Survival Prediction by an Integrated Learning Criterion on Intermittently Varying Healthcare Data

Jianfei Zhang[1], **Lifei Chen**[2], **Alain Vanasse**[3], **Josiane Courteau**[3], **Shengrui Wang**[1]
[1]PROSPCTUS Lab, Department of Computer Science, University of Sherbrooke, Canada
[2]School of Mathematics and Computer Science, Fujian Normal University, China
[3]PRIMUS Research Group, Department of Family Medicine, University of Sherbrooke, Canada
{jianfei.zhang, alain.vanasse, josiane.courteau, shengrui.wang}@usherbrooke.ca clfei@fjnu.edu.cn

Abstract

Survival prediction is crucial to healthcare research, but is confined primarily to specific types of data involving only the present measurements. This paper considers the more general class of healthcare data found in practice, which includes a wealth of intermittently varying historical measurements in addition to the present measurements. Making survival predictions on such data bristles with challenges to the existing prediction models. For this reason, we propose a new semi-proportional hazards model using locally time-varying coefficients, and a novel complete-data model learning criterion for coefficient optimization. Experiments on the healthcare data demonstrate the effectiveness and generalizability of our model and its promise in practical applications.

Introduction

Survival prediction in healthcare research examines the time that elapses from the beginning of follow-up until the event (also "failure") of interest occurs. This event may be adverse, for instance, biological death, readmission; or sometimes be beneficial, such as hospital discharge and healing of a wound. Basically, the goal of survival prediction is to generate prognostic models for understanding disease processes (Ghassemi et al. 2015), exploring interaction between prognostic factors (Khosla et al. 2010), and predicting how new patients will behave in the context of known data (Hong and Hauskrecht 2015). The employment of survival prediction would allow clinicians to answer patients' queries as *when and how probable that a certain disease will recur*, *how long they are likely to live*, and *how well they will respond to a specific therapy*. These can be crucial in healthcare, to assess lifestyle appropriateness and make early decisions on treatment and sometimes on end-of-life care.

Usually, one predicts the clinical outcome (typically, survival time) for patients using measurements of various prognostic factors collected during the follow-up period (e.g., 1-year study). These factors include demographic variable such as age, gender, race, etc, and diagnostic information like blood test results, tumor size, hemoglobin level, etc (Lin et al. 2011). Exploring the simultaneous effects of these factors on survival time is of practical interest for researchers and clinicians, and has been the subject to much research. Broadly, such research attempts to qualify the effects by learning regression coefficients for factors from their present measurements (PresMeas). For example, (non-)parametric models (Jenkins 2005) and semi-parametric Cox proportional hazards models (Cox 1972; Vinzamuri and Reddy 2013; Yu et al. 2008) generally aim at time-independent coefficients, while Aalen additive hazards models (Gaïffas and Guilloux 2012) and logistic regression models (Lin et al. 2011; Liu et al. 2010) estimate coefficients that vary with time, i.e., time-varying coefficients.

Nowadays, the general healthcare data collected by ever-increasing electronic health record (EHR) databases includes a wealth of unaligned historical measurements (HistMeas) due to the patients' checkups at various time intervals (Beirne, Clarkson, and Worthington 2007). For example, of two patients, say p_1 and p_2, diagnosed with COPD (Chronic Obstructive Pulmonary Disease) routinely scheduled for review, p_1 may undergo a monthly checkup while p_2 is examined bimonthly. This means we will acquire a COPD data for which p_2 has intermittent measurements of prognostic factors in contrast to p_1's monthly measurements. Such COPD data is a typical type of intermittently varying (IV) data which is quite general in healthcare. Besides, COPD patients are probably readmitted to hospital frequently due to the need for an emergency treatment; this, as a consequence, makes COPD data involve more intermittent measurements made at unexpected emergency time intervals.

IV data has by far outpaced the processing and analytical capacities of the aforementioned models, mainly because: 1) the observation of PresMeas for a patient indicates (s)he was alive all the time in history, and therefore estimating the survival probabilities at historical measurement time snapshots does not make sense, as stated in (Kalbfleisch and Prentice 2011); 2) the process of estimating coefficients at each snapshot is conducted on only part of patients in the context of IV HistMeas; this may lead to a biased estimate that cannot adequately fit the whole body of data, especially without information exploited from HistMeas for coefficient optimization. Much work such as (Moghaddass and Rudin 2014; Chen et al. 2014; Cortese and Andersen 2010) has demonstrated the significance of HistMeas to survival prediction. This in turn suggests the need for a study on HistMeas in IV data. To our knowledge, there is as yet no documented work concerned with such study in survival prediction.

For this concern, we propose in this paper a simple yet effective semi-proportional hazards (SPH for short) model that uses locally time-varying coefficients, thereby relaxing the proportional hazards assumption for the sake of practicality, while retaining that model's simplicity. To acquire those coefficients, we develop an integrated model learning criterion that includes an objective function based on maximum likelihood of failure and censoring, and, simultaneously, an optimization constraint based on the hazard trajectory explored from HistMeas. Besides, a regularization is applied to prevent overfit arising from model learning. We investigate SPH model on an IV dataset derived from a COPD data collected from Centre Hospitalier Universitaire de Sherbrooke (CHUS). The major contributions of this paper include:

- A new prediction model against IV data, where survival prediction on such data has not been well studied previously.
- A novel criterion for a complete-data learning, which makes full use of present measurements (PresMeas) and historical measurements (HistMeas) in conjunction.
- An application to a healthcare problem of interest by predicting the risk and survival of COPD.

Preliminaries and Related Work

This section will formalize the notion of IV data and then briefly review a recent line of work on survival prediction.

Intermittently Varying Data

Given an IV dataset composed of N time-to-failure labeled individuals, denoted by $\mho = \{(y_i, \delta_i, \mathcal{Z}_i \cup \mathbf{x}_i)\}_{i=1}^{N}$, one may observe some individuals fail at K distinct failure times, $0 < t_1 < t_2 < \cdots < t_K$; some other individuals may drop out at failure times, and the remaining individuals may be still alive right after t_K. For individual i, the failure indicator, δ_i, takes value 1 if failure occurs and 0 otherwise. Accordingly, the observed time y_i represents his/her failure time T_i if $\delta_i = 1$, and (right-)censoring time C_i otherwise. PresMeas of the V prognostic factors made at y_i are recorded as $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iV})$. Prior to failure/dropout, a set of HistMeas, $\mathcal{Z}_i = \{\mathbf{z}_i[\tau] : 0 < \tau < y_i\}$, are made at different time snapshots, where $\mathbf{z}_i[\tau] = (z_{i1}[\tau], \ldots, z_{iV}[\tau])$. This notation can apply to time-independent data as well, as long as $z_{iv}[\tau] \equiv x_{iv}$ when factor v does not change over time.

Figure 1: An example of IV data. A yellow dot means the patient is still alive while a red dot indicates a failure

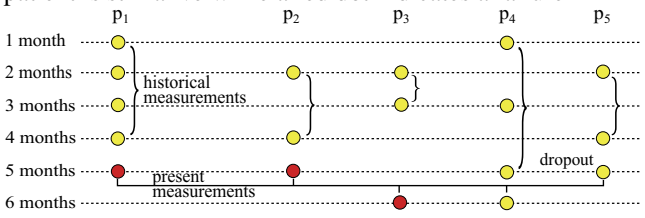

Figure 1 shows an example of IV data in which measurement time snapshots are different over the five patients during the 6-month follow-up. Specifically, each patient has a few HistMeas at different time snapshots from other patients; that is, the measurements are irregularly made and thus unaligned. Patients p_1, p_2 and p_3 experienced a failure while p_4 and p_5 were censored. In this circumstance, we have failure times $t_k \in \{5 \text{ months}, 6 \text{ months}\}$. All PresMeas were made at the failure times. Patients p_1 and p_2 were tied due to having the same failure time (5 months). Patient p_4 was at risk (i.e., still alive) at the end of follow-up, whereas p_5 dropped out at 5 months.

Survival Prediction

The goal of survival prediction is to forecast the failure time T (i.e., how long to survive) of each individual $\mathbf{x}$ from a certain population. Usually, one adopts a survivor function $S(t|\mathbf{x}) = \Pr(T \geq t|\mathbf{x})$ to identify the probability of being still alive at time t (this refers to an observed failure time in training data, e.g., 5 or 6 months in Figure 1). This function depends fully on the hazard function

$$h(t|\mathbf{x}) = \lim_{\Delta t \downarrow 0^+} \frac{\Pr(t \leq T \leq t + \Delta t | T \geq t; \mathbf{x})}{\Delta t},$$

which assesses the instantaneous rate of failure at t, conditional on survival to that time. It can be seen that the greater the value of $h(t)$, the greater the risk of failure at t.

The Cox model (Cox 1972) determines the hazard in a multiplicative manner: $h(t|\mathbf{x}) = h_0(t) \exp[f(\mathbf{x})]$, where $h_0(t)$ represents an unspecified baseline hazard in the context of $\mathbf{x} = (0, \ldots, 0)$ and the link function $f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$. The $\boldsymbol{\beta}$ means a vector of time-independent regression coefficients. By contrast, the Aalen model (Aalen 1989) assumes an additive hazard such that $h(t|\mathbf{x}) = h_0(t) + f(\mathbf{x})$. The logistic regression model estimates the probability of surviving beyond t by means of $\Pr(T \geq t|\mathbf{x}) = (1 + \exp[\mathbf{x}^\top \boldsymbol{\beta} + \mathbf{x}i])^{-1}$ with $\mathbf{x}i$ a threshold. In practice, the prognostic factors' effects (e.g., the effect of a treatment) may change over time with longer follow-up (Fisher and Lin 1999). To accommodate such situations, there was a surge of interest in learning time-varying coefficients $\boldsymbol{\beta}(t)$ instead of $\boldsymbol{\beta}$; examples include (Song and Wang 2013; Sun, Sundaram, and Zhao 2009; Lin et al. 2011).

Returning to the example presented in Figure 1, the red dot for p_1 at 5 months tells us the survival probabilities satisfying $S(1 \text{ month}|\mathrm{p}_1) = \cdots = S(4 \text{ months}|\mathrm{p}_1) \equiv 1$, revealing the difficulties of making use of HistMeas in prediction. Alternatively, those models may learn coefficients by approximating hazards, rather than survival probabilities, for the five patients at each month. In doing so, the estimate of coefficients may fit p_1 well, but not four others who have intermittent measurements, e.g., the measurements for p_2 are available at only 2 months and 4 months.

Our Approach

This section will introduce a new semi-proportional hazards model and its learning approach.

Semi-Proportional Hazards Model

To allow the effect of prognostic factors to vary with time, our approach assigns K classes of coefficients, denoted by

$\mathbf{B} \triangleq (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K)^\top$, for identifying the contributions of factors to risk at different failure times, where the coefficients at t_k are given by $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kV})$. The link function of individual i can thus be redefined as

$$\begin{cases} f_{\mathbf{B}}(\mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\beta}_{k|t_k=y_i} & \text{for PresMeas} \\ f_{\mathbf{B}}(\mathbf{z}_i[\tau]) = \mathbf{z}_i[\tau]^\top \boldsymbol{\beta}_{k|t_k=y_i} & \text{for HistMeas} \end{cases}$$

By this approach, individuals who fail (or drop out) at the same time are specified with the same coefficients, as they may obtain the same (or similar) therapeutic effect from a treatment (Grundy et al. 1999); on the other hand, the HistMeas and PresMeas for an individual are assigned the same coefficients, since the interaction between factors for a given individual does not change. This *locally time-varying-coefficient* setting partly relaxes the proportional hazards assumption. At this point, the hazards are seemingly proportional between tied individuals (and between histories of an individual), and non-proportional between non-tied individuals. Hence, we call such hazards *semi-proportional*.

With the above hazard function, the survival probability of individual $\mathbf{x}_i$ at t can be calculated by

$$S(t|\mathbf{x}_i; \mathbf{B}) = \big(\exp[-H_0(t)]\big)^{\exp[f_{\mathbf{B}}(\mathbf{x}_i)]}.$$

Here, the cumulative baseline hazard $H_0(t) = \int_0^t h_0(u)du$ can be rewritten by a Breslow's estimator (Breslow 1974) in the presence of tied failure times, as follows:

$$H_0(t) = \sum\nolimits_{t_k \le t} h_0(t_k) = \sum\nolimits_{k:t_k \le t} \frac{\big|\mathcal{D}_k\big|}{\sum_{j \in \mathcal{R}_k} \exp\big[f_{\widehat{\mathbf{B}}}(\mathbf{x}_j)\big]}.$$

which is a step function with jumps at failure times. The risk set $\mathcal{R}_k \triangleq \{\forall i : y_i \ge t_k\}$ includes those individuals at risk of failure at t_k, and $\mathcal{D}_k \triangleq \{\forall i : T_i = t_k\}$ contains those who fail at t_k, with $\big|\mathcal{D}_k\big|$ the cardinality. We denote by $\mathcal{R}_k^+ = \mathcal{R}_k / \mathcal{C}_k$ the set of individuals who are still alive right after t_k.

Objective Function based on PresMeas

The only important question so far unaddressed is how to learn $\mathbf{B}$. A straightforward way provided by Cox-type models is to maximize the so-called partial likelihood (Cox 1975; Sun, Sundaram, and Zhao 2009). We thus rewrite such likelihood for the individuals who fail, in the form:

$$L_{\mathcal{D}}(\mathbf{B}|\mho) = \prod_{\substack{t_k:\text{failure} \\ \text{time}}} \prod_{\substack{i \text{ fails}}} \frac{\Pr(i \text{ fails at } t_k)}{\Pr(j \in \mathcal{R}_k \text{ fails at } t_k)} \qquad (1)$$

$$= \prod_{k=1}^{K} \frac{\exp\big[\sum_{i \in \mathcal{D}_k} f_{\mathbf{B}}(\mathbf{x}_i)\big]}{\Big(\sum_{j \in \mathcal{R}_k} \delta_j \exp[f_{\mathbf{B}}(\mathbf{x}_j)]\Big)^{|\mathcal{D}_k|}}.$$

The key idea here is to compare the risk of failure between the individuals who fail and those who are still at risk. Since the tied individuals are in the minority, Eq. 1 employs the Peto's modification on Efron approximation (Hertz-Picciotto and Rockhill 1997).

It is worthy of note that Eq. 1 embodies the idea behind most existing methods (Vinzamuri, Li, and Reddy 2014; Yu et al. 2008), claiming that censoring data are non-informative due to unobserved failure times and it is thus desirable to eliminate these "useless" data from the model, so long as we keep track of the risk set $\mathcal{R}$. However, censoring data do actually provide some information: failures occurred after the censoring times. Needless to say, a mass of significant information would vanish were we to exclude this data. In view of this, we wish to assess the partial likelihood of censoring that can be derived from Eq. 1, yielding

$$L_{\mathcal{C}}(\mathbf{B}|\mho) = \prod_{\substack{t_k:\text{failure} \\ \text{time}}} \prod_{\substack{i \text{ drops} \\ \text{out}}} \frac{\Pr(i \text{ drops out at } t_k)}{\Pr(j \in \mathcal{R}_k \text{ drops out at } t_k)}$$

$$= \prod_{k=1}^{K} \frac{\exp\big[\sum_{i \in \mathcal{C}_k} f_{\mathbf{B}}(\mathbf{x}_i)\big]}{\Big(\sum_{j \in \mathcal{R}_k} (1-\delta_j) \exp[f_{\mathbf{B}}(\mathbf{x}_j)]\Big)^{|\mathcal{C}_k|}},$$

with $\mathcal{C}_k = \{\forall i : C_i = t_k\}$ those who drop out at t_k.

An optimal $\mathbf{B}$ should allow the values of both $L_{\mathcal{D}}$ and $L_{\mathcal{C}}$ to be large. The learning procedure thus comes down to maximizing the two likelihoods in conjunction. To simplify the algebraic manipulations, we perform such maximization by minimizing the negative log partial likelihoods:

$$\ell_{\mathcal{D}}(\mathbf{B}) = \sum_{k=1}^{K} \left(\big|\mathcal{D}_k\big| \log \sum_{j \in \mathcal{R}_k} \delta_j e^{f_{\mathbf{B}}(\mathbf{x}_j)} - \sum_{i \in \mathcal{D}_k} f_{\mathbf{B}}(\mathbf{x}_i) \right)$$

$$\ell_{\mathcal{C}}(\mathbf{B}) = \sum_{k=1}^{K} \left(\big|\mathcal{C}_k\big| \log \sum_{j \in \mathcal{R}_k} (1-\delta_j) e^{f_{\mathbf{B}}(\mathbf{x}_j)} - \sum_{i \in \mathcal{C}_k} f_{\mathbf{B}}(\mathbf{x}_i) \right).$$

Now, the objective function to be minimized falls out as

$$\ell(\mathbf{B}) = \ell_{\mathcal{D}}(\mathbf{B}) + b\ell_{\mathcal{C}}(\mathbf{B}), \qquad (2)$$

where the tuning parameter, b, in effect, trades off the two likelihoods in prediction, given by Laplace smoothing such that $b = \frac{|\mathcal{C}_k|+1}{|\mathcal{C}_k|+|\mathcal{D}_k|+K}$.

Optimization Constraint based on HistMeas

The major drawback to the objective function in Eq. 2 is that only PresMeas are taken into account. As a result, the learning process may yield an inaccurate estimate for the otherwise HistMeas. Exploiting the information contained in HistMeas and imposing this information on Eq. 2 as a constraint can be a naive approach to addressing this problem. However, predicting the cumulative incidence and survival probability by using HistMeas has turned out to be no longer feasible (Kalbfleisch and Prentice 2011), because the observation of $\mathcal{Z}_i$ tells us that individual i is alive at $\tau(< t)$. Thus, we have

$$S(t|\mathcal{Z}_i) = \Pr(T \ge t | \{\mathbf{z}_i[\tau] : 0 < \tau < t\}) \equiv 1.$$

Moreover, the baseline survivor function has no simple interpretation, as argued in (Wang 2004). We therefore turn to the instantaneous hazards that are still obtainable.

The hazards in parametric models (Jenkins 2005) are usually assumed to be drawn from some specific distributions,

Table 1: Changes in the hazards with time, in the context of different survival distributions

Distribution	hazard function
Exponential	constant
Weibull (shape parameter p)	constant if $p = 1$; increasing if $p > 1$; decreasing if $p < 1$
Gamma (shape parameter α)	constant if $\alpha = 1$; concave, increasing if $\alpha > 1$; convex, decreasing if $\alpha < 1$
Log-Normal	increasing and then decreasing

as shown in Table 1. In practice, however, hazard functions under those strong assumptions should not be expected to be suitable for all possible data. Were it not for those assumptions, one might see that the hazard at failure time can be interpreted as maximal. In other words, for the history at τ, the more similar to the PresMeas, the higher the hazard at τ. For a pairwise HistMeas of individual i, say $\mathbf{z}_i[\ddot{\tau}]$ and $\mathbf{z}_i[\dot{\tau}]$, then, the corresponding hazards satisfy

$$h(t|\mathbf{z}_i[\ddot{\tau}]) \geq h(t|\mathbf{z}_i[\dot{\tau}])$$
$$\text{if } \Delta_i[\ddot{\tau}, \dot{\tau}] \ = Sim(\mathbf{z}_i[\ddot{\tau}]) - Sim(\mathbf{z}_i[\dot{\tau}]) \geq 0,$$

where $Sim(\cdot)$ represents the similarity between $\mathbf{x}_i$ and its HistMeas. This study utilizes an inner product metric (Alipanahi et al. 2008) as the similarity measure. Some algebraic manipulations on these hazards at t_k yield the following:

$$\Delta_i[\ddot{\tau}, \dot{\tau}]\big(\mathbf{z}_i[\dot{\tau}] - \mathbf{z}_i[\ddot{\tau}]\big)^\top \boldsymbol{\beta}_{k|t_k=y_i} \leq 0.$$

When we sort all $|\mathcal{Z}_i|$ HistMeas by their similarity to $\mathbf{x}_i$, only $|\mathcal{Z}_i|-1$ pairs (with removal of redundant pairs) are required. For late ease of use, we denote by Q_i a collection of these pairs and then group them by tied failure times. The grouped HistMeas for those tied individuals thus satisfy

$$G(\boldsymbol{\beta}_k) = \sum_{i:y_i=t_k} \sum_{[\ddot{\tau},\dot{\tau}]\in Q_i} \Delta_i[\ddot{\tau}, \dot{\tau}]\big(\mathbf{z}_i[\dot{\tau}] - \mathbf{z}_i[\ddot{\tau}]\big)^\top \boldsymbol{\beta}_k \leq 0.$$

The key to our study is then to minimize Eq. 2 in the context of $G(\boldsymbol{\beta}_k) \leq 0$ for all k. In other words, we aim to maximize the likelihood of PresMeas (including failure and censoring) subject to the constraint derived from the pairwise hazards of HistMeas. In doing so, the coefficients can be optimized so as to fit all measurements well. The rationale for this approach lies partly in the fact that clinicians diagnose patients and provide them with treatments on the basis of medical records (i.e., histories) in addition to the present symptom (Hong and Hauskrecht 2015).

Complete-data Model Learning

Prior to minimizing the objective function in Eq. 2, it is prudent to consider that such minimization may lead to overfitting and poor generalizability of the prediction model. For this reason, one should add a 'capacity' to overcome the possible overfitting tendency, as suggested in (Vinzamuri, Li, and Reddy 2014; Yang et al. 2011). For this capacity, we employ a "ridge" penalty criterion (Verweij and Van Houwelingen 1994) over $\|\boldsymbol{\beta}_k\|_2^2$ to prevent overfit arising. Additionally, some unknown external factors in clinical trials may act on patients such that some of them fail far too early or live far too long. These patients are so-called outliers that may lead the learning process to generate inaccurate coefficients. To reduce sensitivity to these outliers, we attempt to smooth the coefficients so that they vary smoothly across consecutive time points, as described in (Lin et al. 2011). To sum up, the optimal coefficients can be formally defined as the solution to the following problem:

$$\min_{\mathbf{B}} J(\mathbf{B}) \ = \ell(\mathbf{B}) + \lambda_1 \sum_{k=1}^{K} \|\boldsymbol{\beta}_k\|_2^2 + \lambda_2 \sum_{k=1}^{K-1} \|\boldsymbol{\beta}_{t_{k+1}} - \boldsymbol{\beta}_{t_k}\|_2^2$$
$$\text{s.t. } G(\boldsymbol{\beta}_k) \leq 0 \ \ \forall k = 1, 2, \ldots, K \tag{3}$$

Note that, the constraint given by Eq. 3 is fundamental to our approach. By means of this constraint, the underlying hazard trajectory can be explored to further calibrate the coefficients. Yet, the existing models optimize their own objective functions without such a constraint.

A closer look at the above nonlinear programming problem reveals that J is convex and differentiable, derived from the statements presented in (Moghaddass and Rudin 2014), and then the optimal estimate, $\widehat{\mathbf{B}}$, can be a dual solution satisfying the KKT (Karush-Kuhn-Tucker) conditions:

$$J(\widehat{\mathbf{B}}) = \min_{\mathbf{B}} J(\mathbf{B}) = \min_{\mathbf{B}} \max_{\boldsymbol{\mu}} \Big(J(\mathbf{B}) + \sum\nolimits_{k=1}^{K} \mu_k G(\boldsymbol{\beta}_k)\Big)$$
$$\mu_k G(\widehat{\boldsymbol{\beta}}_k) = 0 \ \ \& \ \ \frac{\partial J}{\partial \boldsymbol{\beta}_k}|_{\boldsymbol{\beta}_k=\widehat{\boldsymbol{\beta}}_k} = (0, \ldots, 0)^V,$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)$ with $\mu_k \geq 0$ for all k. We apply the EM algorithm based on downhill simplex method (Navon, Phua, and Ramamurthy 1988) to implement this optimization. The iterative learning process begins with a warm-start $\boldsymbol{\beta}_{t_1} = (0.5, \ldots, 0.5)^V$ and runs until convergence, i.e., until the change of coefficients between two successive iterations is smaller than 10^{-3}.

Experiments

We analyzed and evaluated our approach by comparative experiments on an IV COPD data.

Data and Pre-processing

The original data involves the hospitalization records, but not sufficient prognostic factors, for COPD patients in CHUS during 2012–2013. We extracted and stimulated the measurements on 35 factors for 451 patients based on the original data rather than used it directly. Of those patients, 427 (The failure rate of COPD is usually not up to 94.8%; a high rate is to allow the models which discard the censoring data to perform effectively.) experienced a failure (demise or readmission) and the remaining 24 were considered as censoring. Of those factors, 16 are in binary format, indicating the presence or absence of a particular symptom or sign. Data imputation was used to remedy the omissions: we filled in the missing continuous values through a linear regression presented in (Kim, Golub, and Park 2004), and chose to use

the most frequent value as the default fill-in for the otherwise missing binary values. In order to obtain a meaningful set of coefficients, all continuous values were normalized within the 0 to 1 range via min-max-scaling. Yet, converting binary variables to continuous variables is not required.

Setup

Comparative experiments were designed to study the behavior of our approach against three state-of-the-art models:

- MTLR: a logistic regression model (non-Cox), which builds the survival function through multi-task learning (Lin et al. 2011). The regularizers of this model were estimated via tenfold cross-validation (10CV) in our study.
- KW-Cox: a Cox-type model based on a kernel-weighted partial likelihood. The Epanechnikov kernel function with a bandwidth value of 1.5 was adopted in the experiment, based on the results presented in (Liu et al. 2010).
- SE-Cox: a Cox-type model based on a smoothing empirical likelihood. The kernel bandwidth was set as $N^{-0.2} = 0.3$, as suggested in (Sun, Sundaram, and Zhao 2009).

Note that CV can be also a way of choosing the bandwidth parameters. Since the three models were not inherently designed for the scenario of ties, we modified their likelihood functions to render them applicable to the COPD data. In addition, to provide deeper insight into the functionality of SPH, three reduced versions were designed for comparison.

- SPH-H disposes of all HistMeas via the removal of Eq. 3;
- SPH-C discards all censoring data, thereby excluding the likelihood $\ell_{\mathcal{C}}(\mathbf{B})$ from Eq. 2;
- SPH-S requires the coefficient for each factor to be a constant. Hence, the smoothing is not necessary for $J(\mathbf{B})$.

We rewrote three metrics below for model evaluation.

- Survival AUC (S-AUC), which evaluates performance on the binary COPD classification task. It qualifies the ability of a model to answer the question: *Is COPD likely to recur in patient i within one year*? The AUC (the area under the ROC curve) is redefined as ($\mathbf{1}$ is the indicator function):

$$\text{S-AUC} = \frac{\sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \mathcal{D}_k} \sum_{\mathbf{x}_j \in \mathcal{R}_k^+} \mathbf{1}_{S(t_k|\mathbf{x}_i) < S(t_k|\mathbf{x}_j)}}{\sum_{k=1}^{K} |\mathcal{C}_k| \cdot |\mathcal{R}_k^+|}.$$

- Survival Concordance Index (S-CI), which gives an estimate of how well the output of a model matches the relative time-to-failure for all pairs of patients that can actually be ordered, that is, how accurately the model can answer the question: *Is COPD more likely to recur in patient i or patient j*? In our case, it formally becomes (The n_{pair} is the number of comparable pairs of patients.)

$$\text{S-CI} = \frac{1}{n_{pair}} \sum_{y_i > T_j} \sum_{k: t_k > T_j} \mathbf{1}_{S(t_k|\mathbf{x}_i) > S(t_k|\mathbf{x}_j)}.$$

- Survival Mean Square Error (S-MSE), which measures the quality of survival probability predictions, i.e., prediction accuracy. S-MSE answers the question: *How accurate is the diagnosis that COPD will recur in patient i*? We referred to (Vinzamuri, Li, and Reddy 2014) and modified the MSE for our use, as follows:

$$\text{S-MSE} = \frac{1}{N} \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \mathcal{D}_k \cup \mathcal{C}_k} \left(\delta_i - h_0(t_k) e^{f_{\mathbf{B}}(\mathbf{x}_i)}\right)^2.$$

In all experiments we report generalized 10CV results, over 100 replicates, in the form of *mean* $\pm$ *standard deviation*. The regularization parameters of SPH, λ_1 and λ_2, were selected by another 10CV on the training data.

Results

Figure 2 shows the changes in cumulative hazards (left) and survival probabilities (right) with time when different models are applied to predict a COPD patient who was readmitted to hospital within 25 days. Two interesting differences can be observed: 1) SPH and MTLR are able to produce more smooth curves of cumulative hazards and survival probabilities than SE-Cox and KW-Cox, because the use of regularization in SPH and MTLR allows the coefficients to vary smoothly over time; 2) Superior to the other models, SPH clearly shows that the patient was at quite a high cumulative hazard as early as about the 10th day, especially given that the survival probability at that time drops down to only 14.7%. This patient could thus be issued a flareup, indicating that COPD would soon recur, and offered advice on timely treatment. In clinical trials, this is crucial for COPD patients who are likely to have an acute exacerbation on those days.

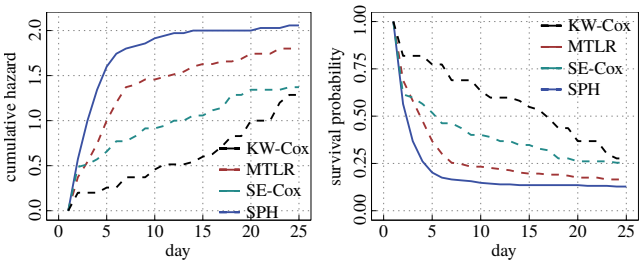

Figure 2: Change in cumulative hazard (left) and survival probability (right) for a COPD patient over 25 days (i.e., the period from discharge to readmission). Note: *the survival probability*= $\exp$(*the negative cumulative hazard*)

In Figure 3, we evaluate S-AUC and S-CI to compare the prediction ability of our approach against others. SPH and MTLR perform better than KW-Cox and SE-Cox since they make full use of data collected from the censoring COPD patients and can thus acquire the exact coefficients to qualify the interaction between the prognostic factors over all patients, not just those patients who have an exact time to readmission. The performances of KW-Cox and SE-Cox are characterized by large variances partly due to the difficulties in choosing their bandwidth parameters (Liu et al. 2010; Sun, Sundaram, and Zhao 2009). Overall, thanks to the findings on the hazard trajectory and the learning criterion leveraging such findings, SPH achieves significantly better performance in terms of S-AUC and S-CI, which demonstrates that our model can more effectively predict the progression

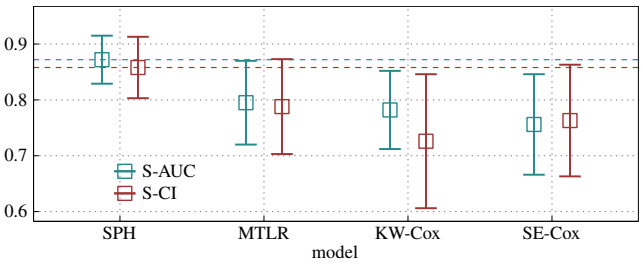

Figure 3: Comparison of different models' performances in terms of S-AUC and S-CI

of COPD. It thus appears that SPH is a good aid to clinicians to help with early diagnosis and treatment.

To gain a better understanding of the merits of our approach, we compared the performance of SPH and its three reduced versions in the context of a varying train/test ratio. It can be seen from Figure 4 that SPH outperforms all of the reduced versions, especially in the training-data-deficient case, testifying to the powerful generalizability of SPH. This reveals SPH's potential capacity to adapt to healthcare data, since the basic argument is that identification of survival time for a patient may be quite expensive or time-consuming in practice, and the known data for analysis is generally much less extensive than the unknown data, i.e., small training data and large test data (Vinzamuri, Li, and Reddy 2014). The comparison between SPH and SPH-H indicates that a mass of useful information is concealed in the HistMeas. As the percentage of training patients grows, we see a downward trend to SPH-C, mainly because more and more censoring patients would be discarded and thus the coefficients generated by SPH-C give rise to overfitting on failure data and underfitting on censoring data.

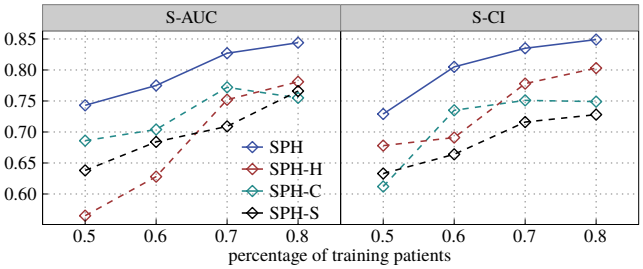

Figure 4: Comparison of SPH and its reduced versions in terms of S-AUC and S-CI, and changes in their performance with varying percentage of training patients

Figure 5 shows the changes in coefficients with respect to the two factors FEV_1 (forced expiratory volume per second) and pack-years (of smoking) during the 52-week follow-up. SPH-S uses time-independent coefficients; in contrast, SPH can yield more smoothly varying coefficients and, more importantly, prevent the factors' effects on prognosis from reversing since, unlike SPH-H and SPH-C, the coefficients generated by SPH for each factor do not switch between positive and negative values. From the medical point of view, FEV_1 has a protective effect against COPD and thus should always be assigned negative coefficients.

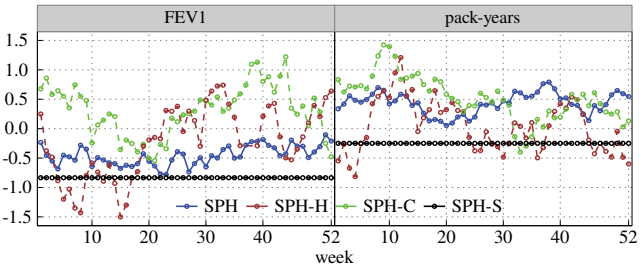

Figure 5: Comparison of changes in coefficients with time

Table 2 presents S-MSE of predicted survival probabilities of patients at 52 weeks and also the 25% lower quantile (13 weeks), median (26 weeks) and 75% upper quantile (39 weeks) of the follow-up period. It can be seen from the table that SPH scores a clear win over other models, yielding much more accurate (lower errors) predictions on survival probability. The semi-proportional hazards approach makes our model more flexible in handling the underlying relationship between tied patients and also the combination of HistMeas and PresMeas. In view of the outstanding performance of SPH in terms of S-MSE, we assert that the use of our approach can enhance the confidence of survival prediction.

Table 2: S-MSE of predicted survival probabilities in various periods of follow-up. **Bold** emphasis indicates superior performance of one model over the others. The numbers in parentheses are the corresponding standard errors

Model	13 weeks	26 weeks	39 weeks	52 weeks
MTLR	4.34(.197)	5.36(.202)	3.41(.370)	6.81(.186)
KW-Cox	3.97(.140)	4.77(.208)	6.12(.363)	6.32(.295)
SE-Cox	5.32(.233)	4.63(.166)	4.26(.183)	5.09(.262)
SPH	**2.56(.083)**	**3.19(.113)**	**3.24(.074)**	**3.88(.128)**
SPH-H	3.27(.178)	3.93(.176)	4.33(.259)	4.28(.243)
SPH-C	5.11(.108)	4.73(.249)	3.90(.355)	4.79(.299)
SPH-S	4.25(.196)	5.82(.253)	4.65(.332)	5.85(.274)

Conclusions

IV data poses a variety of challenges that the existing survival prediction models cannot handle. In this paper, we proposed an effective semi-proportional hazards model with locally time-varying coefficients for a task of survival prediction on IV data. Our main goal was to learn and optimize the coefficients. For this purpose, we designed an integrated complete-data model learning criterion in which the failure and censoring data were encompassed by the objective function and, simultaneously, the historical data were used to build an optimization constraint. Comparative experiments on a COPD data have demonstrated the outstanding performance of our approach and its capacity to undertake survival prediction on general IV healthcare data. We will conduct

further research on IV data including sequence pattern mining in order to explore the underlying patterns of prognostic factors from patients' historical data.

Acknowledgments

We would like to thank Mireille Courteau for useful discussions. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the program PAFI of Centre de recherche du CHUS, and National Natural Science Foundation of China under Grant No. 61175123. Jianfei Zhang was also financed by China Scholarship Council (CSC).

References

Aalen, O. O. 1989. A linear regression model for the analysis of life times. *Stat. Med.* 8(8):907–925.

Alipanahi, B.; Biggs, M.; Ghodsi, A.; et al. 2008. Distance metric learning vs. fisher discriminant analysis. In *AAAI*, 598–603.

Beirne, P. V.; Clarkson, J. E.; and Worthington, H. V. 2007. Recall intervals for oral health in primary care patients. *The Cochrane Library*.

Breslow, N. 1974. Covariance analysis of censored survival data. *Biometrics* 89–99.

Chen, Q.; May, R. C.; Ibrahim, J. G.; Chu, H.; and Cole, S. R. 2014. Joint modeling of longitudinal and survival data with missing and left-censored time-varying covariates. *Stat. Med.* 33(26):4560–4576.

Cortese, G., and Andersen, P. K. 2010. Competing risks and time-dependent covariates. *Biom. J.* 52(1):138–158.

Cox, D. R. 1972. Regression models and life tables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 34:187–220.

Cox, D. R. 1975. Partial likelihood. *Biometrika* 62(2):269–276.

Fisher, L. D., and Lin, D. Y. 1999. Time-dependent covariates in the Cox proportional-hazards regression model. *Annu. Rev. Public Health* 20(1):145–157.

Gaïffas, S., and Guilloux, A. 2012. High-dimensional additive hazards models and the lasso. *Electron. J. Stat.* 6:522–546.

Ghassemi, M.; Pimentel, M. A.; Naumann, T.; Brennan, T.; Clifton, D. A.; Szolovits, P.; and Feng, M. 2015. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In *AAAI*, 446–453.

Grundy, S. M.; Pasternak, R.; Greenland, P.; Smith, S.; and Fuster, V. 1999. Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations: a statement for healthcare professionals from the american heart association and the american college of cardiology. *J. Am. Coll. Cardiol.* 34(4):1348–1359.

Hertz-Picciotto, I., and Rockhill, B. 1997. Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics* 1151–1156.

Hong, C., and Hauskrecht, M. 2015. Multivariate conditional anomaly detection and its clinical application. In *AAAI*, 4239–4240.

Jenkins, S. P. 2005. *Survival analysis*. Chapter 3.

Kalbfleisch, J. D., and Prentice, R. L. 2011. *The statistical analysis of failure time data*, volume 360.

Khosla, A.; Cao, Y.; Lin, C. C.-Y.; Chiu, H.-K.; Hu, J.; and Lee, H. 2010. An integrated machine learning approach to stroke prediction. In *KDD*, 183–192.

Kim, H.; Golub, G. H.; and Park, H. 2004. Imputation of missing values in dna microarray gene expression data. In *CSB*, 572–573.

Lin, H.-c.; Baracos, V.; Greiner, R.; and Chun-nam, J. Y. 2011. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *NIPS*, 1845–1853.

Liu, M.; Lu, W.; Shore, R. E.; and Zeleniuch-Jacquotte, A. 2010. Cox regression model with time-varying coefficients in nested case–control studies. *Biostatistics* 11(4):693–706.

Moghaddass, R., and Rudin, C. 2014. The latent state hazard model, with application to wind turbine reliability. *Ann. Appl. Stat.*

Navon, I. M.; Phua, P. K.; and Ramamurthy, M. 1988. Vectorization of conjugate-gradient methods for large-scale minimization. In *Supercomputing*, 410–418.

Song, X., and Wang, C.-Y. 2013. Time-varying coefficient proportional hazards model with missing covariates. *Stat. Med.* 32(12).

Sun, Y.; Sundaram, R.; and Zhao, Y. 2009. Empirical likelihood inference for the Cox model with time-dependent coefficients via local partial likelihood. *Scand. J. Stat.* 36(3):444–462.

Verweij, P. J., and Van Houwelingen, H. C. 1994. Penalized likelihood in Cox regression. *Stat. Med.* 13(23-24):2427–2436.

Vinzamuri, B., and Reddy, C. K. 2013. Cox regression with correlation based regularization for electronic health records. In *ICDM*, 757–766.

Vinzamuri, B.; Li, Y.; and Reddy, C. K. 2014. Active learning based survival regression for censored data. In *CIKM*, 241–250.

Wang, W. 2004. Proportional hazards regression models with unknown link function and time-dependent covariates. *Stat. Sinica* 14(3):885–906.

Yang, Y.; Shen, H. T.; Ma, Z.; Huang, Z.; and Zhou, X. 2011. $\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*, 1589–1594.

Yu, S.; Fung, G.; Rosales, R.; Krishnan, S.; Rao, R. B.; Dehing-Oberije, C.; and Lambin, P. 2008. Privacy preserving Cox regression for survival analysis. In *KDD*, 1034–1042.