2016 IEEE 16th International Conference on Data Mining

Predicting COPD Failure by Modeling Hazard in Longitudinal Clinical Data

Jianfei Zhang[1], Shengrui Wang[1,*], Josiane Courteau[2], Lifei Chen[3,1], Aurélien Bach[4,1], Alain Vanasse[3,5]
[1]*Département d'Informatique, Université de Sherbrooke, Canada*
[2]*Département de Médecine de Famille, Université de Sherbrooke, Canada*
[3]*Department of Mathematics and Computer Sciences, Fujian Normal University, China*
[4]*cole Nationale Suprieure d'Informatique pour l'Industrie et l'Entreprise (ENSIIE), France*
[5]*Centre de Recherche du Centre Hospitalier Universitaire de Sherbrooke (CHUS), Québec, Canada*
[*]*corresponding author*
Emails: firstname.lastname@usherbrooke.ca

***Abstract*—Chronic obstructive pulmonary disease (COPD) accounts for the highest rate of hospital readmissions and is the third leading cause of death in Canada, the United States and worldwide. Predicting COPD failure provides a prognostic warning of death or readmission, and is crucial to early intervention and decision-making. The aim of this study is to perform COPD failure prediction on longitudinal data. To address the inappropriate estimation of Cox hazard in current approaches, we propose a new representation of hazard to capture the relationship between survival probability and time-varying risk factors in a concise but effective way. To optimize model parameters, we design and maximize a new joint likelihood that comprises two components used to estimate survival status separately for failure and censored patients. A regularized optimization is performed on the joint likelihood to prevent overfitting arising from model learning. Our approach is applied to a real-life COPD data set and outperforms the current state-of-the-art prediction models in terms of the survival AUC, concordance index and Birer score metrics; this reveals that the great promise of our approach for clinical prediction.**

I. Introduction

COPD is preventable and treatable, but not fully reversible. It accounts for the highest rate of readmissions to hospital and the third leading cause of death in Canada, the United States and worldwide [1]. Many of these patients could be adequately managed if warning of adverse clinical events, e.g., biological death and hospital readmission which are generally called failure, was provided early, before their worsening. This is the motivation for the work reported here, whose aim is to perform survival analysis for predicting failure for COPD patients. Such failure prediction is important for health service providers, because it provides useful information for clinicians to use in deciding when to modulate the treatment according to the severity of the patient's disease, so as to prevent avoidable hospital admission or early death.

Failure prediction in survival analysis examines the risk of failure over the time elapsing from the beginning of follow-up until the event of interest occurs. Basically, the goal of failure prediction is to generate prognostic models for understanding disease processes [2], exploring interactions between risk factors [3, 4], and predicting how new patients will behave in the context of known data [5, 6]. The use of a failure prediction allows clinicians to answer patients' queries on probable outcomes in time. Based on the predictive outcomes, for example, a doctor may tell a 70-year-old patient that her survival probability is 60% one year after discharge and 50% at 3 years. Such predictions can be crucial in the choice of treatments, lifestyle modifications and sometimes end-of-life care measures.

Thus far, failure prediction has often been performed on longitudinal clinical data, which have become ubiquitous with the growing use of real-time monitoring and tests in clinical trials [7]. The main characteristic of such data is that, while some disease-specific risk factors are static, others may vary with time. Each of these time-varying risk factors comprises a sequence of observations. A considerable number of approaches have been developed to make the traditional predictive models applicable to such longitudinal data with the time-varying risk factors. Examples include alternatives to the Cox model [8], joint models [9], competing risk models [10], etc. Generally, given a sequence of observations for a time-varying factor, these approaches estimate a Cox hazard at a certain time based on the regression analysis solely on the observation for the factor at that time.

However, none of these approaches provides an effective solution to understanding the clinical relationship between survival probability and time-varying risk factors, because they perform prediction under the "unrealistic" tacit assumption that the hazard to survival at a certain time is dependent only on the observation at that time, that is, independent of the previous observations. The major clinical reason for this "unrealistic" assumption is that the hazard to survival is bound up with all clinical conditions. For example, the observed measurement (i.e., observation) for SpO2 (a COPD-specific risk factor associated with a clinical test on blood oxygen level) may be beyond the normal level on the first day and then return to normal on the second day. SpO2 is a typical time-varying risk factor and estimates of the hazard on the second day should take into account the observations for SpO2 on both days, rather than only on the second day (i.e., the most

2374-8486/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDM.2016.20

recent observation). The recent work presented in [4, 6] also implied that the previous observations are associated with the hazard and therefore useful for predicting survival probability. For these reasons, modeling an effective representation of hazard to accurately capture the relationship between survival probability and all observations for time-varying factors is necessary for failure prediction.

In light of the above arguments, this study aims to derive a new representation of hazard for failure prediction by mining the time-varying risk factors in longitudinal data. In view of the flexibility of the Cox model in survival analysis, we develop a Cox-like model, which incorporates a novel accumulative hazard, while retaining the model's simplicity. To optimize the model parameters, i.e., the regression coefficients of risk factors, we design a new joint likelihood to address the ubiquitous (right-)censoring concealed in clinical data, thus allowing the model to make use of both failure and censored data. Furthermore, a regularized optimization is performed to prevent overfitting arising from model learning. We investigate our approach, i.e., accumulative hazard based joint likelihood (AHJ for short), on 503 patients diagnosed with COPD who were hospitalized at Centre Hospitalier Universitaire de Sherbrooke (CHUS) between the fiscal years 2012 and 2013.

To summarize, the contributions of this paper include:

- A novel representation of hazard, i.e., accumulative hazard, which captures the relationship between survival probability and time-varying factors of longitudinal data in a concise but effective way.
- A newly designed joint likelihood comprising two components that estimate survival status separately for failure and censored patients; this allows us to make full use of the data and thus reduces "information loss" in the learning procedure.
- Comprehensive experimental testing of the proposed model for prediction and classification on a cohort of real-life COPD patients, yielding results which demonstrate that our approach significantly outperforms the current state-of-the-art models and reveal its promise in clinical applications.

The remainder of the paper is structured as follows. section II describes the failure prediction and the Cox-like model. section III presents our approach and section IV provides the experimental results and evaluation of our approach. Finally, we conclude in section V.

II. Preliminaries and Related Work

This section sets out the clinical premises for failure prediction and introduces the Cox model.

A. Time-to-failure prediction in survival analysis

In healthcare applications of interest, readmission is typically measured from an index hospital discharge, which is discharge from the first hospitalization for a particular clinical condition; subsequent admissions within a specified time period after index discharge are considered readmissions. Without loss of generality, we have used the following definitions:

- Failure for a patient is defined as either the first readmission after index discharge within the follow-up period, or death within the follow-up period if this patient was not readmitted to hospital prior to death. We are interested in combined events (death and readmission), a major issue of investigation in care trajectory studies [11].
- Survival indicates a patient still at risk of failure, that is, a patient who has not yet experienced a failure. Survival time is equivalent to failure time, since the time when a patient fails determines how long (s)he survives.
- Censoring means that failure has not occurred by the end of follow-up (supposing that no patients drop out of follow-up). In other words, a censored patient remains alive at the end of follow-up. Censoring time is thus the length of follow-up for this patient.

Failure prediction aims to forecast the probability that a patient, described by $\mathbf{x}=(x^{(1)},x^{(2)},\cdots,x^{(V)})\in\mathbb{R}^V$, a V-dimensional vector of observations for V risk factors, is still at risk of failure (that is, survival) at a prespecified time point t, given as $S(t|\mathbf{x}):=\Pr(y\geq t|\mathbf{x})$. Here, y is a scalar outcome (e.g., failure or censoring time). This probability is usually called survival probability. Usually, one can assess the survival probability at t, $S(t)$ through the corresponding hazard, which is defined as conditional on survival to time t, as follows:

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

B. Cox proportional hazards model

The Cox proportional hazards model is widely adopted in clinical studies and is used extensively in survival analysis. It defines the hazard at t given $\mathbf{x}$ for a patient as

$$h(t|\mathbf{x})=h_0(t)\exp\{\boldsymbol{\beta}\cdot\mathbf{x}\}, \tag{1}$$

where $h_0(t)$ represents an arbitrary baseline hazard in the context of $\mathbf{x}:=(0,\ldots,0)^V$, and $\boldsymbol{\beta}\in\mathbb{R}^V$ is the vector of parameters, i.e., regression coefficients, to be estimated for the Cox model. This model is known as a semi-parametric model because the baseline hazard function is treated non-parametrically, making the hazards proportional. In other words, there is no requirement to specify the distribution of the survival times.

The Cox model is in the generalized linear model (GLM) family. We can see that the parameters have a multiplicative effect on the hazard value which makes this approach different from other members of this family, such as Aalen's additive model [12] with the hazard $h(t|\mathbf{x})=h_0(t)+\exp\{\boldsymbol{\beta}\cdot\mathbf{x}\}$ and the logistic regression model [13], where the probability of

surviving beyond t takes the form

$$\Pr(y \geq t|\mathbf{x}) = \frac{1}{1+\exp\{\boldsymbol{\beta}\cdot\mathbf{x}\}} \quad (2)$$

There are alternatives to the Cox model that extend to the time-varying factors, generally using the hazard in the form

$$h(t|\mathbf{x}^t) = h_0(t)\exp\{\boldsymbol{\beta}\cdot\mathbf{x}^t\}, \quad (3)$$

with $\mathbf{x}^t$ the V observations for V risk factors at time t. Examples include the time-dependent Cox model (TDCox) [8] and the competing risk model (CR) [10].

All these models estimate the hazard at a given time point, say time t, using only the observations for factors at t while leaving out those observations before t. In contrast, we will introduce a new representation of hazard that takes all observations into account.

III. Our Approach

In this section, we present an approach for failure prediction by modeling accumulative hazard, a new representation of hazard, in longitudinal data. We first outline the longitudinal data structure.

A. Observed longitudinal data structure

In order to tackle general clinical estimation problems, we define the structure of sequence of observations in longitudinal data within the predefined time frame as $\mathcal{X}^t = \{\mathbf{x}^\tau : 0 \leq \tau \leq t\}$, where t is the duration of the time frame. In real-world situations, one can observe patient i's factors only at the observed time points of ϕ_i^t up to time t. Thus, we have patient i's sequence of observations up to time t as $\mathcal{X}_i^t = \{\mathbf{x}_i^\tau : \forall\tau \in \phi_i^t\}$, which may be time-varying or static if $\mathbf{x}_i^\tau \equiv \mathbf{x}_i$. In a clinical setting, the maximum duration of the time frame is given as y, which is either a failure time (i.e., survival time) or a censoring time. (This censoring situation may be due to drop-out or the end of follow-up.) In light of the above definitions, the observed data structure can be written as $(\mathcal{X}^y, \epsilon)$, where ϵ is the failure indicator, equal to 1 if the patient fails and 0 otherwise.

We define the data set as $\mathcal{D}$ which consists of the censored patients $\mathcal{C} \triangleq \{i : \epsilon_i = 0\}$ and the failure patients $\mathcal{F} \triangleq \{i : \epsilon_i = 1\}$. Given a time t, $\mathcal{R}^t \triangleq \{i \in \mathcal{D} : y_i \geq t\}$ contains those patients still at risk at t, and $\mathcal{F}^t \triangleq \{i \in \mathcal{F} : y_i = t\}$ contains those who fail at t while $\mathcal{F}^{<t} \triangleq \{i \in \mathcal{F} : y_i < t\}$ prior to t.

B. Representation of accumulative hazard

Before developing a representation of hazard, let's look more closely at the hazard used in the time-varying Cox models, i.e., $h(t|\mathbf{x}^t)$ given by Equation 3. Without loss of generality, our hazard would take the form of $h(t|\mathcal{X}^t)$, i.e., the hazard at t given the sequence of observations up to t. To model the hazard at time t as a function of such sequence, we postulate an accumulative hazard given $\mathcal{X}^t$ in the form[1]

$$h(\mathcal{X}^t) \triangleq h(t|\mathcal{X}^t) \triangleq \int_0^t \gamma(u,t)h_0(u)\exp\{\boldsymbol{\beta}\cdot\mathbf{x}^u\}\mathrm{d}u.$$

Here, $h_0(*)$ is the baseline hazard and $\gamma(u,t)$ stands for the decay ratio of the hazard. By such decay, we can model the amount of the hazard produced by the observations at u remaining at time $t(\geq u)$. This can be an exponential function of time:

$$\gamma(u,t) \triangleq \exp\{\xi(u-t)\}. \quad (4)$$

Simply, we take the ξ value of 1 and thus $0 < \gamma(*) \leq 1$. Note that such positive decay ratio indicates that the hazard will shrink over time but not vanish.

Since the measurements are made at discrete times, the accumulative hazard in the discrete form can be given as

$$\begin{aligned} h(\mathcal{X}^t) &\triangleq \sum\nolimits_{\tau\in\phi^t} \gamma(\tau,t)h_0(\tau)\exp\{\boldsymbol{\beta}\cdot\mathbf{x}^\tau\}\delta^\tau \\ &= h(\mathcal{X}^{t-\delta^{\tau_1}}) + \gamma(t,t)h_0(t)\exp\{\boldsymbol{\beta}\cdot\mathbf{x}^t\}\delta^t \\ &= h(\mathcal{X}^{t-\delta^{\tau_2}}) + \gamma(t-\delta^{\tau_1},t)h_0(t-\delta^{\tau_1}) \\ &\quad \exp\left\{\boldsymbol{\beta}\cdot\mathbf{x}^{t-\delta^{\tau_1}}\right\}\delta^{\tau_1} + h_0(t)\exp\{\boldsymbol{\beta}\cdot\mathbf{x}^t\}\delta^t. \end{aligned} \quad (5)$$

Each τ (i.e., $\tau_1, \tau_2, \ldots$) and the specific t in Equation 5 are the time points corresponding to the sequence of observations. δ^τ is the duration between time τ and its previous observed time point.

With the hazard defined by Equation 5, the survival probability of a given patient at t, $\Pr(y \geq t|\mathcal{X}^t)$, can be calculated by the survival function

$$S(\mathcal{X}^t) = \exp\left\{-\int_0^t h(\mathcal{X}^u)\mathrm{d}u\right\}.$$

To provide the baseline hazard for regression, we formulate an estimate of our baseline accumulative hazard, based on the Breslow's estimator [14], in the context of ties between failure times, as follows:

$$h_0(\mathcal{X}^t) = \frac{\left|\mathcal{F}^t\right|}{\sum_{i\in\mathcal{R}^t}\left(|\phi_i^t|^{-1}\sum_{\tau\in\phi_i^t}\gamma(\tau,t)\exp\{\boldsymbol{\beta}\cdot\mathbf{x}_i^\tau\}\right)}.$$

When optimal coefficients $\widehat{\boldsymbol{\beta}}$ and an optimal baseline accumulative hazard $\widehat{h}_0$ are found, for a test patient $\mathcal{X}_{test}^t$, the survival probability is predicted as

$$S(\mathcal{X}_{test}^t) = \exp\left\{-\sum\nolimits_{i:\mathcal{F}^t\cup\mathcal{F}^{<t}} \widehat{h}_0(\mathcal{X}_{test}^{y_i})\exp\{\widehat{\boldsymbol{\beta}}\cdot\mathbf{x}_{test}^t\}\right\}.$$

[1]For the sake of readability and simplicity, from here on, we have intentionally simplified $t|\mathcal{X}^t$ to $\mathcal{X}^t$, and dropped the subscript i in the concise format unless otherwise stated.

C. Estimation of regression coefficients

The only important question left unaddressed so far is how to estimate $\boldsymbol{\beta} \in \mathbb{R}^V$. In a proportional hazards scheme, a straightforward way of computing $\boldsymbol{\beta}$ is to maximize the partial likelihood [15], in which the baseline hazard has been dropped out. Our baseline accumulative hazard at t, $h_0(\mathcal{X}^t)$, cannot be dropped out because it incorporates all these hazards over the observed time points, and consequently, estimating $\boldsymbol{\beta}$ by maximizing a complex conditional hazard becomes intractable. Instead, we turn to the likelihood of a sequence of failure statuses (i.e., inverse survival statuses). For patient i, the sequence can be described without loss of generality as

$$s_i \triangleq [0,\ldots,0,\epsilon_i],$$

where ϵ_i indicates the failure status at time y_i. The length of s_i is given by the number of possible observed time points, i.e., $|\phi_i^{y_i}|$. Basically, in a time-to-failure prediction setting, the probability (actually, likelihood) that s_i appears given a $\boldsymbol{\beta}$ will be in two categories:

$$\Pr(s_i|\boldsymbol{\beta}) = \begin{cases} \Pr([0,\ldots,0,1]|\boldsymbol{\beta}), & \text{if } \epsilon_i = 1 \\ \Pr([0,\ldots,0,0]|\boldsymbol{\beta}), & \text{otherwise (i.e., } \epsilon_i = 0) \end{cases}$$

Next, we estimate $\boldsymbol{\beta}$ by maximizing the above probability (i.e., likelihood).

Before we move on to the maximization, let's look more closely at the idea behind most current Cox-based methods [2, 3, 8, 16], which claims that censored patients are implicitly informative and that it is thus desirable to eliminate these "useless" data from the model, keeping track of only $\Pr([0,\ldots,0,1]|\boldsymbol{\beta})$. However, censored patients do actually provide some information: failures will occur after the censoring times, although the exact failure times are unobserved. Consequently, a mass of significant information will vanish if we exclude such censored data. For this reason, we wish to assess the likelihood for censored patients. That is, $\Pr([0,\ldots,0,0]|\boldsymbol{\beta})$ will be maximized as well in the maximization procedure. Now, we define a joint likelihood in a multiplicative manner such that

$$\ell_{joint}(\mathcal{D}|\boldsymbol{\beta}) \triangleq \ell(\mathcal{F}|\boldsymbol{\beta}) \times \ell(\mathcal{C}|\boldsymbol{\beta}),$$

where the two components $\ell(\mathcal{F}|\boldsymbol{\beta})$ and $\ell(\mathcal{C}|\boldsymbol{\beta})$ are independent of each other and can be given by the product of the probabilities, that is,

$$\begin{aligned}
\ell(\mathcal{F}|\boldsymbol{\beta}) &= \prod_{i:\epsilon_i=1} \Pr(s_i|\boldsymbol{\beta}) = \prod_{i:\epsilon_i=1} \Pr([0,\ldots,0,\epsilon_i]|\boldsymbol{\beta}) \\
&= \prod_{i \text{ fails}} \Pr(i \text{ survives prior to } y_i|\boldsymbol{\beta}) \times \Pr(i \text{ fails at } y_i|\boldsymbol{\beta}) \\
&= \prod_{i \text{ fails}} \exp\Big\{-\int_0^{<y_i} h(\mathcal{X}_i^t)\mathrm{d}t\Big\}\Big(1-\exp\Big\{-\int_0^{y_i} h(\mathcal{X}_i^t)\mathrm{d}t\Big\}\Big) \\
&= \prod_{i\in\mathcal{F}} \prod_{t\in\phi_i^{<y_i}} \exp\{-h(\mathcal{X}_i^t)\delta^t\}\Big(1-\exp\{-h(\mathcal{X}_i^{y_i})\delta^t\}\Big)
\end{aligned}$$

and

$$\begin{aligned}
\ell(\mathcal{C}|\boldsymbol{\beta}) &= \prod_{i:\epsilon_i=0} \Pr(s_i|\boldsymbol{\beta}) = \prod_{i:\epsilon_i=0} \Pr([0,\ldots,0,\epsilon_i]|\boldsymbol{\beta}) \\
&= \prod_{i \text{ is censored}} \Pr(i \text{ survives at } y_i|\boldsymbol{\beta}) \\
&= \prod_{i \text{ is censored}} \exp\Big\{-\int_0^{y_i} h(\mathcal{X}_i^t)\mathrm{d}t\Big\} \\
&= \prod_{i\in\mathcal{C}} \prod_{t\in\phi_i^{y_i}} \exp\{-h(\mathcal{X}_i^t)\delta^t\}.
\end{aligned}$$

Here, $\phi^{<y_i}$ means the observed time points prior to y_i. To simplify the algebraic manipulations, the maximization can be performed by minimizing the following negative log joint likelihood

$$-\log\ell_{joint}(\mathcal{D}|\boldsymbol{\beta}) = -\log\ell(\mathcal{F}|\boldsymbol{\beta}) - \log\ell(\mathcal{C}|\boldsymbol{\beta}). \tag{6}$$

D. Optimization for regularized coefficients

Prior to minimizing the objective function in Equation 6, it is prudent to consider that such minimization may lead to overfitting and poor generalizability of the prediction model. Hence, one should add a 'capacity' to overcome the possible overfitting tendency, as suggested in [17]. For this capacity, we employ a "ridge" penalty criterion over $\|\boldsymbol{\beta}\|_2^2$ to prevent overfitting arising. To sum up, we have the objective function of the regularized likelihood, as follows:

$$\mathcal{L}(\boldsymbol{\beta}) \triangleq -\log\ell_{joint}(\mathcal{D}|\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_2^2. \tag{7}$$

This function is strictly convex and differentiable when $\lambda \geq 0$. The optimal coefficients that perform well on previously unseen test patients can be formally defined as the solution to the minimization problem, i.e.,

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}).$$

Algorithm 1 Gradient Descent Learning

Choose an arbitrary start point $\boldsymbol{\beta}^{(m)}$ when $m = 0$
repeat
  Compute a descent direction $\frac{\partial\mathcal{L}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(m)}}$
  Search $\eta^* = \operatorname{argmin}_{\eta\geq 0}\mathcal{L}\Big(\boldsymbol{\beta}^{(m)} + \eta\frac{\partial\mathcal{L}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(m)}}\Big)$
  Set $m = m+1$
  Update $\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m-1)} + \eta^*\frac{\partial\mathcal{L}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(m)}}$
until $\mathcal{L}(\boldsymbol{\beta}^{(m)}) - \mathcal{L}(\boldsymbol{\beta}^{(m-1)}) < \varepsilon$
return $\boldsymbol{\beta}^{(m-1)}$

We apply the gradient descent algorithm, shown in Algorithm 1, to implement minimization. The iterative process can start with $\boldsymbol{\beta}^{(0)} := (0.5,\cdots,0.5)^V$, and then compute a descent direction. The iteration runs until a user-defined criterion that determines the convergence of Algorithm 1 when the change of regularized likelihood between the two successive iterations is smaller than a threshold ε. Here, we set $\varepsilon = 10^{-4}$ and limit the maximum iterations to 100.

IV. EXPERIMENTS

In this section, we analyze our approach and evaluate its predictive power by comparative experiments on a real-life COPD data set.

A. COPD data and clinical setup

We were able to collect from a regional hospital EMR the data related to 503 COPD patients who were (re)admitted to CHUS between the fiscal years 2012 and 2013. Table I provides an overview on the demographic characteristics of the patients. Of these patients, 63% were readmitted within 1 year for COPD or other diseases such as digestive disorders, infectious illnesses, neoplasms, etc. Given that frequent readmission constitutes the most important characteristic of COPD patients, we focused our study on the combined events of readmission and death within one year. The data set contains a total of 328 failure patients, consisting of 10 patients who died and 318 who were readmitted. Note that, we thought of the first event of interest as a failure, and consequently, many of these 318 readmitted patients also died within one year. Due to the 1-year follow-up after index discharge, failure is specified as 1-year failure in this study and those who were censored have a 365-day censoring time.

Table I
DEMOGRAPHICS, I.E., POPULATION, OF 503 COPD PATIENTS HOSPITALIZED AT CHUS BETWEEN THE YEARS 2012 AND 2013.

Risk Factor	Observation	Number of patients (%)
gender	Female, Male	259 (51.5%), 244
age (years)	$\geq$75, <75	232 (46.1%), 271
admitting service	Family Medicine	215 (42.7%)
	Internal Medicine	109 (21.7%)
	Pulmonology	155 (30.8%)
	others	24 (4.8%)
# of 1-year readmissions	0 (no readmission)	185 (37%)
	1, 2, $\geq$3	141, 77, 100 (total 63%)
failure of interest	readmission, death	318, 10

The COPD-specific risk factors shown in Table II, ranging from demographics and healthcare information, clinical test and diagnosis to medication for treatment, were selected by the research group at CHUS. We employed a weighted effect coding [18] to create dummy factors from the binary (B) and categorical (C) factors so that they could be directly entered into a regression. All numerical (N) factors were normalized to the range [0,1]. Only time-varying factors involve a few missing values (i.e., observations) because of a small number of patients without information on these factors. We filled in the missing numerical values by means of a linear regression introduced in [19]. In effect coding, we can choose to use "0" as the default fill-in over dummy factors for the missing values on categorical and binary factors.

B. Competing models

Comparative experiments were designed to study the behavior of our approach, AHJ, against the following state-of-the-art competing approaches:

- TDCox: The extended Cox model presented in [8], which is able to work with time-dependent risk factors. The R package `survival` supplies the function `tmerge` to assist with this process.
- GMM: The generalized method of moments [20]. When time-varying factors are involved, GMM can provide efficient estimates for marginal correlated logistic regression models. The R package `GMM` was used, with the setting that the outcome time $y \geq t$ in the logistic regression model (e.g., in Equation 2) was converted to a binary label.
- RSF+TD: A non-parametrically random survival forest [21], which constructs an ensemble of cumulative risk functions by combining the results of many binary survival trees. Similar to the CART (classification and regression tree) approach, these trees are grown by recursive splitting of tree nodes. We executed the widely used implementation from the package `randomForestSRC` (new version published in May 2016) under its default settings, in conjunction with the `tmerge` function to account for time-varying factors.
- CR: A competing risk model proposed in [10] based on the Cox model, which estimates the non-homogeneous Markov failure process and the cause-specific hazard and predicts the transition probabilities of failure status by combining all models. Technically, it is in the class of joint models and addresses the time-varying factors through a multi-state approach. CR was implemented via the package `CIPred`, provided by the authors, under the default settings.
- SVR: The censored support vector regression introduced in [3, 22], which learns an SVM classifier on survival data with censoring. It was trained in our experiments by transforming the hazard ratios as the regression targets of the linear SVM, making it well suited for the task of classification. We did such transformation according to [3], which presents a detailed procedure of transforming the survival times as the target "hazard value". The package `e1071` [23] was used in the experiments to implement SVR.

C. Performance metrics

To evaluate the power of predictability, we employed three performance metrics, the area under the ROC curve (AUC), the concordance index (CI) and the Birer score (BS), and adapted them for our use.

1) Survival AUC: SAUC provides a probability measure of predictive ability at a given time point, that is, it qualifies the ability of a model to address the question "whether patient i would be likely to die or be readmitted to hospital within one year after index discharge?". It can be defined as

$$\text{SAUC} = \frac{1}{|\mathcal{F}(t_o)| \times |\mathcal{C}(t_o)|} \sum_{i \in \mathcal{F}(t_o)} \sum_{j \in \mathcal{C}(t_o)} \mathbb{1}_{S(t_o;\mathbf{x}_i) < S(t_o;\mathbf{x}_j)},$$

Table II

COPD-SPECIFIC RISK FACTORS (IN TERMS OF DEMOGRAPHICS AND HEALTHCARE INFORMATION, CLINICAL TEST AND DIAGNOSIS, AND MEDICATION FOR TREATMENT) USED FOR FAILURE PREDICTION. NUMERICAL, CATEGORICAL AND BINARY FACTORS ARE ANNOTATED N, C AND B, RESPECTIVELY. THE CATEGORICAL FACTORS HAVE MORE THAN TWO LEVELS. OF A TOTAL OF 42 FACTORS, 9 ARE TIME-VARYING (SHOWN IN ***italic***) AND THE REMAINING 33 ARE STATIC. ALL MEDICATIONS ARE ADMINISTERED AS LONG AS THE PATIENTS ARE HOSPITALIZED, AND THUS THE CHANGE OF THESE TIME-VARYING FACTORS IS PARTLY INFLUENCED BY THE TREATMENT.

Demographics and healthcare (14 static factors)												**Test and diagnosis (6 static factors and 9 time-varying factors)**											
demogr.			healthcare visit			historical hospitalization			hospitalization			comorbidities			respiration			cell			dyspnea		
gender	B		physiologist	B		all-cause length of stay (LOS)	N		hospital branch	B		mental health	B		***SpO2***	N		***white cells***	N		***pre***	C	
age	N		nutritionist	B		COPD-cause LOS	N		admitting service	C		asthma	B		***oxygen***	B		***eosinophils***	N		***post***	C	
rural	B		therapist	B		# of all-cause admissions	N		CT-scan	B		diabetes	B		***cough***	B		***lymphocyte***	N				
			social worker	B		# of COPD-cause admissions	N					cardiovascular diseases	B					***neutrophils***	N				
												pulmonary hypertension	B										
												Charlson comorbidity index	N										

Medication for treatment (13 static, binary (B) factors) ACE/ARA:angiotensin-converting enzyme inhibitor / angiotensin II receptor antagonist

SABA:short-acting bronchodilators	LAAC:long-acting anticholinergic	ICS:inhaled corticosteroids	Antibiotics	Vaccines	Beta Blockers	ACE/ARA
LABA:long-acting bronchodilators	LTRA:leukotriene receptor antagonist	Corticosteroids	Tamiflu	Statin	Benzodiazepine	

where $\mathbb{1}_*$ is the indicator function and t_o represents the length of follow-up period (e.g., 1 year). Recall that $\mathcal{F}$ and $\mathcal{R}$ are the sets of failure and censored patients, respectively, SAUC measures the accuracy of comparing the survival probability between these two groups of patients. The rationale for the use of SAUC is that in clinical decision making, the clinicians and researchers are often more interested in evaluating the relative risk of failure between patients, than in the absolute survival times of these patients.

2) Survival CI: SCI is a generalization of the concept of SAUC. It gives an estimate of how accurately a model can answer the question "which one of patients i and j is more likely to die or be readmitted to hospital?". Taking the case of ties into consideration, we formally define SCI as

$$\text{SCI} = \frac{1}{n_{pair}} \sum\nolimits_{i \in \mathcal{F}, y_i \le y_j} \mathbb{1}_{S(\mathcal{X}_i^{y_i}) \le S(\mathcal{X}_j^{y_j})},$$

where n_{pair} is the number of comparable pairs of patients. (Refer to [24] for more details regarding this metric.) Similar to SAUC, SCI takes values from 0.5 (completely random) to 1.0 (perfect prediction).

3) Survival BS: SBS is essentially a mean square error (MSE) of the survival probability forecasts. It measures the quality of survival probability predictions, i.e., prediction accuracy. SBS allows us to measure the power of the model to answer the question "how accurate is the diagnosis that the disease will recur in patient i?" It can be calculated for an overall error measure across all time points, as follows:

$$\text{SBS} = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \epsilon_i - S(\mathcal{X}_i^{y_i})\right)^2,$$

which can take on values only in the range $[0,1]$. A smaller SBS means higher accuracy of a prognosis.

Note that these three metrics are highly independent of each other. This means that a model which performs very well on one of the metrics may not do well on the other two. A sophisticated prediction model should achieve high SAUC and SCI with low SBS.

The regularization parameter, λ, can be set manually. As usual, it should tend to zero as the number of patients tends to infinity. In each of the subsequent experiment categories, the value was chosen for our model by another S5CV with the highest corresponding metric used in the experiment category.

D. Model evaluations

The experiments to assess the models' ability to predict COPD failure fall into the following three categories:

- Survival probability prediction ($SurPred$);
- Survival high-risk and low-risk classification ($SurCls$);
- Indwelling contrast ($IdwCnt$);

1) Survival probability prediction ($SurPred$): This category includes two sets of experiments:

- We reported the stratified 5CV (S5CV) results on the 1-year follow-up data (all data) for the three metrics SAUC, SCI and SBS, over 100 replicates, in the form of *mean* $\pm$ *standard deviation*. Figure 1 plots the results of AHJ against the others on the training and test sets.
- We investigated how well the models perform on different data that are intercepted by the 25% lower quantile (3 months), median (6 months), 75% upper quantile (9 months) and whole period (12 months) of the follow-up. Table III shows the comparison of the different models' performance on these data. Note that:
  - We enforced administrative censoring at each end of follow-up, which made the failure data smaller. To make full use of such small-size data, we turned to a leave-one-out bootstrap [25] approach to evaluate the prediction performance. The bootstrapping analyzes the COPD subpopulation repeatedly, where each subpopulation is a random patient with replacement from the entire data. Interestingly, such leave-one-out cross-validation (LOOCV) generally produces the variance-zero results, mainly because SAUC and SCI are derived from global pairwise comparisons between survival probabilities.

– The reason for this investigation is that the follow-up period in practice depends heavily on the data collection. This poses a great challenge to prediction models, as they are required to be well suitable to various types of clinical data with different distributions of failure times.

Figure 1 shows that AHJ achieves significantly better performance in terms of SAUC, SCI and SBS, revealing that it can more effectively predict the survival time for COPD patients. Specifically, AHJ yields a nearly 10% average improvement in SAUC in comparison to the other models, with a 7% improvements in SCI and a 10% improvements in SBS. We can also see that AHJ has an extremely low SBS on the test set, indicating the high accuracy of predicting the absolute survival probabilities for COPD patients in different predefined time frames. TDCox can obtain high accuracy in terms of SAUC and SCI, because the patients used for maximizing the partial likelihood in the regression model are ordered by their survival times. However, it yields a high SBS because the TDCox model cannot explore latent correlation between previous risks and current risk [6]. In contrast, GMM designs a logistic regression framework to capture the correlation that may be present in the response (i.e., survival probability) and possibly in the risk factors. The RSF conservation-of-events principle [21], which asserts that the sum of the estimated CHFs over observed time points equals the total number of deaths, does not apply to our estimator. Although CR learns a likelihood by comparing the risk over all patients, it cannot achieve as high prediction accuracy as AHJ.

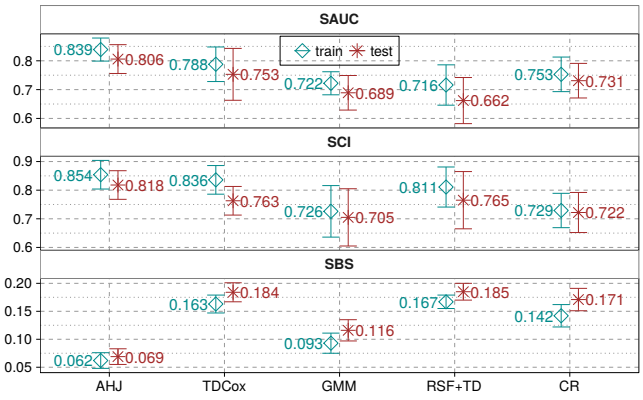

Figure 1. Comparison of the models' $SurPred$ performances, in terms of SAUC, SCI and SBS, on the S5CV training and test patients during 1-year follow-up.

It can be seen from Table III that AHJ scores a clear win over the other models on the data for 9- and 12-month follow-up. As discussed previously, the distribution of survival times changes with the different follow-up durations, and the proportion of censoring will greatly exceed failure in the context of a short-term follow-up, e.g., the 3-month and 6-month cases in the table. This results in poor performances for the competing models, which discard censored patients and maximize the likelihood only for failures in model learning. In contrast, AHJ pays particular attention to the underlying relationship between survival probability and factors, and thus performs more effectively in the case of insufficient failure data (e.g., 3 and 6 months). Over all models shown in the table, roughly speaking, SAUC and SCI grow gradually while SBS diminishes as the follow-up period getting longer, mainly due to the fact that more and more patients die or are readmitted to hospital and become "useful" for model learning. The use of accumulative hazard makes our model more flexible in handling the varying distributions of survival times. In view of the outstanding performance of AHJ in terms of SBS, we assert that the use of our approach can enhance the confidence of failure prediction.

2) Survival high-risk and low-risk classification ($SurCls$): The two sets of experiments in this category are as follows:

- We formulated the COPD prediction problem as predicting high-risk COPD over a predefined time frame, making it a binary classification problem.
  - We applied a 5CV (not S5CV) and segregated the test set into high-risk and low-risk groups, for a fair comparison, based on the median of the training risk ratio scores. Generally, the exponent of the regression coefficients, e.g., $\exp\{\boldsymbol{\beta}\cdot\mathbf{x}\}$ in the TDCox model, is the two-group hazard ratio. This ratio at time t in our model can be approximated by $\sum_{\tau\in\phi^t}\gamma(\tau,t)\exp\{\boldsymbol{\beta}\cdot\mathbf{x}^t\}\delta^\tau$. We chose two criteria commonly used for binary classification in medicine: sensitivity and specificity [26]. The harmonic mean between specificity and sensitivity is also often used and sometimes referred to as an F-measure. We presented the results in Table IV in terms of sensitivity, specificity and a non-weighted harmonic mean, i.e., F1-measure.
  - The reason for this experiment is that classifying COPD patients with high and low risk is clinically crucial, because the high-risk patients would be offered aggressive treatment while those who are low-risk may be offered active surveillance without immediate treatment [27, 28].
- For the sake of investigation, Figure 3 shows the survival curves for two patients, where
  - The high-risk patient is a 79-year-old male discharged on March 8, 2013 after 10 days' index admission LOS. He died at 83 days after index discharge and was never readmitted before that.
  - The low-risk patient is an 80-year-old female discharged on April 24, 2012 after 23 days' index admission LOS. She was still alive 1 year after index discharge and also never readmitted before that.

Table III
COMPARISON OF THE MODELS' $SurPred$ PERFORMANCES, IN TERMS OF SAUC, SCI AND SBS, ON THE LOOCV TEST PATIENTS DURING DIFFERENT FOLLOW-UP PERIODS. **BOLD** EMPHASIS INDICATES SUPERIOR PERFORMANCE OF ONE MODEL OVER THE OTHERS AND ⋆ INDICATES SIGNIFICANCE WITH A PAIRED T-TEST AT $p \leq 0.05$ LEVEL (THIS APPLIES TO ALL SUBSEQUENT TABLES).

t_o (months)	3	6	9	12	3	6	9	12	3	6	9	12
Metric	SAUC				SCI				SBS			
AHJ	.765	**.813**⋆	**.827**⋆	**.842**⋆	.724	**.857**⋆	**.845**⋆	**.870**⋆	.193	.144⋆	**.185**	**.128**⋆
TDCox	.721	.681⋆	.719	.786⋆	.695	.738⋆	.817	.822	.284	.171⋆	.212	.180⋆
GMM	**.779**⋆	.723	.759⋆	.790	**.755**	.728	.785⋆	.819	**.161**	**.137**⋆	.201⋆	.218
RSF+TD	.686	.657	.735⋆	.753	.672	.663	.726	.743⋆	.335	.159	.228	.146⋆
CR	.712	.731⋆	.705	.784⋆	.646	.710	.782⋆	.762⋆	.276	.197	.232	.208⋆

Table IV shows the binary classification accuracies yielded by the different models. AHJ (winner) and SVR (runner-up) perform considerably better than the other competing models, primarily due to their individual characteristics, such as the full use of failure and censored data in AHJ and the robust classification ability of the state-of-the-art SVM-like classifier underlying SVR [29]. Specifically, they achieve higher sensitivities, indicating that more actual high-risk patients are predicted to be in high risk. Similarly, the higher specificity indicates more low-risk patients identified successfully. The performances of TDCox, GMM and RSF are characterized by low accuracies and high variances, mainly due to the difficulties in determining the hazard ratio scores. Thanks to the findings on the representation of hazard and the learning criterion that leverages these findings, AHJ acquires coefficients that make the regression model fit the data better.

Table IV
COMPARISON OF THE MODELS' $SurCls$ PERFORMANCES, IN TERMS OF SPECIFICITY, SENSITIVITY AND F1-MEASURE, ON THE 5CV TEST PATIENTS DURING 1-YEAR FOLLOW-UP.

Criterion	specificity	sensitivity	F1-measure
AHJ	**.839**±.042⋆	**.882**±.033⋆	**.860**±.034⋆
TDCox	.792±.026⋆	.821±.031⋆	.823±.024⋆
GMM	.730±.048⋆	.767±.042⋆	.748±.040⋆
RSF+TD	.810±.033	.775±.043	.804±.034
CR	.826±.025	.806±.037	.828±.024
SVR	.835±.044⋆	.866±.031⋆	.837±.034⋆

It can be seen clearly from Figure 3 that none of these models but AHJ can predict the extremely low survival probability for the high-risk patient at 12 weeks. Moreover, AHJ clearly distinguishes between the two patients as early as at 3 weeks, after which the high-risk patient gets worse and worse while the low-risk patient has a high probability of surviving for a long time. The high-risk patient could thus be issued a warning, indicating that death or readmission would soon occur, and offered advice on early hospitalization. In clinical trials, this is crucial for COPD patients who are likely to have an acute exacerbation of COPD (AECOPD) on those days. The curves yielded by TDCox are quite close to each other at all times; the severity of COPD for the high-risk patient seems to be underestimated. Significant differences between the two patients cannot be found by GMM until around 150 days later, and what's worse, it still predicts high survival probability at 11 weeks for the patient who has already died. GMM, RSF+TD and CR perform ineffectively because of a considerably time delay in death prediction, though they can forecast death for the high-risk patient within 1 year. It thus appears that AHJ is a good aid to clinicians to help with early diagnosis for high-risk patients.

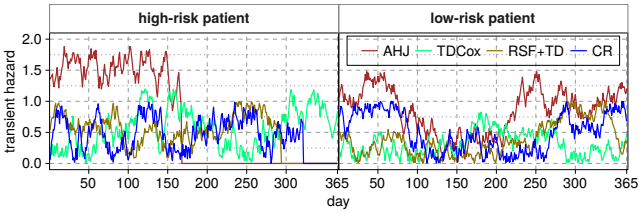

Figure 2. The hazards yielded by the models at various time points. The hazard is assumed to be zero when the survival probability decreases to zero. GMM and SVR do not include a hazard in their learning process.

One might wonder whether the models' performances or survival curves are bound up with the hazard settings. To examine this, Figure 2 shows the hazard curves for the two patients. We observe that AHJ generates hazard values in a wider range than the others, and the hazard distributions of the two patients are not clearly different. In fact, the transient hazard does not change in step with cumulative risk, which is a negative logarithm of the survival probability.

3) Indwelling Contrast ($IdwCnt$): Two sets of experiments are presented in this category:

- We compared AHJ to its three reduced versions, non-AH, AH-failure and AHJ-reg, to investigate three inherent merits of AHJ: accumulative hazard, joint likelihood and regularization. Specifically, non-AH estimates the hazard in a non-accumulative way, where Equation 5 becomes
$$h(\mathcal{X}^t) = h_0(t)\exp\{\boldsymbol{\beta}\cdot\mathbf{x}^t\}.$$
In other words, this version of AHJ will estimate the hazard using only the most recent observations at a given time point. Without extra effort, the corresponding baseline hazard and survival function can be obtained based on non-AH. In the AH-failure version, the joint likelihood ℓ_{joint} is no longer used. Rather, only the likelihood for failure patients, $\ell_{\mathcal{F}}$, is estimated while the censored patients are discarded from the data. The

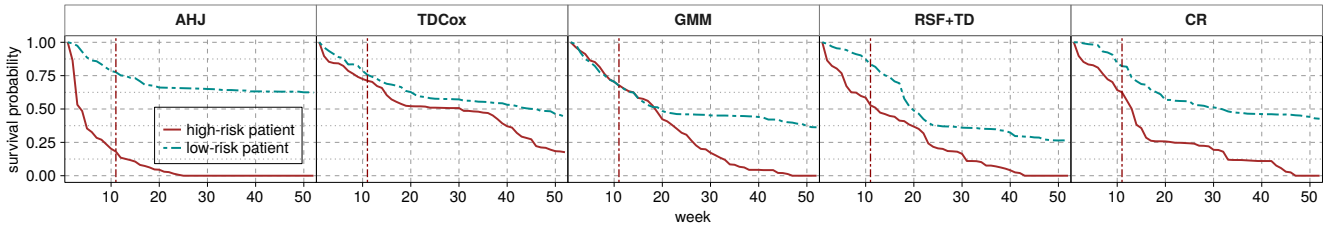

Figure 3. Comparison of changes in 1-year (52 weeks) survival probability predicted by the models for the two individual patients from the high-risk and low-risk groups. SVR is a classifier rather than a prediction model and therefore cannot yield a survival probability curve. The vertical dash is drawn at 83 days (the 12th week), the failure time of the high-risk patient.

difference between AHJ and AHJ-reg is that the reduced version does not regularize the coefficients and therefore straightforwardly optimizes the negative log joint likelihood. We employed an S5CV in this experiment. The comparison between AHJ and its three reduced models is shown in Table V.

- In the second set of experiments, we compared the performances of AHJ and the other competing models on the data of 503 patients for static factors only. The results are shown in Table VI. In this case, the `tmerge` function is no longer used for Cox and RSF. The reason for this experiment was to determine to what extent AHJ, which is primarily designed for time-varying longitudinal data, also displays predictive ability on traditional data with only static factors.

Table V
COMPARISON OF AHJ AGAINST THE THREE REDUCED VERSIONS, IN TERMS OF SAUC, SCI AND SBS, ON THE S5CV TEST PATIENTS DURING 1-YEAR FOLLOW-UP.

Metric	SAUC	SCI	SBS
AHJ	**.862**±.032⋆	**.827**±.045⋆	**.145**±.014⋆
non-AH	.753±.033⋆	.716±.029⋆	.239±.027
AH-falure	.815±.052⋆	.762±.054⋆	.191±.020⋆
AHJ-reg	.776±.036	.785±.042	.224±.018⋆

Looking at Table V, we observe that AHJ outperforms the three reduced models across all three metrics. The comparison between AHJ and AH-failure reveals a mass of "useful" information concealed in the censored data. AHJ's ability to use this information assures its generalizability to clinical data, given that identification of survival time for a patient may be quite expensive and time-consuming in practice, and hence the known (failure) data for analysis is generally much less extensive than the unknown (censored) data. For example, a large number of COPD patients recorded in EMR with certain death times beyond the 1-year mark in our study must, technically, be treated as censored data because of the data deficiency after 1 year. The ineffectiveness of AHJ-reg usually results from the coefficient learning giving rise to overfitting on failure data and even underfitting on censored data. More importantly, AHJ achieves an improvement of over 10% on SAUC and SCI in comparison to non-AH, testifying to the need for an accumulative hazard in failure prediction on longitudinal clinical data.

Table VI
COMPARISON OF THE MODELS' $IdwCnt$ PERFORMANCES, IN TERMS OF S5CV SAUC, SCI AND SBS, ON THE COPD DATA INVOLVING ONLY 33 STATIC FACTORS SHOWN IN TABLE II.

Metric	SAUC	SCI	SBS
AHJ	**.795**±.023⋆	**.776**±.028	**.108**±.016⋆
Cox	.734±.032⋆	.731±.040	.211±.027
GMM	.680±.036	.725±.031⋆	.280±.033⋆
RSF	.727±.048	.693±.026	.261±.023
CR	.761±.025⋆	.737±.018⋆	.167±.019

Looking more closely at Table VI, we note that the performances of all of the models are not as good as those shown in Figure 1, indicating that the 9 time-varying factors (in Table II) are crucial to prediction. The changing observations for these factors contain abundant information about the relationship between survival probability and time-varying risk factors. Although all of the prediction models yield SAUC and SCI values below 80%, AHJ can still be a valuable clinical aid based on its overall performance.

V. CONCLUSIONS

In this paper, we have proposed a new model of hazard to address failure prediction on longitudinal data. In clinical use, this model can better capture the relationship between survival probability and time-varying risk factors in a Cox regression scheme. In addition, we have proposed a joint likelihood to optimize for learning model parameters so that the optimal parameters yielded by our approach can fit both failure and censored data well. Experimental results show that the proposed approach is able to effectively take advantage of longitudinal COPD data and build predictive models more accurate than existing approaches, including the conventional Cox model, the logistic regression model, the Cox-based joint model, the random forest and the censored support vector regression, for making predictions on real-life COPD data. In fact, our approach can be considered as a general paradigm in longitudinal data analysis. For example, our representation of hazard could be adapted to a standard logistic regression model or an additive regression model.

ACKNOWLEDGMENTS

We would like to thank Carol Harris for improving the paper and Mireille Courteau for helping with primary study on COPD. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) to Shengrui Wang under Grant No. 396097-2010, the program PAFI of Centre de Recherche du CHUS (CRCHUS) to Alain Vanasse, and National Natural Science Foundation of China (NSFC) to Lifei Chen under Grant No. 61672157.

REFERENCES

[1] WHO. (2015, March) Chronic obstructive pulmonary disease (COPD). http://www.who.int/mediacentre/factsheets/fs315/en/.

[2] B. Vinzamuri and C. K. Reddy, "Cox regression with correlation based regularization for electronic health records," in *ICDM*, 2013, pp. 757–766.

[3] A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, and H. Lee, "An integrated machine learning approach to stroke prediction," in *KDD*, 2010, pp. 183–192.

[4] R. Moghaddass and C. Rudin, "The latent state hazard model, with application to wind turbine reliability," *Ann. Appl. Stat.*, vol. 9, no. 4, pp. 1823–1863, 2014.

[5] C. Hong and M. Hauskrecht, "Multivariate conditional anomaly detection and its clinical application." in *AAAI*, 2015, pp. 4239–4240.

[6] J. Zhang, L. Chen, A. Vanasse, J. Courteau, and S. Wang, "Survival prediction by an integrated learning criterion on intermittently varying healthcare data," in *AAAI*, 2016, pp. 72–78.

[7] Y. Mao, W. Chen, Y. Chen, C. Lu, M. Kollef, and T. Bailey, "An integrated data mining approach to real-time clinical monitoring and deterioration warning," in *KDD*, 2012, pp. 1140–1148.

[8] L. D. Fisher and D. Y. Lin, "Time-dependent covariates in the Cox proportional-hazards regression model," *Annu. Rev. Public Health*, vol. 20, no. 1, pp. 145–157, 1999.

[9] L. Wu, W. Liu, G. Y. Yi, and Y. Huang, "Analysis of longitudinal and survival data: joint modeling, inference methods, and issues," *J. Probab. Stat.*, vol. 2012, 2011.

[10] G. Cortese and P. K. Andersen, "Competing risks and time-dependent covariates," *Biom. J.*, vol. 52, no. 1, pp. 138–158, 2010.

[11] P. Yadav, M. Steinbach, L. Pruinelli, B. L. Westra, C. Delaney, V. Kumar, and G. J. Simon, "Forensic style analysis with survival trajectories," in *ICDM*, 2015, pp. 1069–1074.

[12] O. O. Aalen, "A linear regression model for the analysis of life times," *Stat. Med.*, vol. 8, no. 8, pp. 907–925, 1989.

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

[14] N. Breslow, "Covariance analysis of censored survival data," *Biometrics*, pp. 89–99, 1974.

[15] D. R. Cox, "Partial likelihood," *Biometrika*, vol. 62, no. 2, pp. 269–276, 1975.

[16] B. Vinzamuri, Y. Li, and C. K. Reddy, "Active learning based survival regression for censored data," in *CIKM*, 2014, pp. 241–250.

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

[18] C. A. W. Craig A. Wendorf, "Primer on multiple regression coding: Common forms and the additional case of repeated contrasts," *Understanding Statistics*, vol. 3, no. 1, pp. 47–57, 2004.

[19] H. Kim, G. H. Golub, and H. Park, "Imputation of missing values in dna microarray gene expression data," in *CSB*, 2004, pp. 572–573.

[20] T. L. Lalonde, A. Q. Nguyen, J. Yin, K. Irimata, and J. R. Wilson, "Modeling correlated binary outcomes with time-dependent covariates," *J Data Sci*, vol. 11, pp. 715–38, 2013.

[21] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Ann. Appl. Stat.*, pp. 841–860, 2008.

[22] P. K. Shivaswamy, W. Chu, and M. Jansche, "A support vector approach to censored targets," in *ICDM*, 2007, pp. 655–660.

[23] D. Meyer and F. T. Wien, "Support vector machines," *The Interface to libsvm in package e1071*, 2015.

[24] H. Steck, B. Krishnapuram, C. Dehing-oberije, P. Lambin, and V. C. Raykar, "On ranking in survival analysis: Bounds on the concordance index," in *NIPS*, 2008, pp. 1209–1216.

[25] B. Efron and R. Tibshirani, "Improvements on cross-validation: the 632+ bootstrap method," *J. Amer. Statist. Assoc.*, vol. 92, no. 438, pp. 548–560, 1997.

[26] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and roc: a family of discriminant measures for performance evaluation," in *AAAI*, 2006, pp. 1015–1021.

[27] 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," in *KDD*, 2015, pp. 1721–1730.

[28] S. Basu Roy, A. Teredesai, K. Zolfaghar, R. Liu, D. Hazel, S. Newman, and A. Marinez, "Dynamic hierarchical classification for patient risk-of-readmission," in *KDD*, 2015, pp. 1691–1700.

[29] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Stat. Comput.*, vol. 14, no. 3, pp. 199–222, 2004.