2017 31st International Conference on Advanced Information Networking and Applications Workshops

Sequential Representation of Clinical Data for Full-fitting Survival Prediction

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

***Abstract*—Survival prediction on time-to-event data associated with patients is crucial in clinical research. Cox-type regression models are widely used for such prediction, but their performance for practical survival prediction suffers due to their use of a maximum partial likelihood estimator, which undermines the effectiveness and robustness of such models. To address this problem, we propose to maximize a new full likelihood that fits the model to all of the data for both failed and censored patients. We also represent time-to-event data by a new sequencing structure, which allows the proposed likelihood to be estimated by predicting event occurrence across the unit time intervals in practice. Furthermore, the likelihood is regularized to prevent overfitting from arising in the model learning step. We investigate the new approach via experimental studies on real-life clinical data and its superior performance compared to other popular state-of-the-art models reveals the great promise of our approach for clinical prediction.**

1 Introduction

Time-to-event data arise when interest is focused on the time elapsing from the beginning of observation until some particular event is experienced; in clinical research such data are known generically as survival data. In such contexts, one is generally concerned with time prediction (e.g., before what time the patient should be provided with a warning against an adverse clinical event such as biological death, injury, onset of disease, hospital readmission, etc.) or risk groups (e.g., which group of patients is more likely to experience the event). Survival prediction, an active field in this area, can aid in the choice of treatments, lifestyle modifications and, sometimes, end-of-life care measures [1].

A considerable number of approaches have been developed to perform survival prediction. Many common approaches, including the (non-)parametric models and the most widely used semi-parametric models – Cox-type regression models [2], make use of measurements of risk factors collected from a *homogeneous population* and then learn models' regression coefficients for risk factors to describe the simultaneous effects of these factors on survival probability. Such survival models can be used to issue warnings regarding the health condition of a patient, such as that her survival probability is 70% 1 year after diagnosis and 40% at 2 years.

The Cox-type models arise in a natural way from the Cox proportional hazards model [2] and, logically, learn the coefficients via the popular maximum partial likelihood estimator (MPLE) [3], by which the learning can be done in a computationally efficient way. For practical clinical trials, however, MPLE turns out to be deficient in effectiveness and robustness, mainly because 1) it is fairly sensitive to missing measurements of risk factors and small (or moderate) sample size of data; 2) it considers the censored data as non-informative, resulting in substantial information loss in prediction [4]. Given that the sample size of clinical data is often small or moderate, and a substantial amount of data may be censored or missing, this poses a great challenge to MPLE-based learning inferences.

In this paper, we aim at a learning inference based on a new maximum likelihood estimator that accounts for learning regression coefficients without suffering from the aforementioned issues. For this purpose, we propose to maximize a new full likelihood that can fit the model to all of the time-to-event data. Moreover, such data is represented by a sequencing structure, which allows the proposed likelihood to be effectively estimated. We perform a regularization on the full likelihood to prevent overfitting arising. We investigate the performance of our approach, which we have called *sequencing* of time-to-event data for *full-fitting survival* prediction (SFS), on a clinical dataset collected from Centre Hospitalier Universitaire de Sherbrooke (CHUS).

To summarize, the contributions of this paper include:

- a novel representation of time-to-event data, i.e., sequencing structure, which allows the likelihood to be more efficiently estimated by predicting event occurrence across the unit time intervals;
- a newly designed full likelihood, which fits the model to all of the data for both failed and censored patients and thus reduces information loss in practical use;
- prediction tests on a cohort of real-life patients, which demonstrate the effectiveness and promise of our approach in clinical applications.

978-1-5090-6231-7/17 $31.00 © 2017 IEEE
DOI 10.1109/WAINA.2017.90

2 PRELIMINARIES

In this section, we provide the basic knowledge necessary for the understanding of subsequent sections, and then state the problem to be addressed in this paper.

2.1 Time-to-event Clinical Data

In survival prediction, the time-to-event data for patient i can be summarized as $(\mathbf{x}_i, y_i, \epsilon_i)$.

- The V-dimensional vector, $\mathbf{x}_i := (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(V)}) \in \mathbb{R}^V$, contains the measurements of V risk factors;
- The event indicator ϵ_i equals 1 if i failed (i.e., experienced the event) and 0 otherwise.
- The observed time, y_i, denotes either survival time T_i or censoring time C_i, i.e.,

$$y_i = \begin{cases} T_i, & \text{if } \epsilon_i = 1 \\ C_i, & \text{otherwise } (\epsilon_i = 0) \end{cases} \tag{1}$$

- Survival means that i is still at risk of experiencing the event: that is, i has not yet failed. In this paper, we denote by $\mathcal{E}_1 = \{\forall i, \ni \epsilon_i = 1\}$ the set of all patients who failed during the period of follow-up.
- Censoring means that i dropped out of the study or has not experienced the event by the end of the observation period. Hence, for all (right-)censored patients, their unobserved exact survival times are longer than the observed censoring times, i.e., $T_i \geq C_i, \forall i \in \mathcal{E}_0$, where the set, given by $\mathcal{E}_0 = \{\forall i, \ni \epsilon_i = 0\}$, comprises those who did not experience the event throughout the period of follow-up.

2.2 Survival Modeling

A survival function is commonly used to forecast the probability of surviving up to time t, or more generally, the probability that the event has not occurred prior to t. This probability is called the *survival probability*, computed by

$$S(t|\mathbf{x}_i) = \Pr(T_i \geq t) = \exp\left\{-\int_0^t h(u|\mathbf{x}_i)\mathrm{d}u\right\}, \tag{2}$$

where the hazard function $h(*)$ yields the instantaneous event occurrence rate, defined as

$$h(t|\mathbf{x}_i) = \lim_{\mathrm{d}t \to 0} \frac{\Pr(t \leq T_i \leq t + \mathrm{d}t | T_i \geq t)}{\mathrm{d}t}. \tag{3}$$

The numerator of this expression is the conditional probability that the event will occur in the interval $[t, t+\mathrm{d}t)$ given that it has not occurred before, and the denominator is the width of the time interval.

The Cox-type models consider that risk factors are multiplicatively related to the hazard, as follows:

$$h(t|\mathbf{x}_i; \mathbf{w}) = h_0(t)\exp\{\mathbf{w} \cdot \mathbf{x}_i\}, \tag{4}$$

- $\mathbf{w} \in \mathbb{R}^V$ represents a vector of regression coefficients, i.e., model parameters, describing how the hazard varies in response to the risk factors;
- $h_0(t)$ is the baseline hazard, describing how the risk of event per time unit changes over time. It is positive and independent of $\mathbf{w}$ and therefore treated non-parametrically, making the models *semi-parametric* and the hazards *proportional*.

2.3 Problem Statements

The key to learning a Cox-type model lies in finding the values of coefficients $\mathbf{w}$ that maximize the probability of the observed data, i.e., maximize the likelihood of $\mathbf{w}$ given the observed data. As is generally the case, to estimate the likelihood we have to write the probability (or probability density) of the observed data as a function of $\mathbf{w}$. In particular, given N patients, the Cox-type models aim to maximize the *partial likelihood* (PL) in the form

$$\begin{aligned} L\big(\mathbf{w}; \{(\mathbf{x}_i, y_i, \epsilon_i)\}_{i=1}^N\big) &= \prod_{i=1}^N \frac{\Pr\big(i \text{ failed at } y_i; \mathbf{w}\big)}{\Pr\big(j \in \mathcal{R}(y_i) \text{ failed at } y_i; \mathbf{w}\big)} \\ &= \prod_{i \in \mathcal{E}_1} \frac{\exp\{\mathbf{w} \cdot \mathbf{x}_i\}}{\sum_{j \in \mathcal{R}(y_i)} \exp\{\mathbf{w} \cdot \mathbf{x}_j\}}, \end{aligned} \tag{5}$$

where $\mathcal{R}(t) \triangleq \{\forall i, \ni y_i \geq t\}$ is the risk set, comprising all patients at risk of experiencing the event just prior to t. Since h_0 would be present in both nominator and denominator of Eq. 5, it has been canceled out and thus no assumptions about the shape of the baseline hazard need to be made; this yields an efficient computation.

For practical clinical research, however, the MPLE suffers from the following weaknesses:

- The risk set $\mathcal{R}(*)$ in the denominator of Eq. 5 means that the PL depends only on the ranking of observed survival times, i.e., the inequality $T_j \geq T_i$ for $\mathbf{x}_j$ and $\mathbf{x}_i$, rather than on their actual numerical values. As a consequence, for finite samples, the bias in the precision of estimating $\mathbf{w}$ as a result of using PL can be rather substantial in the context of small or moderate sample size and missing measurements of risk factors, among other possible situations [5].
- The numerator of Eq. 5 reveals that the PL models only failed patients (in $\mathcal{E}_1$) explicitly, whereas censored cases contribute information pertinent only to the risk set (i.e., the denominator, not the numerator). In other words, censored patients are treated as non-informative [4, 6]; this would lead to substantial information loss in prediction [7].

Indeed, it is well known that the sample size of clinical data is often small or moderate [8], and a substantial amount of data may be censored or missing [9]. To overcome these weaknesses, we develop a regression model whose coefficients can be derived from a new maximum likelihood estimator that fits the survival prediction model to the full dataset. This is the reason for the name *full-fitting* survival prediction.

3 OUR APPROACH

Our approach involves: 1) a new sequencing structure for time-to-event data; 2) a new full likelihood of model regression coefficients used for survival modeling; 3) a model learning

procedure aimed at finding the optimal regression coefficients; and 4) the final survival prediction based on the optimal coefficients learned by our approach.

3.1 Sequencing of Time-to-Event Data

Before we move on to the likelihood, let's look more closely at the time-to-event data $(\mathbf{x}_i, y_i, \epsilon_i)$ for all i. The MPLE ranks all patients by T_is (indicated by y_is and ϵ_is) and then estimates the PL based on that ranking. Cox-type models perform such an MPLE calculation and concentrate only on the explicit occurrence of the event by y_i, but this approach is not particularly effective, as discussed previously.

In contrast, we are interested in the occurrence of the event for i over the period y_i, which, without loss of generality, can be encoded as a binary sequence

$$z_i \triangleq \underbrace{0,\ldots,0,\epsilon_i}_{\tau,2\tau,\ldots,y_i}, \tag{6}$$

where τ is the unit interval of observed time. The entries of this event sequence take values of either 0 or 1, indicating whether the event occurs within the recorded durations $\tau, 2\tau, \ldots, y_i$. The sequence also reveals the survival status for patient i at each time interval over the period between the beginning of observation and the final observed time y_i. Obviously, the sequence length depends on the time when the event occurs and therefore varies from patient to patient. Note that, by such sequencing, we are able to effectively estimate the newly developed likelihood described in the next subsection.

3.2 Full Likelihood

Given N event sequences, the likelihood of coefficients $\mathbf{w}$ is equal to the probability of these sequences given $\mathbf{w}$, i.e.,

$$L(\mathbf{w}; z_1, \ldots, z_N) = \prod_{i=1}^{N} p(z_i; \mathbf{w}),$$

where

$$p(z_i;\mathbf{w}) = \begin{cases} p(\underbrace{0,\ldots,0,1}_{\tau,2\tau,\ldots,y_i};\mathbf{w}), & \text{if } \epsilon_i = 1 \\ p(\underbrace{0,\ldots,0,0}_{\tau,2\tau,\ldots,y_i};\mathbf{w}), & \text{otherwise } (\epsilon_i = 0). \end{cases}$$

Under Eq. 1 and Eq. 6, we have

$$L(\mathbf{w}; z_1, \ldots, z_N) = \prod_{i\in\mathcal{E}_1} p(\underbrace{0,\ldots,0,1}_{\tau,2\tau,\ldots,y_i};\mathbf{w}) \times \prod_{i\in\mathcal{E}_0} p(\underbrace{0,\ldots,0,0}_{\tau,2\tau,\ldots,y_i};\mathbf{w}).$$

The $L(*)$ depends on the actual survival or censoring times of patients, rather than the ranks of observed times used in the PL. Moreover, in addition to the failed patients (in $\mathcal{E}_1$), the censored patients (in $\mathcal{E}_0$) also contribute information to the likelihood. In other words, the model using our likelihood can fit to all of the data for both failed and censored patients.

Alternatively, one might define a likelihood by estimating the contribution based on all of the data only at the observed times y_i for all $i = 1,2,\ldots,N$. Such a likelihood could be called full likelihood as well, in some sense, and discussed in [8, 10]. However, this likelihood is rarely employed in existing work, due to intractability of estimating the density function of survival time in practice.

3.2.1 Likelihood for Censored Patients

The probability of observing the event sequence for censored patient i can be given by the product of conditional probabilities over the observed time, as follows:

$$p(\underbrace{0,\ldots,0,0}_{\tau,2\tau,\ldots,y_i};\mathbf{w}) = p(\underbrace{0}_{\tau};\mathbf{w}) \times p(\underbrace{0}_{2\tau}|\underbrace{0}_{\tau};\mathbf{w}) \times \cdots \times p(\underbrace{0}_{y_i}|\underbrace{0,0\ldots,0}_{\tau,2\tau,\ldots,y_i-\tau};\mathbf{w})$$

It is well known that the occurrence of the event at present is independent of previous survival statuses; we thus have

$$p(\underbrace{0}_{t}|*;\mathbf{w}) = p(\underbrace{0}_{t};\mathbf{w}),$$

and

$$p(\underbrace{0,\ldots,0,0}_{\tau,2\tau,\ldots,y_i};\mathbf{w}) = p(\underbrace{0}_{\tau};\mathbf{w}) \times p(\underbrace{0}_{2\tau};\mathbf{w}) \times \cdots \times p(\underbrace{0}_{y_i};\mathbf{w}).$$

Note that $\underbrace{0,\ldots,0,0}_{\tau,2\tau,\ldots,y_i}$ means that $p(\underbrace{0}_{y_i};\mathbf{w}) > 0$; i.e., patient i survived beyond y_i. That is, the event did not occur for i prior to $y_i - \tau$. Hence, the following inference holds:

$$p(\underbrace{0}_{y_i};\mathbf{w}) > 0 \implies p(\underbrace{0}_{t};\mathbf{w}) \equiv 1, \forall t < y_i.$$

Using the above inference, we continue the manipulations:

$$\begin{aligned} p(\underbrace{0,\ldots,0,0}_{\tau,2\tau,\ldots,y_i};\mathbf{w}) &= p(\underbrace{0}_{y_i};\mathbf{w}) = \Pr(T_i \geq y_i;\mathbf{w}) \\ &= \exp\Big\{-\int_0^{y_i} h(t|\mathbf{x}_i;\mathbf{w})\mathrm{d}t\Big\} \\ &= \exp\Big\{-\sum\nolimits_{t\in\{\tau,2\tau,\ldots,y_i\}} h(t|\mathbf{x}_i;\mathbf{w})\mathrm{d}t\Big\} \\ &= \prod\nolimits_{t\in\{\tau,2\tau,\ldots,y_i\}} \exp\{-h(t|\mathbf{x}_i;\mathbf{w})\tau\}. \end{aligned} \tag{7}$$

As Eq. 7 shows, the likelihood can be computed as probabilities of event across the unit time intervals, rather than across the observed intervals characterized by varying width and random distinct observed times y_i for all $i = 1,2,\ldots,N$. Hence, our approach would be less sensitive to the different distributions of observed times, and thus more suitable for various clinical data.

3.2.2 Likelihood for Failed Patients

Similarly, the probability of the event sequence for failed patient i can be given by

$$p(\underbrace{0,\ldots,0,1}_{\tau,2\tau,\ldots,y_i};\mathbf{w}) = p(\underbrace{0}_{y_i-\tau};\mathbf{w}) \times p(\underbrace{1}_{y_i};\mathbf{w}),$$

where, according to Eq. 7, we obtain

$$p(\underbrace{0}_{y_i-\tau};\mathbf{w}) = \prod\nolimits_{t\in\{\tau,2\tau,\ldots,y_i-\tau\}} \exp\{-h(t|\mathbf{x}_i;\mathbf{w})\tau\},$$

and $p(\underbrace{1}_{y_i};\mathbf{w})$ represents the probability of i experiencing the event right after y_i, given by

$$\begin{aligned} p(\underbrace{1}_{y_i};\mathbf{w}) &= 1 - \Pr(T_i > y_i;\mathbf{w}) \\ &= 1 - \exp\{-h(y_i|\mathbf{x}_i;\mathbf{w})\tau\}. \end{aligned}$$

3.3 Model Learning

As discussed previously, the goal of model learning is to estimate the regression coefficients $\mathbf{w}$. For this purpose, one can resort to a maximum likelihood estimator.

3.3.1 Objective

Based on the full likelihood of coefficients given N event sequences, our learning objective is to maximize this likelihood. To simplify the algebraic manipulations on the likelihood involving an exponential function (the hazard given by Eq. 4), we can perform maximization by minimizing the *negative natural logarithm* of the likelihood. In doing so, the learning problem can be in the form:

$$\min_{\mathbf{w}}\Big(-\ln L(\mathbf{w};z_1,\ldots,z_N)\Big) = \min_{\mathbf{w}}\left(\sum_{i\in\mathcal{E}_1}-\ln p(\underbrace{0,\ldots,0,1}_{\tau,2\tau,\ldots,y_i};\mathbf{w})+\sum_{i\in\mathcal{E}_0}-\ln p(\underbrace{0,\ldots,0,0}_{\tau,2\tau,\ldots,y_i};\mathbf{w})\right).$$

3.3.2 Regularized Objective

Since the above minimization may lead to overfitting, we employed the *ridge* penalty [11], a specific case of the *elastic net* penalty, to prevent overfitting from arising. Briefly, the optimal coefficients that perform well on previously unseen patients can be formally defined as the solution to the regularization problem:

$$\widehat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} \underbrace{-\ln L(\mathbf{w},z_1,\ldots,z_N)+\theta\|\mathbf{w}\|_2^2}_{\ell(\mathbf{w})}.$$

Here, $\ell(\mathbf{w})$ turns out to be strictly convex and differentiable when the regularization constant $\theta \geq 0$.

3.4 Survival Prediction

When an optimal $\widehat{\mathbf{w}}$ is found using our approach, patient i's survival probability at t can be calculated by

$$S(t;\mathbf{x}_i) = S_0(t)^{\exp\{\widehat{\mathbf{w}}\cdot\mathbf{x}_i\}}$$

$$S_0(t) = \exp\Big\{-\int_0^t h_0(u)\mathrm{d}u\Big\} = \prod_{u\in\{\tau,2\tau,\ldots,t\}}\exp\{-h_0(u)\}\tau,$$

where, without loss of generality, the unit time interval τ can take the value 1. The Weibull distribution is commonly used with proportional hazards (PH) models. Accordingly, we used a Weibull-based baseline hazard function, i.e.,

$$h_0(t) \sim Weibull(\lambda,k) = k\lambda^k t^{k-1},$$

where the scale of the distribution is determined by λ and the shape by k, as shown in Fig. 1. Simply, we can set $\lambda = 1$. According to [8] and our finding that k taking values between 0.5 and 1 will not yield any significant change in prediction accuracy, we set $k = 0.8$, making the baseline hazard decrease with increasing t. In fact, this setting is often more realistic than the assumption of a constant hazard function (as in the exponential case).

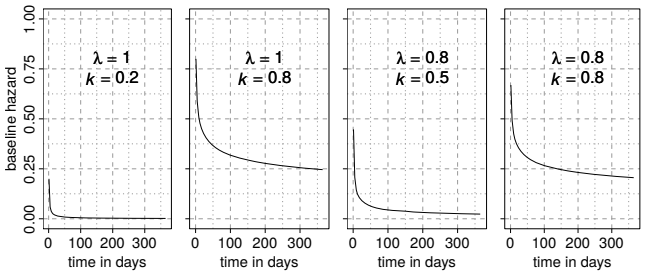

Fig. 1. Change in Weibull-based baseline hazard with different scale and shape parameters.

4 Experiments

In this section, we analyze our approach by comparative experiments on real-life clinical data.

4.1 Survival Metrics

For use in evaluating the predictive power of our approach, we redefined three metrics: the area under the ROC curve (AUC), the concordance index (CI) and the Brier score (BS).

- Survival AUC (SAUC) provides a probability measure of predictive ability during the prespecified period of follow-up t_o. In this paper, we defined it as

  $$\mathrm{SAUC}(t_o) = \frac{1}{\mathcal{E}_1\times\mathcal{E}_0}\sum\nolimits_{i\in\mathcal{E}_1}\sum\nolimits_{j\in\mathcal{E}_0}\mathbb{1}_{S(t_o;\mathbf{x}_i)<S(t_o;\mathbf{x}_j)}.$$

  It measures the *binary classification accuracy*, i.e., the accuracy of a comparison between the survival times for failed and censored patients. The reason for the use of SAUC is that in clinical decision making, the clinicians are often more interested in evaluating patients' relative risk for the event than in their absolute survival times.
- Taking the case of ties into consideration, we formally defined survival CI (SCI) as

  $$\mathrm{SCI} = \frac{1}{n_{ij}}\sum\nolimits_{\forall i,j,\ni\epsilon_i=1\&y_i<y_j}\mathbb{1}_{S(T_i;\mathbf{x}_i)<S(T_i;\mathbf{x}_j)},$$

  where n_{ij} is the number of comparable pairs of patients. Similar to SAUC, SCI takes values from 0.5 (as good as a random predictor) to 1.0 (perfect prediction accuracy).
- Survival BS (SBS) measures the quality of survival predictions, i.e., prediction accuracy. It can be calculated for an overall error measure across all observed times, i.e.,

  $$\mathrm{SBS} = \frac{1}{N}\sum\nolimits_{i=1}^{N}\big(1-\epsilon_i-S(y_i;\mathbf{x}_i)\big)^2,$$

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

A sophisticated survival prediction model should achieve high SAUC and SCI with low SBS. These metrics are useful because they allow us to describe a model's ability to answer different questions, as follows:

- SAUC: Is a patient likely to experience the event within a certain period of follow-up?

- SCI: Which one of two patients is more likely to experience the event?
- SBS: How accurate is the prediction that the event will occur in a certain patient?

4.2 Competing Models

Comparative experiments were designed to study the behavior of our approach against the following popular state-of-the-art competing models that are incorporated into R:

- EN-Cox: The Cox proportional hazards regression model. We used the elastic net penalized Cox model, i.e., EN-Cox, which can be learned using the `cocktail` function in the R package `fastcox` [12].
- GLM: The generalized linear model embodying a Logit link function, i.e., a logistic regression model. The package `glm` was used in our experiments.
- RSF: The Cox-type non-parametrically random survival forest [13] model. It was implemented via the widely used package `randomForestSRC` under its default settings.
- CR: The Cox-type competing risk model. We executed the implementation via the package `CIPred` [14] under the default settings.

4.3 Prediction on COPD Data

4.3.1 Data and Pre-processing

We were able to collect from a regional hospital EMR the data for 503 chronic obstructive pulmonary disease (COPD) patients who were (re)admitted to CHUS in fiscal years 2012 and 2013. Of these patients, 328 (65.2%) failed (died or were readmitted to hospital) within 1 year and the other 175 (34.8%) patients were censored. Due to the 1-year observation period, failure is specified as 1-year failure in the study and those who were censored have a 365-day censoring time. The COPD-specific risk factors shown in Table 1 were selected by Dr. Vanasse (co-author) and his research group at CHUS. We employed a weighted effect coding to create dummy factors from the binary factors so that they could be directly entered into a regression. All numerical factors were normalized to the range [0,1]. We filled in the missing numerical values by means of a linear regression method introduced in [15]. In effect coding, one can choose to use "0" as the default fill-in over dummy factors for the missing values on binary factors.

4.3.2 Results

Table 2 reports the stratified 5 cross-validation (S5CV) results on the observed data within a 1-year period. It can be seen that SFS yields a nearly 8% average improvement in SAUC in comparison to the other models, with a 4% improvement in SCI and a 4% decline in SBS. The extremely low SBS achieved by SFS indicates the high accuracy with which it predicts the absolute survival probabilities for COPD patients. EN-Cox performs poorly, mainly due to the MPLE that does not efficiently fit all data. Similarly, GLM treats censored data as non-informative and also yields poor results. The RSF

Table 1: The COPD-specific risk factors. Of these factors, 17 (top) are associated with demographics and healthcare information, clinical test and diagnosis, and 12 (bottom) with medication for treatment. Numerical factors are shown in **bold** and binary in *italics*.

gender	**index length of stay (LOS)**	*mental health*
age	*index COPD*	*pulmonary hypertension*
rural	*visit physiologist*	**Charlson CI**
	visit social worker	*cardiovascular diseases*
	visit therapist	*cough*
	visit nutritionist	*asthma*
		diabetes
		oxygen

SABA	*LAAC*	*Beta Blockers*	*Antibiotics*	*Vaccines*	*ICS*
LABA	*LTRA*	*Corticosteroids*	*Tamiflu*	*Statin*	*ACE/ARA*

SABA: short-acting bronchodilators
LAAC: long-acting anticholinergic
LABA: long-acting bronchodilators
LTRA: leukotriene receptor antagonist
ICS: inhaled corticosteroids
ACE/ARA: angiotensin-converting enzyme inhibitor/angiotensin II receptor antagonist

conservation-of-events principle [13] does not apply to this real-life data. Although CR learns a likelihood by comparing the risk over all patients, it cannot achieve the high prediction accuracy of SFS. Overall, as an application to assist clinicians in forecasting the progression of disease for different patients, SFS turns out to be a candidate choice.

Table 2: Comparison of the models' S5CV performance, in terms of SAUC, SCI and SBS, for a 1-year observation period, in the form "mean (standard deviation)"

metric	SFS	EN-Cox	GLM	RSF	CR
SAUC	.778(.033)	.681(.028)	.635(.047)	.727(.034)	.754(.029)
SCI	.733(.034)	.672(.035)	.650(.035)	.749(.026)	.696(.022)
SBS	.083(.008)	.132(.006)	.167(.014)	.096(.012)	.115(.007)

We investigated how well the models perform on different data collected over various periods of observation, because such periods in practice depend heavily on the data collection. According to the suggestions from our clinical research group, we tested the models' SAUC performances on the data intercepted by the 25% lower quantile (3 months), the median (6 months), and the 75% upper quantile (9 months) of the original 1-year period of follow-up, as Table 3 shows. Accordingly, we enforced administrative censoring at the end of each observation period, making the data smaller. To make full use of such small-size data, we used a leave-one-out bootstrap cross-validation (LOOCV) to evaluate the prediction performance.

Table 3 shows that SFS scores a clear win over the other models. The distribution of survival times changes with the different observation durations, and the proportion of censoring will greatly exceed failure in the context of a short-term observation period, e.g., the 3-month and 6-month cases. This results in poor performances for the competing models, which maximize their likelihoods only for failures. In contrast, SFS pays particular attention to both failed and censored patients,

and thus performs more effectively in the case of insufficient failed patients (e.g., 3 and 6 months).

Table 3: Comparison of the models' LOOCV performances, in terms of period SAUC, for various periods of observation

model	$t_o = 3$ months	$t_o = 6$ months	$t_o = 9$ months
SFS	.658(.028)	.725(.018)	.768(.042)
EN-Cox	.580(.034)	.631(.036)	.657(.025)
GLM	.611(.036)	.660(.028)	.671(.037)
RSF	.686(.023)	.742(.019)	.735(.025)
CR	.634(.027)	.655(.034)	.689(.022)

Fig. 2 shows the survival probability curves for two patients in different *risk groups*: the failed patient is a 79-year-old male (in the high-risk group) who died at 83 days and the censored patient (in the low-risk group) is an 80-year-old female who was still alive after 1 year. It can be seen clearly that none of these models but SFS can predict the extremely low survival probability for the failed patient at 12 weeks. Moreover, SFS clearly distinguishes between the two patients as early as at 3 weeks. This high-risk patient could thus be issued a warning and offered advice on early treatment. In clinical trials, this is crucial for patients who are likely to have an acute exacerbation at a specific time.

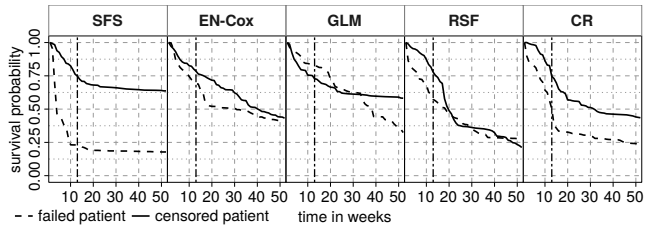

Fig. 2. Comparison of changes in 1-year (52-week) survival probability predicted by the models for the two patients. The vertical dash is drawn at 83 days (the 12th week)

5 Conclusions

In order to overcome the weaknesses of practical survival prediction via Cox-type models using the MPLE, we have proposed the new approach of performing full-fitting survival prediction on time-to-event clinical data. In clinical applications, the proposed approach can better fit the model to the full dataset for both failed and censored patients, thanks to its use of the new full likelihood. In addition, we have proposed a sequencing structure to represent the time-to-event data, making likelihood estimation easy and effective. Experimental results show that the proposed approach is able to build predictive models more accurate than the popular state-of-the-art Cox-type models, for making predictions on real-life clinical data.

Acknowledgments

We would like to thank Carol Harris for improving the paper significantly and Mireille Courteau for helping with primary study on COPD. This work has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) to Shengrui Wang under Grant No. 396097-2010, the National Natural Science Foundation of China (NSFC) to Lifei Chen under Grant No. 61672157, and the program PAFI of Centre de Recherche du CHUS (CRCHUS) to Alain Vanasse. Shengrui Wang is also partly supported by Natural Science Foundation of China (NSFC) under Grant No. 61170130.

References

[1] J. Zhang, S. Wang, J. Courteau, L. Chen, and A. Vanasse, "Predicting COPD failure by modeling hazard in longitudinal clinical data," in *ICDM*, 2016.

[2] D. R. Cox, "Regression models and life tables," *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 34, pp. 187–220, 1972.

[3] ——, "Partial likelihood," *Biometrika*, vol. 62, no. 2, pp. 269–276, 1975.

[4] F. Siannis, J. Copas, and G. Lu, "Sensitivity analysis for informative censoring in parametric survival models," *Biostatistics*, vol. 6, no. 1, pp. 77–91, 2005.

[5] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data*, 2011, vol. 360.

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

[7] 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.

[8] J.-J. Ren and M. Zhou, "Full likelihood inferences in the Cox model: an empirical likelihood approach," *Ann. Inst. Stat. Math.*, vol. 63, no. 5, pp. 1005–1018, 2011.

[9] 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.

[10] H. Zhu, "Likelihood approaches for proportional likelihood ratio model with right-censored data," *Stat. Med.*, vol. 33, no. 14, pp. 2467–2479, 2014.

[11] P. J. Verweij and H. C. Van Houwelingen, "Penalized likelihood in Cox regression," *Stat. Med.*, vol. 13, no. 23-24, pp. 2427–2436, 1994.

[12] Y. Yang and H. Zou, "A cocktail algorithm for solving the elastic net penalized Cox's regression in high dimensions," *Statistics and its Interface*, vol. 6, no. 2, pp. 167–173, 2012.

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

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

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