Received: 29 August 2019 | Accepted: 5 June 2020

DOI: 10.1111/rssa.12611

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Royal Statistical Society
Series A
Statistics
in Society

ORIGINAL ARTICLES

Dynamic survival prediction combining landmarking with a machine learning ensemble: Methodology and empirical comparison

Kamaryn T. Tanner[1] | **Linda D. Sharples**[1] | **Rhian M. Daniel**[2] | **Ruth H. Keogh**[1]

[1]London School of Hygiene and Tropical Medicine, London, UK

[2]Cardiff University, Cardiff, UK

Correspondence
Kamaryn T. Tanner, London School of Hygiene and Tropical Medicine, Department of Medical Statistics, Keppel Street, London WC1E 7HT, UK.
Email: kamaryn.tanner1@lshtm.ac.uk

Abstract

Dynamic prediction models provide predicted survival probabilities that can be updated over time for an individual as new measurements become available. Two techniques for dynamic survival prediction with longitudinal data dominate the statistical literature: joint modelling and landmarking. There is substantial interest in the use of machine learning methods for prediction; however, their use in the context of dynamic survival prediction has been limited. We show how landmarking can be combined with a machine learning ensemble—the Super Learner. The ensemble combines predictions from different machine learning and statistical algorithms with the goal of achieving improved performance. The proposed approach exploits discrete time survival analysis techniques to enable the use of machine learning algorithms for binary outcomes. We discuss practical and statistical considerations involved in implementing the ensemble. The methods are illustrated and compared using longitudinal data from the UK Cystic Fibrosis Registry. Standard landmarking and the landmark Super Learner approach resulted in similar cross-validated predictive performance, in this case, outperforming joint modelling.

KEYWORDS

cystic fibrosis, dynamic prediction, landmark model, machine learning ensemble, survival analysis, time-dependent covariates

1 | INTRODUCTION

Predictive models for time-to-event outcomes are used widely in medicine to identify individuals at elevated risk, to inform treatment strategies and to update patients about their prognosis. For people with life-shortening conditions, it is natural for them to want to know their short-term and long-term survival prospects as time goes on. Answering these questions requires going from a static prediction at diagnosis time to predictions that can be updated dynamically as new information becomes available. Data obtained longitudinally through electronic health records, patient registries and established cohorts have brought opportunities to develop dynamic prediction models for large numbers of individuals. Recent examples of dynamic prediction models focussed on 10-year cardiovascular disease risk based on electronic health records (Paige et al., 2018), survival for people with cystic fibrosis (CF) using registry data (Keogh et al., 2018), intervention-free survival for patients with aortic stenosis based on a cohort (Andrinopoulou et al., 2015) and survival based on breast cancer recurrence data in a French cohort (Lafourcade et al., 2018).

Two techniques for dynamic survival prediction with longitudinal data dominate the statistical literature: joint modelling and landmarking. Joint models specify a joint distribution for longitudinal covariates and the time-to-event process. The most commonly used class of joint models is based on shared random effects (Rizopoulos, 2012; Tsiatis & Davidian, 2004) but the longitudinal process may also be linked to the survival process via shared latent classes (Hickey et al., 2016; Proust-Lima et al., 2014). In the landmarking approach, a survival model is fitted from a series of time origins (landmarks) as a function of predictors measured up to the landmark time. Joint models are flexible and provide consistent predictions when correctly specified (Jewell & Nielsen, 1993; Rizopoulos et al., 2017), and development of statistical software has made the analysis feasible (Hickey et al., 2016; Philipson et al., 2018; Rizopoulos, 2010, 2016). However, this approach can be computationally complex, particularly for large datasets (Rizopoulos et al., 2017). In contrast, landmarking is straightforward to implement and computationally simple, but uses less information from the longitudinal covariate(s) (Rizopoulos et al., 2017; van Houwelingen & Putter, 2012). That the landmarking model is not fitted using a full likelihood has also been cited as a drawback (van Houwelingen & Putter, 2012). In both landmarking and joint modelling, the survival model is typically a proportional hazards model (Ferrer et al., 2018). Recent papers have compared the two approaches using simulation studies. Rizopoulos et al. (2017) demonstrated that joint models tend to outperform landmarking when the effect of time is correctly specified in the longitudinal submodel but as misspecification increases, the differences become smaller and landmarking can even outperform joint modelling but, in contrast, a misspecified association structure between the longitudinal and survival processes did not substantially affect the relative performance of the two analysis methods. Ferrer et al. (2018) agreed that joint models outperform landmarking for a correctly specified joint model but noted that landmarking is less sensitive to a misspecified longitudinal process. Suresh et al. (2017) found that joint modelling provided better performance than landmarking for an illness-death model but the difference was quite small. Maziarz et al. (2017) reported that partly conditional models (landmarking-style models) offered comparable performance to joint models and were more computationally efficient.

Machine learning algorithms are gaining in popularity as tools for clinical prediction. Nonparametric machine learning methods assume no knowledge about the data generating process and include artificial neural networks, support vector machines, decision trees and their relatives, random forests, boosting and bagging (Breiman, 2001). Statistical models, such as generalised additive models (GAM) and penalised regression, are often included in the broader category of statistical learning which encompasses a broad range of methods for learning from data (James et al., 2013). Ensemble approaches combine several algorithms which can include methods of both of the above types. In this paper, we will include these ensemble techniques in the definition of machine learning even though some of the component algorithms are parametric statistical models. Recent applications of machine learning methods for clinical prediction include: assessment of delirium risk (Wong et al., 2018), 3-year survival for CF patients (Alaa & van der Schaar, 2018), mortality in coronary artery disease (Steele et al., 2018) and time to revision surgery after knee replacement (Aram et al., 2018). These studies and others have compared machine learning methods to traditional methods and found mixed results with some finding superior performance for parametric statistical models and others finding improved results with various machine learning algorithms. A review of comparisons of machine learning and logistic regression for clinical prediction found no evidence of benefit to machine learning (Christodoulou et al., 2019).

Most machine learning examples in the survival literature are designed to make a static survival prediction using baseline covariates, and there is a gap in knowledge about how such methods can be applied in the context of dynamic survival prediction. The aim of this paper is to describe a machine learning approach for dynamic prediction of time-to-event outcomes using longitudinal data. In particular, we show how the landmarking concept formulated in discrete time and fit to a single stacked dataset yields a binary outcome model that can be used in conjunction with machine learning. We exploit the fact that many machine learning algorithms have been designed for binary outcomes and, via an ensemble, the results of multiple algorithms may be combined. The focus is on the Super Learner ensemble for which software for implementation is available in R (van der Laan et al., 2007). We motivate and illustrate the methodology using longitudinal data from the UK Cystic Fibrosis Registry. The machine learning approach is compared empirically with a traditional landmarking analysis based on Cox regression and with joint modelling. Predictive performance of the different approaches is evaluated using cross-validation and through application to a holdout dataset.

The paper is organised as follows. We introduce the motivating application in Section 2. Section 3 describes current standard dynamic prediction techniques and Section 4 details our method for dynamic prediction using a machine learning ensemble combined with landmarking. We describe the procedures for model assessment in Section 5 and report results from the analysis of the UK Cystic Fibrosis Registry dataset in Section 6. Finally, we conclude with a discussion in Section 7.

2 | MOTIVATING APPLICATION: SURVIVAL IN CYSTIC FIBROSIS

We illustrate techniques through development of a dynamic survival prediction model for people with CF in the United Kingdom using data from the UK CF Registry. Cystic fibrosis is an inherited chronic and progressive disease with a median survival of 47 years of age (UK Cystic Fibrosis Registry, 2018). In the United Kingdom, there is universal newborn screening for CF and most cases will be diagnosed within months of birth (Lim et al., 2014). Established in 1995, the UK Cystic Fibrosis Registry contains more than 100,000 annual review records from over 12,000 CF patients (Taylor-Robinson et al., 2017), representing over 99% of people with CF in the United Kingdom (UK Cystic

Fibrosis Registry, 2018). The registry includes demographic information, genotype, measures of lung health, prevalence/incidence of complications or infections over the past year and treatment information. These data are generally collected in an outpatient clinic at the patient's annual review and represent systematically collected data as opposed to encounter-based data (Taylor-Robinson et al., 2017). The registry dataset is typical of complex longitudinal datasets that contain right-censored, left-truncated time-to-event data, a mixture of binary, continuous and categorical variables measured at baseline and/or at regular intervals, and some degree of measurement error and missingness. All of these characteristics require choices to be made in implementing the framework for prediction.

The key outcomes of this study are 2-year and 5-year survival of individuals aged 20–50 years. We did not attempt predictions longer than 5 years because at older ages, the small number of individuals in the dataset may lead to a model that is not well-calibrated (Keogh et al., 2018). The event of interest is a composite outcome of death or lung transplantation. This composite outcome was chosen because lung transplantation occurs when patients are considered to have short life expectancy and there are different risk factors for death post-transplant. In this study, we restrict attention to 3 baseline covariates and 16 time-dependent covariates found to be important in the prediction of survival for people with CF in recent studies (Aaron et al., 2015; Alaa & van der Schaar, 2018; Keogh et al., 2018; Liou et al., 2001). A complete listing can be found in Table 1.

3 | STANDARD DYNAMIC PREDICTION TECHNIQUES

3.1 | Dynamic prediction

Static prediction models provide information about survival from the time origin using baseline covariates measured at or before this origin. In dynamic prediction, the aim is to estimate the probability of survival to some time horizon t_{hor} conditional on survival to time s, $s < t_{\text{hor}}$ and conditional on covariates measured up to time s. More generally, a survivor curve showing predictions to multiple time horizons may be obtained. In contrast to static prediction, dynamic models enable predictions to be updated at new times s conditional on new measurements (if available) and on the patient having survived to time s (Proust-Lima & Blanche, 2016; van Houwelingen & Putter, 2012).

Let T_i^* and C_i denote, respectively, the event time and the censoring time for an individual i, $(i = 1, \ldots, n)$. The observed time is $T_i = \min(T_i^*, C_i)$ and $\delta_i = I(T_i^* \leqslant C_i)$ is an indicator of whether the individual experienced an event ($\delta_i = 1$) or was censored ($\delta_i = 0$). Let X_i denote a set of time-fixed covariates and $\mathcal{Y}_i(s)$ denote the longitudinal covariate history up to time s. The conditional survival probability of interest is

$$\pi_i(t_{\text{hor}}|s) = \Pr(T_i^* \geqslant t_{\text{hor}}|T_i^* > s, \mathcal{Y}_i(s), X_i) \tag{1}$$

3.2 | Joint modelling

A shared random effects joint model consists of a model for the survival process and a model for the longitudinal process linked via the shared random effects. The original work on joint models focused on a single longitudinal process modelled using a linear mixed effects model (Hogan & Laird, 1997; Rizopoulos, 2011; Tsiatis & Davidian, 2004) but extensions have been proposed for binary and categorical longitudinal responses (Rizopoulos, 2016) as well as to accommodate multiple longitudinal variables (Chi & Ibrahim, 2006; Lin et al., 2002; Rizopoulos, 2016), and we focus on this setting here. Consider k longitudinal responses and let $Y_{ki}(t)$ denote the value of the kth response for individual i at

TABLE 1 Covariates from the UK Cystic Fibrosis Registry dataset used in the prediction models. The first 3 time-fixed variables are collected at baseline and the remaining 16 time-dependent variables are collected at each annual review

Category	Variable	Type
Baseline measure	Gender	Binary
	Genotype	Categorical
	Age at diagnosis (years)	Numeric
Lung function	Forced expiratory volume in 1 s as percentage of predicted (FEV1%)	Numeric
	FEV1% slope as estimated from mixed effects model	Numeric
	Forced vital capacity as percentage of predicted (FVC%)	Numeric
Respiratory infection in the past year	*Burkholderia cepacia*	Binary
	Methicillin-resistant *Staphylococcus aureus*	Binary
	Pseudomonas aeruginosa	Binary
	Staphylococcus aureus	Binary
Comorbidities	Cystic fibrosis-related diabetes (CFRD)	Binary
	Pancreatic insufficiency	Binary
Other health indicators	Weight (kg)	Numeric
	Body mass index (BMI)	Numeric
	Days (past year) in hospital for IV antibiotics	Categorical
	In hospital without antibiotics (past year)	Categorical
	Use of oxygen therapy (past year)	Binary
	Use of non-invasive mechanical ventilation (past year)	Binary
Calendar time	Calendar year at measurement time	Numeric

time t. We observe $Y_{ki}(t)$, realisations at specific times of the true but unobserved longitudinal process, $m_{ki}(t)$, and the difference is considered a form of measurement error (Barrett et al., 2015; Hickey et al., 2016; Rizopoulos, 2016). A multivariate linear mixed effects model for the longitudinal response is

$$Y_{ki}(t) = m_{ki}(t) + \epsilon_{ki}(t) \tag{2}$$

$$Y_{ki}(t) = X_{ki}^{\top}(t)\beta_k + Z_{ki}^{\top}(t)b_{ki} + \epsilon_{ki}(t) \tag{3}$$

where $X_{ki}(t)$ is the design matrix for the fixed effects β_k, $Z_{ki}(t)$ is the design matrix for random effects b_{ki} and $\epsilon_{ki}(t)$ are independent normally distributed residuals conditional on the model covariates and random effects. The random effects b_{ki} are independent of the error term and are assumed to follow a multivariate normal distribution with mean 0 and variance–covariance matrix D. D describes the covariance between different longitudinal measures taken at the same time. As described by Rizopoulos (2016), this model for the longitudinal responses may be extended to accommodate binary or categorical measures using different link functions.

The model for the longitudinal measures is linked to the survival process through a hazard model, with the hazard at time t assumed to depend on some function of the $\{m_{ki}(u) : u \leqslant t\}$, the unobserved true longitudinal response values up to time t and a vector of time-fixed covariates, X_i. Because the hazard depends on $m_{ki}(t)$, not $Y_{ki}(t)$, the random effects b_{ki} represent the association between the two submodels. A simple form for the hazard model assumes that the hazard at time t depends only on current values of $m_{ki}(t)$:

$$h_i(t) = h_0(t)\exp\left\{\gamma^{\top}X_i + \sum_k \alpha_k m_{ki}(t)\right\} \tag{4}$$

where $h_0(t)$ is the baseline hazard at time t. Options for specification of the baseline hazard include leaving it unspecified (as in Cox regression), modelling it as a Weibull or other distribution, or using a flexible form such as B-splines or a piecewise constant function (Rizopoulos, 2011). The model in (4) can also be extended to incorporate exogenous time-dependent predictors that are not incorporated as responses in the longitudinal model by replacing X_i with a covariate vector depending on time.

Estimation of the joint model parameters can be performed by maximum likelihood or by Bayesian Markov chain Monte Carlo in which the predicted survival probabilities are Monte Carlo estimates based on the posterior predictive distribution of the survival process (Rizopoulos, 2016). If the baseline hazard is unspecified, non-parametric maximum likelihood is used with cumulative incidence modelled as a step function (Rizopoulos, 2010; Wulfsohn & Tsiatis, 1997). Because numerical integration is typically required, for large datasets such as those based on electronic health records, the computation is challenging and may even be infeasible (Paige et al., 2018).

3.3 | Landmarking

In the landmarking approach, the dynamic prediction at a given landmark time is based on a model fitted only on those patients still at risk at the landmark time. Considering landmark time s and a clinically relevant prediction time period, v, predictions of survival to time $t_{hor} = s + v$, conditional on survival and predictors up to time s, can be obtained based on the Cox proportional hazards model (Cox, 1972),

$$h_{s,i}(t|X_i, \mathcal{Y}_i(s), s, v) = h_0(t|s)\exp\{\gamma_s^{\top}X_i + \alpha_s Y_i(s)\}, \quad s < t \leqslant s + v \tag{5}$$

where $h_{si}(t)$ is the hazard for individual i at landmark time s and $Y_i(s)$ is the value of the time-dependent predictors at s. As an extension, values of time-dependent measurements made prior to s may also be included. The required dataset, which we call a sliding landmark dataset, only contains individuals still at risk at s and all individuals are administratively censored at $s + v$.

The basic landmarking approach incorporates longitudinal predictors in the sliding landmark dataset by taking the most recently measured value prior to landmark time s. To avoid loss of longitudinal information due to this 'last observation carried forward' method, predictions of longitudinal measurements based on mixed modelling can be used (Paige et al., 2018). Either way, predictors $Y_i(s)$ appear in the dataset as time-fixed covariates at time s. The proportional hazards model in (5) can be fitted separately for each landmark time $s = \{s_1, \ldots, s_L\}$ leading to L separate models; however, this is likely to be inefficient. van Houwelingen (2007) proposed fitting a combined model across multiple landmark times with efficiency gained through assumptions such as common log hazard ratio parameters across landmark times. An example of a combined model, sometimes called a supermodel, is

$$h_i(t|X_i, \mathcal{Y}_i(s), s, v) = h_{0,s}(t|s)\exp\{\gamma^{\top}X_i + \alpha Y_i(s)\} \tag{6}$$

where $h_{0,s}$ is the baseline hazard for predictions made from landmark time s. The model may be expanded to account for time-varying effects or landmark-dependent effects by letting the parameters depend on $t - s$ or s and we refer to van Houwelingen and Putter (2012) for other extensions to the supermodel. The supermodel in (6) can be fitted to a dataset formed by vertically stacking the data created at each landmark time—the sliding landmark datasets—into one landmark super dataset using a Cox regression model with the baseline hazard stratified by landmark time s. For inference, robust standard errors are required through use of the sandwich estimator.

Estimates of the predicted survival probabilities are obtained from this model using

$$\hat{\pi}_i(s+v|s) = \exp\{-\hat{H}_{0,s,v}\exp(\gamma^\top X_i + \alpha Y_i(s))\} \tag{7}$$

and Breslow's estimate of the cumulative baseline hazard functions, $H_{0,s,v}$ (Breslow, 1972). Landmarking is a less computationally intensive approach compared with joint modelling and can be implemented with standard software.

4 | DYNAMIC PREDICTION USING A MACHINE LEARNING ENSEMBLE

4.1 | Discrete-time survival analysis

Most machine learning algorithms were originally conceived for predicting binary outcomes (classification problems) or continuous outcomes. Their use for prediction of time-to-event outcomes, particularly in the presence of right-censoring, generally requires either modification of the algorithm or manipulation of the data into a form suitable for applying techniques for binary outcomes. Some machine learning algorithms have been adapted for right-censored data and these include survival trees (Bou-Hamad et al., 2011; Segal, 1988, 1997), random survival forests (Ishwaran et al., 2008, 2014), support vector machines (Van Belle et al., 2011), artificial neural networks (Ripley et al., 2004), and extended random forest and gradient boosting (Hothorn et al., 2006). Early work by Liestol et al. (1994) and Biganzoli et al. (1998) explored discrete-time models for survival analysis in neural networks. However, extensions to accommodate time-dependent measures of predictors are still limited and we are aware of only one, *Dynamic-Deep Hit*, a custom deep learning approach designed to learn survival distributions given longitudinal data, created for dynamic prediction (Lee et al., 2019).

In this work, we take an approach that enables us to exploit a large library of machine learning algorithms that are capable of estimating the conditional probability of a binary outcome without one-by-one modification for right-censored data. For this, the data are transformed into a discrete-time format such that each individual has a record in each of a series of short time intervals at which they are at risk. We first outline the discrete-time approach in general terms, starting with the simplified setting of a single time origin, before extending to the dynamic setting. For discrete-time survival analysis, the follow-up time is divided into a sequence of d adjoining time periods from landmark time $s = a_0$ to the end of the prediction horizon $s + v = a_d$: $(a_0, a_1], (a_1, a_2], \ldots, (a_{d-1}, a_d]$. In discrete-time, the 'hazard' in period $(a_{l-1}, a_l]$ is the conditional probability of an individual having an event in time period $(a_{l-1}, a_l]$ given that the individual was event-free up to time a_{l-1} and given time-fixed covariates X_i and the values of the time-dependent covariates at the time origin $\mathcal{Y}_i(a_0)$ (Allison, 1982):

$$P_{li}(X_i, \mathcal{Y}_i(a_0)) = \Pr\{a_{l-1} < T_i \leqslant a_l | T_i > a_{l-1}, X_i, \mathcal{Y}_i(a_0)\} \tag{8}$$

Because these are conditional probabilities, the conditional survival probability is

$$S_{li}(X_i, \mathcal{Y}_i(a_0)) = \Pr(T_i > a_l) = \prod_{j=1}^{l} (1 - P_{ji}(X_i, \mathcal{Y}_i(a_0))) \tag{9}$$

This discrete-time formulation is general and any method for computing the probability of a binary event could be applied. We refer to Tutz and Schmid (2016) for a comprehensive overview of discrete time-to-event modelling. The most common fully parametric statistical approach to estimating the conditional probabilities in (8) is a logistic regression model (Cox, 1972; D'Agostino et al., 1990):

$$\operatorname{logit} P_{li}(X_i, \mathcal{Y}_i(a_0)) = \theta_l + \gamma^{\top} X_i + \alpha Y_i(a_0) \tag{10}$$

where θ_l ($l = 1, 2, \ldots, l$) is a set of parameters capturing the baseline hazard in each discrete interval. The model can be fitted in a pooled way across all discrete time periods. An alternative parameterisation is the complementary log–log model, the discrete-time equivalent of a Cox proportional hazards model (Tutz & Schmid, 2016). In the formulation in (10), the coefficients for X_i and $Y_i(a_0)$ are assumed constant across time periods (i.e. proportional hazards), but more generally they could be allowed to be time dependent. The baseline hazard may be constrained to some particular shape, restricted via groupings or, in the most general case, left to take on any shape by allowing a separate parameter in each time period (Singer & Willett, 1993). Non-parametric machine learning approaches may also be used to estimate the conditional probabilities in (8) (Malley et al., 2012). Methods for fitting novel semi-parametric models to discrete time-to-event data are provided in a tutorial paper by Berger and Schmid (2018).

This discrete-time analysis can be extended to the dynamic prediction setting with time-dependent covariates by adapting landmarking for use in the discrete-time setting. The landmark dataset for each landmark time $s_1, \ldots, s_L$ is separately discretised, as outlined above, and values of time-dependent covariates at the landmark time are used. Figure 1 illustrates the discretisation of a landmark super dataset.

The discrete-time equivalent of the landmark supermodel for dynamic prediction in Equation (6) is:

$$\operatorname{logit}\ P_{s,li}(X_i, \mathcal{Y}_i(s), s) = \theta_{s,l} + \gamma^{\top} X_i + \alpha Y_i(s) \tag{11}$$

This model makes the strict assumption that the parameters, γ and α, do not depend on landmark time. Using the relationship from (9), the corresponding predicted v-year survival probability is:

$$\hat{\pi}_i(t_{\text{hor}} = s + v | X_i, \mathcal{Y}_i(s), s) = \prod_{l=1}^{d} (1 - \hat{P}_{s,li}(X_i, \mathcal{Y}_i(s), s)) \tag{12}$$

For non-parametric machine learning methods such as random forests, model-free estimates of the conditional probabilities of the binary response for each discrete interval are estimated and, as in (12), are multiplied to obtain the predicted v-year survival probability.

Details of how to prepare a discrete-time landmark super dataset are provided in the Supplementary Information. The analyses using the Super Learner ensemble described below are all based on the application of models or machine learning algorithms for binary outcomes to the discretised landmark super dataset.

4.2 | Super learner

The hazard in Equation (8) can be estimated using fully parametric models (e.g. logistic regression), semi-parametric methods (e.g. GAMs) or machine learning algorithms that estimate the conditional

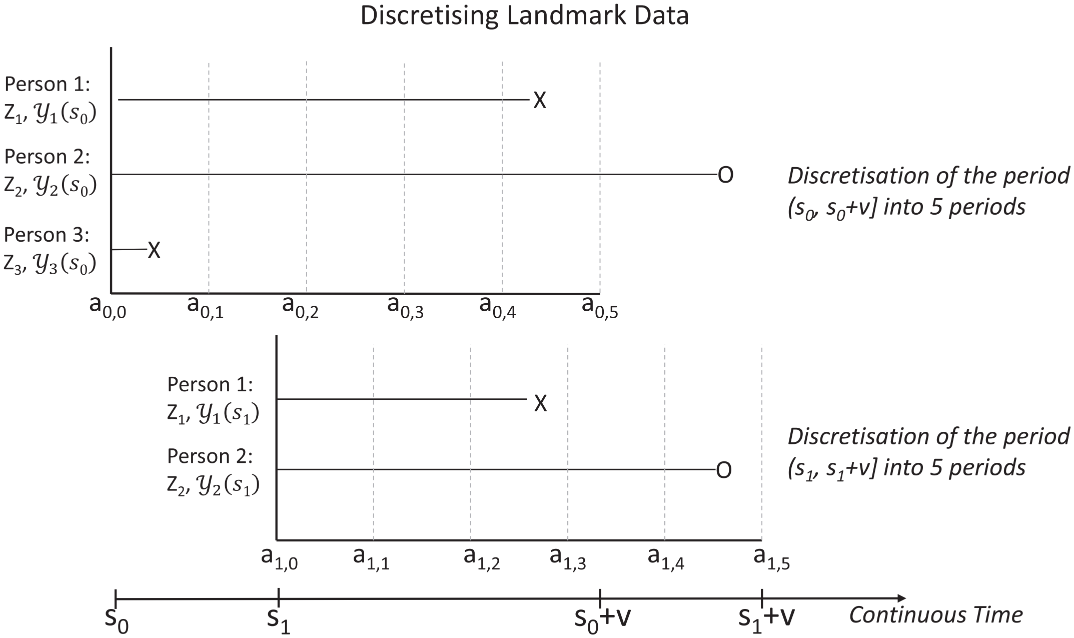

FIGURE 1 Graphical representation of the discretisation of landmark data. Event and censoring times are indicated by X and O symbols, respectively. At landmark times s_0, and s_1, the prediction period is divided into five discrete intervals and the event status of each person in each period is recorded. For landmark times s_0 and s_1, Person 1 will have an event indicator equal to one in time periods $(a_{0,4}, a_{0,5}]$ and $(a_{1,2}, a_{1,3}]$, respectively, and an event indicator equal to zero for all other periods that they were at risk. Because Person 2 is at risk of an event at the start of all five periods for both s_0 and s_1, they will have a record for each interval with an event indicator equal to zero. Person 3 will only have one record, the record corresponding to the event in period $(a_{0,0}, a_{0,1}]$

probability of a binary outcome (Malley et al., 2012). The performance of different machine learning algorithms and statistical models will differ across applications, depending on the features of the dataset and the interrelations between covariates and between the covariates and the outcome (van der Laan & Rose, 2011). Ensemble learning, using a combination of different algorithms or statistical models rather than just one, is designed to remove the dilemma of how to choose a single method. In an ensemble learning system, a library of independently fitted (or ‘trained’) algorithms are each used to predict the target—here, the conditional survival probability—then the ensemble combines the predictions from the component algorithms based on predefined rules or the results of another algorithm.

The Super Learner is a machine learning ensemble developed by van der Laan et al. (2007) that has been implemented in R and SAS. It is in the class of ensembles described as ‘stacking’ algorithms, which employ a separate learner to optimally combine the predictions from the library of algorithms (LeDell, 2016; Wolpert, 1992). The Super Learner is underpinned by theory showing that, given a bounded loss function, it will perform asymptotically as well as the best individual algorithm and asymptotically as well as the optimal combination of learners (van der Laan & Dudoit, 2003; van der Laan et al., 2007).

Here we outline the general use of the Super Learner, as described by Polley et al. (2011), before extending to the discrete-time survival setting (Polley & van der Laan, 2011) and finally to the dynamic discrete-time survival setting. The inputs to the Super Learner algorithm are the data on outcomes and predictors, a user-specified list of algorithms (or ‘learners’) to use and a loss function for quantifying the prediction error. Table S1 provides a brief description of each algorithm and model used in our application to CF data, which are: random forest, gradient boosting, support vector machine, generalised linear models (GLM), lasso, elastic-net GLM, Bayesian GLM and GAMs. These algorithms were selected to provide error 'diversity' while being accessible to statisticians and analysts (see Section 7 for a discussion of error diversity). Additional algorithms are available and, in theory, almost any algorithm could be used within the Super Learner framework by creating a wrapper function to facilitate data exchange between the code for the algorithm and the Super Learner framework. When choosing a list of algorithms, a diverse set of algorithms should be selected for optimal performance (Brown et al., 2005). Candidate learners may include both different algorithms and multiple versions of the same algorithm with different tuning parameters or different subsets of data (Brown et al., 2005; LeDell, 2016; Polley et al., 2011).

The Super Learner involves fitting each of the specified algorithms to the data in a cross-validation procedure, and deriving the optimal combination of the algorithms as defined by the combination that minimises the prediction error. The procedure is described in detail below.

The Super Learner algorithm uses the following steps (Polley et al., 2011):

1. The data are split randomly into V folds (e.g. $V = 10$). By removing each fold in turn as a test set and retaining the remaining $V - 1$ folds as a training set, this provides V training sets and V corresponding test sets.
2. Each of the Q individual algorithms is fitted to each of the V training sets and predicted outcomes are obtained from each algorithm for each individual in the corresponding V test sets.
3. A vector of length Q of optimal weights for each algorithm, ω, is determined by finding the values of ω that minimise the expected value of the selected loss function $\mathcal{L}$ over the n individuals. Let Π be the vector of length n of actual outcomes, $\Pi \in \{0,1\}$ and Ψ be the Q x n matrix of predicted conditional survival probabilities where Ψ_{ij} is the predicted outcome for individual i using algorithm j. By restricting to a convex combination of ω and selecting a squared error loss function, we formulate the minimisation problem to determine the optimal weights as a non-negative least squares problem:

$$\begin{aligned} &\min \frac{1}{n}(\omega\Psi - \Pi)^2 \\ &\text{s.t. } \omega \geqslant 0, \sum \omega = 1 \end{aligned} \tag{13}$$

In practice, we first solve for the non-negative weights and then rescale so they sum to one.

4. The final prediction function is then obtained by fitting each learner to the complete data and using the weights ω, from the previous step, to combine them. In other words, the predicted conditional survival probability for individual i is a weighted combination of the predicted conditional survival probabilities from each individual algorithm.

4.3 | Super learner landmark approach for dynamic prediction

To use the Super Learner ensemble for dynamic survival prediction allowing for time-dependent predictors, we formulate the problem as one of predicting binary outcomes in short time periods defined through the discrete-time landmark super dataset-up outlined in Section 4.1. We also require a loss function to quantify the prediction error and determine the final Super Learner prediction function.

In a standard setting of predicting binary outcomes, common choices for loss functions include the squared error loss function and the negative log loss function (Polley et al., 2011). For survival problems, a squared error loss weighted by the inverse probability of censoring, akin to the Brier score, may be used. In discrete-time survival analysis, these loss functions are defined with respect to the conditional probability of having an event in an interval, (i.e. the hazard) (Polley & van der Laan, 2011). In our motivating example, however, we are interested in predicting survival probabilities over 2 and 5 years. Because our interest is in measuring the loss on the conditional survival function, a more appropriate loss function is a squared error loss on the v-year survival probability (Polley & van der Laan, 2011). This is obtained through an inverse probability of censoring weighted (IPCW) squared error loss function where the IPCWs are used to adjust the loss to account for the information lost due to censoring by re-weighting the individuals who were not censored. This loss function is the squared difference between actual v-year survival and predicted v-year survival, multiplied by an IPCW and then summed across all observations present in each landmark time risk set R_s:

$$\begin{aligned}\mathcal{L}_{v-year,IPCW} = \sum_{s=s_1,\ldots,s_L} \frac{1}{n(s)} \sum_{i\in R_s} \{ & (0-\hat{\pi}_i(s+v|X_i,\mathcal{Y}_i(s),s,v))^2 I(T_i \leqslant s+v,\delta_i=1)(1/\hat{G}_s(T_i)) \\ & +(1-\hat{\pi}_i(s+v|X_i,\mathcal{Y}_i(s),s,v))^2 I(T_i > s+v)(1/\hat{G}_s(s+v))\}\end{aligned} \tag{14}$$

Here, $\widehat{G}_s(t)$ represents the Kaplan–Meier estimate of the censoring distribution at time t, estimated using the data for risk set R_s. This equation shows that we are computing a sum across landmark times in which each individual contributes one value per discrete interval per landmark time that he or she is at risk. Because this loss function compares actual and predicted v-year survival as opposed to the conditional probability of an event in a single interval, it is a more appropriate measure of performance. We use the v-year IPCW squared error loss to determine the weights in the final Super Learner prediction function by minimising $\mathcal{L}_{v-\text{year},IPCW}$ for an algorithm-weighted combination subject to the algorithm weights being non-negative. Details on the implementation using R and availability of sample code are given in Section 6.2.6.

5 | ASSESSMENT OF PREDICTIVE PERFORMANCE

Measures for assessing performance of a predictive model fall into three categories: discrimination (how well the model distinguishes between two subjects with different survival times), calibration (how well the predicted survival probabilities match the observed survival probabilities) and overall performance (a distance-based measure of predictive accuracy) (Steyerberg et al., 2010). In a survival context, measures of discrimination must account for censoring and survival time, not just survival status (Pencina et al., 2012). van Houwelingen and Putter (2012, Chap. 3) summarise measures of predictive performance in the context of dynamic prediction. We use the truncated concordance index (C-index) described by Gerds et al. (2013) which truncates an IPCW C-index at a specified time point that may be earlier than the longest follow-up time. A model is well calibrated if there is good agreement between the predicted survival and observed survival at all levels of event risk. We construct calibration plots to assess this.

The Brier score is an overall performance measure; it measures the mean-square error for the case where the actual outcomes are binary and the predictions are a probability. Following Graf *et al.* (1999), in the presence of censoring an IPCW estimator for the Brier score for assessing prediction of survival to time $s + v$ conditional on survival to landmark prediction time s is given by the element in the sum of the v-year IPCW loss in Equation (14) corresponding to landmark time s.

Because the numeric value of the Brier score has little meaning, the percentage reduction in the Brier score relative to the score obtained from a Kaplan–Meier estimate of v-year survival (i.e. without using any covariates) obtained separately from each landmark prediction time provides a more intuitive assessment of overall performance (van Houwelingen & Putter, 2012). A percent reduction of 100% represents perfect prediction and occurs when the absolute Brier score is zero.

In the context of survival analysis, discrimination and overall performance measures are calculated using IPCWs. A Kaplan–Meier estimate of the censoring distribution, $\widehat{G}(t) = P(C > t)$, can be used to obtain IPCWs with individuals who die, censored at the time of death. Note that many IPCW methods require that C_i be independent of both T_i^* and the covariates.

Robust model assessment requires testing on an external dataset (Altman & Royston, 2000). Therefore, the dataset is split between a training (model development) sample and a test sample. To provide external validation, it has been recommended that the split should be non-random, for example it might be divided according to geography or calendar time (Steyerberg & Vergouwe, 2014). The test sample is used to measure predictive performance of the final model and illustrates how the algorithm would be used in practice; cross-validation is used to assess performance in the training sample. More information is given on this below, in the context of the example.

6 | ANALYSIS OF THE UK CYSTIC FIBROSIS REGISTRY DATASET

6.1 | Study population

The study sample consisted of all individuals in the UK CF Registry who had an annual review between 1/1/2005 and 31/12/2015 and were 16 years of age or greater at the time of the review. Furthermore, any individual without genotype information and without at least one measurement of forced expiratory volume in 1 s as percentage of predicted (FEV1%), forced vital capacity as percentage of predicted (FVC%), body mass index (BMI) and weight from one annual review were omitted (3% of individuals). The resulting dataset consisted of 6363 unique individuals with 43,880 annual review records. 962 of these had the composite event of either death or lung transplant. Seven per cent of individuals had only a single visit recorded in the registry while 55% had seven visits or more recorded.

The amount of missing data is relatively low. Less than 1% of respiratory infection and time spent in hospital data were missing and values were filled in using last observation carried forward and by setting the value to zero, respectively. For the critical lung function predictors, FEV1% and FVC%, approximately 12% of all records had missing values.

6.2 | Implementation of dynamic survival prediction methods

In this section, we outline how the traditional landmarking analysis, joint modelling and the Super Learner landmark approach were implemented in this dataset. For all three methods, age was taken as the timescale.

6.2.1 | Creation of training and test datasets

Performance of all three dynamic prediction methods was measured using both 10-fold cross-validation and a holdout sample (test data). We created the test dataset for validation from all patients at 18 randomly selected adult CF centres in the United Kingdom, representing 20% of the annual review records. Figure 2 depicts this division of data and the 10-fold cross-validation procedure. This test dataset contained 14% of the patients and was used to assess generalisability of the predictive model within the United Kingdom. The training dataset was formed from the remaining patient data and was randomly partitioned by patient ID into 10-folds for cross-validation. Each fold contained approximately 540 individuals and the number of events observed ranged from 94 to 119.

6.2.2 | Implementation of the joint model

In CF, poor lung function has the greatest impact on survival and quality of life (Horsley et al., 2015). FEV1%, a measure of airway obstruction, has received the most attention as a gauge of disease severity and as a predictor of survival for CF patients (Keogh et al., 2018; Liou et al., 2001; Szczesniak et al., 2017). Because FEV1% is measured contemporaneously with FVC%, we modelled these two measures using a multivariate mixed effects model in the longitudinal submodel. FEV1% and FVC% were modelled linearly as a function of age with random intercept and slope as in Equation (3). The remaining time-dependent covariates (see Table 1) were treated as exogenous predictors and incorporated along with the time-fixed covariates into the survival submodel of Equation (4) by fitting an extended Cox model with data specified in long format (i.e. one row per observed measurement time, variables identifying the beginning and end of the time interval in each row, and an event indicator per row). The logarithm of the baseline hazard function was modelled using a B-spline with 15 knots. We considered three association structures for the joint model: one assuming the hazard of an event at time *t* depends on the current value of the longitudinal outcomes at time *t*, one assuming that it depends on both the current values and the slope of FEV1% and one assuming that it depends on the cumulative effect of FEV1% up to time *t*.

6.2.3 | Implementation of the landmarking models

For both the traditional landmarking analysis using a Cox regression and the Super Learner landmark approach, FEV1% and FVC% were modelled as a linear function of age with random intercept and slope using a multivariate mixed effects model. The model also included the fixed effects of the other covariates in Table 1. Replacing the observed values of FEV1% and FVC% in the dataset with predicted values offers three advantages: (a) the predicted values account for measurement error; (b) we can predict values at each landmark age instead of carrying forward a value from the last measurement

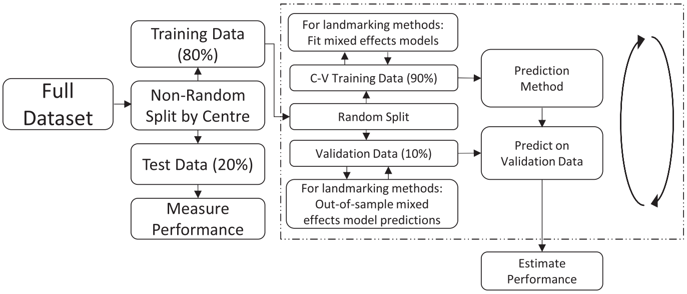

FIGURE 2 Creation of training and test datasets and use of cross-validation for estimating performance. The 'Prediction Method' is either joint model, Cox landmark or Super Learner landmark. Both 10-fold cross-validation and test data are used to measure performance

time and (c) we can predict values even when observed values were missing. To capture information about the trajectory of lung function, we also add the individual's modelled FEV1% slope as a predictor to the landmark models (Keogh et al., 2018; Rizopoulos et al., 2017). In each of the 10 cross-validation steps, we fitted the mixed effects models and predicted values for the longitudinal predictors FEV1%, FVC% and BMI in the nine training folds ('C-V Training Data' in Figure 2). We used out-of-sample predictions of FEV1%, FVC% and BMI in the validation fold ('Validation Data' in Figure 2) (Keogh, 2018). A model was also fitted to the complete training dataset ('Training Data' in Figure 2) and was used to predict on the test dataset ('Test Data' in Figure 2).

A stacked landmark dataset is used in both landmarking approaches. When creating a landmark stacked dataset, each individual contributes one row in the dataset for each landmark age that he or she is at risk. For the Super Learner landmark analysis, the landmark stacked dataset is discretised as described in Section 4.1. At each landmark age a person is at risk, they may contribute as many rows as the number of intervals chosen to discretise the data. For this dataset, we used five discrete intervals. Our investigations showed that larger numbers of discrete intervals made the size of the data unmanageable and resulted in decreased predictive performance, likely due to the large variance in the estimators of the numerous parameters. Since the discrete intervals need not be of equal size, we divided time based on quintiles of event times (Polley & van der Laan, 2011). Table 2 shows a hypothetical simplified landmark super dataset before and after discretisation.

A decision must be made about how to model the main effect of age time). Maximum flexibility is allowed by adding time indicators to each row as dummy variables but this increases the number of parameters to estimate and may sacrifice power and coefficient stability (Singer & Willett, 1993). In this study, we adopted a more parsimonious approach by adding an interaction between the lower bound of the discrete interval and the landmark time indicator, which models the effect of time linearly over each prediction period (as shown in the sample data in Table 2).

6.2.4 | Selection of algorithms and models for super learner

The Super Learner requires a list of candidate algorithms that are appropriate for the problem and that provide diversity and performance. Furthermore, to achieve their best performance, most of these machine learning algorithms require tuning of their hyperparameters, parameters that control some aspect of the how the algorithm functions. To select such a list of algorithm–hyperparameter combinations, we performed some preliminary investigations to look at the effect of different hyperparameter settings on predictive performance of the algorithms. These investigations are further described in the Supplementary Information. Based on this work, we chose to implement the following models and algorithms: gradient boosting, GAMs, GLMs, penalised (ridge, lasso and elastic-net) GLMs, random forest and Bayesian GLMs.

6.2.5 | Method comparison

For all models, dynamic 2-year and 5-year survival predictions were computed at each landmark age for those individuals at risk at that landmark age. The Brier score and C-index were separately calculated at each landmark age that a prediction was made, (20, 21, …, 50) for each analysis method. Because nearly all censoring in the UK CF Registry is administrative censoring, the assumption that C_i is independent of both T_i^* and the covariates, as required for IPCW methods, is reasonable for this dataset. We construct calibration plots by dividing the predicted survival probabilities at a given

TABLE 2 A landmark super dataset is shown before and after discretisation. In the discretised dataset, the survival time has been replaced by an event indicator that only takes the value of 1 in the discrete interval in which an event happened. The two rightmost columns in the discrete dataset represent the covariates used for modelling time, which are an interaction between the lower bound of the discrete time interval and an indicator variable for each landmark time

Landmark super dataset, 5-year horizon

ID	LM time	Survival time	Event	FEV1
1	20	24.0	1	30
2	20	25.0	0	50
1	21	24.0	1	27
2	21	25.5	0	52

Landmark super dataset after discretisation

ID	Discrete interval	Event	FEV1	*I*(*LM* = 20) * time	*I*(*LM* = 21) * time
1	(20.0, 20.8]	0	30	20.0	0
1	(20.8, 21.8]	0	30	20.8	0
1	(21.8, 22.9]	0	30	21.8	0
1	(22.9, 23.7]	0	30	22.9	0
1	(23.7, 25.0]	1	30	23.7	0
2	(20.0, 20.8]	0	50	20.0	0
2	(20.8, 21.8]	0	50	20.8	0
2	(21.8, 22.9]	0	50	21.8	0
2	(22.9, 23.7]	0	50	22.9	0
2	(23.7, 25.0]	0	50	23.7	0
1	(21.0, 22.0]	0	27	0	21.0
1	(22.0, 23.0]	0	27	0	22.0
1	(23.0, 24.4]	1	27	0	23.0
2	(21.0, 22.0]	0	52	0	21.0
2	(22.0, 23.0]	0	52	0	22.0
2	(23.0, 24.4]	0	52	0	23.0
2	(24.4, 25.1]	0	52	0	24.4
2	(25.1, 26.0]	0	52	0	25.1

landmark age into fifths, computing the mean survival probability for each fifth (y-axis) and plotting this against the average Kaplan–Meier survival for the individuals in each fifth (x-axis). In a perfectly calibrated model, the line formed by joining these five points would lie exactly on the line $y = x$.

6.2.6 | Software

All analyses were performed using R v3.4.3 (R Core Team, 2017). The joint model and the longitudinal submodel were fitted using the R package `JMbayes` (Rizopoulos, 2016), the Cox landmark supermodel was fitted using the R package `survival` (Therneau, 2015; Therneau & Grambsch,

2000) and the Super Learner landmark supermodel was fitted using the R package `SuperLearner` (Polley et al., 2018) with a user-defined IPCW squared error loss function. R packages used for the individual algorithms in the Super Learner ensemble are provided in the Supplementary Information. We used the R package `nlme` (Pinheiro et al., 2018) to fit the mixed effects model for FEV1% and FVC% and the Lawson–Hanson algorithm (Lawson & Hanson, 1995) which is implemented in the R package `nnls` (Mullen & van Stokkum, 2012) for the non-negative least squares problem. The Brier score and C-index were computed using the R package `pec` (Mogensen et al., 2012).

R code for using Super Learner to fit a discrete-time landmark supermodel like the one described here is available from https://github.com/KamTan/DynamicPrediction. This code is illustrated using the Mayo Clinic Primary Biliary Cirrhosis dataset publicly available via the R package `survival`. Although we are not permitted to make the UK CF Registry data public, interested researchers may apply for this data by following the instructions provided here: https://www.cysticfibrosis.org.uk/the-work-we-do/uk-cf-registry/apply- for-data-from-the-uk-cf-registry.

6.3 | Results

6.3.1 | Cross-validated performance using the training dataset

We first compare the performance of a joint model, a traditional landmark super model analysis using Cox proportional hazards (‘Cox landmark’) and a discrete-time landmark super model analysis using the Super Learner ensemble (‘SL landmark’) using 10-fold cross-validation on the training dataset.

We investigated Cox landmark models with time-varying effects, with interactions between predictors and with quadratic and cubic terms added for continuous predictors. None of those models offered superior Brier scores or C-index values at a majority of landmark ages as compared to the more parsimonious model including only the linear terms and so the higher order terms were discarded. For the joint model, because no improvement in predictive accuracy or discrimination was achieved with the more complex association structures involving slope or cumulative effects, we chose to relate the hazard of an event at time t only to the current value of FEV1% and FVC% at time t.

Figure 3 shows the composition of the Super Learner prediction function for each of the 10 cross-validation folds. The GAM (2) learner has the largest weight in the Super Learner prediction function at all 10-folds, and overall there is minor variability in which algorithms are chosen and their weights in the prediction function. For example, the Boosting (C) learner receives a weight of 0.15 (15%) in fold 1 but is given a zero weight in both fold 8 and fold 10. Neither the GLM nor the Lasso learners were given a non-zero weight in any of the 10 cross-validation folds.

Table 3 shows the Brier scores for each method for 2-year and 5-year dynamic survival prediction. Figure 4 presents this information as a plot of the percentage reduction in Brier score of each method over Kaplan–Meier reference estimates. For 2-year survival prediction, the SL landmark achieved a 30% reduction in prediction error at 11 of the 31 landmark ages while the Cox landmark and joint model achieved this at only three and two landmark ages, respectively. For 5-year survival prediction, the performance of the Super Learner landmark and the Cox landmark was nearly identical. The reduction in prediction error for the SL landmark ranged from 29% to 52% compared to 28% to 51% for the Cox landmark and 22% to 44% for the joint model. For this dataset, the joint model performance is inferior to that of the two landmarking methods. As seen in Table 3, for all three methods, predictive accuracy decreases at higher landmark ages, particularly over 40 years of age, likely due to less data at older ages.

As shown in Figure 5, the discriminative ability of all three methods for 2-year survival prediction is similar with C-index values exceeding 0.80 at all landmark ages except 43- and 44 years old.

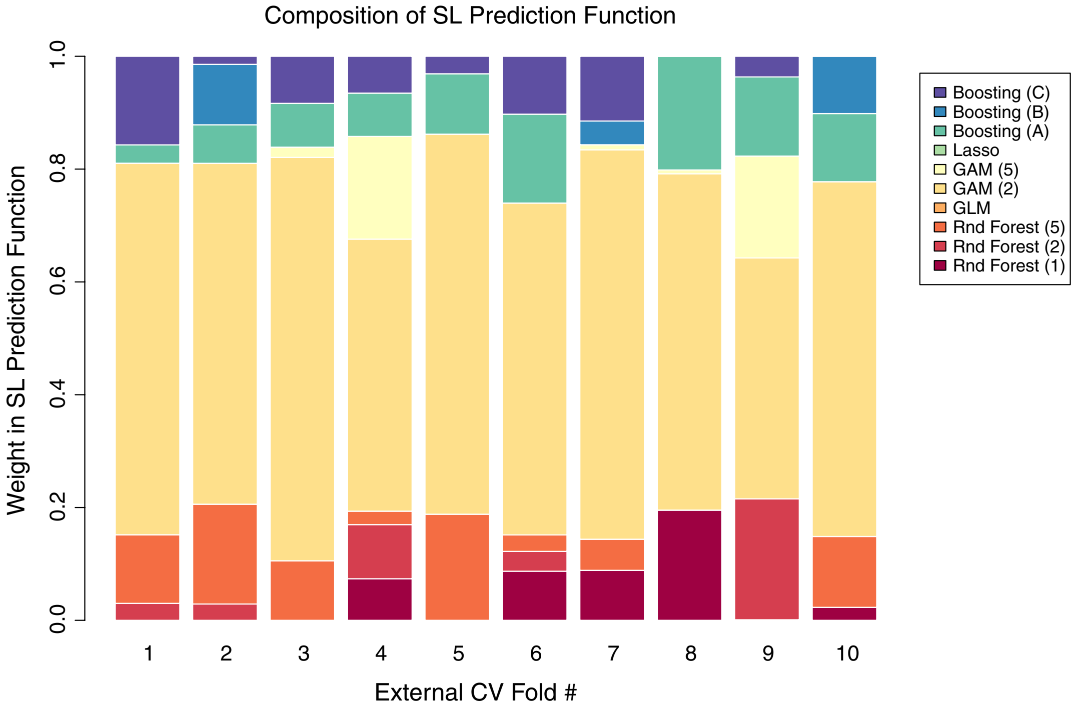

FIGURE 3 At each of the 10 external cross-validation folds, a new Super Learner prediction function is calculated based on the Super Learner internal 10-fold cross-validation loop. The composition of those prediction functions is presented with each learner plotted in a different colour. The alphanumeric character in parentheses after the name of the learner serves to distinguish different configurations of the algorithm based on hyperparameter settings. For example, GAM (2) indicates a GAM with 2 degrees of freedom while GAM (5) indicates 5 degrees of freedom were specified

For 5-year survival prediction, the Cox landmark and SL landmark models have a similar ability to discriminate with C-index measures between 0.85 and 0.90 for landmark ages 20–40 years. The joint model has a slightly lower discriminative ability than either of the two landmark methods.

Figure 6 shows calibration plots for the both the 2-year and 5-year dynamic prediction models at three prediction time points: 25, 35 and 45 years of age. There is good calibration for all three methods. At landmark age 45, where we have data from fewer individuals, the calibration is less satisfactory. The joint model also appears to over-predict 5-year survival at landmark age 35.

Performance was also assessed by fitting each of the three methods to the entire training dataset and then making dynamic survival predictions on the test dataset. Similar performance patterns were seen to those reported on the cross-validated data and details are provided in the Supplementary Information.

6.3.2 | Final model

A final model for prediction of 2-year and 5-year dynamic survival probabilities was computed using the full dataset, which includes both the training data and the test data. The composition of the final model Super Learner prediction function is presented in Table 4 along with the value of the v-year IPCW squared error loss for each algorithm or model. For predicting both 2-year and

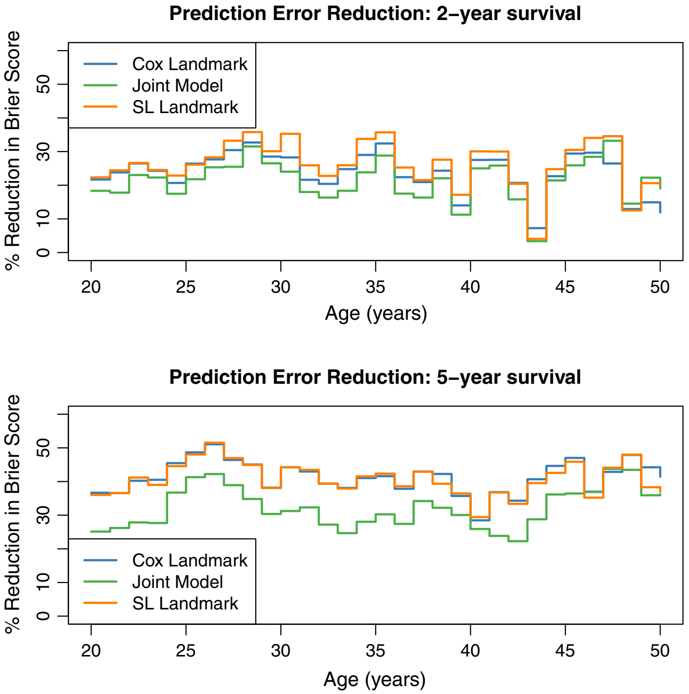

FIGURE 4 The reduction in prediction error (Brier score) over the Kaplan–Meier reference model for 2-year and 5-year dynamic survival prediction using 10-fold cross-validation on the training dataset

TABLE 3 Values of the Brier Score for the three comparison methods for 2-year and 5-year dynamic survival prediction at selected landmark ages. Lower values indicate lower prediction error. These numbers were computed using 10-fold cross-validation on the training data

		Landmark age:						
	Method	**20**	**25**	**30**	**35**	**40**	**45**	**50**
2-year survival	Cox landmark	0.04	0.04	0.05	0.05	0.04	0.05	0.08
	Joint model	0.04	0.05	0.05	0.05	0.04	0.06	0.07
	SL Landmark	0.04	0.04	0.04	0.05	0.03	0.05	0.07
5-year survival	Cox landmark	0.08	0.07	0.08	0.08	0.09	0.09	0.10
	Joint model	0.09	0.08	0.10	0.10	0.09	0.11	0.11
	SL landmark	0.08	0.07	0.08	0.08	0.08	0.09	0.11

5-year survival, GAM predictions make up the majority of the prediction function, 57% and 64%, respectively. Several versions of random forest and boosting algorithms contribute the remainder of the prediction function.

7 | DISCUSSION

We have described a method for adapting the landmarking technique for dynamic survival predictions to be used with a machine learning ensemble that includes statistical models and algorithms for

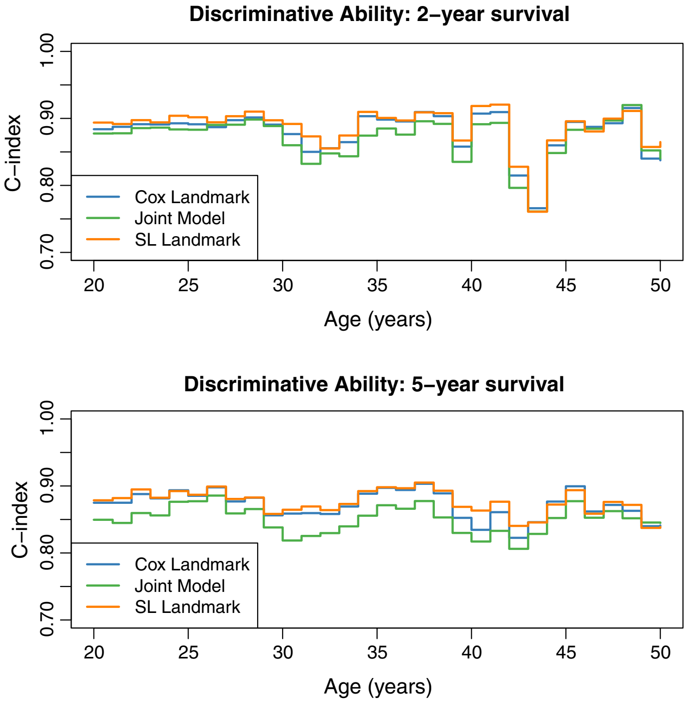

FIGURE 5 Comparison of discriminative ability for the SL landmark, the Cox landmark and the joint model using 10-fold cross-validation on the training data. Discrimination is measured by the C-index, with higher values preferred

estimation of the probability of a binary outcome. A key advantage of the ensemble is that it obviates the need to make an a priori choice as to which statistical model or class of machine learning algorithm will perform best for the problem at hand. The obvious disadvantages are greater complexity and computational cost. Another disadvantage of the ensemble is that it obscures the nature of relationships between predictors and outcomes and, for inference, confidence intervals are not readily available. For the UK CF Registry study population used here, the SL landmark method performed as well as the current state-of-the-art Cox landmark method for dynamic survival prediction both in external cross-validation and on a test dataset, but it did not outperform this simpler approach. Both methods had better predictive performance than the joint model.

Depending on the true data-generating distribution, different algorithms will perform best for different problems. For this dataset and this question, a traditional statistical method performed as well as our ensemble method. Given the high level of discrimination achieved by both methods, there is not much room to make gains by fitting more elaborate models or including more predictors. It is of interest to better understand when there will be performance differences between approaches. For example, certain alternative statistical models or algorithms may be particularly useful when the number of predictors greatly exceeds the number of observations, for cases with interactions among the predictors and when there is non-linearity. Not surprisingly, non-parametric machine learning methods can outperform parametric models when the assumptions of the parametric models are violated. As shown by Gong et al. (2018) and Katzman et al. (2018), machine learning techniques outperformed Cox regression models in a static survival prediction setting when the effects of covariates were nonlinear in the hazard. Lee et al. (2019) report a deep learning approach for dynamic survival prediction using data from the UK CF Registry that offers improved discrimination in their comparisons. Their results are

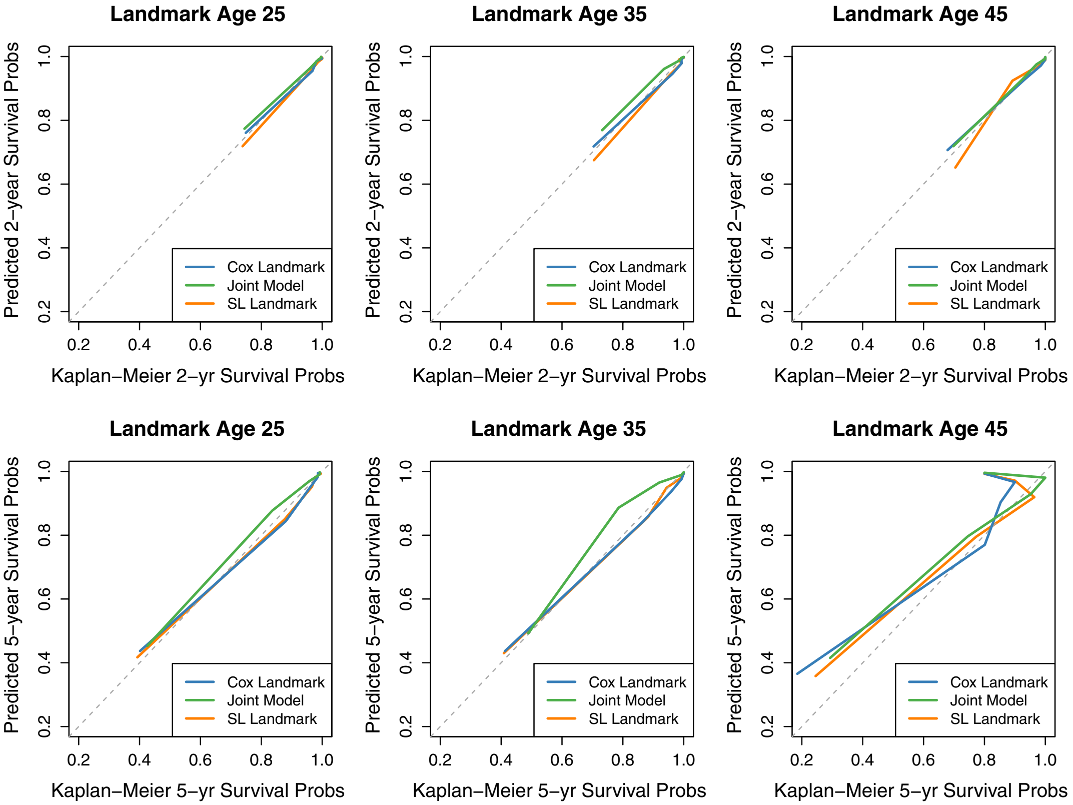

FIGURE 6 Calibration plots for 2-year (top) and 5-year (bottom) survival predictions for three selected landmark ages: 25, 35 and 45. The *y*-axis represents the mean survival probability by quintile group as calculated by each method. The *x*-axis represents the Kaplan–Meier survival probability calculated on the same individuals. The mean probabilities are averaged over the 10 external cross-validation folds. For reference, a 45-degree line is plotted in grey dashes

difficult to compare with ours as the set of predictors used was different, they adopt a competing risks framework and the specifics of the base method implementations are not fully reported in the paper.

This work was motivated by the need for dynamic predictions of survival for patients with life-shortening conditions. We used data from the UK Cystic Fibrosis Registry to illustrate the methods, but they could be applied to other locales and other health-related administrative databases. Because people with CF have regularly scheduled routine care, the longitudinal data are measured at pre-planned, consistent intervals and the amount of missing data is low. Other datasets with more sporadic patient visits or without well-defined data collection may require special modelling techniques, including potentially to accommodate a model for the occurrence of patient visits. Additionally, not all datasets will be immediately generalisable to the overall population of people with the disease. The dynamic survival predictions resulting from this work are relevant for the entire UK population of people with CF as the registry data covers approximately 99% of people with CF in the United Kingdom. Additionally, predictive ability of the SL landmark on a non-random external test dataset suggests the model is not limited in applicability to a specific set of centres in the United Kingdom.

We compared the SL landmark method to a joint model and the Cox landmark method. Instead, these methods could have been added to the Super Learner library of algorithms and each would

TABLE 4 The percent weighting of each learner in the Super Learner prediction function fit to the full dataset for both 2-year and 5-year dynamic prediction of survival. The IPCW squared error loss, the value minimised in the Super Learner regression to compute the prediction function weights, is also included

		2-year survival		5-year survival	
Category	Algorithm and hyperparameters	Weight in SL pred function	IPCW loss	Weight in SL pred function	IPCW loss
Random forest	Ranger: no. trees = 500; min node size = 1	0.0%	0.050	8.1%	0.085
	Ranger: no. trees = 500; min node size = 2	7.4%	0.050	4.2%	0.085
	Ranger: no. trees = 500; min node size = 5	10.5%	0.050	4.9%	0.085
Gradient boosting	xgBoost: eta=0.1, max depth = 3, min obs per node = 10	0.0%	0.047	0.0%	0.083
	xgBoost: eta=0.1, max depth = 4, min obs per node = 10	0.0%	0.048	0.0%	0.083
	xgBoost: eta=0.1, max depth = 4, min obs per node = 10, subsample = 0.5	11.8%	0.050	0.0%	0.087
	xgBoost: eta=0.1, max depth = 4, min obs per node = 1	0.0%	0.049	0.0%	0.085
	xgBoost: eta=0.1, max depth = 6, min obs per node = 10	0.0%	0.050	5.6%	0.085
	xgBoost: eta=0.1, max depth = 6, min obs per node = 1	0.0%	0.050	0.2%	0.087
	xgBoost: eta=0.3, max depth = 6, min obs per node = 10	0.0%	0.056	14.8%	0.093
	xgBoost: eta=0.3, max depth = 6, min obs per node = 1	1.0%	0.054	4.9%	0.095
GLM	glm	0.0%	0.046	0.0%	0.083
	bayesglm	0.0%	0.046	0.0%	0.083
GAM	gam: degrees of freedom = 2	0.0%	0.046	57.3%	0.081
	gam: degrees of freedom = 3	0.0%	0.045	0.0%	0.081
	gam: degrees of freedom = 4	0.0%	0.045	0.0%	0.081
	gam: degrees of freedom = 5	64.3%	0.045	0.0%	0.081
Penalised regression	biglasso: penalty = lasso	0.0%	0.046	0.0%	0.084
	biglasso: penalty = ridge	0.0%	0.059	0.0%	0.133
	glmnet: penalty = 50% lasso / 50% ridge	0.0%	0.046	0.0%	0.083
Reference	mean model (no covariates)	5.1%	0.062	0.0%	0.141

either be selected or not selected in the final combination of algorithms. While any modelling technique can theoretically be included in the Super Learner ensemble, it may not always be practical or necessary. For example, we may not need to add a continuous-time proportional hazards learner because a discretised logistic model converges to the continuous-time Cox proportional hazards model as the time interval gets smaller (Thompson, 1977). To add a joint model to the Super Learner framework would require additional custom code to enable the ensemble to simultaneously work with both discrete-time and continuous-time paradigms and also with models that take fundamentally different approaches to incorporating longitudinal data. However, because the joint model produces dynamic survival predictions at given ages, there is no reason these could not be included as predictions from one of the candidate learners in the regression to determine the algorithm weightings in the Super Learner prediction function. In other words, the predicted survival probabilities from the joint model could be added as another row in the matrix Ψ and then the solution to the minimisation problem of Equation (13) would include the joint model as a candidate.

In pursuit of the best model for dynamic survival prediction, we created custom code in addition to the available `SuperLearner` R package to accommodate a v-year IPCW squared error loss function. This loss function directly measures the difference in predicted and actual conditional survival probabilities. To use the Super Learner without additional custom code, we could use a negative log likelihood loss function or a squared error loss function measured on the conditional hazard. While these are still valid loss functions, they are not directly based on the conditional survival probability we are interested in and, therefore, the resulting Super Learner prediction function may be inferior (Polley & van der Laan, 2011). To investigate how the results obtained using the squared error loss function and log likelihood loss function differ from those using the IPCW squared error loss function, we compared the Brier scores that result from running Super Learner with the same candidate library of algorithms using the three different loss functions. Overall, the performance was nearly identical. A plot comparing the predictive accuracy is available in the Supplementary Information, Figure 3. The difference between the three methods may be greater for a different dataset but this suggests that using a loss function defined on the conditional hazard would not give grossly different results and it has the advantages of being included in the core package in R and not requiring an estimate of the censoring function.

It is interesting to compare the IPCW squared error loss for the learners that were selected for the final prediction function with the loss for the learners that were not selected. From Table 4, we can see the IPCW squared error loss for 5-year survival prediction was 0.081 for all four algorithm-hyperparameter combinations of the GAMs yet a non-zero weight was only allocated to one of them. Furthermore, the GLM learners had a loss of only 0.083 and the lasso had a loss of only 0.084 yet neither of these categories were selected to contribute predictions to the final prediction function. The explanation for this behaviour lies in the attempt to balance bias, variance and covariance which is achieved through error diversity. Simply combining algorithms with the lowest IPCW squared error loss will not necessarily deliver the best performance, especially if their predictions and, therefore their errors, are strongly correlated. In this study, we have encouraged error diversity by including fundamentally different algorithms as well as multiple versions of algorithms with different hyperparameters. Other methods for bringing diversity to the ensemble include repeatedly providing different subsets of the training data to the algorithm (as in bagging), providing a different set of covariates to the algorithm, or by distorting the predictor data in some way (Brown, 2004). Because an ensemble balances the covariance as well as the bias and variance, diversity may be key to performance improvement.

In this study, we fit a single landmark supermodel to a stacked dataset across all landmark ages but we could have fit one separate model per landmark age in a sliding landmark analysis. A sliding landmark analysis allows for a unique Super Learner prediction function with different algorithm weightings at each landmark age. The price for this flexibility is a great deal more parameters to estimate. In this dataset, where there are significantly fewer observations at older ages, there is a benefit to borrowing data from other landmark ages in the landmark supermodel. We found the performance of the landmark supermodel to be superior to that of a sliding landmark analysis.

To take advantage of a vast range of existing algorithmic techniques, we took the approach of discretising continuous time-to-event data. Another option for enabling the application of machine learning to survival outcomes involves augmenting the data to accommodate censoring thereby allowing standard regression-style algorithms to be used. Examples include inverse probability of censoring weighting the data (Goldberg & Kosorok, 2017; Rotnitzky & Robins, 2005; Vock et al., 2016) and pseudo-observations (Andersen & Pohar Perme, 2010; Parner & Andersen, 2010).

A common interest in machine learning is identifying important predictors. When faced with large datasets containing more (possibly collinear) predictors than observations, machine learning has had success in identifying a subset of important predictors. Feature selection may be achieved through an initial screening or filtering step using a correlation-based or information-based ranking of predictors (Guyon & Elisseeff, 2003). Alternatively, wrapper or embedded methods for feature selection build the model and select the variables simultaneously. The lasso regression algorithm is a good example from this category (Tibshirani, 1996, 1997). As feature selection was not the goal of this study, the list of predictors used was based on those used in current models in the literature for predicting CF survival. Because the UK CF Registry contains hundreds of variables per measurement time, this may be an interesting area for further research and was touched upon by Alaa and van der Schaar (2018) in their study of prediction of 3-year mortality for people with CF and also by Lee et al. (2019).

The relative predictive performance of the joint model in this study was inferior to the two landmarking approaches. A likely explanation involves the addition of the time-varying covariates to the survival submodel. Because these covariates can only be measured if the patient is alive (i.e. they are endogenous), they would properly be incorporated into the longitudinal submodel via mixed-effects modelling (Rizopoulos, 2012). However, the large number of predictive time-varying covariates (many of which are binary) in this study makes this infeasible and so, to avoid losing the information contained in these covariates, we included them as if they were exogenous in the survival submodel. When making predictions, the joint model assumes that the value of exogenous time-varying covariates remains constant in the future at the most recently measured level. This assumption is not realistic as we expect the values to change over time with the patient's health. For datasets like this one where there are a large number of time-varying covariates, the landmarking approaches may be more suitable.

The goal of this study was to provide a thorough comparison of the main methods for dynamic survival prediction, to show how a machine learning ensemble method can be used for dynamic survival prediction and to elucidate the numerous considerations and decisions that must be made during the analysis. The performance and completeness in reporting of machine learning methods is mixed. In a recent review by Christodoulou et al. (2019), many machine learning studies failed to report methods for tuning hyperparameters, did not assess calibration, and neglected to clearly report on all decisions and steps taken in the analysis. This difference in reporting standards likely contributes to the existing tension between traditional statistical methods and machine learning. Efforts to improve in this area will help to facilitate an understanding that we do not have to make a choice between machine learning algorithms and statistical models; they are both valuable tools for statisticians and analysts to use in clinical prediction models.

ACKNOWLEDGEMENTS

We thank people with cystic fibrosis and their families for consenting to their data being held in the UK Cystic Fibrosis Registry, and NHS teams in CF centres and clinics for the input of data into the Registry. We also thank the UK Cystic Fibrosis Trust and the Registry Steering Committee for access to anonymised UK Cystic Fibrosis Registry data.

RHK was supported by a MRC fellowship (MR/M014827/1) and a UKRI Future Leaders Fellowship (MR/S017968/1). RMD is supported by a Sir Henry Dale Fellowship from the Wellcome Trust and the Royal Society (grant 107617/Z/15/Z).

REFERENCES

Aaron, S.D., Stephenson, A.L., Cameron, D.W. & Whitmore, G.A. (2015) A statistical model to predict one-year risk of death in patients with cystic fibrosis. *Journal of Clinical Epidemiology*, 68, 1336–1345.

Alaa, A.M. & van der Schaar, M. (2018) Prognostication and risk factors for cystic fibrosis via automated machine learning. *Scientific Reports*, 8, 11242.

Allison, P.D. (1982) Discrete-time methods for the analysis of event histories. *Sociological Methodology*, 13, 61–98. Available from: http://www.jstor.org/stable/270718?origin=crossref.

Altman, D.G. & Royston, P. (2000) What do we mean by validating a prognistic model? *Statistics in Medicine*, 19, 453–473.

Andersen, P.K. & Pohar Perme, M. (2010) Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19, 71–99. Available from: http://journals.sagepub.com/doi/10.1177/0962280209105020.

Andrinopoulou, E.-R., Rizopoulos, D., Geleijnse, M.L., Lesaffre, E., Bogers, A.J.J.C. & Takkenberg, J.J.M. (2015) Dynamic prediction of outcome for patients with severe aortic stenosis: application of joint models for longitudinal and time-to-event data. *BMC Cardiovascular Disorders*, 15, 28. Available from: http://bmccardiovascdisord.biomedcentral.com/articles/10.1186/s12872-015-0035-z

Aram, P., Trela-Larsen, L., Sayers, A., Hills, A.F., Blom, A.W., McCloskey, E.V., Kadirkamanathan, V. & Wilkinson, J.M. (2018) Estimating an individual's probability of revision surgery after knee replacement: a comparison of modeling approaches using a national dataset. *American Journal of Epidemiology*, 187, 2252–2262. https://doi.org/10.1093/aje/kwy121

Barrett, J., Diggle, P., Henderson, R. & Taylor-Robinson, D. (2015) Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and exact likelihood inference. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 77, 131–148.

Berger, M. and Schmid, M. (2018) Semiparametric regression for discrete time-to-event data. *Statistical Modelling*, 18, 322–345.

Biganzoli, E., Boracchi, P., Mariani, L. & Marubini, E. (1998) Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in Medicine*, 17, 1169–1186.

Bou-Hamad, I., Larocque, D. & Ben-Ameur, H. (2011) A review of survival trees. *Statistics Surveys*, 5, 44–71. Available from: http://projecteuclid.org/euclid.ssu/1315833185.

Breiman, L. (2001) Statistical modeling: the two cultures. *Statistical Science*, 16, 199–231.

Breslow, N.E. (1972) Contribution to the discussion of paper by D.R. Cox. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 216–217.

Brown, G. (2004) Diversity in Neural Network Ensembles. Doctor of philosophy, The University of Birmingham.

Brown, G., Wyatt, J., Harris, R. & Yao, X. (2005) Diversity creation methods: a survey and categorisation. *Journal of Information Fusion*, 6, 5.

Chi, Y.Y. & Ibrahim, J.G. (2006) Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, 62, 432–445.

Christodoulou, E., Jie, M., Collins, G.S., Steyerberg, E.W., Verbakel, J.Y. & van Calster, B. (2019) A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110, 12. Available from: http://www.sciencedirect.com/science/article/pii/S0895435618310813.

Cox, D.R. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 187–220. Available from: http://link.springer.com/10.1007/978-1-4612-4380-9_37.

D'Agostino, R.B., Lee, M.-L., Belanger, A., Cupples, L., Anderson, K. & Kannel, W.B. (1990) Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham heart study. *Statistics in Medicine*, 9, 1501–1515.

Ferrer, L., Putter, H. & Proust-Lima, C. (2018) Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment. *Statistical Methods in Medical Research*, 28, 1–18.

Gerds, T.A., Kattan, M.W., Schumacher, M. & Yu, C. (2013) Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*, 32, 2173–2184.

Goldberg, Y. & Kosorok, M.R. (2017) Support vector regression for right censored data. *Electronic Journal of Statistics*, 11, 532–569.Available from: http://arxiv.org/abs/1202.5130.

Gong, X., Hu, M. & Zhao, L. (2018) Big data toolsets to pharmacometrics: application of machine learning for time-to-event analysis. *Clinical and Translational Science*, 11, 305–311.

Graf, E., Schmoor, C., Sauerbrei, W. & Schumacher, M. (1999) Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18, 2529–

Guyon, I. & Elisseeff, A. (2003) An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.

Hickey, G., Philipson, P., Jorgensen, A. & Kolamunnage-Dona, R. (2016) Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Medical Research Methodology*, 16, 117. Available from: http://www.jstor.org/stable/2529876?origin=crossref.

Hogan, J.W. & Laird, N.M. (1997) Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine*, 16, 259–272.

Horsley, A., Cunningham, S. & Innes, J.A. (Eds.) (2015) *Cystic fibrosis*. Oxford: Oxford University Press, Oxford Respiratory Medicine Library, 2nd edn.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A.M. & Van Der Laan, M.J. (2006) Survival ensembles. *Biostatistics*, 7, 355–373.

van Houwelingen, H.C. (2007) Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34, 7085.

van Houwelingen, H.C. & Putter, H. (2012) *Dynamic prediction in clinical survival analysis*. Boca Raton: CRC Press.

Ishwaran, H., Kogalur, U.B., Blackstone, E.H. & Lauer, M.S. (2008) Random survival forests. *Annals of Applied Statistics*, 2, 841–860.

Ishwaran, H., Gerds, T.A., Kogalur, U.B., Moore, R.D., Gange, S.J. & Lau, B.M. (2014) Random survival forests for competing risks. *Biostatistics*, 15, 757–773.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013) *An introduction to statistical learning with applications in R*. New York: Springer.

Jewell, N.P. & Nielsen, J.P. (1993) A framework for consistent prediction rules based on markers. *Biometrika*, 80, 153–164.

Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T. & Kluger, Y. (2018) DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18, 1–12.

Keogh, R. (2018) *GitHub page: ruthkeogh*. Available from: http://github.com/ruthkeogh/landmark_CF.

Keogh, R., Seaman, S., Barrett, J., Taylor-Robinson, D. & Szczesniak, R. (2018) Dynamic prediction of survival in cystic fibrosis: a landmarking analysis using UK patient registry data. *Epidemiology*, 30, 29–37.

van der Laan, M.J. & Dudoit, S. (2003) Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. *U.C. Berkeley Division of Biostatistics Working Paper Series*. Available from: http://www.bepress.com/ucbbiostat/paper130.

van der Laan, M.J. & Rose, S. (2011) *Targeted learning: causal inference for observational and experimental data*. New York: Springer.

van der Laan, M.J., Polley, E.C. & Hubbard, A.E. (2007) Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6, 25.

Lafourcade, A., His, M., Baglietto, L., Boutron-Ruault, M.C., Dossus, L. & Rondeau, V. (2018) Factors associated with breast cancer recurrences or mortality and dynamic prediction of death using history of cancer recurrences: the French E3N cohort. *BMC Cancer*, 18, 171.

Lawson, C. & Hanson, R. (1995) Solving least squares problems. Society for Industrial and Applied Mathematics. Available from: http://epubs.siam.org/doi/abs/10.1137/1.9781611971217.

LeDell, E. (2016) Scalable super learning. In: Buehlmann, P., Drineas, P., Kane, M. & van der Laan, M.J. (Eds.), *Handbook of big data* (pp. 339–358). Boca Raton: CRC Press.

Lee, C., Yoon, J. & van der Schaar, M. (2019) Dynamic-deephit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67, 1–12. Available from: http://ieeexplore.ieee.org/document/8681104/.

Liestol, K., Andersen, P.K. & Andersen, U. (1994) Survival analysis and neural nets. *Statistics in Medicine*, 13, 1189–1200.

Lim, M., Wallis, C., Price, J., Carr, S., Chavasse, R., Shankar, A., Seddon, P. & Balfour-Lynn, I. (2014) Diagnosis of cystic fibrosis in London and South East England before and after the introduction of newborn screening. *Archives of Disease in Childhood*, 99, 197–202.

Lin, H., McCulloch, C.E. & Mayne, S.T. (2002) Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Statistics in Medicine*, 21, 2369–2382.

Liou, T.G., Adler, F.R., Fitzsimmons, S.C., Cahill, B.C., Hibbs, J.R. & Marshall, B.C. (2001) Predictive 5-year survivorship model of cystic fibrosis. *American Journal of Epidemiology*, 153, 345–352.

Malley, J.D., Kruppa, J., Dasgupta, A., Malley, K.G. & Ziegler, A. (2012) Probability machines: consistent probability estimation using nonparametric learning machines. *Methods of Information in Medicine*, 51, 74–81.

Maziarz, M., Heagerty, P., Cai, T. & Zheng, Y. (2017) On longitudinal prediction with time-to-event outcome: comparison of modeling options. *Biometrics*, 73, 83–93.

Mogensen, U.B., Ishwaran, H. & Gerds, T.A. (2012) Evaluating random forests for survival analysis using prediction error curves. *Journal of Statistical Software*, 50, 1.

Mullen, K.M. & van Stokkum, I.H.M. (2012) nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS). Available from: http://cran.r-project.org/package=nnls.

Paige, E., Barrett, J., Stevens, D., Keogh, R.H., Sweeting, M.J., Nazareth, I., Petersen, I. & Wood, A.M. (2018) Landmark models for optimizing the use of repeated measurements of risk factors in electronic health records to predict future disease risk. *American Journal of Epidemiology*, 187, 1530–1538.

Parner, E.T. & Andersen, P.K. (2010) Regression analysis of censored data using pseudoobservations. *The Stata Journal*, 10, 408–422.

Pencina, M.J., D'Agostino, R.B. & Song, L. (2012) Quantifying discrimination of Framingham risk functions with different survival C statistics. *Statistics in Medicine*, 31, 1543–1553.

Philipson, P., Sousa, I., Diggle, P., Williamson, P., Kolamunnage-Dona, R., Henderson, R. & Hickey, G. (2018) joineR: Joint modelling of repeated measurements and time-to-event data. Available from: http://cran.r-project.org/package=nlme.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team (2018) nlme: linear and nonlinear mixed effects models. Available from: http://cran.r-project.org/package=nlme.

Polley, E.C. & van der Laan, M.J. (2011) Super learning for right-censored data. In: *Targeted learning: causal inference for observational and experimental data* (pp. 249–258), chap. 16. New York: Springer.

Polley, E.C., Rose, S. & van der Laan, M.J. (2011) Super learning. In: *Targeted learning: causal inference for observational and experimental data*, chap. 3. New York: Springer.

Polley, E.C., LeDell, E., Kennedy, C. & van der Laan, M. (2018) SuperLearner: super learner prediction. Available from: http://projecteuclid.org/euclid.ssu/1315833185.

Proust-Lima, C. & Blanche, P. (2016) Dynamic predictions. In: Balakrishnan, N., Colton, T., Everitt, B., Piegorsch, W., Ruggeri, F. & Teugels, J. (Eds.) *Wiley StatsRef: statistics reference online*. Wiley Online Library. https://doi.org/10.1002/9781118445112.stat07876.

Proust-Lima, C., Séne, M., Taylor, J.M.G. & Jacqmin-Gadda, H. (2014) Joint latent class models for longitudinal and time-to-event data: a review. *Statistical Methods in Medical Research*, 23, 74–90.

R Core Team (2017) *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available from: http://www.r-project.org/.

Ripley, R.M., Harris, A.L. and Tarassenko, L. (2004) Non-linear survival analysis using neural networks. *Statistics in Medicine*, 23, 825–842.

Rizopoulos, D. (2010) JM: an R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35, 1.

Rizopoulos, D. (2011) Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67, 819–829.

Rizopoulos, D. (2012) *Joint models for longitudinal and time-to-event data, with applications in R*. Boca Raton: Chapman & Hall/CRC Biostatistics Series.

Rizopoulos, D. (2016) The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *Journal of Statistical Software*, 72, 42. Available from: http://arxiv.org/abs/1404.7625.

Rizopoulos, D., Molenberghs, G. & Lesaffre, E. (2017) Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal*, 59, 1261–1276.

Rotnitzky, A. & Robins, J.M. (2005) Inverse probability weighted estimation in survival analysis. In: Armitage, P. & Coulton, T. (Eds.) *Encyclopedia of biostatistics*. New York: Wiley, 2nd edn.

Segal, M.R. (1988) Regression trees for censored data. *Biometrics*, 44, 35–47. Available from: http://www.jstor.org/stable/2531894.

Segal, M.R. (1997) Tree-structured survival analysis. *Epidemiology*, 8, 344–346.

Singer, J.D. & Willett, J.B. (1993) It's about time: using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, 18, 155–195.

Steele, A.J., Denaxas, S.C., Shah, A.D., Hemingway, H. & Luscombe, N.M. (2018) Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS ONE*, 13, e0202344.

Steyerberg, E.W. & Vergouwe, Y. (2014) Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal*, 35, 1925–1931.

Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T.A., Gonen, M., Obuchowski, N., Pencina, M.J. & Kattan, M.W. (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, 21, 128–138.

Suresh, K., Taylor, J.M., Spratt, D.E., Daignault, S. & Tsodikov, A. (2017) Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model. *Biometrical Journal*, 59, 1277–1300.

Szczesniak, R., Heltshe, S.L., Stanojevic, S. & Mayer-Hamblett, N. (2017) Use of FEV1 in cystic fibrosis epidemiologic studies and clinical trials: a statistical perspective for the clinical researcher. *Journal of Cystic Fibrosis*, 16, 318–326.

Taylor-Robinson, D., Archangelidi, O., Carr, S.B., Cosgriff, R., Gunn, E., Keogh, R.H., Mac-Dougall, A., Newsome, S., Schluter, D. K., Stanojevic, S. & Bilton, D. (2017) Data resource profile: the UK cystic fibrosis registry. *International Journal of Epidemiology*, 47, 1–7.

Therneau, T.M. (2015) A package for survival analysis in S. Available from: http://cran.r-project.org/package=survival.

Therneau, T. & Grambsch, P. (2000) *Modeling survival data: extending the Cox model*. New York: Springer-Verlag.

Thompson, W.A. (1977) On the treatment of grouped observations in life studies. *Biometrics*, 33, 463–470.

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Tibshirani, R.J. (1997) The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16, 385–395.

Tsiatis, A.A. & Davidian, M. (2004) Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14, 809–834.

Tutz, G. & Schmid, M. (2016) *Modeling discrete time-to-event data*. Switzerland: Springer International Publishing.

UK Cystic Fibrosis Registry. (2018) UK Cystic Fibrosis Registry Annual Data Report 2017. Tech. rep., Cystic Fibrosis Trust, London.

Van Belle, V., Pelckmans, K., Van Huffel, S. & Suykens, J.A. (2011) Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, 53, 107–118. Available from: http://dx.doi.org/10.1016/j.artmed.2011.06.006.

Vock, D.M., Wolfson, J., Bandyopadhyay, S., Adomavicius, G., Johnson, P.E., Vazquez-Benitez, G. & O'Connor, P.J. (2016) Adapting machine learning techniques to censored time-to-event health record data: a general-purpose approach using inverse probability of censoring weighting. *Journal of Biomedical Informatics*, 61, 119–131.

Wolpert, D.H. (1992) Stacked generalization. *Neural Networks*, 5, 241–259.

Wong, A., Young, A.T., Liang, A.S., Gonzales, R., Douglas, V.C. & Hadley, D. (2018) Development and validation of an electronic health record–based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. *JAMA Network Open*, 1, e181018.

Wulfsohn, M.S. & Tsiatis, A.A. (1997) A joint model for survival and longitudinal data measured with error. *Biometrics*, 53, 330–339.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.
Supplementary Material

How to cite this article: Tanner KT, Sharples LD, Daniel RM, Keogh RH. Dynamic survival prediction combining landmarking with a machine learning ensemble: Methodology and empirical comparison. *J R Stat Soc Series A*. 2020;00:1–28. https://doi.org/10.1111/rssa.12611