DYNAMIC RISK PREDICTION TRIGGERED BY INTERMEDIATE EVENTS USING SURVIVAL TREE ENSEMBLES

BY YIFEI SUN 1, SY HAN CHIOU 2, COLIN O. WU 3, MEGHAN McGARRY 4 AND CHIUNG-YU HUANG 5

1Department of Biostatistics, Columbia University, ys3072@cumc.columbia.edu
2Department of Mathematical Sciences, University of Texas at Dallas, schiou@utdallas.edu
3National Heart, Lung, and Blood Institute, National Institutes of Health, wuc@nhlbi.nih.gov
4Department of Pediatrics, University of California San Francisco, Meghan.McGarry@ucsf.edu
5Department of Epidemiology and Biostatistics, University of California San Francisco, ChiungYu.Huang@ucsf.edu

With the availability of massive amounts of data from electronic health records and registry databases, incorporating time-varying patient information to improve risk prediction has attracted great attention. To exploit the growing amount of predictor information over time, we develop a unified framework for landmark prediction using survival tree ensembles, where an updated prediction can be performed when new information becomes available. Compared to conventional landmark prediction with fixed landmark times, our methods allow the landmark times to be subject-specific and triggered by an intermediate clinical event. Moreover, the nonparametric approach circumvents the thorny issue in model incompatibility at different landmark times. In our framework, both the longitudinal predictors and the event time outcome are subject to right censoring, and thus existing tree-based approaches cannot be directly applied. To tackle the analytical challenges, we propose a risk-set-based ensemble procedure by averaging martingale estimating equations from individual trees. Extensive simulation studies are conducted to evaluate the performance of our methods. The methods are applied to the Cystic Fibrosis Patient Registry (CFFPR) data to perform dynamic prediction of lung disease in cystic fibrosis patients and to identify important prognosis factors.

1. Introduction

Cystic fibrosis (CF) is a genetic disease characterized by a progressive, irreversible decline in lung function caused by chronic microbial infections of the airways. Despite recent advances in diagnosis and treatment, the burden of CF care remains high, and most patients succumb to respiratory failure. There is currently no cure for CF, so early prevention of lung disease for high-risk patients are essential for successful disease management. The goal of this research is to develop flexible and accurate event risk prediction algorithms for abnormal lung function in pediatric CF patients by exploiting the rich longitudinal information made available by the Cystic Fibrosis Foundation Patient Registry (CFFPR).

The CFFPR is a large electronic health record database that collects encounter-based records of over 300 unique variables on patients from over 120 accredited CF care centers in the United States (Knapp et al., 2016). The CFFPR contains detailed information on potential risk factors, including symptoms, pulmonary infections, medications, test results, and medical history. Analyses of CFFPR suggested that the variability in spirometry measurements over time is highly predictive of subsequent lung function decline (Morgan et al., 2016). Moreover, the acquisition of chronic, mucoid, or multidrug-resistant subtypes of Pseudomonas aeruginosa (PA) leads to more severe pulmonary disease, accelerating the decline

Keywords and phrases: Dynamic prediction, landmark analysis, multi-state model, survival tree, time-dependent predictors.
in lung function (McGarry et al., 2020). In this paper, the event of interest is the progressive loss of lung function, defined as the first time that the percent predicted forced expiratory volume in 1 second (ppFEV1) drops below 80% in CFFPR. Since risk factors such as weight and height in pediatric patients can change substantially over time, models with baseline predictors have limited potential for long-term prognosis. Incorporating repeated measurements and intermediate clinical events would reflect ongoing CF management and result in more accurate prediction.

To incorporate the longitudinal patient information in risk prediction, one major approach is joint modeling (see, for example, Rizopoulos, 2011; Taylor et al., 2013). Under the joint modeling framework, a longitudinal submodel for the time-dependent variables and a survival submodel for the time-to-event outcome are postulated; the sub-models are typically linked via latent variables. Such a model formulation provides a complete specification of the joint distribution, based on which the survival probability given the history of longitudinal measurements can be derived. Most joint modeling methods consider a single, continuous time-dependent variable. Although attempts have been made to incorporate multiple time-dependent predictor variables (Proust-Lima, Dartigues and Jacqmin-Gadda, 2016; Wang, Luo and Li, 2017), correct specification of the model forms for all the time-dependent covariates and their associations with the event outcome remains a major challenge. Moreover, it is not clear how existing joint modeling approaches can further incorporate the information on the multiple intermediate events, such as the acquisition of different subtypes of PA, in risk prediction.

Another major approach that can account for longitudinal predictors is landmark analysis, where models are constructed at pre-specified landmark times to predict the event risk in a future time interval. For example, at each landmark time, one may postulate a working Cox model with appropriate summaries of the covariate history up to the landmark time (e.g., last observed values) as predictors and then fit the Cox model using data from subjects who are at risk of the event. The estimation can either be performed using a separate model at each landmark time point or a supermodel for all landmark time points. (van Houwelingen, 2007; van Houwelingen and Putter, 2008, 2011). This way, multiple and mixed type time-dependent predictors can be easily incorporated. Moreover, to better exploit the repeated measurements and to handle measurement errors, one may also consider a two-stage landmark approach (Rizopoulos, Molenberghs and Lesaffre, 2017; Sweeting et al., 2017; Ferrer, Putter and Proust-Lima, 2019): in the first step, mixed-effects models are used to model the longitudinal predictors; in the second step, functions of the best linear unbiased prediction (BLUP) estimator of the random effects are included as predictors of the landmark Cox model. Other than Cox models, Parast, Cheng and Cai (2012) considered time-varying coefficient models to incorporate a single intermediate event and multiple biomarker measurements. Zheng and Heagerty (2005), Maziarz et al. (2017), and Zhu, Li and Huang (2019) further considered the impact of informative observation times of repeated measurements on future risk.

Direct application of the existing landmark analysis method to the CFFPR data may not be ideal for the following reasons: first, imposing semiparametric working models at different landmark times may result in incompatible models and inconsistent predictions (Jewell and Nielsen, 1993; Rizopoulos, Molenberghs and Lesaffre, 2017). In other words, a joint distribution of predictors and event times that satisfies the models at all the landmark times simultaneously may not exist. Second, the specification of how the predictor history affects the future event risk may require deep clinical insight. For example, researchers have shown that various summaries of the repeated measurements, including the variability (Morgan et al., 2016), the rate of change (Mannino, Reichert and Davis, 2006), and the area under the trajectory curve (Domanski et al., 2020), can serve as important predictors of disease risks, while
the last observed value has been commonly used in the statistical literature. Therefore, non-parametric statistical learning methods are appealing in landmark prediction, because they require minimal model assumptions and have the potential to deal with a large number of complicated predictors. Tanner et al. (2021) applied super learners for landmark prediction in CF patients, where discrete time survival analysis were conducted via the use of machine learning algorithms for binary outcomes.

In this paper, we propose a unified framework for landmark prediction using survival tree ensembles, where the landmark times can be subject-specific. A subject-specific landmark can be defined by an intermediate clinical event that modifies patients’ risk profiles and triggers the need for an updated evaluation of future risk. In our application, the acquisition of chronic PA usually leads to accelerated deterioration in the pulmonary function and serves as a natural landmark. When the landmark time is random, the number of observed predictors at the landmark time often varies across subjects, creating analytical challenges in fully utilizing the available information. Moreover, unlike static risk prediction models where baseline predictors are completely observed, the observation of the time-dependent predictors is further subject to right censoring. To tackle these problems, we propose a risk-set-based approach to handle the possibly censored predictors. To avoid the instability issue of a single tree, we propose a novel ensemble procedure based on averaging unbiased martingale estimating equations derived from individual trees. Our ensemble method is different from existing ensemble methods that directly average the cumulative hazard predictions and has strong empirical performances in dealing with censored data.

The rest of this article is organized as follows. In Section 2, we introduce a landmark prediction framework that incorporates the repeated measurements and intermediate events. In Section 3, we propose tree-based ensemble methods to deal with censored predictors and outcomes. We propose a concordance measure to evaluate the prediction performance in Section 4 and define a permutation variable importance measure in Section 5. The proposed methods are evaluated by extensive simulation studies in Section 6 and are applied to the CFFPR data in Section 7. We conclude the paper with a discussion in Section 8.

### 2. Model Setup

In contrast to static risk prediction methods that output a conditional survival function given baseline predictors, dynamic landmark prediction focuses on the survival function conditioning on the predictor history up to the landmark time. Since history information involves complicated stochastic processes, challenges arise as to how to partition the history processes when applying tree-based methods. In what follows, we first introduce a generalized definition of the landmark survival function, starting from either a fixed or subject-specific landmark time. We then express the history information as a fixed-length predictor vector on which recursive partition can be applied.

Denote by $T$ a continuous failure event time and by $T_L$ a landmark time. The landmark is selected a priori and is usually clinically meaningful. We allow $T_L$ to be either fixed or subject-specific. We focus on the subpopulation that is free of the failure event at $T_L$ and predict the risk after $T_L$. Denote by $Z$ the baseline predictors and denote by $\mathcal{H}(t)$ other information observed on $[0, t]$. Our goal is to predict the probability conditioning on all the available information up to $T_L$, that is,

\[ P(T - T_L \geq t \mid T \geq T_L, T_L, \mathcal{H}(T_L), Z) \]

(1)

To illustrate the observed history $\mathcal{H}(t)$, we consider two types of predictors that are available in the CFFPR data. The first type of predictors is repeated measurements of time-dependent variables such as weight and ppFEV1. It is worthwhile to point out that both internal and external time-dependent predictors can be included, as only their history up to $T_L$ will be used. We denote this type of predictors by $W(t)$, a $q$-dimensional vector of time-dependent
variables, and assume \( W(\cdot) \) is available at fixed time points \( t_1, t_2, \ldots, t_K \). The observed history up to \( t \) is \( H_W(t) = \{W(s)dO(s), 0 < s \leq t\} \), where \( O(t) \) is a counting process that jumps by one when \( W(\cdot) \) is measured (i.e., \( dO(t_k) = 1 \) for \( k = 1, \ldots, K \)). The second type of predictors is the timings of intermediate clinical events such as pseudomonas infections. Denote by \( U_j \) the time to the \( j \)th intermediate event, \( j = 1, \ldots, J \). The observed history up to \( t \) is \( H_U(t) = \{I(U_j \leq s), 0 < s \leq t, j = 1, \ldots, J\} \). Collectively, we have a system of history processes \( H(t) = (H_W(t), H_U(t)) \).

In our framework, both \( t_k \) and \( U_j \) can serve as landmark times. Due to the stochastic nature of \( U_j \), the order of \( \{t_k, U_j, k = 1, \ldots, K, j = 1, \ldots, J\} \) can not be pre-determined. As a result, the number of available predictors at a given landmark time can vary across subjects. For illustration, we consider one fixed time point \( t_1 \) at age 7 and one intermediate event chronic PA (cPA), of which the occurrence time is denoted by \( U \). Figure 1 depicts the observed data of two study subjects. At \( t_1 = 7 \), subject 1 has experienced cPA \((U_1 = 4.8 \leq t_1)\), while subject 2 remains free of cPA \((U_1 = 10.2 > t_1)\). Let \( W(t) \) be the body weight measured at time \( t \) (i.e., \( q = 1 \)). The probabilities of interest are given as follows:

(I) At a fixed landmark time \( T_L = t_1 \), we predict the risk at \( t_1 + t, t > 0 \), among those who are at risk, that is, \( T \geq t_1 \). Note that subjects in the risk set may or may not have experienced the intermediate event prior to \( t_1 \). Given \( W(t_1) \) and the partially observed \( U_1 \), the conditional survival probability (1) can be reexpressed as

\[
\begin{align*}
  &P(T \geq t + t_1 \mid T \geq t_1, Z, W(t_1), U_1), &\text{if } U_1 \leq t_1, \\
  &P(T \geq t + t_1 \mid T \geq t_1, Z, W(t_1), U_1 > t_1), &\text{otherwise.}
\end{align*}
\]

In other words, at \( t_1 \), we output the former for subjects who experience the intermediate event prior to \( t_1 \) (subject 1, Figure 1a), while output the latter for others (subject 2, Figure 1b).

(II) At a random landmark time \( T_L = U_1 \), we predict the risk for subjects who have experienced the intermediate event and are free of the failure event. The predictor value \( W(t_1) \) is available only if \( U_1 \geq t_1 \). In this case, we predict

\[
\begin{align*}
  &P(T \geq t + U_1 \mid T \geq U_1, Z, U_1), &\text{if } U_1 \leq t_1, \\
  &P(T \geq t + U_1 \mid T \geq U_1, Z, W(t_1), U_1), &\text{otherwise.}
\end{align*}
\]

Therefore, at \( U_1 \), we output the former for subjects whose \( W(t_1) \) is observed after \( U_1 \) (Subject 1, Figure 1c), while output the latter for others (Subject 2, Figure 1d).
In the above example, the observed predictors vary across subjects. We then represent the history \( \mathcal{H}(T_L) \) as a vector with a fixed length, so that tree-based methods can be applied to estimate the probability in (1). Define the complete predictors \( X = \{W(t_1), \ldots, W(t_K), U_1, \ldots, U_J\} \). The information in \( X \) may not be fully available at a given landmark time. We define the available information up to \( t \) by \( X(t) = \{W(t_1,t), \ldots, W(t_K,t), U_1(t), \ldots, U_J(t)\} \), where

\[
W(t_k,t) = \begin{cases} W(t_k), & \text{if } t_k \leq t, \\ \text{NA}_q, & \text{otherwise}, \\ U_j(t), & \text{if } U_j \leq t, \\ t^+, & \text{otherwise}.
\end{cases}
\]

By a slight abuse of notation, we write \( U_j(t) = t^+ \) if \( U_j > t \) and \( W(t_k,t) = \text{NA}_q \) if \( t_k > t \), with \( \text{NA}_q \) denoting a non-numeric \( q \)-dimensional vector. Here an NA value indicates that the covariate value is collected after the landmark time and thus should not be used for constructing prediction models. In other words, under our setting NA is treated as an attribute rather than missing data, as the target probability is not conditional on \( W(t_k) \) for \( t_k > T_L \). The covariate history \( \mathcal{H}(T_L) \) can then be expressed as \( X(T_L) \), which is a \((qK+J)\)-dimensional vector. This way, a covariate not being observed is predictive of the outcome, and the target survival probability function can be expressed as follows:

\[
S(t \mid a, z, x) = P(T \geq t + a \mid T \geq T_L = a, Z = z, X(T_L) = x).
\]

In the example depicted in Figure 1, we have \( X(t) = \{W(t_1,t), U_1(t)\} \). For \( T_L = t_1 \), the predictor values in Figure 1a and 1b correspond to \( x = (19.7, 4.8) \) and \( x = (21.5, 7^+) \), respectively. For \( T_L = U_1 \), the predictor values in Figures 1c and 1d correspond to \( x = (\text{NA}, 4.8) \) and \( x = (21.5, 10.2) \), respectively. Since the information at \( T_L \) involves left-bounded intervals and non-numeric values, applying semiparametric methods to estimate \( S(t \mid a, z, x) \) is challenging. Hence we propose tree-based methods to handle partially observed predictors.

### 3. Survival trees and ensembles for landmark prediction

At a given landmark time, we build a tree-based model to predict future event risk. Survival trees are popular nonparametric tools for risk prediction. The original survival trees take baseline predictors as input variables and output the survival probability conditioning on the baseline predictors (see, for example, Gordon and Olshen, 1985; Ciampi et al., 1986; Segal, 1988; Davis and Anderson, 1989; LeBlanc and Crowley, 1992, 1993; Zhang, 1995; Molinaro, Dudoit and Van der Laan, 2004; Steingrimsson et al., 2016), and ensemble methods have been applied to address the instability issue of a single tree (Hothorn et al., 2004, 2006; Ishwaran et al., 2008; Zhu and Kosorok, 2012; Steingrimsson, Diao and Strawderman, 2019). However, existing methods may not be directly applied, because the predictors in \( X \) are not completely observed at \( T_L \), and the available predictors \( (T_L, X(T_L)) \) are subject to right censoring. In the absence of censoring, we introduce a partition scheme for subjects who are event-free at the landmark time in Section 3.1. To handle censored data, we propose risk-set methods to estimate the partition-based landmark survival probability in Section 3.2 and propose an ensemble procedure in Section 3.3.

#### 3.1. Partition on partially observed predictors at the landmark time

A tree partitions the predictor space into disjoint subsets termed terminal nodes and assigns the same survival prediction for subjects that enter the same terminal node. In dynamic risk prediction, the population of interest is subjects who remain event-free at the landmark time, that is, those with \( T \geq T_L \). We use \( \mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_M\} \) to denote a partition on the sample space of \( (T_L, Z, X(T_L)) \mid T \geq T_L \), where \( \tau_m, m = 1, \ldots, M \), are the terminal nodes. The terminal
nodes are formed recursively using binary partitions by asking a sequence of yes-or-no questions. Existing implementations of trees usually do not handle mixtures of numeric and nominal variables. Since a variable in \( X (T_L) \) may take either numeric/ordinal values or NA, the conventional partition scheme needs to be extended. When a variable in \( W (t_k) \) is nominal, its counterpart in \( W (t_k, T_L) \) is also nominal and can be split applying existing approaches. In what follows, we focus on the case where the longitudinal marker measurements \( W (t_k) \) are numeric/ordinal.

When \( T_L \) is random, the partition is based on the variables in \((T_L, Z, X (T_L))\). We consider the following situations:

(R1) When \( W (t_k, T_L)'s \) are the splitting variables, conventional splitting rules may not be directly applied because they take NA values when \( t_k > T_L \). Supposed \( W \) is an element of \( W (t_k, T_L) \) and \( c \) is a cutoff value. We consider two possible splits, (a) \( \{ W > c \} \) versus \( \{ W \leq c \text{ or } W = \text{NA} \} \) and (b) \( \{ W > c \text{ or } W = \text{NA} \} \) versus \( \{ W \leq c \} \), and select the split that yields a larger improvement in the splitting criterion. We note that \( W = \text{NA} \) (i.e., \( T_L < t_k \)) is treated as an attribute rather than missing data.

(R2) Conventional splitting rules cannot be directly applied to \( U_j (T_L) \), because \( U_j (T_L) \) can be \( T_L^+ \) and the support of \( U_j (T_L) \) is not an ordered set. To tackle this problem, we employ a set of transformed predictors \( \{ U_1 (T_L) / T_L, ..., U_J (T_L) / T_L \} \), which, together with \( T_L \), contains the same information as \( \{ U_1 (T_L), ..., U_J (T_L), T_L \} \). Then the rule \( U_j (T_L) / T_L > c \) can be applied for \( c \in (0, 1] \), and \( U_j (T_L) / T_L > 1 \) is equivalent to \( U_j (T_L) = T_L^+ \).

The splitting rules in (R1) and (R2) can be implemented by transforming \( X (T_L) \) to a numeric/ordinal vector and passing the transformed predictors into a tree algorithm using conventional partition rules: First, we create two features for each element \( W \) in \( W (t_k, T_L) \). The two features, denoted by \( W^+ \) and \( W^- \), take the same value as \( W \) when \( W \neq \text{NA} \) (i.e., \( t_k > T_L \)), and take extreme values \( M \) and \( -M \) otherwise, where \( M \) is a large positive number outside the possible range of predictor values. Instead of using \( W \), we use \( W^+ \) and \( W^- \) as candidate variables for splitting. The partition \( \{ W^+ > c \} \text{ vs } \{ W^+ \leq c \} \) is equivalent to the type (b) partition \( \{ W > c \text{ or } W = \text{NA} \} \text{ vs } \{ W \leq c \} \), because observations satisfying \( W = \text{NA} \text{ or } W > c \) are assigned to the same child node; similarly, splitting based on \( W^- \) yields type (a) partitions. As an example, if landmark is cPA and \( W \) denotes the observed age/7 weight at the landmark time, then the rule “Is \( W^+ > 20? \)” is equivalent to “Is weight at age 7 greater than 20kg or does the subject have cPA before age 7?”.

Implementing such a partition rule is equivalent to the “missings together” approach (Zhang, Holford and Bracken, 1996) and missingness incorporated in attributes (Twala, Jones and Hand, 2008). Second, for each transformed intermediate event time \( U_j (T_L) / T_L (j = 1, ..., J) \), we replace the value \( 1^+ \) with \( M \). The details are described in Algorithm 1 (see Table 1). The partition scheme also apply to the trivial case of fixed landmark times. Figure 2 illustrates the proposed data preprocessing procedure.

The partition scheme described above guarantees that each individual has a well-defined pathway to determine its node membership. Given such a partition \( T \), one can define a partition function \( l_T (a, z, x) \), which returns the terminal node in \( T \) that contains \((a, z, x)\). We define the following partition-based survival function,

\[
S_T (t \mid a, z, x) = P(T \geq t + a \mid T \geq T_L, (T_L, Z, X (T_L)) \in l_T (a, z, x)).
\]

The probability \( S_T (t \mid a, z, x) \) approximates the target function \( S(t \mid a, z, x) \).

3.2. Partition-based estimation at the landmark time The follow-up of a subject can be terminated due to loss to follow-up or study end. We now consider estimating \( S_T (t \mid a, z, x) \) with censored data. Denote by \( C \) the censoring time and assume \( C \) is independent of \((T_L, T, X) \) given \( Z \). We follow the convention to define \( Y = \min(T, C) \) and
Then we have
\[
\lambda(T > t) = \frac{P(T > t)}{P(T > t - 1)}
\]
and replace the value \( M \) if the predictor is not observed. For partially observed intermediate events (e.g., \( U(T_L) \)), we first divide the event time by \( T_L \) and replace the value \( 1^+ \) with \( M \). After preprocessing, zero-variance features and duplicate features can be removed before running the tree algorithm.

\[ \Delta = I(T \leq C) \]. We further define \( Y_L = \min(T_L, Y) \) and \( \Delta_L = I(T_L \leq T, T_L \leq C) \). Note that \( \Delta_L = 1 \) is the at-risk indicator at \( T_L \). For subjects who are free of the failure event at \( T_L \), one can observe \( T_L \) and \( X(T_L) \) when \( \Delta_L = 1 \). The training data are \( \{(Y_i, \Delta_i, Y_{li}, \Delta_{li}, X_i(Y_{li}), Z_i), i = 1, \ldots, n\} \), which are assumed to be independent identically distributed replicates of \( (Y, \Delta, Y_L, \Delta_L, X(Y_L), Z) \).

Define \( N(t, a) = \Delta I(Y - a \leq t) \). Denote by \( \lambda(t \mid a, z, x) \) the landmark hazard function, that is, \( \lambda(t \mid a, z, x)dt = P(T - T_L \in [t, t + dt]) | T - T_L \geq t, T_L = a, Z = z, X(T_L) = x \). The survival function \( S(t \mid a, z, x) \) and the hazard function \( \lambda(t \mid a, z, x) \) have a one-to-one correspondence relationship: \( S(t \mid a, z, x) = \exp\{- \int_0^t \lambda(u \mid a, z, x)du \} \). For \( t > 0 \), we have

\[
E\{N(dt, Y_L) - I(Y - Y_L \geq t)\lambda(t \mid a, z, x)dt \mid \Delta_L = 1, Y_L = a, Z = z, X(Y_L) = x\} = E\{N(dt, a) - I(T \geq a + t, C \geq a + t)\lambda(t \mid a, z, x)dt \mid C \geq a, T \geq T_L = a, Z = z, X(T_L) = x\}
\]

\[
= E\{I(T - T_L \in [t, t + dt]) - I(T - T_L \geq t)\lambda(t \mid a, z, x)dt \mid C \geq a, T \geq T_L = a, Z = z, X(T_L) = x\}
\]

\[
= \exp\{- \int_0^t \lambda(u \mid a, z, x)du \} \times P(C \geq a + t \mid C \geq a, Z = z) = 0.
\]

Then we have

\[
\lambda(t \mid a, z, x)dt = \frac{E\{N(dt, Y_L) \mid \Delta_L = 1, Y_L = a, Z = z, X(Y_L) = x\}}{E\{I(Y - Y_L \geq t) \mid \Delta_L = 1, Y_L = a, Z = z, X(Y_L) = x\}}.
\]

Conditioning on \( \Delta_L = 1 \) and \( (Y_L, Z, X(Y_L)) \), subjects with \( Y - Y_L \geq t \) can be viewed as a representative sample of the population with \( T - T_L \geq t \) for each \( t > 0 \). Heuristically, the numerator and denominator in (4) can be estimated using partition-based estimators in the subsample with \( \Delta_L = 1 \). Given a partition \( \mathcal{T} \), the function \( S_T(t \mid a, z, x) \) in (3) can be
estimated by the following estimator,

\[
\hat{S}_T(t \mid a, z, x) = \exp \left\{ - \int_0^t \sum_{i=1}^n \Delta_{Li} I((Y_{Li}, Z_i, X_i(Y_{Li})) \in l_T(a, z, x)) N_i(du, Y_{Li}) \right\}.
\]

(5)

When a new subject is event-free at the landmark time \(T_{L_0}\) with predictors \(Z_0\) and \(X_0(T_{L_0})\), the predicted survival probability based on a single tree is \(\hat{S}_T(t \mid T_{L_0}, Z_0, X_0(T_{L_0}))\). In practice, the partition \(T\) can be constructed via a recursive partition algorithm, and the split-complexity pruning can be applied to determine the size of the tree (LeBlanc and Crowley, 1993). More details of the tree algorithm are given in the Supplementary Materials.

3.3. Survival tree ensembles based on martingale estimating equations

Since the prediction based on a tree is often unstable, ensemble methods such as bagging (Breiman, 1996) and random forests (Breiman, 2001) have been commonly applied. The original random forests perform the prediction for a new data point by averaging predictions from a large number of trees, which are often grown sufficiently deep to achieve low bias (Friedman, Hastie and Tibshirani, 2001). However, for censored data, a large tree may result in a small number of observed failures in the terminal nodes, leading to increased estimation bias of survival or cumulative hazard functions. Existing survival forests inherit from the original random forest and directly average the cumulative hazard prediction from individual trees. Therefore, the node size parameter needs to be carefully tuned to achieve accurate prediction. On the other hand, if the target estimate can be expressed as the solution of an unbiased estimating equation, a natural way is to solve the averaged estimating equations. In what follows, we propose an ensemble procedure based on averaging martingale estimating equations.

For \(b = 1, \ldots, B\), we draw the \(b\)th bootstrap sample from the training data. Let \(T = \{T_b\}_{b=1}^B\) be a collection of \(B\) partitions constructed using the bootstrap datasets. Each partition is constructed via a recursive partition procedure where at each split, \(m\) predictors are randomly selected as the candidate variables for splitting, and \(m\) is smaller than the number of predictors. Let \(l_b\) be the partition function based on the partition \(T_b\). The tree-based estimation from \(T_b\) can be obtained from the following estimating equation,

\[
\sum_{i=1}^n w_{bi} I((Y_{Li}, Z_i, X_i(Y_{Li})) \in l_b(a, z, x)) \Delta_{Li} \{N_i(dt, Y_{Li}) - I(Y_i - Y_{Li} \geq t) \lambda(t \mid a, z, x) \} dt = 0,
\]

where \(w_{bi}\) is the frequency of the \(i\)th observation in the \(b\)th bootstrap sample. Note that when \(w_{bi} = 1\) for all \(i = 1, \ldots, n\), solving the above estimating equation yields the estimator in (5). To perform prediction using all the trees, we consider the following averaged estimating equation,

\[
\sum_{i=1}^n w_i(a, z, x) \{N_i(dt, Y_{Li}) - I(Y_i - Y_{Li} \geq t) \lambda(t \mid a, z, x) \} dt = 0,
\]

where \(w_i(a, z, x) = \sum_{b=1}^B w_{bi} I((Y_{Li}, Z_i, X_i(Y_{Li})) \in l_b(a, z, x)) \Delta_{Li} / B\). Solving the averaged estimating equation yields

\[
\hat{S}_T(t \mid a, z, x) = \exp \left\{ - \int_0^t \sum_{i=1}^n w_i(a, z, x) N_i(ds, Y_{Li}) \right\}.
\]

The estimator \(\hat{S}_T(t \mid a, z, x)\) can be viewed as an adaptive nearest neighbour estimator (Lin and Jeon, 2006), where the weight assigned to each observation comes from random forests. The risk prediction algorithm is given in Table 1.
Algorithm 1: The survival tree ensemble algorithm

**Input**: \((Y_i, \Delta_i, Y_{Li}, \Delta_{Li}, X_i(Y_{Li}), Z_i), i = 1, \ldots, n\)

**Output**: \(\hat{S}_T(t \mid a, z, x)\)

1. **Function Transform**\((a, x)\):
   
   ```
   /* The function transforms the longitudinal predictors \(X(T_L)\) (i.e., \(x\)) to a vector \(\bar{X}(T_L)\) (i.e., \(\bar{x}\)) such that the rules in (R1) and (R2) on \(X(T_L)\) can be achieved using a conventional tree-structured rule on \(\bar{X}(T_L)\)
   */
   \(\bar{x} \leftarrow \text{vector(size:} 2Kq + J)\);
   for \(l = 1\) to \(Kq + J\) do
   
   if \(1 \leq l \leq Kq\) and \(x[l] \neq \text{NA}\) : // observed longitudinal measurements
   then
   \(\bar{x}[2l-1] \leftarrow x[l]\);
   \(\bar{x}[2l] \leftarrow x[l]\);
   else if \(1 \leq l \leq Kq\) and \(x[l] = \text{NA}\) : // unobserved longitudinal measurements
   then
   \(\bar{x}[2l-1] \leftarrow M\);
   \(\bar{x}[2l] \leftarrow -M\);
   /* The split \(\hat{x}[2l-1] > c\) vs. \(\hat{x}[2l] \leq c\) corresponds to \(x[l] > c\) or \(x[l] = \text{NA}\) vs. \(x[l] \leq c\) or \(x[l] = \text{NA}\) */
   else if \(Kq + 1 \leq l \leq Kq + J\) and \(x[l] = a +\) : // unobserved intermediate events
   then
   \(\bar{x}[Kq + l] \leftarrow M\);
   else
   \(\bar{x}[Kq + l] \leftarrow x[l]/a\); // observed intermediate events
   end
   /* The split \(\hat{x}[Kq + l] > c\) vs. \(\hat{x}[Kq + l] \leq c\) corresponds to \(x[l]/a > c\) vs. \(x[l]/a \leq c\) */
   end
   return \(\bar{x}\);

20. **Data Preprocessing**

21. if \(i = 1\) to \(n\) do
22.     \(\bar{X}_i(Y_{Li}) \leftarrow \text{Transform}(Y_{Li}, X_i(Y_{Li}))\);
23. end
24. \(\bar{x} \leftarrow \text{Transform}(a, x)\);

28. **Survival probability prediction**

29. Initialize the weights: \(w_i \leftarrow 0, i = 1, \ldots, n\);
30. for \(b = 1\) to \(B\) do
31.     Draw the \(b\)th bootstrap sample from the training data;
32.     Construct a partition \(\mathcal{T}_b\) using subjects with \(\Delta_L = 1\) in the \(b\)th bootstrap sample:
33.     • The predictors are \(\{Y_{Li}, Z, \bar{X}(Y_{Li})\}\) and the censored outcome is \((Y - Y_{Li}, \Delta)\);
34.     • At each split, a random selection of \(m\) predictors are used as the candidate splitting variables;
35.     for \(i = 1\) to \(n\) do
36.         if \(\Delta_{Li} = 1\) and \((Y_{Li}, Z, \bar{X}(Y_{Li}))\) and \((a, z, \bar{x})\) are in the same terminal node then
37.             \(w_i \leftarrow w_i + w_{Li}\);
38.             /* \(w_{Li}\) is the frequency of the \(i\)th observation in the \(b\)th bootstrapped sample */
39.         end
40.     end
41.     Predict the survival probability using
42.     
43.     \[
44.     \hat{S}_T(t \mid a, z, x) = \exp \left\{ - \sum_{i=1}^{n} \frac{w_i \Delta_i I(Y_i - Y_{Li} \leq t)}{\sum_{j=1}^{n} w_j I(Y_j - Y_{Lj} \geq Y_i - Y_{Li})} \right\}
45.     \]
46. end
4. Evaluating the landmark prediction performance

To evaluate the performance of the predicted risk score, we extend the cumulative/dynamic receiver operating characteristics (ROC) curves (Heagerty, Lumley and Pepe, 2000), which has been commonly used when a risk score is based on baseline predictors. We note that ROC and concordance indices for dynamic prediction at a fixed landmark time has been studied in the literature (Rizopoulos, Molenberghs and Lesaffre, 2017; Wang, Luo and Li, 2017). Here we consider a more general case where the landmark time $T_L$ is random and subject to right censoring. For $t > 0$, subjects with $0 \leq T - T_L < t$ are considered as cases and subjects with $T - T_L \geq t$ are considered as controls. The ROC curve then evaluates the performance of a risk score that discriminates between subjects who have experienced the events prior to $T_L + t$ and those who do not.

Let $g(t, T_L, Z, X(T_L))$ denote a risk score based on $(T_L, Z, X(T_L))$, with a larger value indicating a higher chance of being a case. For each $t > 0$, the true positive rate and false positive rate at a threshold of $c$ are defined as follows,

\[
\text{TPR}_t(c) = P(g(t, T_L, Z, X(T_L)) > c \mid 0 \leq T - T_L < t, T_L \leq \tau_0),
\]

\[
\text{FPR}_t(c) = P(g(t, T_L, Z, X(T_L)) > c \mid T - T_L \geq t, T_L \leq \tau_0),
\]

where $\tau_0$ is a pre-specified constant. The ROC curve is defined as $\text{ROC}_t(p) = \text{TPR}_t(\text{FPR}_t^{-1}(p))$.

Following the arguments of McIntosh and Pepe (2002), it can be shown that $g(t, a, z, x) = 1 - S(t \mid a, z, x)$ yields the highest ROC curve, which justifies the use of the proposed time-dependent ROC curve. Moreover, the area under the ROC curve is equivalent to the following concordance measure,

\[
\text{CON}_t(g)
\]

\[
= P(g(t, T_{L1}, Z, X_1(T_{L1}))) > g(t, T_{L2}, Z, X_2(T_{L2})) \mid 0 \leq T_2 - T_{L2} < t \leq T_1 - T_{L1}, T_{L1} \leq \tau_0, T_{L2} \leq \tau_0) + 0.5P(g(t, T_{L1}, Z, X_1(T_{L1}))),
\]

\[
= g(t, T_{L2}, Z, X_2(T_{L2})) \mid 0 \leq T_2 - T_{L2} < t \leq T_1 - T_{L1}, T_{L1} \leq \tau_0, T_{L2} \leq \tau_0),
\]

where $(T_{L1}, Z_1, X_1(T_{L1}))$ and $(T_{L2}, Z_2, X_2(T_{L2}))$ are independent pairs of observations, and the second term accounts for potential ties in the risk score.

In practice, one usually builds the model on a training dataset and evaluates its performance on an independent test dataset that are also subject to right censoring. To simplify notation here, we construct the estimator for $\text{CON}_t(g)$ using the observed data introduced in Section 3.2, although $\text{CON}_t$ evaluated using the test data should be used in real applications. Define $d_{ij}(t) = g(t, Y_{i,j}, Z_j, X_{j}(Y_{i,j}))-g(t, Y_{i,k}, Z_i, X_i(Y_{i,k}))$. The $\text{CON}_t(g)$ measure can be estimated by

\[
\widehat{\text{CON}}_t(g)
\]

\[
= \frac{\sum_{i \neq j} \{I(d_{ij}(t) > 0) + 0.5I(d_{ij}(t) = 0)\} \Delta_i I(0 \leq Y_{i} - Y_{j} < t < Y_{i} - Y_{j} \leq \tau_0, Y_{i,j} \leq \tau_0)}{S_C(Y_{j} \mid Z_j)S_C(Y_{i} + t \mid Z_i)},
\]

where $\widehat{S}_C(t \mid z)$ is an estimator for the conditional censoring distribution $S_C(t \mid z) = P(C \geq t \mid Z = z)$. For example, survival trees or forests can be applied to estimate $S_C(t \mid z)$; when censoring is completely random, the Kaplan-Meier estimator can also be applied. Under regularity conditions, we show that $\widehat{\text{CON}}_t(g)$ consistently estimates $\text{CON}_t(g)$ in the Supplementary Materials. In practice, one can either use the concordance at a given time point $t$ or an integrated measure to evaluate the overall concordance on a given time interval $[t_L, t_U]$. In the latter case, a weighted average of concordance on a grid of time points in $[t_L, t_U]$ can be reported. For example, one may assign equal weights to all the time
points. Another example is a weight proportional to the denominator in \( \hat{C}_\text{ON}(g) \), that is, \( \sum_{i \neq j} \Delta_j I(0 \leq Y_j - Y_{Lj} \leq t < Y_i - Y_{Lij}, Y_{Lij} \leq \tau_0, Y_{Lj} \leq \tau_0) / \hat{S}_C(Y_j | Z_j) \hat{S}_C(Y_{Lij} + t | Z_i) \), to avoid potentially unstable estimation for very small or large time points.

5. Permutation variable importance Variable importance is a useful measure for understanding the impact of predictors in tree ensembles and can be used as a reference for variable selection (Breiman, 2001). In the original random forests, each tree is constructed using a bootstrap sample of the original data, and the out-of-bag (OOB) data can be used to estimate the OOB prediction performance. The permutation variable importance of a predictor is computed as the average decrease in model accuracy on the OOB samples when the respective feature values are randomly permuted.

To study variable importance in dynamic risk prediction using censored data, we consider an extension of variable importance. Following the original random forests, the OOB prediction for each training observation is made based on trees constructed without using this observation. Applying the same arguments as in Section 4, the prediction based on trees built without the \( i \)th subject is

\[
\hat{S}_{-i}(t | a, Z, x) = \exp \left\{ - \int_0^t \frac{\sum_{k=1}^n w_{k,-i}(a, z, x)N_k(ds, Y_{Lk})}{\sum_{k=1}^n w_{k,-i}(a, z, x)I(Y_k - Y_{Lk} \geq s)} \right\}
\]

where \( w_{k,-i}(a, z, x) = \sum_{k=1}^B w_{bk} I((Y_{Lk}, Z_k, X_k(Y_{Lk})) < l_b(a, z, x), w_{bk} = 0) \Delta_{Lb} \). Define \( d_{ij}(t) = -\hat{S}_{-i}(t | Y_{Lij}, Z_i, X_i(Y_{Lij})) + \hat{S}_{-j}(t | Y_{Lij}, Z_j, X_j(Y_{Lij})) \). The OOB concordance at \( t \) can be calculated by applying (6). To compute variable importance for a predictor, we permute this predictor and calculate the OOB concordance after permutation. We repeat the permutation multiple times (e.g., 100 times) and define the variable importance as the average difference in OOB concordances over all the permutations.

Permuting variables is straightforward in the case of fixed landmark times (i.e., \( T_L = t_k \)), where we randomly shuffle the observed values of the predictor among individuals who remained under observation at the landmark time, that is, subjects with \( \Delta_L = 1 \). When the landmark time is random (i.e., \( T_L = U_j \)), we propose different permutation procedures for calculating variable importance according to the type of the variable: (1) If the variable of interest is completely observed at the landmark time \( T_L \) (e.g., the value of \( T_L \) and baseline covariates \( Z \)), we randomly shuffle its values among subjects with \( \Delta_L = 1 \). (2) If the variable of interest is an intermediate event time \( U_{j'} (j' \neq j) \), we propose to permute its relative value to the landmark time, that is, \( U_{j'}(T_L) / T_L \), among subjects with \( \Delta_L = 1 \). Note that the reason to permute the ratio \( U_{j'}(T_L) / T_L \) instead of the untransformed intermediate event time directly is to avoid incompatible pairs of the intermediate event time and the landmark time under permutation. (3) If the variable of interest is an element of \( W(t_k) \), we permute its values among subjects with complete observations at the landmark time (i.e., \( T_L \geq t_k \) and \( \Delta_L = 1 \)). This is because \( W(t_k) \) is not used in prediction for subjects with \( T_L < t_k \).

6. Simulation Simulation studies were conducted to assess the performance of the proposed methods in estimating the landmark survival probability in (1) under scenarios when the landmark time is fixed or random. In both cases, we generated the time-independent predictors \( Z = (Z_1, \ldots, Z_{10}) \) from a multivariate normal distribution with \( E(Z_i) = 1, \) \( \text{Var}(Z_i) = 1, \) and \( \text{Cov}(Z_i, Z_j) = 0, \) for \( i, j = 1, \ldots, 10 \). The longitudinal predictors were observed intermittently at time points specified below.

In the first set of simulations, the longitudinal predictors \( W(\cdot) \) were measured at fixed landmark time points \( t_k = k \) for \( k = 1, \ldots, K \). The longitudinal predictors \( W(t) = (W_1(t), \ldots, W_{10}(t)) \) were generated from \( W_i(t) = a_i F(b_i t) / t, \) \( i = 1, \ldots, 10, \) where \( a_i \) and
where $b_i$ are independent standard uniform random variables and $F(x) = 1 - \exp(-x^2)$. The probability in (1) can be expressed as $P(T \geq t_k + t \mid T \geq t_k, W(t_1), \ldots, W(t_k), Z)$ for $0 = t_0 < t < t_{k+1} - t_k$, $k \geq 1$, and $P(T \geq t \mid Z)$ for $0 < t < t_1$. For $k \geq 1$, we assume the hazard function of $T$ on $(t_k, t_{k+1})$ depends on the history of $W(\cdot)$ up to $t$ only through its value at $t_k$. For $t \in (t_k, t_{k+1})$ and $k \geq 0$, we consider the following hazard functions for $T$,

(I) $\lambda(t \mid \mathcal{H}_W(t), Z) = t^2 \exp\left\{-5 + \sum_{j=1}^{10} \alpha_{kj}W_j(t_k) + \sum_{j=1}^{10} \beta_{kj}W_j(t_k)Z_j + \sum_{j=1}^{3} Z_j^2\right\}$, $\alpha_{kj} = \beta_{kj} = 2I(k = 1) + 4I(k \geq 2)$ for $1 \leq j \leq 3$, and $\alpha_{kj} = \beta_{kj} = 0$ for $4 \leq j \leq 10$;

(II) $\lambda(t \mid \mathcal{H}_W(t), Z) = 0.1t^2 + \exp\left\{-5 + \sum_{j=1}^{10} \alpha_{kj}W_j(t_k) + \sum_{j=1}^{10} \beta_{kj}W_j(t_k)Z_j + \sum_{j=1}^{3} Z_j^2\right\}$, $\alpha_{kj} = \beta_{kj} = I(k = 1) + 2I(k \geq 2)$ for $1 \leq j \leq 3$, and $\alpha_{kj} = \beta_{kj} = 0$ for $4 \leq j \leq 10$.

The closed-form expressions of the true landmark survival probabilities under Models (I) and (II) are given in the Supplementary Materials. When evaluating the prediction performance of different approaches, we focus on the landmark probability at $t_2 = 2$.

In the second set of simulations, we consider the case where both an intermediate event and longitudinal markers are present. Specifically, we generated the event times from irreversible multi-state models with three states: healthy, diseased, and death. We assume that all subjects started in the healthy state, disease onset is an intermediate event, and death is the event of interest. We generated the time to the first event, denoted by $D$, from a uniform distribution on $[0,5]$. Define the disease indicator, $I$, where $I = 1$ indicates the subject moves from the healthy state to the disease state at time $D$, and $I = 0$ indicates the subject moves from the healthy state to death at time $D$. The disease indicator was obtained via a logistic regression model, $\log P(I = 1 \mid Z, W(D)) = \sum_{j=1}^{3} W_i(D) + \sum_{j=1}^{3} Z_j + \gamma$, where $\gamma$ is a frailty variable following a gamma distribution with mean $1$ and variance $0.5$, $W(t) = (W_1(t), \ldots, W_{10}(t))$ were generated from $W_j(t) = a_j(1 - \exp(-0.04t^2))$, and $a_j$ follows a uniform distribution on $[-1,1]$ for $j = 1, \ldots, 10$. Given a subject had developed the disease at time $D$, i.e., $I = 1$, the residual survival time, $R$, was generated from the following models,

(III) $\log R = -5 + \sum_{j=1}^{3} W_j(D) + \sum_{j=1}^{3} Z_j^2 + \sum_{j=1}^{3} W_j(D)Z_j + \log(1 + D) + \gamma + \epsilon$, where $\epsilon$ is a standard normal random variable;

(IV) $\log R = -5 + \sum_{j=1}^{3} W_j(D) + \sum_{j=1}^{3} Z_j^2 + \sum_{j=1}^{3} W_j(D)Z_j + \log(1 + D) + \gamma + I(Z_1 > 0)\epsilon_1 + I(Z_1 \leq 0)\epsilon_2$, where $\epsilon_1$ and $\epsilon_2$ are independent normal random variables with variances $1$ and $0.25$, respectively;

(V) The hazard function of $R$ is $t^2 \exp[-5 + \sum_{j=1}^{3} \{2I(1 \leq D < 2) + 4I(D \geq 2)\} \{W_j(D) + W_j(D)Z_j + Z_j^2\}]$.

When $I = 1$, the time to death is $T = D + R$, and the time to the intermediate event is $U = D$; when $I = 0$, the time to death is $T = D$ and the intermediate event does not occur. Under Models (III)-(V), we consider the following three scenarios depending on how the landmark time and the longitudinal markers are observed:

(A) The landmark is the intermediate event, i.e., $T_L = U$, and $W(\cdot)$ is observed intermittently at $t_k = k, k = 1, \ldots, 5$. The target probability in (1) is

$$P(T \geq U + t \mid T \geq U, W(t_1, U), \ldots, W(t_K, U), Z).$$

(B) The landmark time is fixed at $T_L = a$, and $W(\cdot)$ is observed at the intermediate event. The target probability is

$$\begin{cases} P(T - a \geq t \mid T \geq a, U, W(U), Z), & U \leq a, \\ P(T - a \geq t \mid T \geq a, U > a, Z), & U > a. \end{cases}$$

(C) The landmark is the intermediate event, and $W(\cdot)$ is observed at the intermediate event. The target probability is $P(T \geq U + t \mid T \geq U, W(U), Z)$. 
scenario (A) is motivated by the CFFPR data, where the longitudinal marker is regularly monitored. Scenarios (B) and (C) are motivated by applications where markers are observed when a disease is diagnosed. In Scenario (B), we set \( a = 2 \). Due to the complicated relationship between event times and longitudinal markers, deriving the closed-form expression of the true probability under Models (III)-(V) is challenging. We outline the the Monte Carlo method used to approximate the true probabilities in the Supplementary Materials.

For all scenarios, the censoring time was generated from an independent exponential distribution with rate \( c \), where \( c \) was chosen to achieve either a 20\% or 40\% rate of censoring at the baseline. We simulated 500 training datasets with sample sizes of 200 and 400 at baseline. The results for large sample sizes (\( n = 5000 \)) are included in the Supplementary Materials. The trees were constructed with a minimum terminal node size of 15. When a single tree was used for prediction, the size of the tree was determined by split-complexity pruning via ten-fold cross-validation. To grow the trees in the ensemble method, we randomly selected \( \left\lfloor \sqrt{p} \right\rfloor \) variables at each splitting step and did not prune the trees. We applied the log-rank splitting rule in the ensemble method in order to compare the martingale-based ensemble approach with the default ensemble approach in random survival forests. Each fitted model was evaluated on independent test data with 500 observations. The evaluating criteria were the integrated mean absolute error (IMAE), the integrated mean squared error (IMSE), and the integrated Brier score (IBS), and were estimated using repeated measurements from subjects who are at-risk at the landmark time. Scenarios (B) and (C) are motivated by applications where markers are observed and were estimated using repeated measurements from subjects who are at-risk at the landmark time. The simulation results are summarized in Tables 2 and 3, in which the proposed ensemble method used to approximate the true probabilities in the Supplementary Materials.

The simulation results are summarized in Tables 2 and 3, in which the proposed ensemble method outperforms the others based on the four evaluation criteria we considered. As expected, when the sample size increases from 200 to 400, the IMAEs, the IMSEs, and the IBSs
of the proposed methods decrease, while the integrated concordance increases. The conventional random survival forest approach performs similarly to the proposed ensemble method when \( n = 200 \) but loses its edge as the sample size increases. On the other hand, the simple landmark Cox model and the two-stage landmark approach yield similar results under Model (I), but the former yields better results under Model (II). When comparing the tree models with the Cox models, we observe that a small IMSE does not necessarily accompany by a large integrated concordance. We conjecture this is because the concordance measure only depends on the order of the predicted survival probability and is less sensitive in terms of risk calibration. In summary, the proposed ensemble method has strong performances and serves as an appealing tool for dynamic risk prediction.

### Table 2

<table>
<thead>
<tr>
<th>Scenario (I)</th>
<th>IMAE ((\times 1000))</th>
<th>IMSE ((\times 1000))</th>
<th>IBS ((\times 1000))</th>
<th>ICON ((\times 1000))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n)</td>
<td>cen</td>
<td>Tr</td>
<td>E1</td>
</tr>
<tr>
<td>Scenario (I)</td>
<td>200</td>
<td>20%</td>
<td>213</td>
<td>196</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>215</td>
<td>202</td>
<td>222</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>20%</td>
<td>201</td>
<td>172</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>205</td>
<td>178</td>
<td>212</td>
</tr>
<tr>
<td>Scenario (II)</td>
<td>200</td>
<td>20%</td>
<td>73</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>76</td>
<td>79</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>64</td>
<td>59</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>62</td>
<td>64</td>
<td>83</td>
</tr>
</tbody>
</table>

### 7. Application to the Cystic Fibrosis Foundation Patient Registry Data

Understanding the risk factors associated with the progressive loss of lung function is crucial in managing CF. The risk for lung disease depends on patient characteristics, and certain patient groups, such as Hispanic patients, are at increased risk of severe disease for reasons not yet known (McGarry et al., 2019). Our goal is to build prediction models for the development of moderate airflow limitation, defined as the first time that ppFEV1 drops below 80% in CFPPR. Our analysis focused on 5,398 pediatric CF patients who were diagnosed before age one between 2008 and 2013; among them, 419 were Hispanic. The data were subjected to right-censoring due to loss of follow-up or administrative censoring. A total of 4,507 failure events were observed with a median follow-up time of 7.71 years.

The rich information in the CFPPR data renders the possibility of a comprehensive evaluation of important risk factors. We considered baseline predictors including gender, ethnicity, maternal education status (\( \geq 16 \) years of education vs. else), insurance status, geographic location (West, Midwest, Northeast, and South), and mutation class (severe, mild, and unknown). Since baseline factors may have limited predictability, we further included repeated measurements and intermediate events as predictors. The longitudinal measurements, ppFEV1, percent predicted forced vital capacity (ppFVC), weight, and height, were assessed regularly throughout the study and were annualized at integer ages via last observation carried forward. The intermediate events include different subtypes of PA (initial acquisition,
### Table 3

Summaries of integrated mean absolute error (IMAE) ($\times 1000$), integrated mean squared error (IMSE) ($\times 1000$), integrated Brier score (IBS) ($\times 1000$), and integrated concordance (ICON) ($\times 1000$) of different methods in the second set of simulations. cen is the censoring percentage; Tr: the proposed tree with concordance splits; E1: the proposed survival tree ensemble; E2: the original random survival forest; C1: the landmark Cox model; C2: the two stages landmark Cox model.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>n</th>
<th>cen</th>
<th>Tr</th>
<th>E1</th>
<th>E2</th>
<th>C1</th>
<th>Tr</th>
<th>E1</th>
<th>E2</th>
<th>C1</th>
<th>Tr</th>
<th>E1</th>
<th>E2</th>
<th>C1</th>
<th>Tr</th>
<th>E1</th>
<th>E2</th>
<th>C1</th>
</tr>
</thead>
<tbody>
<tr>
<td>(III-A)</td>
<td>200</td>
<td>20%</td>
<td>192</td>
<td>167</td>
<td>178</td>
<td>310</td>
<td>84</td>
<td>79</td>
<td>72</td>
<td>79</td>
<td>237</td>
<td>98</td>
<td>79</td>
<td>86</td>
<td>260</td>
<td>816</td>
<td>902</td>
<td>900</td>
</tr>
<tr>
<td>400</td>
<td>20%</td>
<td>266</td>
<td>194</td>
<td>198</td>
<td>405</td>
<td>114</td>
<td>85</td>
<td>93</td>
<td>319</td>
<td>140</td>
<td>81</td>
<td>87</td>
<td>323</td>
<td>610</td>
<td>906</td>
<td>906</td>
<td>530</td>
<td></td>
</tr>
<tr>
<td>40%</td>
<td>20%</td>
<td>173</td>
<td>154</td>
<td>169</td>
<td>255</td>
<td>68</td>
<td>61</td>
<td>69</td>
<td>185</td>
<td>70</td>
<td>67</td>
<td>78</td>
<td>210</td>
<td>904</td>
<td>909</td>
<td>909</td>
<td>544</td>
<td></td>
</tr>
<tr>
<td>40%</td>
<td>20%</td>
<td>204</td>
<td>178</td>
<td>191</td>
<td>323</td>
<td>85</td>
<td>76</td>
<td>84</td>
<td>239</td>
<td>74</td>
<td>64</td>
<td>76</td>
<td>241</td>
<td>882</td>
<td>917</td>
<td>915</td>
<td>523</td>
<td></td>
</tr>
<tr>
<td>(III-B)</td>
<td>200</td>
<td>20%</td>
<td>234</td>
<td>216</td>
<td>219</td>
<td>249</td>
<td>94</td>
<td>81</td>
<td>89</td>
<td>122</td>
<td>143</td>
<td>132</td>
<td>133</td>
<td>161</td>
<td>544</td>
<td>713</td>
<td>700</td>
<td>614</td>
</tr>
<tr>
<td>40%</td>
<td>20%</td>
<td>245</td>
<td>228</td>
<td>215</td>
<td>268</td>
<td>99</td>
<td>83</td>
<td>87</td>
<td>138</td>
<td>144</td>
<td>133</td>
<td>133</td>
<td>169</td>
<td>545</td>
<td>652</td>
<td>645</td>
<td>584</td>
<td></td>
</tr>
<tr>
<td>40%</td>
<td>20%</td>
<td>240</td>
<td>206</td>
<td>219</td>
<td>236</td>
<td>95</td>
<td>65</td>
<td>70</td>
<td>92</td>
<td>132</td>
<td>116</td>
<td>126</td>
<td>133</td>
<td>572</td>
<td>764</td>
<td>760</td>
<td>646</td>
<td></td>
</tr>
<tr>
<td>(III-C)</td>
<td>200</td>
<td>20%</td>
<td>199</td>
<td>180</td>
<td>181</td>
<td>206</td>
<td>73</td>
<td>62</td>
<td>63</td>
<td>91</td>
<td>142</td>
<td>132</td>
<td>134</td>
<td>161</td>
<td>543</td>
<td>717</td>
<td>706</td>
<td>615</td>
</tr>
<tr>
<td>40%</td>
<td>20%</td>
<td>248</td>
<td>191</td>
<td>195</td>
<td>411</td>
<td>109</td>
<td>83</td>
<td>81</td>
<td>324</td>
<td>127</td>
<td>81</td>
<td>83</td>
<td>329</td>
<td>680</td>
<td>906</td>
<td>871</td>
<td>526</td>
<td></td>
</tr>
<tr>
<td>40%</td>
<td>20%</td>
<td>168</td>
<td>150</td>
<td>165</td>
<td>258</td>
<td>74</td>
<td>59</td>
<td>66</td>
<td>187</td>
<td>138</td>
<td>113</td>
<td>128</td>
<td>143</td>
<td>550</td>
<td>901</td>
<td>879</td>
<td>510</td>
<td></td>
</tr>
<tr>
<td>(IV-A)</td>
<td>200</td>
<td>20%</td>
<td>183</td>
<td>163</td>
<td>170</td>
<td>311</td>
<td>80</td>
<td>69</td>
<td>69</td>
<td>238</td>
<td>94</td>
<td>80</td>
<td>81</td>
<td>261</td>
<td>844</td>
<td>902</td>
<td>876</td>
<td>519</td>
</tr>
<tr>
<td>40%</td>
<td>20%</td>
<td>253</td>
<td>233</td>
<td>256</td>
<td>292</td>
<td>121</td>
<td>100</td>
<td>110</td>
<td>150</td>
<td>160</td>
<td>147</td>
<td>147</td>
<td>187</td>
<td>522</td>
<td>783</td>
<td>678</td>
<td>623</td>
<td></td>
</tr>
<tr>
<td>40%</td>
<td>20%</td>
<td>200</td>
<td>197</td>
<td>187</td>
<td>325</td>
<td>90</td>
<td>72</td>
<td>79</td>
<td>241</td>
<td>83</td>
<td>64</td>
<td>76</td>
<td>240</td>
<td>885</td>
<td>915</td>
<td>876</td>
<td>510</td>
<td></td>
</tr>
<tr>
<td>(IV-B)</td>
<td>200</td>
<td>20%</td>
<td>199</td>
<td>180</td>
<td>181</td>
<td>206</td>
<td>73</td>
<td>62</td>
<td>63</td>
<td>91</td>
<td>142</td>
<td>132</td>
<td>134</td>
<td>161</td>
<td>543</td>
<td>717</td>
<td>706</td>
<td>615</td>
</tr>
<tr>
<td>40%</td>
<td>20%</td>
<td>184</td>
<td>153</td>
<td>156</td>
<td>190</td>
<td>65</td>
<td>49</td>
<td>53</td>
<td>72</td>
<td>130</td>
<td>115</td>
<td>125</td>
<td>132</td>
<td>581</td>
<td>770</td>
<td>770</td>
<td>647</td>
<td></td>
</tr>
<tr>
<td>40%</td>
<td>20%</td>
<td>186</td>
<td>152</td>
<td>157</td>
<td>188</td>
<td>66</td>
<td>46</td>
<td>50</td>
<td>72</td>
<td>134</td>
<td>117</td>
<td>127</td>
<td>135</td>
<td>580</td>
<td>712</td>
<td>707</td>
<td>620</td>
<td></td>
</tr>
<tr>
<td>(IV-C)</td>
<td>200</td>
<td>20%</td>
<td>269</td>
<td>245</td>
<td>243</td>
<td>263</td>
<td>114</td>
<td>98</td>
<td>103</td>
<td>150</td>
<td>136</td>
<td>137</td>
<td>167</td>
<td>520</td>
<td>828</td>
<td>833</td>
<td>649</td>
<td></td>
</tr>
<tr>
<td>40%</td>
<td>20%</td>
<td>299</td>
<td>276</td>
<td>275</td>
<td>297</td>
<td>128</td>
<td>112</td>
<td>117</td>
<td>155</td>
<td>159</td>
<td>146</td>
<td>146</td>
<td>185</td>
<td>552</td>
<td>789</td>
<td>789</td>
<td>631</td>
<td></td>
</tr>
<tr>
<td>40%</td>
<td>20%</td>
<td>252</td>
<td>209</td>
<td>212</td>
<td>244</td>
<td>101</td>
<td>81</td>
<td>90</td>
<td>147</td>
<td>111</td>
<td>125</td>
<td>151</td>
<td>552</td>
<td>904</td>
<td>897</td>
<td>694</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(V-A)</td>
<td>200</td>
<td>20%</td>
<td>212</td>
<td>185</td>
<td>190</td>
<td>349</td>
<td>89</td>
<td>75</td>
<td>78</td>
<td>311</td>
<td>133</td>
<td>112</td>
<td>115</td>
<td>333</td>
<td>688</td>
<td>832</td>
<td>824</td>
<td>534</td>
</tr>
<tr>
<td>40%</td>
<td>20%</td>
<td>229</td>
<td>208</td>
<td>215</td>
<td>368</td>
<td>103</td>
<td>90</td>
<td>92</td>
<td>350</td>
<td>115</td>
<td>100</td>
<td>103</td>
<td>316</td>
<td>707</td>
<td>815</td>
<td>820</td>
<td>526</td>
<td></td>
</tr>
<tr>
<td>40%</td>
<td>20%</td>
<td>193</td>
<td>166</td>
<td>179</td>
<td>377</td>
<td>75</td>
<td>66</td>
<td>69</td>
<td>310</td>
<td>107</td>
<td>105</td>
<td>105</td>
<td>362</td>
<td>828</td>
<td>859</td>
<td>843</td>
<td>534</td>
<td></td>
</tr>
<tr>
<td>(V-B)</td>
<td>200</td>
<td>20%</td>
<td>234</td>
<td>208</td>
<td>207</td>
<td>239</td>
<td>92</td>
<td>81</td>
<td>82</td>
<td>70</td>
<td>161</td>
<td>150</td>
<td>151</td>
<td>175</td>
<td>564</td>
<td>672</td>
<td>677</td>
<td>572</td>
</tr>
<tr>
<td>40%</td>
<td>20%</td>
<td>180</td>
<td>158</td>
<td>154</td>
<td>188</td>
<td>54</td>
<td>42</td>
<td>42</td>
<td>70</td>
<td>161</td>
<td>150</td>
<td>151</td>
<td>175</td>
<td>564</td>
<td>672</td>
<td>677</td>
<td>572</td>
<td></td>
</tr>
<tr>
<td>40%</td>
<td>20%</td>
<td>168</td>
<td>136</td>
<td>141</td>
<td>173</td>
<td>53</td>
<td>31</td>
<td>35</td>
<td>55</td>
<td>160</td>
<td>135</td>
<td>143</td>
<td>161</td>
<td>566</td>
<td>715</td>
<td>710</td>
<td>599</td>
<td></td>
</tr>
<tr>
<td>(V-C)</td>
<td>200</td>
<td>20%</td>
<td>177</td>
<td>141</td>
<td>149</td>
<td>178</td>
<td>63</td>
<td>33</td>
<td>39</td>
<td>62</td>
<td>147</td>
<td>120</td>
<td>128</td>
<td>146</td>
<td>604</td>
<td>691</td>
<td>685</td>
<td>589</td>
</tr>
</tbody>
</table>

Non-continuous cell values represent the integrated mean absolute error (IMAE) ($\times 1000$), integrated mean squared error (IMSE) ($\times 1000$), integrated Brier score (IBS) ($\times 1000$), and integrated concordance (ICON) ($\times 1000$) for different methods in the second set of simulations. cen is the censoring percentage; Tr: the proposed tree with concordance splits; E1: the proposed survival tree ensemble; E2: the original random survival forest; C1: the landmark Cox model; C2: the two stages landmark Cox model.

mucoid, chronic, multidrug-resistant), methicillin-sensitive staphylococcus aureus (MSSA),
methicillin-resistant staphylococcus aureus (MRSA), as well as the diagnoses of CF-related
diabetes (CFRD) and pancreatic insufficiency.

For landmark prediction, we considered the following fixed and random landmark times:

(LM1) The landmark time is age 7, with the target prediction interval [7, 22].
(LM2) The landmark time is age 12, with the target prediction interval [12, 22].
(LM3) The landmark time is the acquisition of chronic Pseudomonas aeruginosa (cPA), and
the target prediction interval is from the time of acquiring cPA to age 22.

The fixed landmark ages 7 and 12 correspond to middle childhood and preadolescence. The
random landmark event cPA was considered because patients with cPA are more likely to de-
velop increased inflammation, leading to an accelerated loss in lung function (Kamata et al.,
2017). The median time to cPA in our dataset was 15.5 years. While PA is a frequent pathogen
in cystic fibrosis, early PA is eradicated in the majority of patients through inhaled and in-
travenous antibiotics (Döring et al., 2004; Heltshe et al., 2018). When PA is not eradicated
after initial PA, it converts to chronic PA. Since initial PA is frequently eradicated and does
not have the long-term impact on pulmonary function, we chose the landmark time of cPA
(Harun et al., 2016). A model using initial PA as the landmark is reported in the Supple-
mentary Materials. To perform risk prediction, we used Model (LM1) to obtain the future
risk for a patient who is event-free at age 7. An updated prediction can be carried out using
Model (LM2) if the patient remains event-free at age 12. Upon converting to cPA, the pre-
dicted risk can be updated using Model (LM3). Although landmark prediction models can
be constructed at multiple time points, we suggest practitioners focusing on the predicted
probabilities given by the landmark model that is most close in time. To illustrate the event
history and landmark times, we show the occurrence of PA during the follow-up period for
a random sample of 50 patients in the Supplementary Materials. At each landmark time, the
exact timing of an intermediate event is available only if it has occurred before the landmark
time.

The data were partitioned into a training set (60%) and a test set (40%). The landmark
prediction models were built using the training data and were evaluated on the test data via
the proposed concordance measure. Results from the simple landmark Cox model and the
two-stage approach were also reported for comparison. Similar to the conventional landmark
prediction models, the Cox models were constructed using subjects who remained event-
free and uncensored at each landmark time. In the Cox models, the $j$th partially observed
intermediate events was incorporated via two predictors, $U_j I(U_j \leq T_L)$ and $I(U_j > T_L)$.
In the simple landmark Cox model, the partially observed repeated measurements at $t_k$ were
expressed using $W(t_k) I(t_k \leq T_L)$ and $I(t_k > T_L)$ under Model (LM3). The models were
built in the way described in the simulation section.

The concordance measures are summarized in Table 4. For landmark prediction at ages 7
and 12, we reported the average $\text{CON}_t$ at 50 equally spaced time points on the target predic-
tion intervals. For the prediction at the landmark age of 12, both Models (LM1) and (LM2)
can be applied: (LM1) used history up to age 7 while (LM2) used history up to age 12. As
expected, incorporating additional information between ages 7 and 12 results in an increase
of average concordance from 0.711 to 0.739 in our ensemble model. For the landmark model
at cPA, we used the concordance at a time horizon of 5 years after cPA as the evaluation
criterion. Since we focus on the risk prior to age 22, predicting the 5-year risk for individuals
who acquired the chronic form of PA after age 17 is not feasible. Therefore, the concordance
was evaluated in the subsample of subjects who developed cPA before age 17. The ensemble
method yielded better performances compared to its competitors.

In an attempt to identify important predictors in the ensemble models, we computed the
permutation variable importance with 100 permutations. Under the fixed landmark time mod-
els, we permuted all of the repeated measurements of a marker simultaneously to evaluate the
TABLE 4
Concordance measures evaluated using the test data in the CFFPR analysis. The column Landmark gives the left bound of the target prediction interval. The integrated concordance is reported for Models (LM1) and (LM2) over the interval where the risk prediction is performed, where \( \text{CON}_{[a,b]} = \frac{1}{50} \sum_{j=1}^{50} \text{CON}_t/j \) and \( t_j = a + (b - a)j/50 \). The concordance at year 5 after cPA is reported for Model (LM3).

<table>
<thead>
<tr>
<th>Landmark</th>
<th>Measure</th>
<th>Model</th>
<th>Tree</th>
<th>Ensemble</th>
<th>Cox-1</th>
<th>Cox-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age 7</td>
<td>( \text{CON}_{[0,15]} )</td>
<td>LM1</td>
<td>0.654</td>
<td>0.748</td>
<td>0.695</td>
<td>0.699</td>
</tr>
<tr>
<td>Age 12</td>
<td>( \text{CON}_{[0,10]} )</td>
<td>LM2</td>
<td>0.667</td>
<td>0.739</td>
<td>0.674</td>
<td>0.674</td>
</tr>
<tr>
<td>Age 12</td>
<td>( \text{CON}_{[0,10]} )</td>
<td>LM1</td>
<td>0.620</td>
<td>0.711</td>
<td>0.611</td>
<td>0.669</td>
</tr>
<tr>
<td>cPA</td>
<td>( \text{CON}_5 )</td>
<td>LM3</td>
<td>0.763</td>
<td>0.813</td>
<td>0.788</td>
<td>-</td>
</tr>
</tbody>
</table>

Note: Cox-1 stands for the simple one-stage landmark Cox models, and Cox-2 stands for the two-stage landmark Cox models.

The overall impact of the longitudinal marker. The results of the permutation variable importance are summarized in Figure 3, where ppFEV1, ppFVC, weight, and height are identified as the top four important predictors for both landmark ages 7 and 12 (Figures 3a and 3b). Following them, intermediate events related to PA and staphylococcus aureus are moderately important. We note that mucoid PA and MSSA became more important at age 12 when compared to age 7. This could be due to the fact that these intermediate events are less common before age 7. When using cPA as the landmark, the repeated measurements after the acquisition of cPA were not used in prediction, and thus the number of observed repeated measurements varied across subjects. Unlike baseline variables and intermediate events of which the permutations were performed among subjects who experienced cPA, the permutation of a marker at a specific time point (e.g., age 7) was performed among subjects who experienced cPA after the time point. To this end, we plot the variable importance for predictors at different ages separately, so that the variable importance measures in each plot are based on permuting the same set of subjects. Figure 4a shows the importance of baseline variables and intermediate events, and the timing of cPA plays an important role in predicting future event risk. We note that baseline factors such as ethnicity have a relatively low variable importance. However, this does not mean that ethnicity does not affect the risk of lung function decline. We conjecture that the effect of ethnicity was predominantly mediated through spirometry measurements. Therefore, one should be cautious when interpreting the variable importance. When applying the proposed method in health disparity research, one can further build separate models for Hispanic and non-Hispanic patients. Additional analyses in different ethnic groups are included in the Supplementary Materials.

Our models identify repeated measurements of weight and height as important variables in landmark prediction. To provide more insight into how historical weight measurements affect future risk in our ensemble model, we present the predicted event-free probabilities for hypothetical patients with different weight trajectories in Figure 5. Specifically, we consider Hispanic and non-Hispanic male patients whose weights were in the 10th, 50th, and 90th weight-for-age percentiles of the corresponding ethnicity-gender subgroup. For all six patients, the intermediate event times were fixed at the median derived from the Kaplan-Meier estimates, while categorical predictors and continuous predictors were fixed at the reference levels and mean values, respectively. At the landmark age 7, patients with the 10th percentile weight trajectories had the highest predicted risk, followed by patients with 90th percentile and those with 50th percentile weight trajectories (Figure 5a). At the landmark age 12, the predicted risk in patients with 10th percentile weight remains the highest, followed by patients with 50th percentile and those with 90th percentile weight trajectories (Figure 5b). A
(a) Variable importance when the landmark time is age 7 and the risk prediction interval is $[7, 22]$.

(b) Variable importance when the landmark time is age 12 and the risk prediction interval is $[12, 22]$.

Fig 3: The permutation variable importance in the CFFPR data analysis, when the landmark times are ages 7 and 12. The boxplots show the decreases in OOB concordances from the 100 permutations and are ranked in descending order according to the mean value.

A possible explanation for the predicted curves of the 50th and 90th weight percentiles to flip between landmark age 7 and 12 is a higher degree of survivor bias at the later landmark age; in other words, the event-free individuals at landmark age 12 tended to have better lung function than those at landmark age 7. So if overweight is associated with a higher risk of lung function decline, the 90th percentile patients are more likely to fail before age 12. As a result, we have a group of healthier overweight subjects at the landmark age 12, and thus their risk beyond age 12 can be lower than individuals with 50th percentile weight. As for underweight individuals, the degree of selection bias may not be large enough to compensate for the poor lung function and thus remains the highest risk group at landmark age 12. To summarize, the landmark prediction models built at age 7 and age 12 are for different survivor populations, and thus can not be directly compared.
8. Discussion  In this paper, we proposed a unified framework for tree-based risk prediction with updated information. Compared to semiparametric methods, our methods can handle a large, growing number of predictors over time and do not impose strong model assumptions. Furthermore, the landmark times at which a prediction is performed are allowed to be subject-specific and defined by intermediate clinical events. Notably, our ensemble procedure averages the unbiased martingale estimation equations instead of survival probabilities and avoids the potential bias arising due to small terminal node sizes.

Our discussion has focused on the case where the time-dependent variables $W(\cdot)$ are observed at fixed time points $t_1, \ldots, t_K$. The proposed method can also be applied to the case where repeated measurements are collected at irregular time points, such as hospitalizations. When building prediction models, one can consider using up to $K$ repeated measurements as predictors, where $K$ is a fixed integer. Denote by $V_1 < V_2 < \cdots < V_K$ the potential observation times of $W(\cdot)$. At time $t$, the available information can be expressed using $X(t) = \{W(V_1, t), \ldots, W(V_K, t), V_1(t), \ldots, V_K(t)\}$, where $W(V_k, t) = W(V_k)$ and $V_k(t) = V_k$ if $V_k \leq t$, while $W(V_k, t) = \text{NA}_q$ and $V_k(t) = t^+$ otherwise. In other words, the random measurement times $V_1, \ldots, V_K$ are treated as intermediate event times. In this way, our framework can incorporate irregularly observed covariate information.
Fig 5: Survival predictions for patients in different weight groups. Curves on the left show the repeated weight measurements before the landmark time. Curves on the right show the survival predictions over the interval of interests for the corresponding groups. The weight groups were chosen by the percentiles and ethnicity, \(\cdots\): 10th percentile, non-Hispanic; \(\cdots\): 50th percentile, non-Hispanic; \(\cdots\): 90th percentile, non-Hispanic; \(\cdots\): 10th percentile, Hispanic; \(\cdots\): 50th percentile, Hispanic; \(\cdots\): 90th percentile, Hispanic.

**SUPPLEMENTARY MATERIAL**

**Supplement to Dynamic Risk Prediction Triggered by Intermediate Events Using Survival Tree Ensembles.** The supplementary document contains details of tree algorithms, a proof of consistency of \(\hat{\text{CON}}_1(g)\), additional simulation results, and additional results in the analysis of CFFPR data. The additional simulation results include derivations of the closed-form and empirical estimates of true landmark probabilities, simulation results with the sample size 5000, simulation results on variable importance, and a summary of running time for the proposed ensemble method. The additional data results include more discussion on Models LM1 and LM2, landmark analysis at initial PA, summary statistics of landmark Cox models, and tree diagrams. Example codes used in simulation are available at Github.

Acknowledgements The authors would like to thank the referees, the Associate Editor and the Editor for their constructive comments that improved the quality of this paper. The authors thank the CF Foundation for the use of the CFFPR data to conduct this study. Additionally, we would like to thank the patients, care providers, and care coordinators at CF centers throughout the U.S. for their contributions to the CFFPR. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung and Blood Institute (NHLBI), the National Institutes of Health (NIH) and the U.S. Department of Health and Human Services. Sun’s research was supported by NIH (R21HL156228). Wu’s research was supported in part by the Intramural Research Program of the NHLBI/NIH. Huang’s research was supported by NIH (R01CA193888). McGarry’s research was supported by NIH (1K23HL133437-01A1) and CF Foundation Therapeutics (MCGARR16A0).

REFERENCES


