Complication Risk Profiling in Diabetes Care: A Bayesian Multi-Task and Feature Relationship Learning Approach

Bin Liu, Ying Li, Soumya Ghosh, Zhaonan Sun, Kenney Ng, and Jianying Hu, *Fellow, IEEE*

Abstract—Diabetes mellitus, commonly known as diabetes, is a chronic disease that often results in multiple complications. Risk prediction of diabetes complications is critical for healthcare professionals to design personalized treatment plans for patients in diabetes care for improved outcomes. In this paper, focusing on Type 2 diabetes mellitus (T2DM), we study the risk of developing complications after the initial T2DM diagnosis from longitudinal patient records. We propose a novel multi-task learning approach to simultaneously model multiple complications where each task corresponds to the risk modeling of one complication. Specifically, the proposed method strategically captures the relationships (1) between the risks of multiple T2DM complications, (2) between different risk factors, and (3) between the risk factor selection patterns, which assumes similar complications have similar contributing risk factors. The method uses coefficient shrinkage to identify an informative subset of risk factors from high-dimensional data, and uses a hierarchical Bayesian framework to allow domain knowledge to be incorporated as priors. The proposed method is favorable for healthcare applications because in addition to improved prediction performance, relationships among the different risks and among risk factors are also identified. Extensive experimental results on a large electronic medical claims database show that the proposed method outperforms state-of-the-art models by a significant margin. Furthermore, we show that the risk associations learned and the risk factors identified lead to meaningful clinical insights.

Index Terms—Healthcare Analytics; Diabetes; Risk Prediction; Multi-task Learning; Correlated Shrinkage

1 Introduction

Diabetes mellitus, commonly known as diabetes, is a chronic disease that affects nearly half a billion people around the globe [1], [2]. In the United States alone, more than 23 million people were diagnosed with diabetes as of 2017, with another 7.2 million undiagnosed [3]. Type 2 diabetes mellitus (T2DM) is the most common form of diabetes, and it accounts for more than 90% of all diabetes cases [3]. T2DM is characterized by hyperglycemia—abnormally elevated blood glucose (blood sugar) levels, and is almost always associated with a number of complications [4]. Over time, the chronic elevation of blood glucose levels caused by T2DM leads to blood vessel damage which in turn leads to associated complications, including kidney failure, blindness, stroke, heart attack, and in severe cases even death. Meanwhile, the cost of diabetes care has been increasing over the past decades and the annual cost reaches $327 billion in US as of 2017 [3], [5], [6]. T2DM management requires continuous medical care with multifactorial risk-reduction strategies beyond glycemic control [7]. Mitigating the risk of complications is of significance for T2DM management. On the one hand, T2DM complications include severe diseases, such as kidney failure and heart attack, and thus require expensive medical procedures. On the other hand, nearly 75% of all diabetes care expenditures are spent on treatment of diabetes complications [5], [8]. Risk profiling of T2DM complications is critical for healthcare professionals to appropriately adapt personalized treatment plans for patients in diabetes care, improving care quality and reducing cost.

The recent abundance of the electronic health records (EHRs) and electronic medical claims data has provided an unprecedented opportunity to apply predictive analytics to improve T2DM management. In this paper, we study the risk profiling of T2DM complications from longitudinal patient medical records: *what is the probability that a patient will develop complications within a time window after the initial T2DM diagnosis?* In the literature, EHRs and claims data have been leveraged for a wide range of healthcare applications including disease onset prediction [9], [10], [11], [12], [13], disease progression [14], [15], patient stratification [16], [17], hospital readmission prediction [18], [19], and mortality prediction [20], [21]. However, there are unique difficulties that arise when performing data-driven risk prediction and profiling of T2DM complications from patient medical records:

- The first challenge stems from the need to effectively capture correlations between multiple T2DM complications. Considering that different complications are the manifestations of a common underlying condition—hyperglycemia, modeling complications as independent of one another leads to suboptimal models.

- *Bin Liu, Ying Li, Zhaonan Sun and Jianying Hu are with the Center for Computational Health, IBM Thomas J. Watson Research Center, 1101 Kitchawan Rd, Yorktown Heights, NY 10598. E-mail: bin.liu1@ibm.com, {liying, zsun, jyhu}@us.ibm.com*
- *Soumya Ghosh and Kenney Ng are with the Center for Computational Health, IBM T.J. Watson Research Center, 75 Binney Street, Cambridge, MA 02142. E-mail: {ghoshso,kenney.ng}@us.ibm.com*

Manuscript received June xx, 2018; revised November xx, 2018.

- Patient medical record data contain rich information about relationships among medical concepts and risk factors, pertinent to T2DM. However, developing statistical methods that can effectively exploit this information is challenging.
- Further, when using patient medical record data for risk prediction and profiling, each patient is typically represented by a high-dimensional feature vector while only a small subset of the predictors are actually relevant. It is essential to be able to identify the subset of predictors that are useful for predictive analysis to facilitate model transparency and interpretability.
- Finally, it is desirable for the model to have the ability to leverage T2DM domain knowledge. Such clinical domain knowledge is often available or partially available, and incorporating it into the analysis can lead to more accurate inferences.

In this paper, we address these challenges by developing methods for simultaneously modeling multiple complications for risk profiling in diabetes care. We begin by formulating T2DM complication risk prediction as a Multi-Task learning (MTL) [22] problem with each complication corresponding to one task. MTL jointly learns multiple tasks using a shared representation so that knowledge obtained from one task can help the other tasks. We then develop a novel MTL model to capture task relationships driven by the underlying disease and the dependencies between information-rich features (risk factors). Further, assuming that similar T2DM complications have similar contributing risk factors, we endow our models with the ability to perform correlated shrinkage through a novel *correlated Horseshoe* distribution. This allows us to identify subsets of risk factors for different complications while accounting for associations between the complications. We call the proposed method **T**ask **RE**lationship and **F**eature relationship **LE**arning with correlated **S**hrinkage (TREFLES). We formulate TREFLES in a hierarchical Bayesian framework, allowing us to easily capture domain knowledge through carefully chosen priors.

Finally, we assess our proposed innovations through extensive experiments on patient level data extracted from a large electronic medical claims database. The results show that the proposed approach consistently outperforms previous models by a significant margin and demonstrate the effectiveness of the simultaneous modeling framework over modeling each complication independently. Furthermore, we show that the risk associations learned and the risk factors identified lead to meaningful clinical insights.

In summary, our key contributions are as follows:

- We provide a systematic study on risk profiling by simultaneously modeling of multiple complications in chronic disease care using T2DM as a case study, although the methodology will generalize to other chronic diseases as well.
- We design a novel model, *TREFLES*, that jointly captures relationships between risks, risk factors, and risk factor selection learned from the data with the ability to incorporate domain knowledge as priors.
- We demonstrate the effectiveness of TREFLES in both predictive capability and clinical interpretability via a comprehensive study of T2DM complications using a large electronic medical claims database.

The proposed method is favorable for healthcare applications beyond diabetes care. It provides a powerful tool for not only improving predictive performance, but also for recovering clinically meaningful insights about relationships among different risks and risk factors.

2 Related Work

2.1 Healthcare Predictive Analytics with Longitudinal Patient Records

From an applications perspective, our work falls into the category of studies that apply predictive analytics and use longitudinal patient records to improve the practice of healthcare management. With abundance of the EHRs and medical claims data, building predictive models from those data has attracted significant attention from both academia and industry. One of the most active research focus is risk prediction, in which EHRs or medical claims are leveraged to predict patients' risks of adverse health events. As such many researches have focused on predicting the onset of different diseases such as heart failure [9], [12], [13], chronic obstructive pulmonary disease [11], [12], and lung disease [23]. Besides disease risk prediction, longitudinal patient records are used to hospital readmission prediction [18], [19], mortality prediction [20], [21], and risk stratification [16], [17]. Beyond direct risk prediction tasks, longitudinal patient records have been applied to study the disease progression [14], [15] of chronic diseases, and to identify patient phenotypes [24], [25], which can facilitate predictive analytics. Deep learning has attracted a lot of attention for healthcare predictive analytics [12], [26], [27], [28]. One major criticism on the black-box deep models for healthcare, as pointed by Caruana *et al.* [29], lies in the difficulties to understand and interpret the models. For this reason, our proposed TREFLES model builds on logistic regression as a baseline to facilitate model transparency and interpretability. Yadav *et al.* [30] presents a comprehensive survey on EHR data mining.

Our work is related to chronic disease prediction and prevention using longitudinal patient records. Prevention and management of chronic diseases are complicated due to their complications and comorbidities. In particular, we focus on diabetes, which is one of the most important and common chronic diseases. Recently, there have been some work on predictive analytics for diabetes and its complications. Razavian *et al.* [10] shows that claims data can be leveraged to predict T2DM onset. EHRs are also used to predict gestational diabetes in early pregnancy [31]. Oh *et al.* [32] applied EHRs to capture the trajectories of T2DM patients and found that different trajectories can lead to different risk patterns. Bertsimas *et al.* [33] exploits electronic medical records for personalized diabetes treatment recommendation. The most relevant work to ours is Liu *et al.* [34], which applies multi-task learning survival analysis to predict the onset of T2DM complications. However the multi-task learning method used in Liu *et al.* [34] only models the correlations between task correlations. Different from previous studies, this paper presents a comprehensive study to investigate the risk prediction and profiling of T2DM

TABLE 1: Comparisons between our proposed TREFLES model and other major MTL approaches for healthcare predictive analytics.

Property	MTFL	MTRL	FETR	**TREFLES**
Task relationships		✓	✓	✓
Feature relationships	✓		✓	✓
Fine-grained medical relationships				✓
Correlated shrinkage				✓

complications from patient medical records for diabetes care through a novel multi-task learning model.

2.2 Multi-task Learning

Our work is also related to multi-task learning (MTL) [22], which aims to jointly learn multiple tasks using a shared representation so that knowledge obtained from one task can help other tasks. Recently, MTL models have been widely used in the healthcare domain such as disease progression [15], mortality prediction in acute care [21], risk stratification [35], elderly care [36] and diabetes [34]. Feature relationship learning based approaches (known as MTFL) [37] and task relationship learning based approaches (known as MTRL) [38] are the two most widely used MTL strategies [39]. MTFL assumes that task association is released through a subset of features shared among tasks. The main idea of MTFL approaches is to learn a few features common across the tasks using different sparsity techniques [37], [40], [41]. MTRL assumes that the task association is revealed in the structure of the coefficient matrix; and the coefficient matrix can be modeled using probabilistic models. A widely used probabilistic model is to assume the coefficient matrix generated from a Matrix Variate Normal (MVN) distribution [38]. One major advantage of MTRL is the existence of an explicit parameter that represents the relatedness between tasks. Zhang *et al.* [39] presents a comprehensive survey on MTL. Most similar to our approach is the feature and task relationship learning (FETR) method recently proposed by Zhao *et al.* [42]. Similar to FETR, our proposed TREFLES model is a generalization of both MTRL and MTFL, and simultaneously learns the relationships both between tasks and between features. In healthcare analytics, correlations between features are important to model. Different from FETR, TREFLES captures more fine-grained feature relationships by grouping features into groups according to domain knowledge. Furthermore, TREFLES is able to capture the correlated coefficient shrinkage among tasks through a novel correlated Horseshoe prior. As we shall show in our study, TREFLES is favorable for healthcare applications where we not only obtain better prediction performances, but also derive clinically meaningful insights about the relationships among the different complications and among the different risk factors.

Table 1 shows the comparisons between our proposed TREFLES model and other major MTL approaches in the context of healthcare predictive analytics.

3 Simultaneous Modeling of Multiple Complications for Risk Profiling

In this section, we first formulate the problem of diabetes complications risk profiling, and then introduce the proposed approach to simultaneously model multiple complications, addressing the aforementioned challenges.

TABLE 2: Mathematical Notations

Symbol	Description
N, M, K	number of subjects, features, and complications
i, j, k	index of subjects, features, and complications
$c_{ki} \in \{0, 1\}$	event of complication k for patient i where 1 indicates an observed event and 0 otherwise
y_{ki}	probability (risk) of patient i for complication k
$\mathbf{x}_i \in \mathbb{R}^M$	vector of features for patient i
$\mathbf{w}_k \in \mathbb{R}^M$	vector of coefficients for complication k
$\mathbf{W} \in \mathbb{R}^{M \times K}$	$\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_K]$ is the matrix of coefficients
$\mathbf{w}^j \in \mathbb{R}^K$	$\mathbf{w}^j$ is the j^{th} row of coefficient complication $\mathbf{W}$
$\mathbf{\Omega} \in \mathbb{R}^{K \times K}$	matrix of relatedness between complications
$\mathbf{\Omega}_0 \in \mathbb{R}^{K \times K}$	matrix of prior knowledge about risk association
$z, \mathcal{G}_z$	index and the z^{th} group of features
$\mathbf{W}_z \in \mathbb{R}^{G_z \times K}$	matrix block where features belongs to group $\mathcal{G}_z$
$\mathbf{\Sigma}_z \in \mathbb{R}^{G_z \times G_z}$	correlation matrix between features in group G_z
λ_{jk}, τ	local and global shrinkage parameter in (correlated) Horseshoe prior for w_{jk}

3.1 Diabetes Complications Risk Profiling

The goal is to build an effective approach to predict the risk of a patient developing complication(s) within a follow-up window Δt after the initial T2DM diagnosis. Specifically, as shown in Fig. 1, patients are included when they are initially diagnosed with T2DM and no complication records are observed before the index date. Following [9], [10], [43], for each patient $i \in \{1, \ldots, N\}$ we aggregate the longitudinal patient records until the patient was initially diagnosed with T2DM into a vector of M features (risk factors). Then each patient is represented as a feature vector $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iM}]^\top$. Let there be K complications in consideration indexed by $k \in \{1, \ldots, K\}$. We use $c_{ki} \in \{0, 1\}$ to represent the onset event of patient i developing complication k in the follow-up window Δt and use y_{ki} to represent the event probability (risk). For each complication k we observe a set of complication observations $\mathcal{D}_k = \{\langle \mathbf{x}_i, c_{ki} \rangle\}_{i \in \mathcal{N}_k}$, where $\mathcal{N}_k$ are the patients observed in complication k. The set of all observed complication events are denoted as $\mathcal{D} = \{\mathcal{D}_k\}_{k=1}^K$. Given $\mathcal{D}$, we aim to build a predictive model $y_{ki} = \Pr(c_{ki} | \Theta, \mathbf{x}_i)$, where Θ are the model parameters, to predict the risk that patient i will develop complication k during follow-up window. Table 2 summarizes useful notations used in this paper.

3.2 Learning Associations between Multiple Complications

Given the features (risk factors) $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iM}]^\top$ observed up until the initial T2DM diagnosis for patient i, we model the risk of patient i developing complication k in the follow-up window Δt as:

$$y_{ki} = \Pr(c_{ki} | \Theta, \mathbf{x}_i) = \sigma(\mathbf{w}_k^\top \mathbf{x}_i), \tag{1}$$

where $\mathbf{w}_k$ is the coefficient vector for complication k, and $\sigma(t)$ is a logistic function $\sigma(t) = \frac{1}{1+e^{-t}}$. Then the event onset can be modeled as a draw from a Bernoulli distribution $c_{ki} \sim \text{Bernoulli}(\sigma(\mathbf{w}_k^\top \mathbf{x}_i))$.

To capture and leverage the association between the risks of the different T2DM complications, we formulate the complication risk prediction problem as a multi-task learning

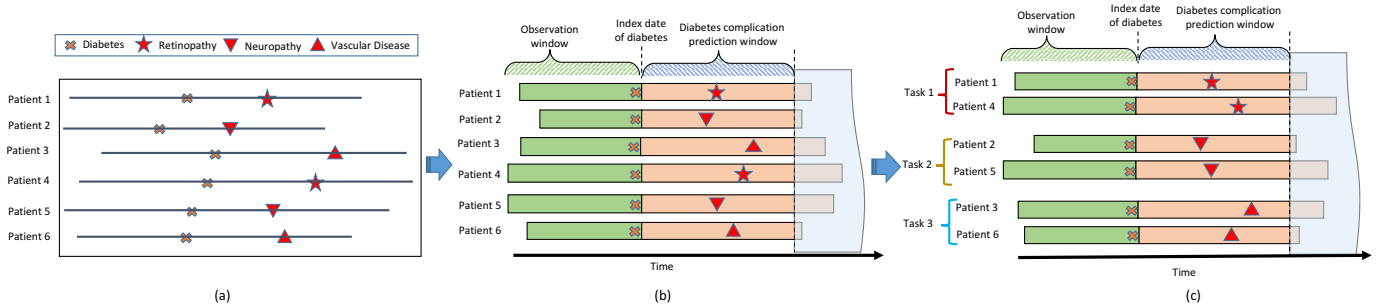

Fig. 1: Illustration of the proposed framework for simultaneous modeling of multiple T2DM complications. (a) Patients are included when they are diagnosed with T2DM (orange cross) for the first time, and no complication records are observed before the index date. (b) For each patient, features (risk factors) are derived from patients' medical records up to the time of the initial T2DM diagnosis. Outcome is evaluated in the follow-up window. (c) Multi-task learning formulation: the predictions of multiple complications in consideration (*e.g.*, retinopathy, neuropathy and vascular disease) are grouped into different tasks where each task models only one complication. Multi-task learning (MTL) is applied to capture the association between the different complications. To simplify the illustration, only positive cases are shown.

problem. As shown in Fig. 1, we group the predictions of multiple complications in consideration (e.g., retinopathy, neuropathy and vascular disease) into different learning tasks. Each task models only one complication risk via Equation (1). Next, we apply multi-task learning to capture the association between different complications.

3.3 Learning Multi-task and Feature Relationships with Correlated Shrinkage

As shown in Fig. 2, we aim to capture three types of dependencies in our multi-task learning framework. First, the complication tasks are related since they all stem from a common underlying condition–hyperglycemia. Second, there are associations between the features since they are derived from and represent the health status of the same set of real patients. Third, similar T2DM complications have similar contributing risk factors that lead to the development of those complications.

3.3.1 Modeling Task and Feature Associations

Let $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K] \in \mathbb{R}^{M \times K}$ denote the matrix of coefficients of all K complications in consideration. To explore the latent association between the risks of T2DM complications, we impose explicit structure on the coefficient matrix $\mathbf{W}$. Specifically, we assume that the coefficient matrix $\mathbf{W}$ follows a Matrix Variate Normal (MVN) distribution[1]:

$$\mathbf{W} \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{\Omega}). \tag{2}$$

The the first term $\mathbf{0}$ is a $M \times K$ matrix of zeros representing the mean of $\mathbf{W}$. The second term $\mathbf{\Sigma}$ is a $M \times M$ symmetric positive definite matrix representing the row-wise covariances of $\mathbf{W}$, *i.e.* the correlations between the features.

1. MVN distribution: the probability density function for the random matrix $\mathbf{X} \in \mathbb{R}^{M \times K}$ that follows the matrix normal distribution with form of $\Pr(\mathbf{X} \mid \mathbf{M}, \mathbf{U}, \mathbf{V}) = \frac{\exp\left(-\frac{1}{2}\operatorname{tr}\left[\mathbf{V}^{-1}(\mathbf{X}-\mathbf{M})^T\mathbf{U}^{-1}(\mathbf{X}-\mathbf{M})\right]\right)}{(2\pi)^{KM/2}|\mathbf{U}|^{K/2}|\mathbf{V}|^{M/2}}$, with mean $\mathbf{M} \in \mathbb{R}^{M \times K}$, row covariance matrix $\mathbf{U} \in \mathbb{R}^{M \times M}_{++}$ and column covariance matrix $\mathbf{V} \in \mathbb{R}^{K \times K}_{++}$. $\mathbb{R}_{++}$ means symmetric positive definite matrix.

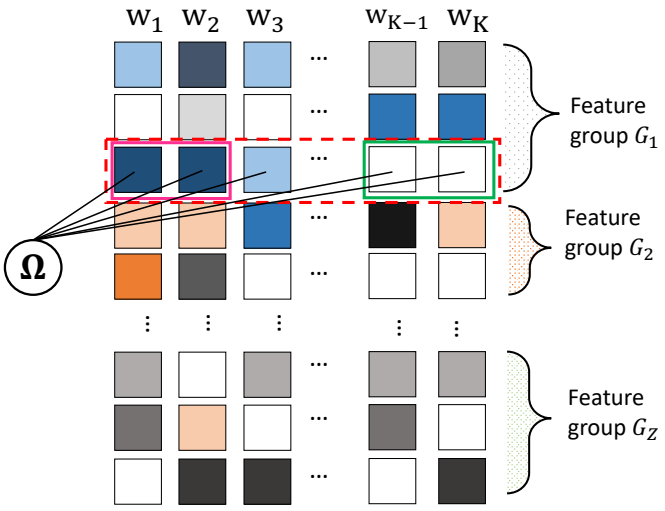

Fig. 2: Illustration of association structure in the coefficient matrix $\mathbf{W}$ of TREFLES model. TREFLES captures the relationships (1) between the risks of multiple T2DM complications (matrix columns), (2) between the different risk factors (matrix rows), and (3) between the risk factor selection patterns, which assumes similar complications have similar contributing risk factors.

The third term $\mathbf{\Omega}$ is a $K \times K$ symmetric positive definite matrix representing the column-wise covariances of $\mathbf{W}$, *i.e.* the correlations between the tasks.

Equation (2) captures both the relationships between tasks through $\mathbf{\Omega}$ and correlations among features through $\mathbf{\Sigma}$. As a result, this formulation is a generalization [42] of the two most widely used MTL strategies: the task relation learning approaches [38], [44] and the feature relationship learning approaches [37], [40]. When $\mathbf{\Sigma}$ is diagonal, we recover task relationship learning, and by setting $\mathbf{\Omega}$ to a diagonal matrix, we recover feature relationship learning.

In healthcare, features can be very fine-grained and domain knowledge is often available to group similar fea-

tures into higher level representations. In this paper, we leverage this knowledge and group the diagnosis features in the patient medical records according to the ontologies of the International Classification of Diseases (ICD) [45]. As a result, we group the features $\{x_j\}_{j=1}^{M}$ into Z groups $\{\mathcal{G}_z\}_{z=1}^{Z}$, where $\mathcal{G}_z$ has G_z features with $\sum_z G_z = M$. Let $\mathbf{w}^j = [w_{j1}, w_{j2}, \ldots, w_{jK}] \in \mathbb{R}^K$ be the j row of coefficient complication matrix $\mathbf{W}$, then $\mathbf{w}^j$ is the j^{th} coefficient across the K tasks. As shown in Fig. 2, we group coefficient matrix $\mathbf{W}$ into Z blocks where each $\mathbf{W}_z \in \mathbb{R}^{G_z \times K}$ is a matrix block where feature j belongs to group $\mathcal{G}_z$, namely, $\mathbf{W}_z = \{\mathbf{w}^j\}_{j \in \mathcal{G}_z}$. We assume $\mathbf{W}_z$ follows a MVN distribution:

$$\mathbf{W}_z \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma}_z, \mathbf{\Omega}) \tag{3}$$

where $\mathbf{0}$ is the mean, $\mathbf{\Sigma}_z$ is the correlations between features, and $\mathbf{\Omega}$ is the correlations between tasks. The zero mean indicates *a-priori* the features are assumed to have no effect. As a result, Equation (3) captures both the relationships between T2DM complications and the relationships between features. Then we have,

$$\Pr(\mathbf{W}_z | \mathbf{0}, \mathbf{\Sigma}_z, \mathbf{\Omega}) = \frac{\exp\left(-\frac{1}{2}\text{tr}\left[\mathbf{\Omega}^{-1}\mathbf{W}_z^\top \mathbf{\Sigma_z}^{-1}\mathbf{W}_z\right]\right)}{(2\pi)^{KG_z/2}|\mathbf{\Sigma_z}|^{K/2}|\mathbf{\Omega}|^{G_z/2}}. \tag{4}$$

3.3.2 Correlated Shrinkage

Patient medical record data are usually high-dimensional with a large numbers of potentially relevant features. We are interested in identifying an informative subset of coefficients, which reflect the contributing risk factors responsible for the development of a specific complication, by shrinking irrelevant coefficients towards zero. Sparsity-promoting priors are widely used in this context. Perhaps, the most popular example is the Laplacian prior which gives rise to the Lasso [46] ℓ_1 regularizer. However, such a prior provides uniform shrinkage — it shrinks values close and far from zero alike. The Horseshoe prior [47] provides an attractive alternative. As shown in Fig. 3, compared with Laplacian prior, the Horseshoe prior maintains an infinitely tall spike at zero, while exhibiting Cauchy-like heavy tails. As a consequence, it shrinks small values to zero more strongly than the Laplace prior, while its heavy tails allow some coefficients to escape completely un-shrunk. This property allows the Horseshoe prior to be more robust to large signals while providing strong shrinkage towards zero to noise. We can place a Horseshoe prior on w_{jk} to promote sparsity on the j^{th} coefficient of task k by setting,

$$\begin{aligned} w_{jk}|\lambda_{jk}, \tau &\sim \mathcal{N}(0, \lambda_{jk}^2\tau^2), \\ \lambda_{jk} &\sim \text{C}^+(0, 1), \\ \tau &\sim \text{C}^+(0, b_0) \end{aligned} \tag{5}$$

where $\text{C}^+(0,1)$ and $\text{C}^+(0, b_0)$ are half-Cauchy distributions, λ_{jk} is called the local shrinkage parameter, τ is the global shrinkage parameter, and b_0 is a global hyperparameter.

However, the vanilla Horseshoe prior fails to capture correlations among tasks. Recall that in our MTL setting, we assume that similar T2DM complications (tasks) should have similar contributing features. Note that $\mathbf{w}^j = [w_{j1}, w_{j2}, \ldots, w_{jk} \ldots, w_{jK}] \in \mathbb{R}^K$ is the j^{th} coefficient across the K tasks. Ideally, pairs of $w_{jk}, k \in \{1, \ldots, K\}$ would have more similar shrinkage if their tasks (k) are positively correlated.

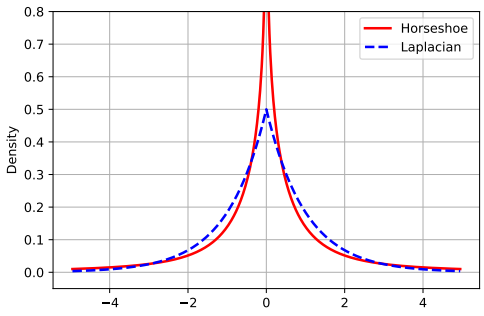

Fig. 3: The Horseshoe prior and Laplacian prior. Compared with Laplacian prior, Horseshoe prior maintains an infinitely tall spike at zero and exhibiting Cauchy-like heavy tails.

To do so, we introduce a novel *correlated Horseshoe* prior.

Definition 1. Correlated Horseshoe. A *correlated Horseshoe* prior is a set of K Horseshoe priors $\boldsymbol{\pi} = [\pi_{\text{HS}}(w_1|\lambda_1, \tau), \pi_{\text{HS}}(w_2|\lambda_2, \tau), \ldots, \pi_{\text{HS}}(w_K|\lambda_K, \tau)]$ on a set of K variables $\mathbf{w} = [w_1, w_2, \ldots, w_K] \in \mathbb{R}^K$ such that: (1) $\boldsymbol{\pi}$ preserves the correlations between the variables in $\mathbf{w}$, and (2) each $\pi_{\text{HS}}(w_k|\lambda_k, \tau)] \in \boldsymbol{\pi}$ itself is a Horseshoe prior.

Specifically, for each risk factor $j \in \{1, \ldots, M\}$ across K tasks, $\mathbf{w}^j = [w_{j1}, w_{j2}, \ldots, w_{jK}]$, we construct the correlated Horseshoe prior by employing a Gaussian copula [48] to couple the local shrinkage parameters λ_{jk} together via the task correlations reflected in $\mathbf{\Omega}$, while forcing the marginals of λ_{jk} to retain their half-Cauchy distributions.

A copula is a multivariate probability distribution with uniform marginal on $(0, 1)$ for each of its variable; it links a set of marginal distributions together to form another joint distribution accordingly to Sklar's theorem [49]. Sklar's theorem states that every K-dimensional multivariate cumulative distribution function $H(v_1, v_2, \ldots, v_K)$ can be expressed with its marginals $\{F_k(v_k)\}_{k=1}^{K}$ and a copula function $C(\cdot)$, namely,

$$H(v_1, v_2, \ldots, v_K) = C\left[F_1(v_1), F_2(v_2), \ldots, F_K(v_K)\right]. \tag{6}$$

Conversely, for any univariate distribution functions $F_k(v_k)$ and copula $C(\cdot)$, the function $H(v_1, v_2, \ldots, v_K)$ defines a K-dimensional distribution function with marginals $F_1(v_1), F_2(v_2), \ldots, F_K(v_K)$.

Sklar's theorem allows us to separate the modeling of the marginal distributions $F_k(v_k)$ from the dependence structure through the copula function $C(\cdot)$. In our case, we need marginals $F_k(v_k)$ to be half-Cauchy and their joint distribution to be a multivariate normal distribution with correlations parameterized with $\mathbf{\Omega}$. We resort to the widely used Gaussian copula [48], [50], which is defined as

$$C\left\{F_k(v_k)\right\} = \Phi_{\mathbf{\Omega}}(\Phi^{-1}(F_1(v_1)), \ldots, \Phi^{-1}(F_K(v_K)), \tag{7}$$

where Φ^{-1} is the inverse of a standard normal distribution and $\Phi_{\mathbf{\Omega}}$ is a zero mean K-dimensional multivariate normal distribution with covariance matrix $\mathbf{\Omega}$. Note that the

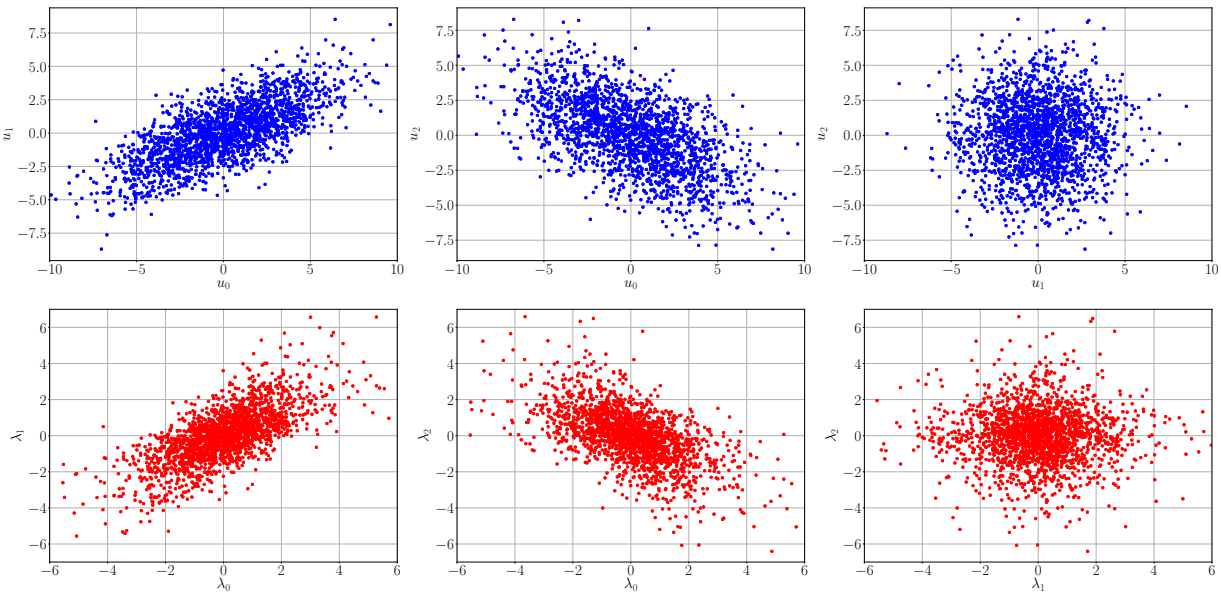

Fig. 4: Illustration of correlated Horseshoe in which the local shrinkage parameters λ_ks (1) retain marginal half-Cauchy distributions, and (2) preserve correlations between the variables. Top: scatter plots of random variables $[u_0, u_1, u_2]$ from a 3-dimensional multivariate normal distribution. Bottom: the corresponding scatter plots of local shrinkage parameters $[\lambda_0, \lambda_1, \lambda_2]$ in log-space.

cumulative distribution $F_k(v_k)$ is uniformly distributed on $(0, 1)$, so $\Phi^{-1}(F_k(v_k))$ corresponds a normal variable. The above process can be reverted. Given the copula function and an instance of the joint distribution, we can get the corresponding random variables with marginals $F_k(v_k)$ and preserves the correlations in the joint distribution.

Let $\mathbf{u}^j = [u_{j1}, u_{j2}, \ldots, u_{jK}] \in \mathbb{R}^K$ be a K-dimensional vector that follows a multivariate normal distribution

$$[u_{j1}, u_{j2}, \ldots, u_{jk} \ldots, u_{jK}] \sim \mathcal{MN}(\mathbf{0}, \mathbf{\Omega}), \tag{8}$$

Observe that $\mathbf{u}^j$ preserves the correlations between tasks through $\mathbf{\Omega}$ and $u_{jk} \sim \mathcal{N}(0, \Omega_{kk})$. Next, we need to ensure that λ_{jk} follows the half-Cauchy distribution. We use inverse transform sampling [51] to guarantee half-Cauchy marginals. Inverse transform sampling is based on the result that given a uniform random variable $a \sim U(0, 1)$, we can generate another random variable b with a cumulative distribution function (cdf) F, by setting $b = \mathrm{F}^{-1}(a)$, as long as F is invertible. Now, if $b \sim \mathrm{C}^+(0, 1)$, then $\mathrm{F}(b) = \frac{2}{\pi}\tan^{-1}(b)$ and since, $\Phi(u_{jk}) \sim U(0, 1)$, where $\Phi(u_{jk})$ is the cdf of u_{jk}, $\mathrm{F}^{-1}(\Phi(u_{jk}))$ follows a half-Cauchy distribution. The correlated Horseshoe is thus completely specified as,

$$\begin{aligned}
&\mathbf{u}^j \sim \mathcal{MN}(\mathbf{0}, \mathbf{\Omega}), \quad \Phi(u_{jk}) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{u_{jk}}{\sqrt{2\Omega_{kk}}}\right)\right], \\
&\lambda_{jk} = \mathrm{F}^{-1}(\Phi(u_{jk})) = \tan\left(\frac{\pi\Phi(u_{jk})}{2}\right) \quad \forall k \in \{1, \ldots, K\}, \\
&w_{jk}|\lambda_{jk}, \tau \sim \mathcal{N}(0, \lambda_{jk}^2\tau^2), \quad \tau \sim \mathrm{C}^+(0, b_0).
\end{aligned} \tag{9}$$

We emphasize that λ_{jk}s are correlated via the latent variables $\mathbf{u}^j$, allowing us to preserve task correlations. At the same time their marginal half-Cauchy behavior retains the desirable properties of the Horseshoe distribution. Fig. 4 shows an example of the correlated Horseshoe, in which $\mathbf{u}$ is sample from zero mean multivariate normal distribution:

$$\mathbf{u} \sim \mathcal{MN}(\mathbf{0}, \mathbf{\Omega}), \quad \mathbf{\Omega} = \begin{bmatrix} 10 & 5 & -5 \\ 5 & 5 & 0 \\ -5 & 0 & 7 \end{bmatrix}.$$

As shown in Fig. 4 (top), u_0 and u_1 are positively correlated, u_0 and u_2 is negatively correlated, and u_1 and u_2 are independent. The local shrinkage parameters λ_ks of the correlated Horseshoe, as shown in Fig. 4 (bottom), can well preserve the correlations between variables in $\mathbf{u}$, from which λ_ks are derived.

3.3.3 *Capturing Domain Knowledge*

In order to utilize available domain knowledge, we impose an Inverse-Wishart prior distribution on $\mathbf{\Omega}$

$$\mathbf{\Omega} \sim \mathcal{IW}(\delta\mathbf{\Omega}_0, \nu). \tag{10}$$

The Inverse-Wishart distribution is a conjugate prior for the multivariate Gaussian distribution. $\mathbf{\Omega}_0$ is a known symmetric positive definite matrix that contains all prior knowledge about the risk associations. δ and ν are two tuning parameters. When domain knowledge on risk associations is available, the prior distribution can leverage the information and help improve the estimation of $\mathbf{\Omega}$. When domain knowledge about risk associations is not available, we can set $\mathbf{\Omega}_0$ to be the identify matrix $\mathbf{I}$.

3.4 Probabilistic Model

Based on above discussion, we summarize our Task RElationship and Feature relationship Learning with correlated Shrinkage (TREFLES) model in a hierarchical Bayesian framework. In TREFLES, $\left\{\{\mathbf{\Sigma}_z\}_{z=1}^Z, \mathbf{\Omega}, \mathbf{U}\right\}$ are latent variables. Summarizing:

1. Task and feature relations

1-a. Prior on risk association $\mathbf{\Omega} \sim \mathcal{IW}(\delta\mathbf{\Omega}_0, \nu)$

1-b. For each feature group $z \in \{1, \ldots, Z\}$:

$$\mathbf{W}_z \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma}_z, \mathbf{\Omega})$$

2. Event of patient i with complication k

$$\text{Draw onset event } c_{ki} \sim \text{Bernoulli}(\sigma(\mathbf{w}_k^\top \mathbf{x}_i))$$

3. Correlated Horseshoe prior

3-a. For each risk factor $j \in \{1, \ldots, M\}$ sample

$$\mathbf{u}^j \sim \mathcal{MN}(\mathbf{0}, \mathbf{\Omega})$$

3-b. For each task $k \in \{1, \ldots, K\}$:

$$\Phi(u_{jk}) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{u_{jk}}{\sqrt{2\Omega_{kk}}}\right)\right],$$
$$\lambda_{jk} = \text{F}^{-1}(\Phi(u_{jk})) = \tan\left(\frac{\pi\Phi(u_{jk})}{2}\right),$$
$$w_{jk}|\lambda_{jk}, \tau \sim \mathcal{N}(0, \lambda_{jk}^2\tau^2), \quad \tau \sim \text{C}^+(0, b_0).$$

3.5 Prediction

Note that in Equation (9), we have $w_{jk}|\lambda_{jk}, \tau \sim \mathcal{N}(0, \lambda_{jk}^2\tau^2)$ and λ_{jk} is a function of u_{jk}, which is sampled from $\mathcal{MN}(\mathbf{0}, \mathbf{\Omega})$. An equivalent non-centered reparameterization is given by $\tau\lambda_{jk} \cdot w_{jk}$, where $w_{jk} \sim \mathcal{N}(0, 1)$. Here, we use this equivalent parameterization for computational convenience. Let $\mathbf{\Lambda} \in \mathbb{R}^{M\times K}$ be a matrix with element λ_{jk}, then we can reparameterize the matrix of coefficients as

$$\boldsymbol{\beta} = \tau\mathbf{\Lambda} \circ \mathbf{W}, \tag{11}$$

where $\circ$ represents a pointwise (Hadamard) product between $\mathbf{\Lambda}$ and $\mathbf{W}$. Finally, we model the risk of complication k for patient i as, $y_{ki} \mid \beta_k, \mathbf{x}_i = \sigma(\boldsymbol{\beta}_k^\top \mathbf{x}_i)$.

4 Parameter Estimation for TREFLES Model

Let $\Theta = \{\{\mathbf{W}_z, \mathbf{\Sigma}_z\}_{z=1}^Z, \mathbf{\Omega}, \mathbf{U}, \tau\}$ denote all parameters to be estimated, and $\Phi = \{\mathbf{\Omega}_0, \delta, \nu\}$ denote all hyperparameters. For each task k we observe a set of complication events $\mathcal{D}_k = \{\langle \mathbf{x}_i, c_{ki}\rangle\}_{i\in\mathcal{N}_k}$, where $\mathcal{N}_k$ represents the patients observed for complication k. The observed complication events are denoted as $\mathcal{D} = \{\mathcal{D}_k\}_{k=1}^K$. Given $\{\mathcal{D}, \Phi\}$ the posterior distribution,

$$\begin{aligned}
&\Pr(\Theta|\mathcal{D}, \Phi) \\
&\propto \Pr(\tau)\Pr(\mathbf{\Omega}) \prod_{k=1}^K \prod_{i=1}^{\mathcal{N}_k} \Pr(c_{ki}|\boldsymbol{\beta}_k, \mathbf{x}_i) \prod_{z=1}^Z \Pr(\mathbf{W}_z) \prod_{j=1}^M \Pr(\mathbf{u}_j) \\
&\propto \frac{2b_0}{\pi(b_0^2+\tau^2)} |\mathbf{\Omega}|^{-\frac{\nu+K+1}{2}} \exp\left(-\frac{\delta}{2}\text{tr}(\mathbf{\Omega}_0\mathbf{\Omega}^{-1})\right) \\
&\times \prod_{k=1}^K \prod_{i=1}^{\mathcal{N}_k} \Pr(c_{ki}|\boldsymbol{\beta}_k, \mathbf{x}_i) \prod_{z=1}^Z \frac{\exp\left(-\frac{1}{2}\text{tr}\left[\mathbf{\Omega}^{-1}\mathbf{W}_z^\top\mathbf{\Sigma}_z^{-1}\mathbf{W}_z\right]\right)}{(2\pi)^{KG_z/2}|\mathbf{\Sigma}_z|^{K/2}|\mathbf{\Omega}|^{G_z/2}} \\
&\times \prod_{j=1}^M \frac{\exp\left(-\frac{1}{2}\mathbf{u}^j\mathbf{\Omega}^{-1}(\mathbf{u}^j)^\top\right)}{(2\pi)^{K/2}|\mathbf{\Omega}|^{1/2}}.
\end{aligned} \tag{12}$$

We estimate the parameters via maximizing the log posterior $\ell(\Theta) = \log\Pr(\Theta|\mathcal{D}, \Phi)$, then we have

$$\begin{aligned}
&\ell(\Theta) = \\
&\sum_{k=1}^K \sum_{i=1}^{\mathcal{N}_k} \Big\{ c_{ki}\log\sigma(\boldsymbol{\beta}_k^\top\mathbf{x}_i) + (1-c_{ki})\log(1-\sigma(\boldsymbol{\beta}_k^\top\mathbf{x}_i)) \Big\} \\
&+ \sum_{z=1}^Z \Big\{ -\frac{1}{2}\text{tr}\left[\mathbf{\Omega}^{-1}\mathbf{W}_z^\top\mathbf{\Sigma}_z^{-1}\mathbf{W}_z\right] - \frac{K}{2}\log|\mathbf{\Sigma}_z| \\
&- \frac{G_z}{2}\log|\mathbf{\Omega}| \Big\} + \sum_{j=1}^M \Big\{ -\frac{1}{2}\mathbf{u}_j^\top\mathbf{\Omega}^{-1}\mathbf{u}^j - \frac{1}{2}\log|\mathbf{\Omega}| \Big\} \\
&- 2\log(b_0^2+\tau^2) - \frac{\nu+K+1}{2}\log|\mathbf{\Omega}| - \frac{1}{2}\text{tr}(\delta\mathbf{\Omega}_0\mathbf{\Omega}^{-1})
\end{aligned} \tag{13}$$

Objective Function. We rewrite the negative log-posterior $\ell(\Theta)$ to obtain the following objective function $\mathcal{O}(\Theta)$ to minimize:

$$\begin{aligned}
&\mathcal{O}(\Theta) = \\
&\sum_{k=1}^K \sum_{i=1}^{\mathcal{N}_k} \Big\{ -c_{ki}\log\sigma(\boldsymbol{\beta}_k^\top\mathbf{x}_i) - (1-c_{ki})\log(1-\sigma(\boldsymbol{\beta}_k^\top\mathbf{x}_i)) \Big\} \\
&+ \frac{1}{2}\sum_{z=1}^Z \Big\{ \text{tr}\left[\mathbf{\Omega}^{-1}\mathbf{W}_z^\top\mathbf{\Sigma}_z^{-1}\mathbf{W}_z\right] \Big\} + \frac{K}{2}\sum_{z=1}^Z \log|\mathbf{\Sigma}_z| \\
&+ \frac{\xi}{2}\log|\mathbf{\Omega}| + \frac{\delta}{2}\text{tr}(\mathbf{\Omega}_0\mathbf{\Omega}^{-1}) + \frac{1}{2}\sum_{j=1}^M \mathbf{u}^j\mathbf{\Omega}^{-1}(\mathbf{u}_j)^\top \\
&+ 2\log(b_0^2+\tau^2) \\
&\text{s.t. } \mathbf{\Omega} \succeq 0, \mathbf{\Sigma}_z \succeq 0.
\end{aligned} \tag{14}$$

where $\mathbf{X} \succeq 0$ means that the matrix $\mathbf{X}$ is positive semidefinite, and $\xi = 2M + K + \nu + 1$.

Solving the above optimization problem is non-trivial. The optimization problem is not convex since $\log|\mathbf{\Omega}|$ and $\log|\mathbf{\Sigma}_z|$ are concave functions. Therefore we adopt an iterative algorithm to solve the problem [52]. Within each iteration, the blocks $\mathbf{W}_z$, $\mathbf{\Sigma}_z$, $\mathbf{\Omega}$, $\mathbf{U}$, and τ are updated alternatively.

Update $\mathbf{W}_z$ given others: With other parameters fixed, objective function w.r.t $\mathbf{W}_z$ becomes

$$\begin{aligned}
\underset{\{\mathbf{W}_z\}_{z=1}^Z}{\arg\min} \; &\sum_{k=1}^K \sum_{i=1}^{\mathcal{N}_k} \Big\{ -c_{ki}\log\sigma(\boldsymbol{\beta}_k^\top\mathbf{x}_i) \\
&- (1-c_{ki})\log(1-\sigma(\boldsymbol{\beta}_k^\top\mathbf{x}_i)) \Big\} \\
&+ \sum_{z=1}^Z \left\{ \frac{1}{2}\text{tr}\left[\mathbf{\Omega}^{-1}\mathbf{W}_z^\top\mathbf{\Sigma}_z^{-1}\mathbf{W}_z\right] \right\}.
\end{aligned}$$

This is a convex optimization problem with respect to $\mathbf{W}_z$. We use stochastic gradient descent method to update the $\{\mathbf{W}_z\}_{z=1}^Z$. Stochastic gradient descent has been widely used for many machine learning tasks [53]. The main process involves randomly scanning training instances and iteratively updating parameters. In each iteration, we randomly sample an instance $\langle x_i, c_{ki}\rangle$, and we minimize $\mathcal{O}(\Theta)$ using the update rule for $\Theta = \Theta - \epsilon \cdot \frac{\partial\mathcal{O}(\Theta)}{\partial\Theta}$, where ϵ is a learning rate. Note that $\mathbf{w}_k = [\mathbf{w}_{\mathcal{G}_1}, \mathbf{w}_{\mathcal{G}_2}, \ldots, \mathbf{w}_{\mathcal{G}_Z}]^\top$ and

$\mathbf{W}_z = \{\mathbf{w}^j\}_{j\in\mathcal{G}_z}$. Let $\mathbf{w}_k^{\mathcal{G}_z} = [w_{jk}, w_{jk}, \cdots, w_{jk}]^\top, j \in \mathcal{G}_z$ be the k column of $\mathbf{W}_z$, then $\mathbf{w}_k^{\mathcal{G}_z}$ corresponds to the coefficients of features in group $\mathcal{G}_z$ in task k. Given an instance $\langle \mathbf{x}_i, c_{ki}\rangle$, the gradient with respect to $\mathbf{w}_k^{\mathcal{G}_z}$ is

$$\frac{\partial \mathcal{O}}{\partial \mathbf{w}_k^{\mathcal{G}_z}} = -\Big(c_{ki} - \sigma(\boldsymbol{\beta}_k^\top \mathbf{x}_i)\Big)\frac{\partial \boldsymbol{\beta}_k^\top \mathbf{x}_i}{\partial \mathbf{x}_i^{\mathcal{G}_z}} + \big[\boldsymbol{\Sigma}_z^{-1}\mathbf{W}_z\boldsymbol{\Omega}^{-1}\big]_k^{\mathcal{G}_z} \tag{15}$$

where $\mathbf{x}_i^{\mathcal{G}_z}$ is the features in group z, and $[\mathbf{X}]_k$ means the k-th column of matrix $\mathbf{X}$. So we have $\frac{\partial \boldsymbol{\beta}_k^\top \mathbf{x}_i}{\partial \mathbf{x}_i^{\mathcal{G}_z}} = \tau \boldsymbol{\lambda}_k^{\mathcal{G}_z} \circ \mathbf{x}_i^{\mathcal{G}_z}$.

Update $\mathbf{U}$ given others: With other parameters fixed, the objective function w.r.t $\mathbf{U}$ becomes

$$\arg\min_{\mathbf{U}} \sum_{k=1}^{K}\sum_{i=1}^{\mathcal{N}_k}\Big\{ - c_{ki}\log\sigma(\boldsymbol{\beta}_k^\top \mathbf{x}_i) - (1-c_{ki})\log(1-\sigma(\boldsymbol{\beta}_k^\top \mathbf{x}_i))\Big\} + \frac{1}{2}\sum_{j=1}^{M}\mathbf{u}^j\boldsymbol{\Omega}^{-1}(\mathbf{u}^j)^\top$$

To apply SGD, we optimize columns $\mathbf{u}_k$ instead of rows $\mathbf{u}^j$. Note than $\sum_{j=1}^{M}\mathbf{u}^j\boldsymbol{\Omega}^{-1}(\mathbf{u}^j)^\top = \mathrm{tr}(\mathbf{U}\boldsymbol{\Omega}^{-1}\mathbf{U}^\top)$. Given an instance $\langle \mathbf{x}_i, c_{ki}\rangle$, the gradient with respect to $\mathbf{u}_k$ is

$$\frac{\partial \mathcal{O}}{\partial \mathbf{u}_k} = -\Big(c_{ki} - \sigma(\boldsymbol{\beta}_k^\top \mathbf{x}_i)\Big)\frac{\partial \boldsymbol{\beta}_k^\top \mathbf{x}_i}{\partial \mathbf{u}_k} + \big[\mathbf{U}\boldsymbol{\Omega}^{-1}\big]_k \tag{16}$$

Note that $\beta_{jk} = \tau\lambda_{jk}w_{jk}$ and λ_{jk} is a function of u_{jk} with $\lambda_{jk} = \tan\left(\frac{\pi\Phi(u_{jk})}{2}\right), \Phi(u_{jk}) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{u_{jk}}{\sqrt{2\Omega_{kk}}}\right)\right]$. Then we have $\frac{\partial \boldsymbol{\beta}_k^\top \mathbf{x}_i}{\partial \mathbf{u}_k} = \tau\frac{\partial f(\mathbf{u}_k)}{\partial \mathbf{u}_k} \circ \mathbf{x}_i$, where $\left.\frac{\partial f(\mathbf{u}_k)}{\partial \mathbf{u}_k}\right|_{jk} = \frac{\pi}{2}\sec^2\left(\frac{\pi\Phi(u_{jk})}{2}\right)\frac{1}{\sqrt{2\pi\Omega_{kk}^2}}\exp\left(-\frac{u_{jk}^2}{2\Omega_{kk}^2}\right)$.

Update τ given others: With other parameters fixed, the objective function w.r.t τ becomes

$$\arg\min_{\tau} \sum_{k=1}^{K}\sum_{i=1}^{\mathcal{N}_k}\Big\{ - c_{ki}\log\sigma(\boldsymbol{\beta}_k^\top \mathbf{x}_i) - (1-c_{ki})\log(1-\sigma(\boldsymbol{\beta}_k^\top \mathbf{x}_i))\Big\} + 2\log(b_0^2 + \tau^2)$$

The gradients with respect to τ are given by

$$\frac{\partial \mathcal{O}}{\tau} = -\sum_{k=1}^{K}\sum_{i=1}^{\mathcal{N}_k}\Big(c_{ki} - \sigma(\boldsymbol{\beta}_k^\top \mathbf{x}_i)\Big)\boldsymbol{\lambda}_k^\top \mathbf{x}_i + \frac{4\tau}{b_0^2 + \tau^2} \tag{17}$$

where $\boldsymbol{\lambda}_k$ is the k-column of $\boldsymbol{\Lambda}$. This allows τ to be updated using gradient decent.

Update $\boldsymbol{\Omega}$ given others: With other parameters fixed, the objective function w.r.t $\boldsymbol{\Omega}$ becomes

$$\arg\min_{\boldsymbol{\Omega}} \sum_{z=1}^{Z}\left\{\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}^{-1}\mathbf{W}_z^\top\boldsymbol{\Sigma}_z^{-1}\mathbf{W}_z\right]\right\} + \frac{\delta}{2}\mathrm{tr}(\boldsymbol{\Omega}_0\boldsymbol{\Omega}^{-1}) + \frac{\xi}{2}\log|\boldsymbol{\Omega}|, \tag{18}$$

The last term $\log|\boldsymbol{\Omega}|$ can be seen as a penalty on the complexity of $\boldsymbol{\Omega}$, and can be replaced with the constraint $\mathrm{tr}(\boldsymbol{\Omega}) = 1$ [38]. Then above Equation (18) can be reformulated as:

$$\arg\min_{\boldsymbol{\Omega}} \sum_{z=1}^{Z}\left\{\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}^{-1}\mathbf{W}_z^\top\boldsymbol{\Sigma}_z^{-1}\mathbf{W}_z\right]\right\} + \frac{\delta}{2}\mathrm{tr}(\boldsymbol{\Omega}_0\boldsymbol{\Omega}^{-1}) \quad \text{s.t. } \boldsymbol{\Omega} \succeq 0,\ \mathrm{tr}(\boldsymbol{\Omega}) = 1 \tag{19}$$

where $\boldsymbol{\Omega} \succeq 0$ means that the matrix $\boldsymbol{\Omega}$ is positive semidefinite. Equation (19) has an analytical solution:

$$\boldsymbol{\Omega} = \frac{\left(\frac{1}{2}\sum_{z=1}^{Z}\mathbf{W}_z^\top\boldsymbol{\Sigma}_z^{-1}\mathbf{W}_z + \frac{\delta}{2}\boldsymbol{\Omega}_0\right)^{\frac{1}{2}}}{\mathrm{tr}\left[\left(\frac{1}{2}\sum_{z=1}^{Z}\mathbf{W}_z^\top\boldsymbol{\Sigma}_z^{-1}\mathbf{W}_z + \frac{\delta}{2}\boldsymbol{\Omega}_0\right)^{\frac{1}{2}}\right]}. \tag{20}$$

Update $\boldsymbol{\Sigma}_z$ given others: With other parameters fixed, the objective function w.r.t $\boldsymbol{\Sigma}_z$ becomes

$$\arg\min_{\boldsymbol{\Sigma}_z} \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}^{-1}\mathbf{W}_z^\top\boldsymbol{\Sigma}_z^{-1}\mathbf{W}_z\right] + \frac{K}{2}\log|\boldsymbol{\Sigma}_z|. \tag{21}$$

Similar to the case of updating $\boldsymbol{\Omega}$, the last term $\log|\boldsymbol{\Sigma}_z|$ in Equation (21) can be seen as a penalty on the complexity of $\boldsymbol{\Sigma}_z$, and can be replaced with a constraint $\mathrm{tr}(\boldsymbol{\Sigma}_z) = 1$. Then above Equation (21) can be reformulated as:

$$\arg\min_{\boldsymbol{\Sigma}_z} \mathrm{tr}\left[\boldsymbol{\Omega}^{-1}\mathbf{W}_z^\top\boldsymbol{\Sigma}_z^{-1}\mathbf{W}_z\right] \quad \text{s.t. } \boldsymbol{\Sigma}_z \succeq 0,\ \mathrm{tr}(\boldsymbol{\Sigma}_z) = 1. \tag{22}$$

The Equation (22) has an analytical solution:

$$\boldsymbol{\Sigma}_z = \left(\mathbf{W}_z\boldsymbol{\Omega}^{-1}\mathbf{W}_z^\top\right)^{\frac{1}{2}} \Big/ \mathrm{tr}\left[\left(\mathbf{W}_z\boldsymbol{\Omega}^{-1}\mathbf{W}_z^\top\right)^{\frac{1}{2}}\right]. \tag{23}$$

Above iterative algorithm uses a block coordinate update strategy and non-convex relaxations when necessary (Equations (19) and (22)), and it cannot guarantee to converge to the global optimal [54].

5 Experiments

In this section we present empirical evaluations to carefully vet our model on patient level data extracted from a large real-world electronic medical claims database.

5.1 Experimental Setup

T2DM cohort construction. We conduct a retrospective cohort study using the MarketScan Commercial Claims and Encounter (CCAE) database from Truven Health [55]. The data on the patients are contributed by a selection of large private employers' health plans, as well as government and public organizations. We use a dataset of de-identified patients between the years 2011 and 2014. The patient cohort used in the study consisted of T2DM patients selected based on the following criteria:

I. The frequency ratio between Type 2 diabetes visits to Type 1 diabetes visits is larger than 0.5; AND

II-a. The patient has two (2) or more Type 2 diabetes labeled events on different days; OR

II-b. The patient received insulin and/or antidiabetic medication.

TABLE 3: List of the T2DM complications and the number of subjects included in this study.

Complication	Description	Example ICD-9 codes	# Subjects
Retinopathy (RET)	Eye disease caused by damage to the blood vessels in the tissue at the back of the eye (retina)	25050, 25052, 24950, 24951, 36201-36207	7552
Neuropathy (NEU)	Nerve damage most often affecting the legs and feet	25060, 25062, 24960, 24961	11151
Nephropathy (NEP)	Kidney disease or damage	25040, 25042, 24940, 24941	3969
Vascular Disease (VAS)	Vascular diseases including peripheral vascular disease, cardiovascular disease, and cerebrovascular diseases	25070, 25072, 24970, 24971	6735
Cellulitis (CEL)	Serious bacterial skin infection	37313, 37531, 38010-38016	11148
Pyelonephritis (PYE)	Inflammation of the kidney, typically due to a bacterial infection	5900 - 5909	609
Osteomyelitis (OST)	Inflammation or infection of the bone and bone marrow; common in patients with diabetic foot problems	73000-73007, 73009-73017, 73019-73027, 73029	909
Renal (REN)	Renal failure	28521, 585, 5854-5856, 586, 5845-5849	5172
Hyperosmolar state (HHS)	One of two serious metabolic derangements characterized by hyperglycemia, hyperosmolarity, and dehydration without significant ketoacidosis	25020, 25022, 24920, 24921	1077
Ketoacidosis (KET)	A complex disordered metabolic state characterized by hyperglycemia, ketoacidosis, and ketonuria	25010, 25012, 24910, 24911	1617
Sepsis (SEP)	Immune response triggered by an infection that causes injury to the body's own tissues and organs	0380-0389	2559
Shock (SHK)	A critical condition brought on by a sudden drop in blood flow through the body	78550, 78551, 78552, 78559	777

Further, patients who were under 19 years or age or over 64 years or age at the initial T2DM diagnosis are removed.

Study inclusion criteria. We focus on the risk of developing complications in the two year time window immediately following the initial T2DM diagnosis. Following inclusion criteria of predictive models using observational data [10], [43], we selected patients with at least two years of observations before the initial T2DM diagnosis, and no complication records are observed before the index date. Also positive and negative instances are selected with long enough observations in the follow-up time window. Guided by clinical experts and following rules from American Diabetes Association [56], we identified 17 common complications of T2DM. We used 12 of the 17 T2DM complications because the remaining 5 complications did not have enough instances in our dataset. Table 3 shows the complications selected in this study and the corresponding number of patients.

Prediction variables. We use following prediction variables:

- **Patient demographics:** age and gender.
- **Diagnoses:** historical medical conditions encoded as International Classification of Disease (ICD) codes. ICD codes are grouped according to their first three digits and ICD codes appearing in fewer than 200 patients are filtered out. This results in 296 unique ICD features. Patients with less than 10 occurrences of ICD codes are removed.
- **Medications:** medications that were received before the initial T2DM diagnosis date. A total of 19 therapeutic classes related to glucose control, cardiac related drugs, and antibiotics were selected.

This results in a total of 317 features.

5.2 Evaluation Protocol

Baselines. We compare the new **TREFLES** method with following set of strong baselines:

- Single task learning (**STL**): For each task, we use a logistic regression to model the risk of each complication independently.
- Multi-task feature learning (**MTFL**) [37], [40]: MTFL assumes that task association is captured through a subset of features shared among tasks. It learns a few features common across the tasks using group sparsity, *i.e.*, the ℓ_1/ℓ_2-norm regularization on $\mathbf{W}$, which both couples the tasks and enforces sparsity.
- Multi-task relationship learning (**MTRL**) [38]: MTRL assumes that the task association is revealed in the structure of the coefficient matrix $\mathbf{W}$, but it only considers the task correlations in $\mathbf{W}$ neglecting the correlations between features.
- Feature and task relationship learning (**FETR**) [42]: FETR learns the relationships both between tasks and between features simultaneously. It can be seen a special case of our model without feature grouping and correlated shrinkage.

Evaluation metrics. We evaluate the models using AUC (area under the receiver operating characteristic curve). AUC is a standard metric in predictive analytics, it measures how the true positive rate (sensitivity) varies with false positive rate (false alarm).

Training and testing. We used 5-fold cross validation to report results for each model. All the models are implemented with gradient descent optimization and we apply the Adam [57] method to automatically adapt the step size during parameter estimation.

5.3 Incorporating Domain Knowledge

Grouping of features. We group ICD features according to the domain knowledge encoded in the ICD ontologies. Specifically, we group ICD-9 codes together when they have a same parent node (3 digits) in the ICD-9 hierarchy.

TABLE 4: Performance comparisons between the proposed TREFLES model and the baseline approaches for the 12 complications in terms of AUC values. The AUC average and standard deviation (in parenthesis) over the 5-fold cross validation trials are reported. We conducted the Wilcoxon signed rank test for the proposed TREFLES model with each baseline model to perform significance tests. ** (*) indicates that AUC value of TREFLES model is statistically significant different from the corresponding baseline with $p < 0.05$ ($p < 0.1$).

Method	RET	NEU	NEP	VAS	CEL	PYE	OST	REN	HHS	KET	SEP	SHK
STL	0.5397** (0.0108)	0.5889** (0.0092)	0.5905** (0.0096)	0.6581** (0.0096)	0.5983** (0.0049)	0.6222** (0.0263)	0.7574** (0.0468)	0.7351** (0.0110)	0.6186** (0.0323)	0.6558** (0.0240)	0.7611** (0.0152)	0.7794** (0.0410)
MTFL	0.5487** (0.0073)	0.6034** (0.0134)	0.6340** (0.0086)	0.7059** (0.0085)	0.6047** (0.0077)	0.5604** (0.0687)	0.8094** (0.0565)	0.7801** (0.0078)	0.6794** (0.0311)	0.7011** (0.0335)	0.7962** (0.0099)	0.8316** (0.0292)
MTRL	0.5643** (0.0087)	0.6100** (0.0103)	0.6456** (0.0099)	0.7069** (0.0105)	0.6283** (0.0046)	0.6909 (0.0633)	0.8480** (0.0534)	0.7933** (0.0073)	0.6990** (0.0187)	0.7347** (0.0348)	0.8182** (0.0178)	0.8679* (0.0209)
FETR	0.5815** (0.0178)	0.6488** (0.0063)	0.6336** (0.0126)	0.7290** (0.0137)	0.6589** (0.0067)	0.6913 (0.0474)	0.8610** (0.0506)	0.8087** (0.0163)	0.6878 (0.0304)	0.7320** (0.0416)	0.8298** (0.0140)	0.8709* (0.0262)
TREFLES	**0.5985** (0.0150)	**0.6697** (0.0075)	**0.6655** (0.0130)	**0.7478** (0.0091)	**0.6793** (0.0074)	**0.7194** (0.0422)	**0.8828** (0.0571)	**0.8316** (0.0130)	**0.7229** (0.0242)	**0.7626** (0.0341)	**0.8425** (0.0165)	**0.8784** (0.0247)

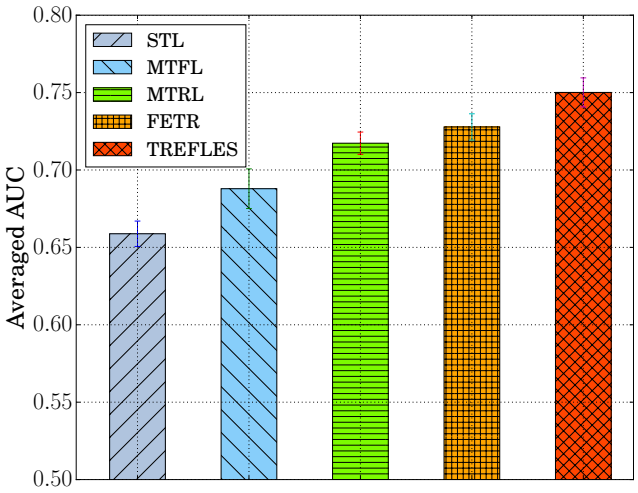

Fig. 5: Performance comparisons between the proposed TREFLES model and the baseline approaches in terms of AUC (averaged over all 12 tasks).

Prior risk association $\mathbf{\Omega}_0$. Note that our model can incorporate prior knowledge on complication associations through $\mathbf{\Omega}_0$. We construct prior associations using the human disease network [58], which provides the Phi-correlations between pairs of diseases. We aggregate the Phi-correlations between pairs of ICD codes under pairs of T2DM complications. This results in a $\mathbf{\Omega}_0$ that represents our prior knowledge about the correlations between the T2DM complications in our study.

5.4 Results

Table 4 shows the AUCs between the proposed TREFLES model and the baseline approaches on all 12 complication risk prediction tasks. The average and standard deviation (in parenthesis) over the 5-fold cross validation trials are reported. We also conducted the Wilcoxon signed rank test for the proposed TREFLES model with each baseline model to perform significance tests. ** (*) indicates that AUC value of TREFLES model is statistically significant different from the corresponding baseline with $p < 0.05$ ($p < 0.1$). Our approach consistently and significantly (in most cases) outperforms the baseline methods on all the 12 tasks. Fig. 5 plots the average AUCs and standard deviations across the 12 tasks for the different methods.

MTL versus STL: We observe that all multi-task learning models (MTFL, MTRL, FETR and TREFLES) consistently and significantly outperform the single task learning method. In particular, our TREFLES model outperforms the single task learning method by 9.1% in AUC on average. This confirms our assumption that directly modeling complications as independent of one another can lead to suboptimal models. Note that the different complications are manifestations of a common underlying condition–hyperglycemia, so their risks should be related. By simultaneously modeling multiple complications, MTL can capture and leverage the associations between complications using a shared representation. As a result, MTL models can significantly outperform STL models in risk prediction of T2DM complications.

TREFLES model versus baseline MTL models: As shown in Fig. 5, our TREFLES model outperforms all baseline MTL models. TREFLES (AUC 0.7501 ± 0.0091) is better than the best baseline model FETR (AUC 0.7278 ± 0.0094) by 2.2% in AUC. We also observe that the task relationship learning based method MTRL (AUC 0.7173 ± 0.0072) is more favorable than the feature relationship learning based method MTFL (AUC 0.6879 ± 0.0128). FETR outperforms MTRL because it simultaneously learns the relationships both between tasks and between features. TREFLES not only captures the relationships between tasks and between features, it also identifies the common contributing risk factors through the correlated coefficient shrinkage mechanism and incorporates domain knowledge through carefully constructed priors. As a result, TREFLES can significantly improve upon FETR.

5.5 Learned Risk Associations

In this section we discuss the estimated risk association matrix $\hat{\mathbf{\Omega}}$ from our TREFLES model. Matrix $\hat{\mathbf{\Omega}}$ represents the relatedness between complications learned from data. We first transfer the covariance matrix $\hat{\mathbf{\Omega}}$ to its correlation matrix $\hat{\mathbf{R}}$, whose elements have a ranges from -1 to 1. We

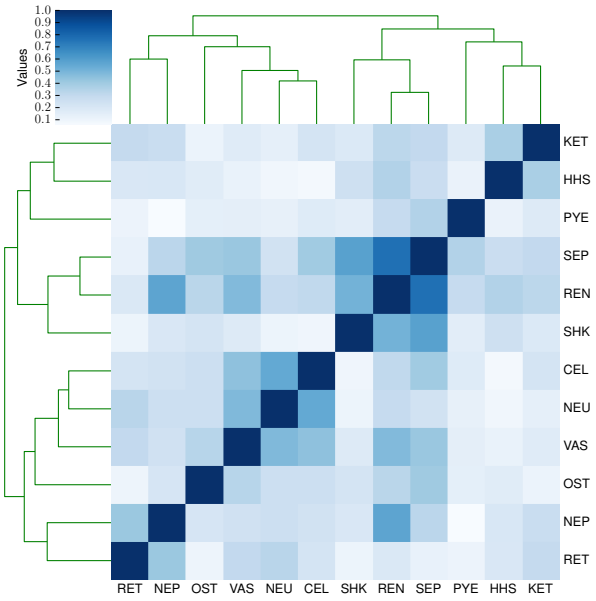

Fig. 6: Heatmap and dendrogram of the hierarchical clustering of the correlation matrix learned by TREFLES.

observe that all the elements in the correlation matrix $\hat{\mathbf{R}}$ learned by TREFLES have positive values. This is because all the complications are manifestations of a common underlying condition–hyperglycemia and they are positively correlated. Then we perform a hierarchical clustering on $\hat{\mathbf{R}}$. Fig. 6 shows the heatmap and the dendrogram of the hierarchical clustering results. Darker colors indicate higher correlation. We can observe clusters between the risk associations of the 12 complications. In particular, CEL, NEU, VAS, OST, NEP and RET form one cluster while the remaining complications of KET, HHS, PYE, SEP, REN and SHK form a second cluster.

The clusters are clinically meaningful. The first cluster of CEL, NEU, VAS, OST, NEP and RET represents the local complications caused by long standing or mismanaged diabetes, and the second cluster of KET, HHS, PYE, SEP, REN and SHK represents complications involving multiple sites or systemic complications. Cluster 2 indicates more severe pathophysiologic manifestations of the disease than the cluster 1.

5.6 Identified Risk Factors

Table 5 shows the top-5 risk factors/predictors (according to their coefficients) for each diabetic complication identified by our model. Most of the risk factors identified by our model are known to be clinically associated with the corresponding diabetic complications (indicated by *) [59]. For example, the medical condition of "Disorders of fluid, electrolyte, and acid-base balance", which consistently appears in the top listing for all the diabetic complications, is indicative of many acid-based and electrolyte disorders that may be due to complications of T2DM and the medications diabetic patients receive [60], [61]. Age is another major known risk factor for retinopathy, neuropathy, nephropathy and vascular disease including cardiovascular disease and the proposed method correctly identifies these associations [62].

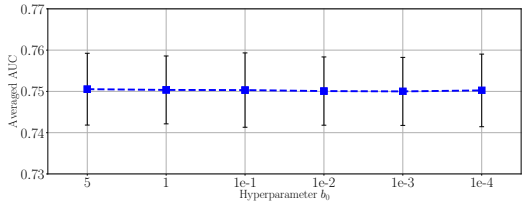

Fig. 7: Impacts of hyperparameter b_0 of TREFLES model on prediction performances.

The underlying mechanism of age as a risk factor could be due to the fact that older adults tend to have long-standing diabetes, and consequently have associated microvascular and macrovascular complications. Insulin treatment is identified as a risk factor for retinopathy, nephropathy, and cellulitis but not for the other complications [63].

5.7 Impacts of hyperparameter b_0 and prior risk association $\mathbf{\Omega}_0$

Fig. 7 shows the prediction results in terms of averaged AUC over all tasks for different values of b_0, which is the hyperparameter for global shrinkage parameter τ. We found TREFLES model is not sensitive to hyperparameter b_0.

Incorporating prior risk association $\mathbf{\Omega}_0$ improves prediction performances—AUC with $\mathbf{\Omega}_0$ is 0.7501 ± 0.0091 vs AUC 0.7495 ± 0.0093 when without $\mathbf{\Omega}_0$—though improvement is not significant. One reason would be the population used in the human disease network study and that used in our study are different. Also, the impact of priors will become negligible when there are enough training data. However, the capability to incorporate domain knowledge becomes important when there are not enough data and reliable domain knowledge is available.

6 Discussion and Conclusion

In this paper, we provided a systematic study on risk profiling by simultaneously modeling multiple complications in chronic disease care using T2DM as a case study. We proposed a novel multi-task learning model, *TREFLES*, that jointly captures relationships between risks, risk factors, and risk factor selection learned from the data with the ability to incorporate domain knowledge as priors. TREFLES is favorable for healthcare applications because in additional to improved prediction performance, clinically meaningful insights about the relationships among different complications and risk factors can also be identified. Extensive experiments on a T2DM patient dataset extracted from a large electronic medical claims database validated the improved prediction performance of TREFLES over current state of the art methods. Also the risk associations learned as well as the risk factors identified by TREFLES lead to meaningful insights that were consistent with clinical findings.

Limitations and Future Research. There are a number of limitations in this work and interesting future research directions. First, we aggregated longitudinal patient records into a vector of risk factor, and each patient was represented by the vector. Such data aggregation neglects the temporal information in the longitudinal data. An interesting future

TABLE 5: Top-5 risk factors (with the highest coefficients) for each complication as identified by the TREFLES model.

Retinopathy (RET)	Neuropathy (NEU)	Nephropathy (NEP)
1.79 Antidiabetic Agents, Insulin*	4.07 Hereditary and idiopathic peripheral neuropathy	1.98 Disorders of fluid, electrolyte, and acid-base balance*
1.42 Disorders of fluid, electrolyte, and acid-base balance*	4.03 Inflammatory and toxic neuropathy	1.29 Heart failure
1.17 Other retinal disorders*	2.42 Chronic ulcer of skin	1.26 Antidiabetic Agents, Insulin
1.12 Age*	2.07 Disorders of fluid, electrolyte, and acid-base balance*	1.26 Nonspecific findings on examination of urine
0.89 Nonspecific findings on examination of urine	1.70 Age*	0.94 Age*
Vascular Disease (VAS)	**Cellulitis (CEL)**	**Pyelonephritis (PYE)**
8.32 Chronic ulcer of skin	3.89 Chronic ulcer of skin*	1.65 Disorders of fluid, electrolyte, and acid-base balance
3.10 Disorders of fluid, electrolyte, and acid-base balance*	2.78 Disorders of fluid, electrolyte, and acid-base balance	1.51 Other disorders of urethra and urinary tract*
2.18 Hereditary and idiopathic peripheral neuropathy	2.51 Bacterial infection in conditions classified elsewhere and of unspecified site*	1.22 Bacterial infection in conditions classified elsewhere and of unspecified site*
2.16 Age*	2.20 Antidiabetic Agents, Insulin	1.11 Calculus of kidney and ureter*
1.88 Atherosclerosis*	2.17 Hereditary and idiopathic peripheral neuropathy*	0.91 Congenital anomalies of urinary system
Osteomyelitis (OST)	**Renal (REN)**	**Hyperosmolar state (HHS)**
3.73 Chronic ulcer of skin*	8.23 Disorders of fluid, electrolyte, and acid-base balance*	4.40 Disorders of fluid, electrolyte, and acid-base balance
1.84 Bacterial infection in conditions classified elsewhere and of unspecified site*	3.04 Heart failure	1.52 Heart failure*
1.56 Open wound of foot except toes alone*	2.71 Hypertensive chronic kidney disease*	1.34 Disorders of mineral metabolism
1.44 Disorders of fluid, electrolyte, and acid-base balance	2.55 Chronic ulcer of skin	1.25 Nondependent abuse of drugs*
1.37 Other and unspecified protein-calorie malnutrition	2.25 Other diseases of lung	1.19 Hypertensive chronic kidney disease
Ketoacidosis (KET)	**Sepsis (SEP)**	**Shock (SHK)**
5.68 Disorders of fluid, electrolyte, and acid-base balance*	6.10 Disorders of fluid, electrolyte, and acid-base balance*	6.10 Disorders of fluid, electrolyte, and acid-base balance*
1.23 Disorders of mineral metabolism	3.46 Bacterial infection in conditions classified elsewhere and of unspecified site*	2.42 Other diseases of lung
1.22 Nonspecific findings on examination of urine*	3.39 Chronic ulcer of skin*	2.06 Heart failure*
1.10 Diseases of pancreas	2.96 Other diseases of lung	1.65 Pneumonia, organism unspecified
1.03 Nondependent abuse of drugs*	2.43 Chronic kidney disease (CKD)	1.55 Certain adverse effects not elsewhere classified*

* indicates that the medical conditions have been mentioned in the clinical literature as the risk factors for the corresponding complications.

work is better feature representation that can capture temporal patient information to improve the risk prediction. Second, different complications could correspond to different severities of diabetes and we can use this knowledge to impose additional constraints on the risk correlations to potentially improve performance. Third, the coefficient shrinkage strategy can be extended to incorporate domain knowledge about the risk factors to potentially improve interpretability. Finally, we only evaluated our methods with one dataset; and the dataset is limited in terms of longitude. A better model evaluation can be achieved by setting different follow-up window sizes and different feature (risk factor) extraction window sizes, given data with long enough records. We are also interested in applying our model to other chronic disease conditions with multiple complications or comorbidities which might benefit from the proposed modeling innovations proposed here.

REFERENCES

[1] World Health Organization, "Global report on diabetes," 2016.
[2] International Diabetes Federation, "Idf diabetes atlas 2017," 2017.
[3] Centers for Disease Control and Prevention, "National diabetes statistics report: estimates of diabetes and its burden in the united states," *National Diabetes Statistics Report*, 2017.
[4] J. M. Forbes and M. E. Cooper, "Mechanisms of diabetic complications," *Physiological reviews*, vol. 93, no. 1, pp. 137–188, 2013.
[5] American Diabetes Association, "Economic costs of diabetes in the u.s. in 2017," *Diabetes Care*, vol. 41, no. 5, pp. 917–928, 2018.
[6] N. R. F. Collaboration *et al.*, "Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4· 4 million participants," *The Lancet*, vol. 387, no. 10027, pp. 1513–1530, 2016.
[7] American Diabetes Association, "Standards of medical care in diabetes 2013," *Diabetes care*, vol. 36, no. Suppl 1, p. S11, 2013.
[8] W. H. Herman, "The economic costs of diabetes: is it time for a new treatment paradigm?" *Diabetes care*, vol. 36, no. 4, pp. 775–776, 2013.
[9] K. Ng, S. R. Steinhubl, S. Dey, W. F. Stewart *et al.*, "Early detection of heart failure using electronic health records," *Circulation: Cardiovascular Quality and Outcomes*, vol. 9, no. 6, pp. 649–658, 2016.
[10] N. Razavian, S. Blecker, A. M. Schmidt, A. Smith-McLallen, S. Nigam, and D. Sontag, "Population-level prediction of type 2 diabetes from claims data and analysis of risk factors," *Big Data*, vol. 3, no. 4, pp. 277–287, 2015.
[11] B. E. Himes, Y. Dai, I. S. Kohane, S. T. Weiss, and M. F. Ramoni, "Prediction of chronic obstructive pulmonary disease (copd) in asthma patients using electronic medical records," *JAMIA*, vol. 16, no. 3, pp. 371–379, 2009.
[12] Y. Cheng, F. Wang, P. Zhang, and J. Hu, "Risk prediction with elec-

tronic health records: A deep learning approach," in *Proceedings of the 2016 SIAM International Conference on Data Mining*, 2016, pp. 432–440.

[13] E. Choi, A. Schuetz, W. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset." *JAMIA*, vol. 24, no. 2, pp. 361–370, 2017.

[14] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 85–94.

[15] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *Proc. of the 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, ser. KDD '11, 2011, pp. 814–822.

[16] X. Wang, F. Wang, J. Hu, and R. Sorrentino, "Towards actionable risk stratification: A bilinear approach," *Journal of biomedical informatics*, vol. 53, pp. 147–155, 2015.

[17] R. Chen, J. Sun, R. S. Dittus, D. Fabbri, J. Kirby, C. L. Laffer, C. D. McNaughton, and B. Malin, "Patient stratification using electronic health records from a chronic disease management program," *IEEE journal of biomedical and health informatics*, 2016.

[18] D. He, S. C. Mathews, A. N. Kalloo, and S. Hutfless, "Mining high-dimensional administrative claims data to predict early hospital readmissions," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 272–279, 2014.

[19] I. Bardhan, J.-h. Oh, Z. Zheng, and K. Kirksey, "Predictive analytics for readmission of patients with congestive heart failure," *Information Systems Research*, vol. 26, no. 1, pp. 19–39, 2014.

[20] Y. P. Tabak, X. Sun, C. M. Nunez, and R. S. Johannes, "Using electronic health record data to develop inpatient mortality predictive model: Acute laboratory risk of mortality score (alarms)," *JAMIA*, vol. 21, no. 3, pp. 455–463, 2013.

[21] N. Nori, H. Kashima, K. Yamashita, H. Ikai, and Y. Imanaka, "Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care," in *Proc. of the 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2015, pp. 855–864.

[22] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, Jul. 1997.

[23] P. Schulam and S. Saria, "A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure," in *NIPS*, 2015, pp. 748–756.

[24] J. C. Ho, J. Ghosh, and J. Sun, "Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 115–124.

[25] C. Liu, F. Wang, J. Hu, and H. Xiong, "Temporal phenotyping from longitudinal electronic health records: A graph based framework," ser. KDD '15, 2015, pp. 705–714.

[26] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun *et al.*, "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, no. 1, p. 18, 2018.

[27] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*.

[28] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," *Journal of the American Medical Informatics Association*, vol. 24, no. 2, pp. 361–370, 2016.

[29] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," ser. KDD '15, 2015, pp. 1721–1730.

[30] P. Yadav, M. Steinbach, V. Kumar, and G. Simon, "Mining electronic health records (ehrs): A survey," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 85:1–85:40, Jan. 2018.

[31] H. Qiu, H.-Y. Yu, L.-Y. Wang, Q. Yao, S.-N. Wu, C. Yin, B. Fu, X.-J. Zhu, Y.-L. Zhang, Y. Xing *et al.*, "Electronic health record driven prediction for gestational diabetes mellitus in early pregnancy," *Scientific reports*, vol. 7, no. 1, p. 16417, 2017.

[32] W. Oh, E. Kim, M. R. Castro, P. J. Caraballo, V. Kumar, M. S. Steinbach, and G. J. Simon, "Type 2 diabetes mellitus trajectories and associated risks," *Big data*, vol. 4, no. 1, pp. 25–30, 2016.

[33] D. Bertsimas, N. Kallus, A. M. Weinstein, and Y. D. Zhuo, "Personalized diabetes management using electronic medical records," *Diabetes care*, vol. 40, no. 2, pp. 210–217, 2017.

[34] B. Liu, Y. Li, Z. Sun, S. Ghosh, and K. Ng, "Early prediction of diabetes complications from electronic health records: A multi-task survival analysis approach," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[35] J. Wiens, J. Guttag, and E. Horvitz, "Patient risk stratification with time-varying parameters: a multitask learning approach," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2797–2819, 2016.

[36] Z. Sun, F. Wang, and J. Hu, "Linkage: An approach for comprehensive risk prediction for care management," in *Proc. of the 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, ser. KDD '15, 2015, pp. 1145–1154.

[37] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[38] Y. Zhang and D.-Y. Yeung, "A convex formulation for learning task relationships in multi-task learning," in *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'10, 2010, pp. 733–742.

[39] Y. Zhang and Q. Yang, "A survey on multi-task learning," *arXiv preprint arXiv:1707.08114*, 2017.

[40] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Advances in neural information processing systems*, 2007, pp. 41–48.

[41] J. Chen, L. Tang, J. Liu, and J. Ye, "A convex formulation for learning a shared predictive structure from multiple tasks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 5, pp. 1025–1038, 2013.

[42] H. Zhao, O. Stretcu, R. Negrinho, A. Smola, and G. Gordon, "Efficient multi-task feature and relationship learning," in *2017 NIPS workshop on Learning with Limited Labeled Data: Weak Supervision and Beyond*, 2017.

[43] J. M. Reps, M. J. Schuemie, M. A. Suchard, P. B. Ryan, and P. R. Rijnbeek, "Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data," *Journal of the American Medical Informatics Association*, vol. 25, no. 8, pp. 969–975, 2018.

[44] Y. Zhang and D.-Y. Yeung, "A regularization approach to learning task relationships in multitask learning," *ACM Trans. Knowl. Discov. Data*, vol. 8, no. 3, pp. 12:1–12:31, 2014.

[45] W. H. Organization *et al.*, "International classification of diseases:[9th] ninth revision, basic tabulation list with alphabetic index," 1978.

[46] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[47] C. M. Carvalho, N. G. Polson, and J. G. Scott, "The horseshoe estimator for sparse signals," *Biometrika*, vol. 97, no. 2, pp. 465–480, 2010.

[48] P. X.-K. Song, M. Li, and Y. Yuan, "Joint regression analysis of correlated data using gaussian copulas," *Biometrics*, vol. 65, no. 1, pp. 60–68, 2009.

[49] M. Sklar, "Fonctions de repartition an dimensions et leurs marges," *Publ. Inst. Statist. Univ. Paris*, vol. 8, pp. 229–231, 1959.

[50] G. Elidan, "Copulas in machine learning," in *Copulae in mathematical and quantitative finance*. Springer, 2013, pp. 39–60.

[51] L. Devroye, "Sample-based non-uniform random variate generation," in *Proceedings of the 18th conference on Winter simulation*. ACM, 1986, pp. 260–265.

[52] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM Journal on imaging sciences*, vol. 6, no. 3, pp. 1758–1789, 2013.

[53] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, 2010, pp. 177–186.

[54] C. Chen, B. He, Y. Ye, and X. Yuan, "The direct extension of admm for multi-block convex minimization problems is not necessarily convergent," *Mathematical Programming*, vol. 155, no. 1-2, pp. 57–79, 2016.

[55] Truven health, "Truven health marketscan research databases," 2018. [Online]. Available: https://truvenhealth.com/markets/life-sciences/products/data-tools/marketscan-databases

[56] American Diabetes Association and others, "Report of the expert committee on the diagnosis and classification of diabetes mellitus," *Diabetes care*, vol. 26, no. suppl 1, pp. s5–s20, 2003.

[57] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[58] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8685–8690, 2007.
[59] W. T. Cade, "Diabetes-related microvascular and macrovascular diseases in the physical therapy setting," *Physical therapy*, vol. 88, no. 11, pp. 1322–1335, 2008.
[60] B. F. Palmer and D. J. Clegg, "Electrolyte and acid–base disturbances in patients with diabetes mellitus," *New England Journal of Medicine*, vol. 373, no. 6, pp. 548–559, 2015.
[61] N. Sotirakopoulos, I. Kalogiannidou, M. Tersi, K. Armentzioiou, D. Sivridis, and K. Mavromatidis, "Acid-base and electrolyte disorders in patients with diabetes mellitus," *Saudi Journal of Kidney Diseases and Transplantation*, vol. 23, no. 1, p. 58, 2012.
[62] S. Shamshirgaran, A. Mamaghanian, A. Aliasgarzadeh, N. Aiminisani, M. Iranparvar-Alamdari, and J. Ataie, "Age differences in diabetes-related complications and glycemic control," *BMC endocrine disorders*, vol. 17, no. 1, p. 25, 2017.
[63] K. K. Viktil, H. S. Blix, T. A. Moger, and A. Reikvam, "Polypharmacy as commonly defined is an indicator of limited value in the assessment of drug-related problems," *British journal of clinical pharmacology*, vol. 63, no. 2, pp. 187–195, 2007.

PLACE PHOTO HERE

Bin Liu received his Ph.D. from Rutgers University. He is a research scientist at IBM Thomas J. Watson Research Center. He is interested in data mining/machine learning, and their intersections with healthcare, business analytics, recommender systems, and privacy/security. He has published in premier journals such as IEEE TKDE, ACM TKDD, ACM TOPS, ACM TIST; and top conferences such as KDD, AAAI, WSDM, USENIX Security. He currently serves on the Editorial Board of the Journal of Business Analytics, and has served as a reviewer for many journals, including IEEE TKDE, ACM TOIS, ACM TKDD, ACM TIST, and ACM TMIS. He has served regularly on program committees of conferences, including KDD, AAAI, CIKM, SDM, and RecSys. He is a member of the ACM and IEEE.

PLACE PHOTO HERE

Ying Li is a research staff member in the IBM T.J. Watson Research Center. She graduated with a Ph.D. degree of Biomedical Informatics at Columbia University. Her research interests involve pharmacovigilance, drug repurposing and medication use related analysis using real world evidence and data mining techniques. She has published several articles in refereed journals and conferences, including Nature Biotechnology, Journal of the American Medical Informatics Association (JAMIA) and Drug Safety journals, and American Medical Informatics Association Annual Symposium.

PLACE PHOTO HERE

Soumya Ghosh is a Research Scientist at IBM research and the MIT-IBM Watson AI institute. His research interests span topics in machine learning, health informatics, and computer vision and are inspired by the challenges of learning from richly structured datasets. His recent work has focused on designing effective priors and efficient learning of Bayesian neural networks as well as on Bayesian nonparametric models and inference algorithms for partitioning images, videos and text. He holds a PhD in computer science from Brown University and joined IBM research after a postdoctoral stint at Disney Research.

PLACE PHOTO HERE

Zhaonan Sun received her Bachelor degree in Statistics from the Renmin University of China. She received her Doctor of Philosophy (PhD) in Statistics from Purdue University in 2014. She is currently a Research Staff Member in the Center for Computational Health at IBM T.J. Watson Research Center. Her research lies in developing statistical and machine learning methods to generate insights in the healthcare domain.

PLACE PHOTO HERE

Kenney Ng is a research staff member in the Center for Computational Health and manager of the Health Analytics Research Group at IBM Research Cambridge. He received B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology. His current research focus is on the development and application of data mining, machine learning, and AI techniques to analyze, model and derive actionable insights from real world health data. His prior research areas include information retrieval, speech recognition, probabilistic modeling, topic modeling, statistical language modeling and neural networks. Before IBM Research, he was a senior software engineer and architect in IBM Software Group for a number of products including IBM Patient Care and Insights (IPCI), eDiscovery Analyzer (eDA), IBM Content Analyzer (ICA), Omnifind Yahoo! Edition (OYE), IBM Classification Module (ICM), and Omnifind Discovery Edition (ODE). Prior to IBM, he was a principal software engineer at iPhrase Technologies and held research positions at the MIT Laboratory for Computer Science, BBN Technologies, and MIT Lincoln Laboratory. He is a member of the Institute of Electrical and Electronics Engineers and the American Medical Informatics Association.

PLACE PHOTO HERE

Jianying Hu (Ph.D.) is IBM Fellow; Global Science Leader, AI for Healthcare; and Program Director of the Center for Computational Health at IBM Research. Prior to joining IBM in 2003 she was with Bell Labs at Murray Hill, New Jersey. Dr. Hu has conducted and led extensive research in machine learning, data mining, statistical pattern recognition, and signal processing, with applications to healthcare analytics and medical informatics, business analytics, and multimedia content analysis. Her recent focus has been on leading research efforts to develop advanced machine learning, data mining and visual analytics methodologies for deriving data-driven insights from real world healthcare data to facilitate learning health systems. Dr. Hu served as Chair of the Knowledge Discovery and Data Mining (KDDM) Working Group of the American Medical Informatics Association (AMIA) from 2014 to 2016. She has published over 120 peer reviewed scientific papers and holds 31 patents. She has served as Associate Editor for the journals IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, Pattern Recognition, and International Journal for Document Analysis and Recognition. Dr Hu currently serves on the Editorial Board of the journal JAMIA Open, Advisory Board of the Journal of Healthcare Informatics Research, the Computational Science Advisory Board of Micheal J. Fox Foundation, the Industry Advisory Board of NJIT, and the Western Pennsylvania HIT Advisory Board. Dr. Hu is a fellow of IEEE, a fellow of the International Association of Pattern Recognition, and an overseas fellow of Royal Society of Medicine. She received the Asian American Engineer of the Year Award in 2013.