Early Prediction of Diabetes Complications from Electronic Health Records: A Multi-task Survival Analysis Approach

Bin Liu, Ying Li, Zhaonan Sun, Soumya Ghosh, Kenney Ng
Center for Computational Health, IBM Thomas J. Watson Research Center
bin.liu1@ibm.com, liying@us.ibm.com, zsun@us.ibm.com, ghoshso@us.ibm.com, kenney.ng@us.ibm.com

Abstract

Type 2 diabetes mellitus (T2DM) is a chronic disease that usually results in multiple complications. Early identification of individuals at risk for complications after being diagnosed with T2DM is of significant clinical value. In this paper, we present a new data-driven predictive approach to predict *when* a patient will develop complications after the initial T2DM diagnosis. We propose a novel survival analysis method to model the time-to-event of T2DM complications designed to simultaneously achieve two important metrics: 1) accurate prediction of event times, and 2) good ranking of the relative risks of two patients. Moreover, to better capture the correlations of time-to-events of the multiple complications, we further develop a multi-task version of the survival model. To assess the performance of these approaches, we perform extensive experiments on patient level data extracted from a large electronic health record claims database. The results show that our new proposed survival analysis approach consistently outperforms traditional survival models and demonstrate the effectiveness of the multi-task framework over modeling each complication independently.

Introduction

Type 2 diabetes mellitus (T2DM) is a chronic disease that affects almost half a billion people around the globe (World Health Organization 2016). It is characterized by hyperglycemia— abnormally elevated blood glucose (blood sugar) levels, and is almost always associated with a number of complications (Forbes and Cooper 2013). 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. T2DM management requires continuous medical care with multifactorial risk-reduction strategies beyond glycemic control (American Diabetes Association and others 2013). Early prediction of T2DM complications is critical for healthcare professionals to appropriately adapt treatment plans for patients.

The recent abundance of the electronic health records (EHRs) has provided an unprecedented opportunity to apply predictive analytics to improve T2DM management. In this paper, we study the early prediction of T2DM complications from historical EHR records: *When* will a patient develop complications after the initial T2DM diagnosis? Given the EHR records of two patients, which patient is more likely to develop complications? In the literature, EHRs have been applied to disease onset prediction (Ng et al. 2016; Razavian et al. 2015), patient stratification (Wang et al. 2015; Chen et al. 2016), readmission prediction (He et al. 2014), and mortality prediction (Tabak et al. 2013). However, there are unique characteristics and challenges to the problem of T2DM complications time-to-event prediction.

One of the main challenges for such time-to-event prediction problems is the existence of censored data in which events of interest are unobserved. Events of interest may not be observed due to the limited duration of the study period or due to losing track of patients during the observation period. As such, predictive models based on standard machine learning approaches, which usually optimize a loss function, cannot be directly applied to analyze censored longitudinal data. Survival analysis (Cox 1972; Miller Jr 2011) is a class of widely used statistical tools to model time-to-event censored data and thus can be adapted to model T2DM complication events. However, traditional survival analysis models suffer from several limitations. The popular Cox model (Cox 1972) does not directly model event probability; instead, it maximizes a partial likelihood objective, which depends only on the relative ordering of the survival time of individuals, not on their actual values. Parametric survival models (Lawless 1998; Mittal et al. 2013) provide another popular alternative. These methods assume that the baseline hazard function follows some distribution, such as Exponential, Weibull or Log-normal. However, the distribution may not be flexible enough to capture the complex event patterns observed in practice. A second challenge stems from the need to effectively capture the correlations between multiple T2DM complications. Considering that the different complications are manifestations of a common underlying condition — hyperglycemia, modeling complications as independent of one another will lead to suboptimal models.

To address these challenges, we present a data-driven approach to predict *when* a patient will develop complication(s) after the initial T2DM diagnosis. Our contributions include a novel survival analysis approach, **RankSvx**, to model time-to-event of T2DM complications. RankSvx si-

Table 1: Mathematical Notations

Symbol	Description
$N, \langle i, j \rangle$	number and indices of patients
M, k	number and index of T2DM complications
t_{ki}	time of event of patient i for complication k
c_{ki}	indicator of censoring for event t_{ki} with $c_{ki} = 1$ means observed and $c_{ki} = 0$ means censoring
$\mathbf{x}_i$	$\mathbf{x}_i$ is the explanatory covariates for patient i
$\mathbf{w}_k, \mathbf{W}$	$\mathbf{w}_k$ is the coefficients for complication k; $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_M]$ is the matrix of coefficients
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	order graph with vertices $\mathcal{V}$ represents patients and edge $\mathcal{E}(i, j)$ indicates event time order $T_i \leq T_j$.
$\mathbf{\Omega}$	matrix of relatedness between complications
$\mathbf{\Omega}_0$	matrix of prior knowledge about risk association

multaneously optimizes two objective functions: a regression function that models the event times of the observed events, and a ranking function that models the relative risks of both the observed and censored events. As a result, the proposed survival approach has the advantage of simultaneously achieving two important desiderata: accurate prediction of event times, as well as an accurate ranking of the relative risks of two patients. Moreover, to better capture the correlations of time-to-events of multiple complications, we further develop a multi-task version of the survival model. The multi-task model allows us to not only capture the relatedness between different complications but also incorporate domain knowledge as prior information.

To assess the performance of our proposed innovations, we perform extensive experiments on patient level data extracted from a large electronic health record claims database. The results show that our new proposed survival analysis approach consistently outperforms traditional survival models and demonstrate the effectiveness of the multi-task framework over modeling each complication independently.

Problem Definition

Our goal is to build an effective data-driven predictive approach to predict when a patient will develop complication(s) after the initial T2DM diagnosis. Specifically, for patient i we observe a set of D risk factors, denoted as $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iD}]^\top$, for an observation window up until the patient was diagnosed with T2DM. Let there be M complications in consideration indexed by $k \in \{1, \cdots, M\}$. We use t_{ki} to represent the time when patient i develops complication k. We use the indicator c_{ki} to represent the censoring of the event t_{ki} where $c_{ki} = 1$ means observed and $c_{ki} = 0$ means censored. We aim to build a predictive model $f(t_{ki}|\Theta, \mathbf{x}_i)$ to predict when patient i will develop complication k. Table 1 shows some important mathematical notations used in this paper.

Multi-task Survival Analysis to Model T2DM Complication Events

In this section, we present our multi-task survival analysis to model time-to-event of T2DM complications.

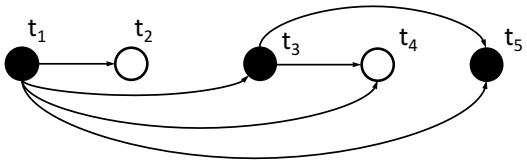

Figure 1: Illustration of event order graph in time-to-event modeling. Filled vertex indicates an observed event and an empty circle denotes a censored observation. An arrowed edge $\mathcal{E}(i, j)$ between two nodes indicates event time order $T_i \leq T_j$. Note that all nodes connected to an observed event node i correspond to the risk set R_i for patient i in the Cox model.

Modeling Single T2DM Complication Events

Before proceeding to consider all the complications, we first look at how to model a single complication event, namely the single-task learning paradigm.

One of the main challenges in survival modeling is the existence of censored data in which the events of interests are not observed due to either the time limitation of the study period or to losing track of the patient during the observation period. Due to the uncertainty caused by the censored data, we decompose the objective function into two parts:

$$\alpha \mathcal{L}_{\text{obs}}\big(t_i, f(\mathbf{x}_i|\Theta)\big) + (1 - \alpha)\mathcal{L}_{\text{cen}}\big(t_i, f(\mathbf{x}_i|\Theta)\big) + g(\Theta). \quad (1)$$

The first term $\mathcal{L}_{\text{obs}}\big(t_i, f(\mathbf{x}_i|\Theta)\big)$ models the observed event, and the loss function can be any used in standard generalized regression models. The second term $\mathcal{L}_{\text{cen}}\big(t_i, f(\mathbf{x}_i|\Theta)\big)$ models the censored data and will be discussed in detail in the following section. The weight term α balances the two loss functions. Finally, $g(\Theta)$ is a regularization term that controls the model complexity.

Modeling censored data as a ranking task. We cast the modeling of censored data as a ranking problem, where the task is to order the event times. First, survival analysis, represented by the Cox model, can be regarded as the modeling of the event order due to the introduction of censored data. It answers the order question: "which one of patients i and j is more likely to develop a disease?". Raykar *et al.* (Raykar et al. 2007) show that Cox's partial likelihood is a lower bound of the *concordance index* (CI), which is one of the most commonly used metrics for survival models. Second, we aim to use the ranking to compliment the modeling of actual event time since the first term of our objective function as shown in Equation (1) already models the event times.

We construct an *event order graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ as shown in Fig. 1. The set of vertices $\mathcal{V}$ represents all the patients, where each filled vertex indicates an observed event time, while an empty circle denotes a censored observation. An arrowed edge $\mathcal{E}(i, j)$ between two nodes indicates event time order $T_i \leq T_j$. Note that all nodes connected to an observed event node i correspond to the risk set R_i for patient i in the Cox model. We aim to correctly rank the relative risks of two patients, which is equal to maximizing the probability of all pairs of patients whose predicted event times are correctly ordered among all patients that can actually be or-

dered. Then we maximize following likelihood

$$\log \prod_{\mathcal{E}_{ij}} \Pr(T_j > T_i|\Theta) = \log \prod_{\mathcal{E}_{ij}} \Pr\left[f(\mathbf{x}_j|\Theta) - f(\mathbf{x}_i|\Theta)\right]. \quad (2)$$

There are multiple choices of functions (*e.g.*, Hinge, Sigmoid and exponential) to model the order between event order $T_j > T_i$. We follow (Raykar et al. 2007; Rendle et al. 2009) and choose the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$. Then we have following loss function for the censored data

$$\mathcal{L}_{\text{cen}}\big(t_i, f(\mathbf{x}_i|\Theta)\big) = \sum_{\mathcal{E}_{ij}} \log \sigma\left[f(\mathbf{x}_j|\Theta) - f(\mathbf{x}_i|\Theta)\right]. \quad (3)$$

It can be shown that the modeling of event orders actually approximates the concordance index (CI), one of the most commonly used metrics for survival models:

$$\text{CI} = \frac{1}{|\mathcal{V}|} \sum_{\substack{T_i \\ \forall c_i=1}} \sum_{T_j > T_i} \mathbf{1}_{f(x_j) > f(x_i)} \quad (4)$$

where $\mathbf{1}(x)$ is an indicator function. CI can be interpreted as the fraction of all pairs of patients whose predicted survival times are correctly ordered among all patients that can actually be ordered.

Combined regression and ranking: We propose a unified framework to combine both regression and ranking to model time-to-event:

$$\begin{aligned} \min \quad & -(1-\alpha) \sum_{\mathcal{E}_{ij}} \log \sigma\left[f(\mathbf{x}_j|\Theta) - f(\mathbf{x}_i|\Theta)\right] \\ & + \alpha \mathcal{L}_{\text{obs}}\big(t_i, f(\mathbf{x}_i|\Theta)\big) + g(\Theta). \end{aligned} \quad (5)$$

The unified framework has the advantage to simultaneously achieve two important desiderata: accurate prediction of event times, and good ranking of the relative risks of two patients. The semi-parametric Cox model, which maximizes the partial likelihood, approximates the ranking of relative risks but often does not perform well in event time prediction. Parametric survival models, which make rigorous statistical assumptions about the survival time, may not be flexible enough to capture the complex event patterns. We aim to use the unified framework to complement each other.

We directly model the survival time for patient i as $f(\mathbf{x}_i|\Theta) = \mathbf{w}^\top \mathbf{x}_i$ where $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iD}]^\top$ is the explanatory covariates vector for patient i. The observed loss function can be any used in standard generalized regression models. In particular, we consider Squared, Poisson, and Log-normal loss functions:

$$\mathcal{L}_{\text{obs}}\big(t_i, f(\mathbf{x}_i|\Theta)\big) = \begin{cases} \sum_i \frac{1}{2}\left(t_i - \mathbf{w}^\top \mathbf{x}_i\right)^2 & \text{Squared} \\ \sum_i \left(e^{\mathbf{w}^\top \mathbf{x}_i} - t_i \mathbf{w}^\top \mathbf{x}_i\right) & \text{Poisson} \\ \sum_i \frac{1}{2}\left(\log(t_i) - \mathbf{w}^\top \mathbf{x}_i\right)^2 & \text{Log-normal} \end{cases}$$

Modeling Multiple T2DM Complication Events via Multi-task Learning

To capture and leverage the association between the risks of the different T2DM complications, we formulate the complications prediction task as a multi-task learning problem. As shown in Fig. 2, 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 and survival analysis is applied to model the time-to-event of the complication. Then we apply multi-task learning to capture the association between the different complications.

Multi-task learning (MTL) (Caruana 1997) aims to jointly learn multiple tasks using a shared representation so that knowledge obtained from one task can help the other tasks. In particular, we adopt the *task relation learning* based MTL approach (Zhang and Yang 2017) due to its flexibility to incorporate prior information. Let $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M]_{D\times M}$ denote the matrix of coefficients for all M of the complications. We aim to explore the hidden association between the risks of T2DM complications. We assume that the risk association is revealed in the structure of the coefficient matrix $\mathbf{W}$. Following (Zhang and Yeung 2010; Sun, Wang, and Hu 2015), we use the covariance matrix of $\mathbf{W}$ to characterize T2DM complication risk association. Specifically, we assume that the coefficient matrix $\mathbf{W}$ follows a Matrix Variate Normal (MVN) distribution:

$$\mathbf{W} \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Gamma}, \mathbf{\Omega}). \quad (6)$$

The the first term $\mathbf{0}$ is a $D \times M$ matrix of zeros representing the mean of $\mathbf{W}$. The second term $\mathbf{\Gamma}$ is a $D \times D$ matrix representing the row-wise covariances of $\mathbf{W}$. In this paper we assume that rows of $\mathbf{W}$ are independent of each other. In other words, the coefficients of different features in the same target are not correlated. Then $\mathbf{\Gamma}$ becomes a diagonal matrix, and we can set $\mathbf{\Gamma} = \mathbf{I}$. The third term $\mathbf{\Omega}$ is a $M \times M$ symmetric positive definite matrix representing the column-wise covariance of $\mathbf{W}$. It is unknown and reflects the risk association between multiple complications. Then we have

$$\Pr(\mathbf{W}|\mathbf{0}, \mathbf{I}, \mathbf{\Omega}) = \frac{\exp\left(-\frac{1}{2}\operatorname{tr}\left[\mathbf{\Omega}^{-1}\mathbf{W}^\top\mathbf{W}\right]\right)}{(2\pi)^{MD/2}|\mathbf{\Omega}|^{D/2}} \quad (7)$$

Further, domain knowledge about risk association is often available or partially available. In order to utilize available domain knowledge, we impose an Inverse-Wishart prior distribution on $\mathbf{\Omega}$

$$\Pr(\mathbf{\Omega}) \sim \mathcal{IW}(\delta\mathbf{\Omega}_0, \nu). \quad (8)$$

The Inverse-Wishart distribution is a conjugate prior for the multivariate variate distribution $\mathbf{\Omega}$. $\mathbf{\Omega}_0$ 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 set $\mathbf{\Omega}$ to be $\mathbf{I}$.

The posterior probability of the parameters can be written as

$$\begin{aligned} & \sum_{k=1}^{M} \frac{1}{N_k}\left[\alpha \sum_{\substack{i \\ \forall c_{ki}=1}} \mathcal{L}_{\text{obs}}(t_{ki}, \mathbf{w}_k^\top \mathbf{x}_i) - (1-\alpha) \sum_{\mathcal{E}_{ij}^k} \log \sigma\left[\mathbf{w}_k^\top(\mathbf{x}_j - \mathbf{x}_i)\right]\right] \\ & + \operatorname{tr}\left[\left(\frac{\lambda_1}{2}\mathbf{W}^\top\mathbf{W} + \frac{\lambda_2}{2}\mathbf{\Omega}_0\right)\mathbf{\Omega}^{-1}\right] + \frac{\lambda_3}{2}\log|\mathbf{\Omega}| + \frac{\eta}{2}\sum_{k=1}^{M}||\mathbf{w}_k||^2 \end{aligned} \quad (9)$$

where $\operatorname{tr}(\cdot)$ and $|\cdot|$ denote the trace and determinant

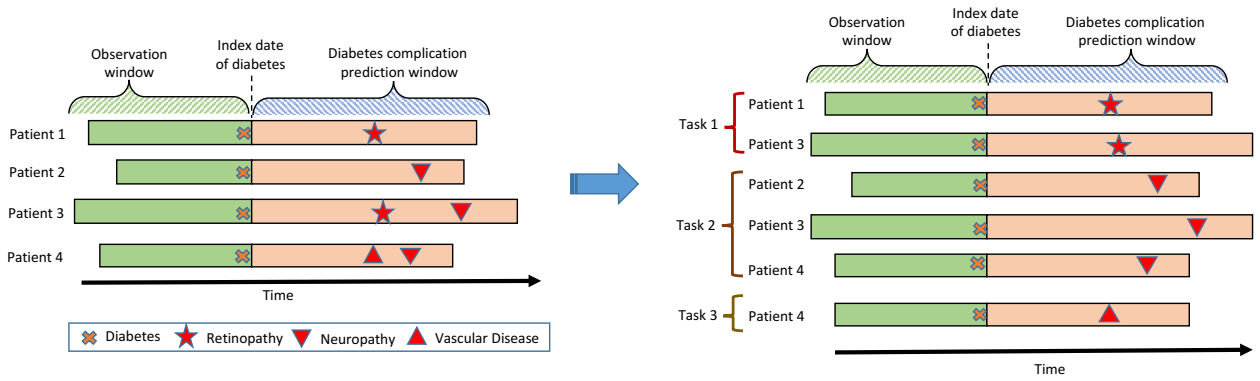

Figure 2: Proposed framework for early prediction of T2DM complications. We aim to predict when a patient will develop complications after the initial T2DM diagnosis. We group the predictions of multiple complications in consideration (*e.g.*, retinopathy, neuropathy and vascular disease) into different tasks where each task models only one complication. Multi-task learning (MTL) is applied to capture the association between the different complications.

of a matrix; λ_1, λ_2, λ_3 and η are tuning parameters; and $\sum_{k=1}^{M} ||\mathbf{w}_k||^2$ is a regularization term to control the model complexity. $1/N_k$ is added to avoid the task imbalance problem when training instances are unbalanced among tasks where N_k is the number of instances in k-th task.

Parameter Estimation

Given time-to-event observations $\mathbf{Y} = \{t_{ki}, c_{ki}\}$ and covariates $\mathbf{X}$, we need to estimate the model parameters $\{\mathbf{W}, \mathbf{\Omega}\}$ via solving the optimization problem in Equation (9). However, log determination ($\log|\mathbf{\Omega}|$) is concave, making the objective function non-convex. Therefore we adopt an iterative algorithm to solve the problem. Within each iteration, the two blocks $\mathbf{W}$ and $\mathbf{\Omega}$ are updated alternatively.

Update W given Ω:

Given $\mathbf{\Omega}$, minimizing Equation (9) becomes minimizing the following function

$$\sum_{k=1}^{M} \frac{1}{N_k} \left[\alpha \sum_{\forall c_{ki}=1} \mathcal{L}_{\text{obs}}\left(t_{ki}, \mathbf{w}_k^\top \mathbf{x}_i\right) - (1-\alpha) \sum_{\mathcal{E}_{ij}^k} \log \sigma \left[\mathbf{w}_k^\top (\mathbf{x}_j - \mathbf{x}_i)\right] \right] + \text{tr}\left[\left(\frac{\lambda_1}{2}\mathbf{W}^\top \mathbf{W}\right)\mathbf{\Omega}^{-1}\right] + \frac{\eta}{2}\sum_{k=1}^{M} ||\mathbf{w}_k||^2. \tag{10}$$

We use stochastic gradient descent to update the parameters. Stochastic gradient descent has been widely used for many machine learning tasks (Bottou 2010). The main process involves randomly scanning training instances and iteratively updating parameters. In each iteration, for complication k, we randomly sample an observed instance and its comparison set $\langle k, i, \mathrm{R}_i^k \rangle$[1], and we maximize $\mathcal{O}(\Theta)$ using the following update rule for Θ: $\Theta = \Theta - \epsilon \cdot \frac{\partial \mathcal{O}(\Theta)}{\partial \Theta}$, where ϵ is a learning rate. For complication k, given observed instance and its comparison set $\langle k, i, \mathrm{R}_i^k \rangle$, the gradient with respect to $\mathbf{w}_k$ is

$$\frac{\partial \mathcal{O}}{\partial \mathbf{w}_k} = \frac{1}{N_k}\left[(1-\alpha)\sum_{j \in R_i^k}\left(\frac{e^{-\mathbf{w}_k^\top(\mathbf{x}_j - \mathbf{x}_i)}}{1+e^{-\mathbf{w}_k^\top(\mathbf{x}_j-\mathbf{x}_i)}}\right)(\mathbf{x}_i - \mathbf{x}_j) + \alpha \frac{\partial \mathcal{L}_{\text{obs}}\left(t_{ki}, \mathbf{w}_k^\top \mathbf{x}_i\right)}{\partial \mathbf{w}_k}\right] + \lambda_1 \Omega_k^{-1}\mathbf{W} + \eta \mathbf{w}_k.$$

The gradient $\frac{\partial \mathcal{L}_{\text{obs}}\left(t_{ki}, \mathbf{w}_k^\top \mathbf{x}_i\right)}{\partial \mathbf{w}_k}$ for different loss functions are as follows:

$$\frac{\partial \mathcal{L}_{\text{obs}}\left(t_{ki}, \mathbf{w}_k^\top \mathbf{x}_i\right)}{\partial \mathbf{w}_k} = \begin{cases} -\left(t_i - \mathbf{w}_k^\top \mathbf{x}_i\right)\mathbf{x}_i & \text{Squared} \\ \left(e^{\mathbf{w}_k^\top \mathbf{x}_i} - t_{ki}\right)\mathbf{x}_i & \text{Poisson} \\ -\left(\log(t_{ki}) - \mathbf{w}_k^\top \mathbf{x}_i\right)\mathbf{x}_i & \text{Log-normal} \end{cases}$$

Update Ω given W:

Given $\mathbf{W}$, minimizing Equation (9) becomes

$$\arg\min_{\mathbf{\Omega}} \text{tr}\left[\left(\frac{\lambda_1}{2}\mathbf{W}^\top\mathbf{W} + \frac{\lambda_2}{2}\mathbf{\Omega}_0\right)\Omega^{-1}\right] + \frac{\lambda_3}{2}\log|\mathbf{\Omega}| \tag{11}$$

The last term $\log|\mathbf{\Omega}|$ is a penalty on the complexity of $\mathbf{\Omega}$, and can be replaced with a constraint $\text{tr}(\mathbf{\Omega}) = 1$ (Zhang and Yeung 2010). Then above Equation (11) can be reformulated as:

$$\arg\min_{\mathbf{\Omega}} \text{tr}\left[\left(\frac{\lambda_1}{2}\mathbf{W}^\top\mathbf{W} + \frac{\lambda_2}{2}\mathbf{\Omega}_0\right)\Omega^{-1}\right] \quad \text{s.t. } \mathbf{\Omega} \succeq 0,\ \text{tr}(\mathbf{\Omega}) = 1 \tag{12}$$

where $\mathbf{\Omega} \succeq 0$ means that the matrix $\mathbf{\Omega}$ is positive semidefinite. Equation (12) has an analytical solution

$$\mathbf{\Omega} = \frac{\left(\frac{\lambda_1}{2}\mathbf{W}^\top\mathbf{W} + \frac{\lambda_2}{2}\mathbf{\Omega}_0\right)^{\frac{1}{2}}}{\text{tr}\left[\left(\frac{\lambda_1}{2}\mathbf{W}^\top\mathbf{W} + \frac{\lambda_2}{2}\mathbf{\Omega}_0\right)^{\frac{1}{2}}\right]}. \tag{13}$$

Experiments

This section presents an empirical evaluation of our model using patient level data extracted from a large electronic health record claims database.

[1] In practice, for each observed instance i in task k we can randomly sample a subset of R_i^k when the total comparison set size $|\mathrm{R}_i^k|$ is large.

Table 2: List of the five T2DM complications in this study.

T2DM Complication (Abbreviation)	Description	Example ICD codes
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, E08311-E0839
Neuropathy (NEU)	nerve damage most often affecting the legs and feet	25060, 25062, 24960, 24961
Nephrology (NEP)	kidney disease characterized by hardening of the glomerulus	25040, 25042, 24940, 24941
Vascular Disease (VAS)	vascular diseases including peripheral vascular disease, cardiovascular disease, and cerebrovascular diseases	25070, 25072, 24970, 24971, E0851, E08621-E08622, E0859
Hyperosmolar (HYPER)	serious condition caused by high blood sugar levels	25020, 25022, 24920, 24921, E0800, E0900, E1100, E1300

Table 3: Data statistics and patient characteristics.

Complication	# instances	# observations	Female ratio	Average age (SD)	19–44 pct.	45–54 pct.	55–64 pct.
RET	5604	1868	35.03%	52.50 (8.58)	17.02%	33.21%	49.50%
NEU	11874	3958	36.97%	52.53 (8.59)	16.97%	33.01%	49.82%
NEP	4074	1358	37.02%	52.52 (8.91)	17.53%	31.44%	50.86%
VAS	2517	839	39.85%	53.17 (8.31)	15.06%	31.55%	53.12%
HYPER	651	217	36.41%	52.00 (8.90)	19.35%	32.72%	47.93%

a SD, standard deviation

Experimental setup and data

We conducted a retrospective cohort study using the MarketScan Commercial Claims and Encounter (CCAE) database from Truven Health. The data on the patients are contributed by a selection of large employers, health plans, and government and public organizations. We used a dataset of deidentified patients between the years 2011 and 2015. 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 have two (2) or more Type 2 diabetes records on different days; OR

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

Finally, patients who were under 19 years old or over 64 years old at first diagnosis of T2DM are removed.

We use following prediction variables:

- **Patient demographics:** Patient demographics include age, gender and weight index. In addition to age as one continuous variable, we also include three binary variables for age intervals of 19–44, 45–54 and 55–64.
- **ICD codes:** We use the historical medical conditions features encoded as International Classification of Disease (ICD) codes. We use group ICD codes according to their first three digits, and filter out ICDs appearing in fewer than 100 patients. As a result we have 359 ICD features.

We further removed patients with less than 20 ICD records.

Five common complications of T2DM, described in Table 2 are used in this study. Table 3 shows some basic statistics of the patient cohort.

Evaluation protocol

We aim to answer the following two questions:

Question 1: How does the performance of our proposed model (**RankSvx**) compare to traditional survival models and regression models? To this end, we compare our proposed model with the following baseline algorithms:

- Cox model (Cox 1972): Cox is the most widely used survival model and is a semi-parametric model as it does not assume any distribution on the baseline function.
- Parametric survival models (Mittal et al. 2013) including Weibull, Log-Logistic, and Log-normal. They make different assumptions about the baseline survival function.
- Regression models (*i.e.*, squared regression, Poisson regression, and Log-normal regression) that directly model the event times but cannot leverage the censored data.

Question 2: How does the performance of the multi-task learning approach compare to the single-task learning approach? To this end, we compare our proposed multi-task version of the model (**MTL-RankSvx**) to our single-task version (**RankSvx**).

We evaluate the models using the following metrics:

Concordance index (CI): CI is one of the most commonly used metrics for survival models. It can be interpreted as the fraction of all pairs of patients, the order of whose predicted response matches the order of their observed response. CI is defined as $\mathrm{CI} = \frac{1}{N_{test}} \sum_{\forall c_i=1}^{t_i} \sum_{t_j>t_i} \mathbf{1}_{f(x_i)<f(x_j)}$ where N_{test} is the number of comparable pairs in the test dataset.

Mean Absolute Error (MAE): MAE is defined as the average of the differences between predicted time values and the actual observed event times $\mathrm{MAE} = \frac{1}{N} \sum_{\forall c_i=1} |t_i - \hat{t}_i|$.

Training and testing We randomly sample 67% of the cohort as training data, and we use the remaining 33% hold out for testing. All the models are implemented with gradient descent optimization and and we apply the Adam (Kingma and Ba 2014) method to automatically adapt the step size

Table 4: Performance comparisons between proposed RankSvx model and previous approaches in terms of concordance index (CI) in single-task learning setting.

Method		RET	NEU	NEP	VAS	HYPER	avg
Survival	Cox	0.5552	**0.6107**	0.6027	0.6092	0.5524	0.5860
	Weibull	0.5066	0.5207	0.5699	0.5399	0.5790	0.5432
	Log-Logistic	0.5108	0.5217	0.5821	0.5483	0.5833	0.5492
	Log-normal	0.5082	0.5241	0.5806	0.5497	0.5799	0.5485
Regression	Squared	0.5205	0.5217	0.5100	0.4830	0.5246	0.5120
	Poisson	0.5643	0.5244	0.5424	0.4510	0.4862	0.5137
	Log-normal	0.4764	0.5709	0.5332	0.5495	0.4481	0.5156
RankSvx	Squared	0.5569	0.5643	0.6164	0.5946	0.5974	0.5859
	Poisson	**0.5650**	0.6078	0.6361	**0.6220**	0.5687	0.5999
	Log-normal	0.5613	0.6026	**0.6405**	0.6111	**0.6050**	**0.6041**

Table 5: Performance comparisons between proposed RankSvx model and previous approaches in terms of mean absolute error (MAE) in single-task learning setting. Errors are measured in months.

Method		RET	NEU	NEP	VAS	HYPER	avg
Survival	Cox	15.1983	16.6252	14.8260	13.0178	9.0817	13.7498
	Weibull	6.6241	7.1144	6.9690	6.6553	6.6980	6.8122
	Log-Logistic	5.8822	6.7371	6.7977	6.5462	6.8298	6.5586
	Log-normal	5.5834	6.7408	6.7068	6.2985	**5.8953**	6.2450
Regression	Squared	6.4013	6.5058	6.4329	6.3967	7.8003	6.7074
	Poisson	6.3484	6.5252	**6.6112**	6.5280	6.3421	6.4710
	Log-normal	6.4669	6.3368	6.2769	**6.1314**	6.6093	6.3643
RankSvx	Squared	5.8519	6.3734	6.7831	6.5283	6.7692	6.4612
	Poisson	5.9522	6.4550	6.7879	6.5900	8.0567	6.7683
	Log-normal	**5.5133**	**6.1209**	6.7235	6.2654	5.9280	**6.1102**

during the parameter estimation procedures. We use grid search for parameter tuning and report the best result for each model.

Prior risk association Ω_0 Note that our model can incorporate prior knowledge on complication association through Ω_0. We construct associations leveraging the human disease network (Goh et al. 2007) 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 Ω_0 that represents our prior knowledge about the correlations between the T2DM complications in our study.

Result Comparisons

In this subsection, we present the comparisons on the two metrics (CI and MAE) between our proposed models and the baseline methods. Note that CI measures relative risk ranking and MAE measures event times prediction accuracy.

Comparing RankSvx with previous approaches Table 4 and Table 5 show the comparisons between our RankSvx model with previous approaches in terms of CI and MAE respectively. From Table 4 we can see that across most complications RankSvx (with different loss functions) outperforms the Cox model, parametric survival models, and regression models. Survival models perform better than regression models as they can handle censored data. The semi-parametric Cox model can achieve better CI performances than their parametric peers. However, as shown in Table 5, parametric survival models can achieve much better event time prediction performance in terms of MAE. This is because the Cox model optimizes the partial likelihood objective, which depends only on the relative ordering of the survival time of individuals but not on their actual values. RankSvx model can simultaneously achieve best performances in both metrics on average. In particular, RankSvx model with Log-normal loss function performs the best.

Comparing MTL-RankSvx with RankSvx We next compare the performance of multi-task learning against single-task learning. As the RankSvx model with Log-normal loss function performs the best, we compare MTL-RankSvx with RankSvx using the Log-normal loss function. We would expect MTL perform better when there are some tasks, whose information is not enough to learn the model, can benefit from the correlation from the other tasks. For this consideration, for each task we respectively use 25%, 50%, 75% and 100% of dataset while keep other tasks unchanged. We compare the performances of MTL-RankSvx and of RankSvx in this setting. Table 6 shows the comparisons between MTL-RankSvx with STL-RankSvx. We can see that MTL-RankSvx can improve STL-RankSvx in most cases. Further, we observe that when the number of training samples is small, the task can better improve its performance through multi-task learning framework. For example, we can observe more improvement of HYPER and VAS, which are the two complications with fewest training samples.

Discussion While we observed that MTL-RankSvx can improve STL-RankSvx in most cases, the improvements seem not to be significant. We found that the learned task

Table 6: Comparisons between MTL-RankSvx and STL-RankSvx. We compare the MTL and STL models by setting different percentage of dataset in each task.

(a) concordance index (CI)

	25%		50%		75%		100%	
	STL	MTL	STL	MTL	STL	MTL	STL	MTL
ERT	0.5468	0.5509	0.5603	0.5604	0.5623	0.5628	0.5613	0.5652
NEU	0.5798	0.5797	0.5895	0.5885	0.5968	0.5959	0.6026	0.6054
NEP	0.6069	0.6093	0.6262	0.6273	0.6343	0.6350	0.6405	0.6425
VAS	0.5821	0.5920	0.6015	0.6036	0.6104	0.6112	0.6111	0.6170
HYPER	0.5406	0.5517	0.5977	0.5993	0.6010	0.6034	0.6050	0.6098
avg	0.5712	0.5767	0.5951	0.5958	0.6010	0.6016	0.6041	0.6080

(b) mean absolute error (MAE)

	25%		50%		75%		100%	
	STL	MTL	STL	MTL	STL	MTL	STL	MTL
ERT	5.5161	5.5072	5.5254	5.5245	5.5161	5.5167	5.5133	5.5273
NEU	6.1282	6.1288	6.1294	6.1351	6.1465	6.1505	6.1209	6.1338
NEP	6.7388	6.7446	6.7110	6.7075	6.7009	6.7020	6.7235	6.7290
VAS	6.2387	6.2447	6.2561	6.2586	6.2537	6.2566	6.2654	6.2723
HYPER	6.1777	6.1534	5.9211	5.9153	5.9351	5.9241	5.9280	5.9183
avg	6.1599	6.1557	6.1086	6.1082	6.1105	6.1100	6.1102	6.1161

association matrices were close to diagonal matrices, indicating that the tasks did not have high association with each other. Since the fundamental idea of multitask learning is to leverage association among multiple tasks, it is expected that MTL may not have significant improvement over STL when the associations are not strong. The finding that the tasks did not have high association with each other is a bit counter intuitive. One reason could lie in the relatively short observation window of the dataset. It is possible that some preexisting complications were treated as new onsets. In this case, associations between different T2DM complications could be reduced.

Related Work

From an applications perspective, our work falls into the category of research that apply predictive analytics and use EHRs to improve the practice of healthcare management (Yadav et al. 2015). Building predictive models from EHR records have attracted significant attention from both academia and industry, and have been used in disease onset prediction (Ng et al. 2016), patient stratification (Wang et al. 2015; Chen et al. 2016), hospital readmission prediction (He et al. 2014), and mortality prediction (Tabak et al. 2013; Nori et al. 2015). More recently, Razavian *et al.* (Razavian et al. 2015) show that EHRs can be leveraged to predict T2DM onset. Oh *et al.* (Oh et al. 2016) applied EHRs to capture the trajectories of T2DM patients and found that different trajectories can lead to different risk patterns. To the best of our knowledge, this paper presents the first study to investigate the early prediction of T2DM complications from EHRs.

Technically, our work is related to survival analysis. Survival analysis (Cox 1972; Miller Jr 2011) is a class of widely used statistical tools to model time-to-event. However traditional survival analysis models have several limitations. The widely used Cox model (Cox 1972) does not directly model the event probability; instead, it maximizes the partial likelihood objective, which depends only on the relative ordering of the survival time of individuals, not on their actual values. It will require another cumbersome procedure to fit a non-parametric survival function after the coefficients of Cox model are determined for prediction purposes. Therefore, Cox based models are limited for the task of predicting the survival time for individual patients (Yu et al. 2011). Parametric survival models (Lawless 1998) are another popular alternative. These methods assume that the baseline hazard function follows some distribution, such as Exponential, Weibull and Log-normal. However, the distribution might not flexible enough to capture the complex event patterns in real practice. As a result, there is a need for machine learning based survival models (Wang, Li, and Reddy 2017) which are free from rigorous statistical assumptions. An important distinction between our method and prior methods is that our proposed approach has the advantage of simultaneously achieving two important desiderata: accurate prediction of event times, as well as an accurate ranking of the relative risks of two patients by simultaneously optimizing both objective functions.

Our work is also related to multi-task learning (MTL) (Caruana 1997), which aims to jointly learn multiple tasks using a shared representation so that knowledge obtained from one task can help other tasks. In particular, our work falls into the category of *task relation learning* based MTL approaches (Zhang and Yang 2017) due to its flexibility to incorporate prior information (Zhang and Yeung 2010; Sun, Wang, and Hu 2015). There are some previous research (Li et al. 2016) that apply multi-task learning for survival analysis, however, they are different from our work in that they study single-task survival analysis through the multi-task leaning framework by decomposing event timeline into multiple time windows.

Conclusion and Future Work

In this paper, we proposed a novel survival analysis approach, in which models were learned from historical EHR records, to predict when a patient will develop complications after being diagnosed with T2DM. The proposed survival approach has the advantage to achieve two important metrics: accurate prediction of event times and good ranking of the relative risks of two patients. Moreover, to better capture the correlations of time-to-events of multiple complications, we further developed a multi-task version of the survival model. Finally, extensive experiments on a T2DM patient dataset extracted from a large healthcare claims database validated the performance of our new proposed survival analysis and demonstrated the effectiveness of the multi-task framework.

There are a number of interesting future research directions. First, we only used basic demographic information and static ICD codes in our evaluation. Incorporating more features or new feature representations can potentially improve prediction performance. Second, it is important to not only predict complication events but also to analyze and identify the important associated risk factors. Finally, we are also interested in adapting our models to other chronic diseases and other types of electronic health record data.

References

[American Diabetes Association and others 2013] American Diabetes Association and others. 2013. Standards of medical care in diabetes 2013. *Diabetes care* 36(Suppl 1):S11.

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

[Caruana 1997] Caruana, R. 1997. Multitask learning. *Mach. Learn.* 28(1):41–75.

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

[Cox 1972] Cox, D. 1972. Regression models and life-tables. *J. of the Royal Statistical Society (B)* 34(2):187–220.

[Forbes and Cooper 2013] Forbes, J. M., and Cooper, M. E. 2013. Mechanisms of diabetic complications. *Physiological reviews* 93(1):137–188.

[Goh et al. 2007] Goh, K.-I.; Cusick, M. E.; Valle, D.; Childs, B.; Vidal, M.; and Barabási, A.-L. 2007. The human disease network. *Proceedings of the National Academy of Sciences* 104(21):8685–8690.

[He et al. 2014] He, D.; Mathews, S. C.; Kalloo, A. N.; and Hutfless, S. 2014. Mining high-dimensional administrative claims data to predict early hospital readmissions. *Journal of the American Medical Informatics Association* 21(2):272–279.

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

[Lawless 1998] Lawless, J. 1998. Parametric models in survival analysis. *Encyclopedia of Biostatistics*.

[Li et al. 2016] Li, Y.; Wang, J.; Ye, J.; and Reddy, C. K. 2016. A multi-task learning formulation for survival analysis. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1715–1724. ACM.

[Miller Jr 2011] Miller Jr, R. G. 2011. *Survival analysis*, volume 66. John Wiley & Sons.

[Mittal et al. 2013] Mittal, S.; Madigan, D.; Cheng, J. Q.; and Burd, R. S. 2013. Large-scale parametric survival analysis. *Statistics in medicine* 32(23):3955–3971.

[Ng et al. 2016] Ng, K.; Steinhubl, S. R.; Dey, S.; Stewart, W. F.; et al. 2016. Early detection of heart failure using electronic health records. *Circulation: Cardiovascular Quality and Outcomes* 9(6):649–658.

[Nori et al. 2015] Nori, N.; Kashima, H.; Yamashita, K.; Ikai, H.; and Imanaka, Y. 2015. Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 855–864.

[Oh et al. 2016] Oh, W.; Kim, E.; Castro, M. R.; Caraballo, P. J.; Kumar, V.; Steinbach, M. S.; and Simon, G. J. 2016. Type 2 diabetes mellitus trajectories and associated risks. *Big data* 4(1):25–30.

[Raykar et al. 2007] Raykar, V. C.; Steck, H.; Krishnapuram, B.; Dehing-Oberije, C.; and Lambin, P. 2007. On ranking in survival analysis: Bounds on the concordance index. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, 1209–1216.

[Razavian et al. 2015] Razavian, N.; Blecker, S.; Schmidt, A. M.; Smith-McLallen, A.; Nigam, S.; and Sontag, D. 2015. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data* 3(4):277–287.

[Rendle et al. 2009] Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, 452–461.

[Sun, Wang, and Hu 2015] Sun, Z.; Wang, F.; and Hu, J. 2015. Linkage: An approach for comprehensive risk prediction for care management. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 1145–1154.

[Tabak et al. 2013] Tabak, Y. P.; Sun, X.; Nunez, C. M.; and Johannes, R. S. 2013. Using electronic health record data to develop inpatient mortality predictive model: Acute laboratory risk of mortality score (alarms). *Journal of the American Medical Informatics Association* 21(3):455–463.

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

[Wang, Li, and Reddy 2017] Wang, P.; Li, Y.; and Reddy, C. K. 2017. Machine learning for survival analysis: A survey. *arXiv preprint arXiv:1708.04649*.

[World Health Organization 2016] World Health Organization. 2016. Global report on diabetes.

[Yadav et al. 2015] Yadav, P.; Steinbach, M.; Kumar, V.; and Simon, G. 2015. Mining electronic health records (ehr): A survey. *Department of Computer Science and Engineering*.

[Yu et al. 2011] Yu, C.-N.; Greiner, R.; Lin, H.-C.; and Baracos, V. 2011. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *Advances in Neural Information Processing Systems*, 1845–1853.

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

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