noisy and irregular longitudinal measurements. It not only leads to outstanding prediction performance, but also able to predict the disease earlier than supervised learning method. This makes sense since CHI can be applied on the time series data from day 1, while supervised learning methods usually need to take a segment of time series as input. We further demonstrate that, in both simulation studies and the application on a real-world dataset of SSI, even without label information, the CHI can be trained using irregular time series data that shows comparable performances with supervised method that uses the label information. It indicates that CHI has better clinical utility, i.e., to train CHI, we need less training data and shorter length of time series for achieving the same level of prediction accuracy. Particularly, when the length of time series for training is short (i.e., in other words, when disease is less progressed, the classes become less separated in the time series data, making the supervised method less suitable) or the noise level in the data is large (i.e., which also makes the classes less separated in the time series data), the CHI method could be a great tool for risk monitoring when useful label information is not available.

It is also of interest to mention some practical suggestions for parameter tuning for CHI. For each experiment we will show in this section, we always randomly split the data into two equal parts, one for training and one for testing. For training, we always use 10-fold cross validation to tune the parameters. For applications where the dataset is large enough to afford expensive parameter tuning, it is recommended that a rough grid search is used (e.g., only a few levels for each parameter are considered). Then around the optimal configuration of the parameters, another finer resolution grid search is used. For same applications where feature selection is not critically important as first priority or sample size is not sufficiently large for an expensive full grid search, it is recommended to use a two step approach. In the first step, the optimal parameters for α and β could be identified by the 10-fold cross validation. Then, with the fixed α and β, the feature selection component is added into the learning and the best parameter for γ could be identified by the 10-fold cross validation. Then is the lambda. More details could be found in our simulation studies and real-world applications.

4.1. Simulation study

4.1.1. Simulation steps

We simulate data following the procedure described below:

Step 1. First, we need to generate the underlying health index curve for each subject group. Without loss of generality, we assume that there are two groups, normal vs. diseased. Denote the two health index functions as $f_1(t)$ and $f_2(t)$ for the two groups, respectively. Both functions should be monotonic; and one group progresses faster than another group, i.e., the group with AD is progressing faster than the normal aging group. This leads to the following generic requirements on the two functions:

$$\begin{cases} f_1(t) > f_2(t) & \text{if } t \in (t_{min}, t_{max}) \\ f_1(t_{1,1}) > f_1(t_{1,2}) & \text{if } t_{1,1} > t_{1,2} \\ f_2(t_{2,1}) > f_2(t_{2,2}) & \text{if } t_{2,1} > t_{2,2} \end{cases}$$

In our simulation study, we choose quadratic forms for both functions: $f_l(t) = A_l t^2 + B_l t$ for $l \in \{1, 2\}$, where $A_l \sim \mathcal{U}(A_{l,min}, A_{l,max})$, $B_l \sim \mathcal{U}(B_{l,min}, B_{l,max})$. Quadratic functions have been found useful for modeling the degradation trajectory in some health conditions including AD [24,19]. Also, it provides sufficient flexibility to simulate different scenarios to investigate the performance and behavior of the developed algorithm.

Step 2. Based on the underlying health index, we could further generate the longitudinal clinical measurements as long as the probabilistic relationships between the features of the health index are established. Here, we assume that $\mathbf{x}_{n,t}^k = \mathcal{N}(f(t), \sigma_k^2)$ for $k \in \{1, \ldots, d\}$. Apparently, the features that have larger σ_k^2 tend to be less useful for estimating the health index. We randomize the variance parameters of the features from a uniform distribution: $\sigma_k \sim U(\sigma_{min}, \sigma_{max})$. *Step 3.* The remaining issue is to decide on the length of observation for each subject (denoted as T_n) and the specific time locations (denoted as $t_1, t_2, \ldots, t_{T_n}$), where measurements of the features are simulated. Let $t_{time} \in \{t_1, t_2, \ldots, t_{T_n}\}$. Here $t_{time} \sim U(t_{min}, t_{max})$ and $T_n \in \{N_{min}, N_{min} + 1, \ldots, N_{max}\}$, both taking uniform distributions.

4.1.2. Simulation results

We conduct several experiments with the simulated data to investigate the performance of our method across different settings. For all the experiments, we set the number of features $d = 100$. For each group, we simulate 50 subjects, while $t_{min} = 3$ and $t_{max} = 7$. Unless notified otherwise, the parameters of the health index functions are set to be $A_{1,min} = 1,\ A_{1,max} = 4,\ B_{1,min} = 1,\ B_{1,max} = 2,\ A_{2,min} = 0,\ A_{2,max} = 1,\ B_{2,min} = 0,\ B_{2,max} = 1$. For each experiment, we always randomly split the dataset into two even parts, one for training and one for testing. Also, due to the space limit, we only evaluate Rule 1 for the final prediction by the Area Under the Curve (AUC) on the testing data for performance comparison. Results for the other two rules show similar patterns.

In the first experiment, we investigate the performance of our method across a range of choices on $\sigma_{min},\ \sigma_{max}$. For each simulation, we repeat it 50 times and derive the mean and standard derivation of the AUC. The results are shown in Table 1.

In order to evaluate the contribution of monotonicity and label information terms, we not only investigate the performance of our full method, but also the performances when we fix $\alpha = 0$ or fix $\beta = 0$. Actually, these two models also correspond to a **SVM model** and a **Model trained without label information**, respectively. As shown in Table 1, for each choice, our method performs well when the data is reasonably noisy, i.e., when σ's for the corresponding features are small. The performance will drop if the features become more noisy. Another observation is that both the label information and the monotonicity constraint contribute significantly to the performance, and it also shows that, enforcing monotonicity alone leads to satisfactory AUC than utilizing label information alone. This makes sense, because that when σ is large, the measurements of the features of the two groups overlap significantly, such that the label information becomes less useful. However, since the longitudinal trend can be captured by the monotonicity constraint, it plays a crucial role to discriminate the two groups. This demonstrates that the CHI method can effectively utilize the degradation characteristics of the underlying disease process for learning and recovering the patient's condition using noisy measurements. As the **Model trained without label information** can provide good performance, it has a great potential for clinical applications where the label information is costly to acquire.

Another experiment is to investigate the performance of our method across a range of choices on $A_{2,max}, B_{2,max}$. Following the same structure as we used in the first experiments, the results are shown in Table 2.

As larger difference between $A_{2,max}$ and $B_{2,max}$ means that the underlying health index functions of the two groups are more separated, Table 2 essentially tells that when the two are well separated and the features are not too noisy, the monotonicity constraint is redundant with the label information, playing insignificant role in this scenario. However, the **Model trained without label information** is still able to match the performance of the supervised **SVM model**.

We further investigated the performance of our method on feature selection. With all the other parameters in the simulation procedure fixed as the default values we have mentioned, we set $\sigma_{min} = 5,\ \sigma_{max} = 10$ to allow the simulation of many noisy features. Fig. 3 shows that our method can effectively identify a subset of features that can achieve a good AUC performance. Note that in the results of Fig. 3, the other parameters except γ are optimized

Table 1
AUC performance across different $\sigma_{min},\ \sigma_{max}$.

$[\sigma_{min}, \sigma_{max}]$	$(\alpha = 0, \beta^*)$	$(\beta = 0, \alpha^*)$	(α^*, β^*)
$[0, 5]$	0.960 ± 0.033	0.984 ± 0.019	0.985 ± 0.017
$[5, 10]$	0.831 ± 0.065	0.831 ± 0.069	0.862 ± 0.049
$[10, 15]$	0.662 ± 0.085	0.674 ± 0.083	0.704 ± 0.089
$[15, 20]$	0.599 ± 0.059	0.610 ± 0.095	0.627 ± 0.079

Table 2
AUC performance across different $A_{2,max},\ B_{2,max}$.

$[A_{2,max}, B_{2,max}]$	$(\alpha = 0, \beta^*)$	$(\beta = 0, \alpha^*)$	(α^*, β^*)
$[2, 1]$	0.629 ± 0.076	0.661 ± 0.055	0.661 ± 0.055
$[4, 2]$	0.850 ± 0.048	0.855 ± 0.035	0.881 ± 0.044
$[7, 2]$	0.892 ± 0.046	0.881 ± 0.064	0.892 ± 0.046
$[9, 3]$	0.939 ± 0.045	0.921 ± 0.044	0.939 ± 0.045

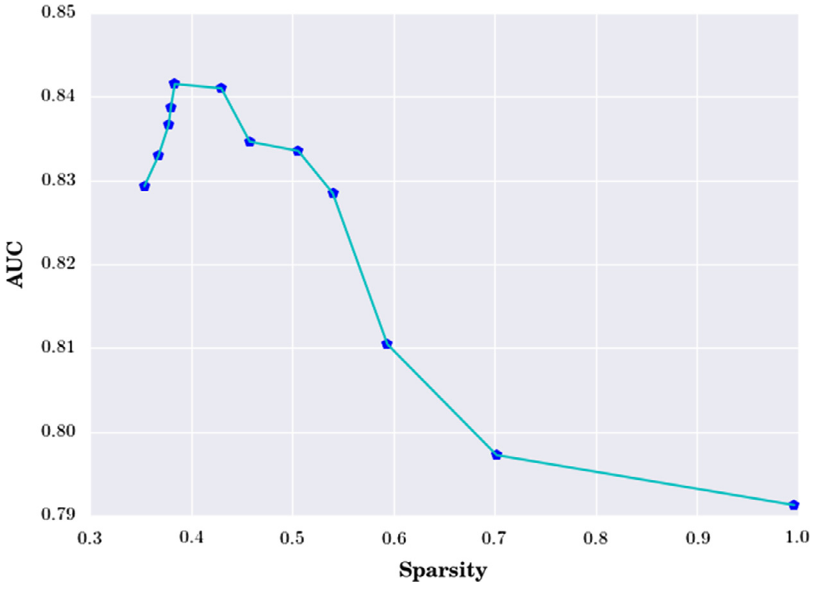

Fig. 3. The AUC of selected features.

by 10-folder cross validation on the training data and the AUC is evaluated on the testing data.

4.2. Longitudinal neuroimaging data from the ADNI

We use the FDG-PET images of 162 subjects (Alzheimer's Disease: 74, Normal aging: 88) downloaded from ADNI (www.loni.ucla.edu/ADNI). For each subject, there are at least three time points and at most seven time points. The data has been preprocessed and the Automated Anatomical Labeling has been used to segment each image into 116 anatomical volumes of interest (AVOIs). We select 90 AVOIs that are in the cerebral cortex in our study. Each AVOI becomes a variable in the application of the proposed health index method here. The measurement data of each region, according to the mechanism of FDG-PET, is the regional average FDG binding counts, representing the degree of glucose metabolism. Extensive evidences in the literature have shown that the glucose metabolism will decline as a function of the aging, while the pathology of neurodegenerative diseases such as AD will further accelerate the declination, providing a perfect application example for implementing and testing the proposed health index method.

Similarly to the simulation studies, we first investigate the contribution of the label information and the monotonicity constraint in the prediction performance. Such a comparison is made across a range of ratios of the training/testing datasets for all three rules, as shown in Table 3, while for each model training, 10-fold cross validation is used to find the optimal parameters. For simplicity, we set $\lambda = \gamma = 0$ in those experiments. An overall observation is that the results are consistent with our simulation studies. For ADNI data, it seems that both the label information and the monotonicity constraint contribute significantly to the total performance. This observation is also valid for all the Rules. Also, Rule 1 is slightly better than Rule 2, and both are better than Rule 3. Also, for most of the cases, the **Model trained without label information** could provide satisfactory prediction result. This is a valuable trait since in clinical practices, it could be very costly to label the individuals. A model that can have clinical utility using less expensive data could be very attractive.

We then investigate the feature selection capability of our method with different γ. Again, with 50% of the data as the training data, we adopt 10-folder cross validation to identify the optimal remaining parameters in the model; then we investigate the feature selection results with different γ. The results are shown in Fig. 4, which demonstrates that our method can lead to efficient feature selection. We further identify the best set of features that lead to the highest AUC, as illustrated by the vertical line in Fig. 4 with the names of the identified features in the legend.

4.3. Longitudinal wound assessment data

We further implement our method on a SSI dataset with longitudinal wound measurements from 857 patients, among which 169 are SSI patients and 539 are normal control. The data include wound measurement variables, for example, wound edge distance, temperature, include exudate amount, etc. Some other physiological variables such as heart rate are also provided in the data. Subjects are measured in time length ranging from 3 days to 21 days.

Table 3
AUC performance on the ADNI data across different ratios of training/testing datasets.

Training	HIRule1		
Size	($\alpha = 0, \beta^*$)	($\alpha^*, \beta = 0$)	(α^*, β^*)
20%	0.888 ± 0.045	0.841 ± 0.064	0.898 ± 0.037
25%	0.878 ± 0.040	0.837 ± 0.054	0.896 ± 0.020
33.3%	0.902 ± 0.033	0.859 ± 0.038	0.915 ± 0.024
50%	0.902 ± 0.043	0.873 ± 0.038	0.915 ± 0.024
66.7%	0.903 ± 0.043	0.891 ± 0.034	0.919 ± 0.033
75%	0.933 ± 0.036	0.903 ± 0.039	0.937 ± 0.033
	HIRule2		
20%	0.853 ± 0.052	0.838 ± 0.048	0.874 ± 0.039
25%	0.881 ± 0.032	0.838 ± 0.031	0.892 ± 0.024
33.3%	0.885 ± 0.033	0.844 ± 0.037	0.900 ± 0.025
50%	0.909 ± 0.022	0.872 ± 0.041	0.921 ± 0.017
66.7%	0.919 ± 0.026	0.873 ± 0.040	0.930 ± 0.022
75%	0.921 ± 0.020	0.852 ± 0.051	0.927 ± 0.019
	HIRule3		
20%	0.714 ± 0.046	0.669 ± 0.049	0.745 ± 0.035
25%	0.694 ± 0.047	0.693 ± 0.045	0.733 ± 0.041
33.3%	0.738 ± 0.041	0.704 ± 0.047	0.759 ± 0.036
50%	0.761 ± 0.044	0.721 ± 0.054	0.778 ± 0.050
66.7%	0.765 ± 0.075	0.743 ± 0.062	0.793 ± 0.067
75%	0.778 ± 0.058	0.746 ± 0.042	0.801 ± 0.043

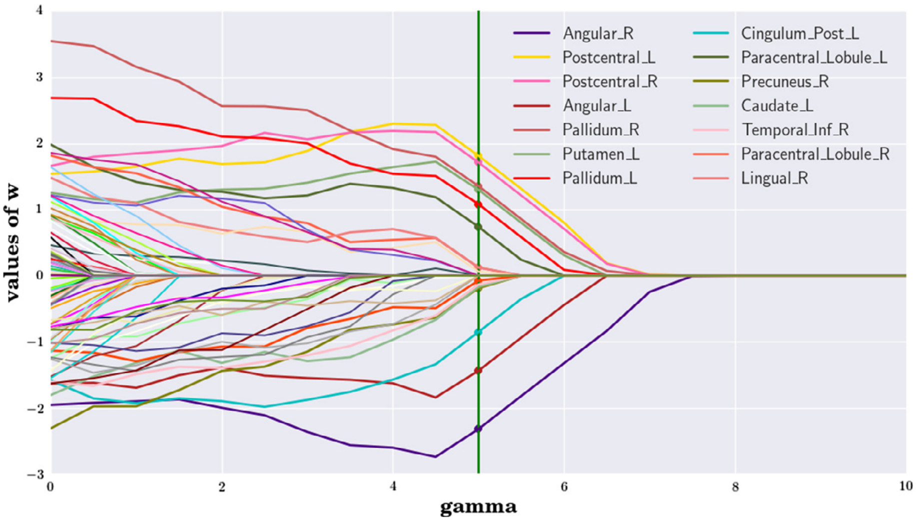

Fig. 4. The solution path of sparse learning.

Many of the longitudinal measurements correlate with the degradation process of the SSI patients, as Fig. 2 clearly demonstrates. While most existing SSI prediction models only use preoperative and operative variables, [20,4,35,14,31], to the best of our knowledge, we are the first team who have developed a systematic treatment to utilize the dynamic wound data of an individual for SSI monitoring and risk prediction. Due to the space limit, we only present a part of our results. Table 4 shows a comparison of different settings of our method that use different number of time points in model training. From Table 4, we can see that the **Model trained without label information** can achieve comparable prediction performance as the **SVM model** that uses label information. Also, it seems that, both the monotonicity constraint and the label information contribute significantly to the total performance.

5. Discussion and conclusions

In this paper, we have developed a novel formulation to monitor the patient condition using their longitudinal clinical measurements that appear to be irregular time series data with label information. Specifically, we focus on the degenerative disease conditions such as Alzheimer's disease, for which the underlying disease degradation process is monotonic. Comparing with many risk score models that have been developed for different healthcare applications but mostly focusing on predicting the likelihood of a certain outcome at a pre-specified time (i.e., 10 years risk or within-30 days readmission risk) based on some static measurements, our formulation is fundamentally new since we take multivariate longitudinal measurements as input and convert them into a health index to capture changes happening over the course of progression. We have developed a novel formulation to tackle the unique learning challenges (i.e., the monotonicity constraint and the homogeneity requirement of health index) and data challenges (i.e., as the data is essentially irregular time series with label information), and have further derived the algorithms to mitigate the challenges associated with the nonsmooth convex optimization by first identifying its dual reformulation as a constrained smooth optimization problem, and then, using an efficient block coordinate descent algorithm with efficient projection to iteratively solve the optimization problem. Extensive numerical studies are performed on both synthetic datasets and real-world applications on Alzheimer's disease and Surgical Site Infection that demonstrate the utility and efficacy of the proposed method. Further research topics, for instance, include extensions of the proposed methods to many other diseases that may have different degradation characteristics from the degenerative diseases. It is also of interest to extend the method to nonlinear health index models.

Table 4
AUC performance for different length of time.

Rule	Length	$\alpha = 0,\ \beta^*$	$\alpha^*,\ \beta = 0$	$\alpha^*,\ \beta^*$
Rule1	5	0.626 ± 0.022	0.596 ± 0.010	0.663 ± 0.027
	10	0.765 ± 0.026	0.702 ± 0.013	0.785 ± 0.029
	15	0.805 ± 0.017	0.759 ± 0.006	0.826 ± 0.011
Rule2	5	0.585 ± 0.021	0.587 ± 0.010	0.618 ± 0.020
	10	0.720 ± 0.029	0.675 ± 0.013	0.746 ± 0.018
	15	0.770 ± 0.019	0.737 ± 0.008	0.789 ± 0.017
Rule3	5	0.613 ± 0.050	0.628 ± 0.017	0.653 ± 0.010
	10	0.810 ± 0.013	0.809 ± 0.017	0.822 ± 0.014
	15	0.839 ± 0.013	0.840 ± 0.011	0.847 ± 0.014

Conflict of interest

No conflict of interest is observed by the authors.

Acknowledgement

The authors acknowledge funding support from the National Science Foundation – United States under Grant CMMI-1505260.

Appendix A. Proof to Theorem 1

Proof. To solve Eq. (6), we first derive its equivalent min–max problem Eq. (7):

$$\min_{\mathbf{v}} \max_{\lambda,\mathbf{p}\geqslant\mathbf{0},\mathbf{q}\geqslant\mathbf{0}} \frac{1}{2}\|\mathbf{v}-\bar{\mathbf{v}}\|^2 - \lambda\langle\mathbf{a},\mathbf{v}\rangle - \langle\mathbf{p},\mathbf{v}\rangle + \langle\mathbf{q},\mathbf{v}-\mathbf{c}\rangle \tag{7}$$

and safely swap min and max due to strong duality. By optimizing $\mathbf{v}$, we get the optimal form for $\mathbf{v} : \mathbf{v}^* = \lambda\mathbf{a} + \mathbf{p} - \mathbf{q} + \bar{\mathbf{v}}$. Plug $\mathbf{v}^*$ into Eq. (7), then we get:

$$\max_{\lambda,\mathbf{p}\geqslant\mathbf{0},\mathbf{q}\geqslant\mathbf{0}} -\frac{1}{2}\|\lambda\mathbf{a}+\mathbf{p}-\mathbf{q}+\bar{\mathbf{v}}\|^2 - \langle\mathbf{q},\mathbf{c}\rangle \tag{8}$$

which is the equivalent dual problem of Eq. (6). $\mathbf{v}^*$ can be recovered by solving Eq. (8). First to simplify Eq. (8), we define

$$f_i(\lambda) := \min_{\mathbf{p}_i \geqslant 0, \mathbf{q}_i \geqslant 0} (\frac{1}{2}(\lambda \mathbf{a}_i + \mathbf{p}_i - q_i + \bar{\mathbf{v}}_i)^2 + \mathbf{c}_i \mathbf{q}_i)$$

and $f(\lambda) := \sum_{i=1}^n f_i(\lambda)$, thus (8) is equivalent to:

$$\begin{aligned} \min_\lambda f(\lambda) =& \min_\lambda \sum_{i=1}^n \min_{\substack{\mathbf{p}_i \geqslant 0 \\ \mathbf{q}_i \geqslant 0}} \left(\frac{1}{2}(\lambda \mathbf{a}_i + \mathbf{p}_i - \mathbf{q}_i + \bar{\mathbf{v}}_i)^2 + \mathbf{c}_i \mathbf{q}_i \right) \\ =& \min_\lambda \sum_{i=1}^n f_i(\lambda) \end{aligned} \tag{9}$$

By solving Eq. (9), we can obtain the optimal value for $\mathbf{p}_i$ and $\mathbf{q}_i$: if $\lambda \mathbf{a}_i + \bar{\mathbf{v}}_i \leqslant 0, \mathbf{p}_i^* = -(\lambda \mathbf{a}_i + \bar{\mathbf{v}}_i)$ and $\mathbf{q}_i^* = 0$; if $\lambda \mathbf{a}_i + \bar{\mathbf{v}}_i > 0, \mathbf{p}_i^* = 0$ and $\mathbf{q}_i^* = \max(\lambda \mathbf{a}_i + \bar{\mathbf{v}}_i - \mathbf{c}_i, 0)$. Then $f_i(\lambda)$ can be simplified by eliminating $\mathbf{p}_i$ and $\mathbf{q}_i$

$$f_i(\lambda) = \begin{cases} 0 & \text{if } \lambda \mathbf{a}_i + \bar{\mathbf{v}}_i \leqslant 0 \\ \frac{1}{2}(\lambda \mathbf{a}_i + \bar{\mathbf{v}}_i - \mathbf{c}_i - \max(\lambda \mathbf{a}_i + \bar{\mathbf{v}}_i - \mathbf{c}_i, 0))^2 & \\ \quad - \frac{1}{2}\mathbf{c}_i^2 + \mathbf{c}_i(\lambda \mathbf{a}_i + \bar{\mathbf{v}}_i) & \text{otherwise} \end{cases}$$

Note that $f_i(\lambda)$ is convex and differential; so is $f(\lambda)$. Therefore, it turns to be a root finding problem. In general, it is hard to find the exact solution. However, in this case, due to the convexity and smoothness, its differential is a nondecreasing piece linear function. By utilizing this structure, we can define efficient algorithm to solve it exactly. The differential of $f_i(\lambda)$ can be computed from

$$\begin{aligned} \nabla f_i(\lambda) &= \begin{cases} 0 & \text{if } \lambda \mathbf{a}_i + \bar{\mathbf{v}}_i \leqslant 0 \\ \mathbf{a}_i \min(\lambda \mathbf{a}_i + \bar{\mathbf{v}}_i - \mathbf{c}_i, 0) + \mathbf{a}_i \mathbf{c}_i & \text{otherwise} \end{cases} \\ &= \mathbf{a}_i \min(\max(\lambda \mathbf{a}_i + \bar{\mathbf{v}}_i, 0), \mathbf{c}_i). \end{aligned}$$

In the following, we basically exam each linear piece of its gradient until meet the zero point. Note that it has up to $2d$ pieces due to the following decomposition:

For $\mathbf{a}_i > 0$, one has

$$\nabla f_i(\lambda) = \begin{cases} 0 & \lambda < -\frac{\bar{\mathbf{v}}_i}{\mathbf{a}_i} \\ \mathbf{a}_i^2 \lambda + \mathbf{a}_i \bar{\mathbf{v}}_i & -\frac{\bar{\mathbf{v}}_i}{\mathbf{a}_i} \leqslant \lambda \leqslant \frac{\mathbf{c}_i}{\mathbf{a}_i} - \frac{\bar{\mathbf{v}}_i}{\mathbf{a}_i} \\ \mathbf{a}_i \mathbf{c}_i & \lambda > \frac{\mathbf{c}_i}{\mathbf{a}_i} - \frac{\bar{\mathbf{v}}_i}{\mathbf{a}_i} \end{cases}$$

For $\mathbf{a}_i < 0$, one has

$$\nabla f_i(\lambda) = \begin{cases} \mathbf{a}_i \mathbf{c}_i & \lambda < \frac{\mathbf{c}_i}{\mathbf{a}_i} - \frac{\bar{\mathbf{v}}_i}{\mathbf{a}_i} \\ \mathbf{a}_i^2 \lambda + \mathbf{a}_i \bar{\mathbf{v}}_i & \frac{\mathbf{c}_i}{\mathbf{a}_i} - \frac{\bar{\mathbf{v}}_i}{\mathbf{a}_i} \leqslant \lambda \leqslant -\frac{\bar{\mathbf{v}}_i}{\mathbf{a}_i} \\ 0 & \lambda > -\frac{\bar{\mathbf{v}}_i}{\mathbf{a}_i} \end{cases}$$

Algorithm 2 essentially exams one piece by one piece in an increasing order. To go through all pieces, it needs $O(p \log p)$ complexity. □

References

[1] A. Beck, L. Tetruashvili, On the convergence of block coordinate descent type methods, SIAM J. Optim. (2013).

[2] A. Beck, L. Tetruashvili, Safer Care for the Acutely Ill Patient: Learning from Serious Incidents, National Patient Safety Agency, 2013.

[3] Y. Bengio, P. Frasconi, Input-output HMMS for sequence processing, IEEE Trans. Neural Networks (1996).

[4] R. Berger et al., Development and validation of a risk-stratification score for surgical site occurrence and surgical site infection after open ventral hernia repair, J. Am. College Surg. (2013).

[5] G. Box, G. Jenkins, Time Series Analysis: Forecasting and Control, rev. ed., Holden-Day, Oakland, California, 1976.

[6] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[7] T.G. Dietterich, Machine learning for sequential data: a review, in: Structural, Syntactic, and Statistical Pattern Recognition, Springer, 2002.

[8] J.T. DiPiro, R.G. Martindale, A. Bakst, P.F. Vacani, P. Watson, M.T. Miller, Infection in surgical patients: effects on mortality, hospitalization, and postdischarge care, Am. J. Heal. Pharm. 55 (8) (1998) 777–781.

[9] J. Durbin, S.J. Koopman, Time Series Analysis by State Space Methods, Oxford University Press, 2001.

[10] E.B. Fox, Dissertation on Bayesian Nonparametric Learning of Complex Dynamical Phenomena, MIT, 2009.

[11] R. Gaynes et al., Surgical site infection (SSI) rates in the united states, 1992–1998: the national nosocomial infections surveillance system basic SSI risk index, Clin. Infect. Diseases (2001).

[12] J.D. Hamilton, Time Series Analysis, Princeton University Press, Princeton, 1994.

[13] J. Ho, J. Ghosh, J. Sun, Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization, KDD, 2014.

[14] V. Ho et al., Differing risk factors for incisional and organ/space surgical site infections following abdominal colorectal surgery, Diseases Colon Rectum (2011).

[15] D.W. Hosmer Jr., S. Lemeshow, Applied Logistic Regression, John Wiley and Sons, 2004.

[16] A. Hye et al., Proteome-based plasma biomarkers for alzheimer's disease, Brain (2006).

[17] T. Jung, K.A.S. Wickrama, An introduction to latent class growth analysis and growth mixture modeling, Soc. Personal. Psychol. Compass (2008).

[18] A.E. Kanters, D.M. Krpata, J.A. Blatnik, Y.M. Novitsky, M.J. Rosen, Modified hernia grading scale to stratify surgical site occurrence after open ventral hernia repairs, J. Am. College Surg. (2012).

[19] N.L. Komarova, C.J. Thalhauser, High degree of heterogeneity in alzheimer's disease progression patterns, LoS Comput. Biol. (2013).

[20] E.H. Lawson, B.L. Hall, C.Y. Ko, Risk factors for superficial vs deep/organ-space surgical site infections: implications for quality improvement initiatives, JAMA Surg. 148 (9) (2013) 849–858.

[21] P.F. Lazarsfeld, N.W. Henry, Latent Structure Analysis, Houghton Mifflin, Boston, 1968.

[22] J. Liu, J. Chen, S. Chen, J. Ye, Learning the optimal neighborhood kernel for classification, in: IJCAI, 2009.

[23] K. Liu, N. Gebraeel, J. Shi, A data-level fusion model for developing composite health indices for degradation modeling and prognostic analysis, IEEE Trans. Autom. Sci. Eng. (2013).

[24] U.M. Mann, E. Mohr, M. Gearing, T.N. Chase, Heterogeneity in alzheimer's disease: progression rate segregated by distinct neuropsychological and cerebral metabolic profiles, J. Neurol. Neurosurg. Psychiat. (1992).

[25] A. McCallum, D. Freitag, F.C. Pereira, Maximum entropy markov models for information extraction and segmentation, in: ICML, 2000.

[26] S. Mueller et al., Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (ADNI), Alzh. Dem. (2005).

[27] J.R. Petrella, R.E. Coleman, P.M. Doraiswamy, Neuroimaging and early diagnosis of alzheimer disease: a look to the future, Radiology 226 (2) (2003) 315–336.

[28] M.B. Priestley, Spectral Analysis and Time Series, Academic Press, 1981.

[29] G. Rabinovici et al., 11c-pib pet imaging in alzheimer disease and frontotemporal lobar degeneration, Neurology (2007).

[30] S. Saria, Dissertation on the Digital Patient: Machine Learning Techniques for Analyzing Electronic Health Record Data, Stanford University, 2011.

[31] L. Saunders, M. Perennec-Olivier, P. Jarno, F. L'Hériteau, A.G. Venier, L. Simon, M. Giard, J.M. Thiolet, J.F. VielRAISIN group, Improving prediction of surgical site infection risk with multilevel modeling, PLoS One 9 (5) (2014) e95295.

[32] U. Thissen, R. Van Brakel, A.P. De Weijer, W.J. Melssen, L.M.C. Buydens, Using support vector machines for time series prediction, Chemomet. Intell. Lab. Syst. (2003).

[33] P. Tseng, S. Yun, A coordinate gradient descent method for nonsmooth separable minimization, Math. Program. (2009).

[34] J. Twisk, T. Hoekstra, Classifying developmental trajectories over time should be done with great caution: a comparison between methods, J. Clin. Epidemiol. (2012).

[35] C. van Walraven, R. Musselman, The surgical site infection risk score (SSIRS): a model to predict the risk of surgical site infections, PLoS One (2013).

[36] X. Wang, D. Sontag, F. Wang, Unsupervised learning of disease progression models, KDD, 2014.

[37] M. Weiner et al., 2014 update of the alzheimer's disease neuroimaging initiative: a review of papers published since its inception, Alzh. Dem. (2012).

[38] M.W. Weiner et al., The alzheimer's disease neuroimaging initiative: a review of papers published since its inception, Alzh. Dem. (2012).

[39] W.A. Woodward, H.L. Gray, A.C. Elliott, Applied Time Series Analysis, CRC Press, 2012.

[40] J. Ye, et al., Heterogeneous data fusion for alzheimer's disease study, KDD, 2008.

[41] L. Yuan, Y. Wang, P.M. Thompson, V.A. Narayan, J. Ye, Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data, Neuroimage (2012).

[42] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, Multimodal classification of alzheimer's disease and mild cognitive impairment, Neuroimage (2011).

[43] J. Zhou et al., Feafiner: biomarker identification from medical data through feature generalization and selection, KDD, 2013.

[44] J. Zhou, J. Liu, V.A. Narayan, J. Ye, Modeling disease progression via fused sparse group lasso, KDD, 2012.

[45] J. Zhou, J. Liu, V.A. Narayan, J. YeAlzheimer's Disease Neuroimaging Initiative, Modeling disease progression via multi-task learning, Neuroimage 78 (2013) 233–248.

[46] J. Zhou, F. Wang, J. Hu, J. Ye, From micro to macro: data driven pheno-typing by densification of longitudinal electronic medical records, KDD, 2014.