Landmark proportional subdistribution hazards models for dynamic prediction of cumulative incidence functions

Qing Liu, 
*University of Pittsburgh, USA*

Gong Tang and Joseph P. Costantino 
*University of Pittsburgh, and NRG Oncology Statistics and Data Management Center, Pittsburgh, USA*

and Chung-Chou H. Chang 
*University of Pittsburgh, USA*

[Received November 2019. Final revision June 2020]

**Summary.** An individualized dynamic risk prediction model that incorporates all available information collected over the follow-up can be used to choose an optimal treatment strategy in realtime, although existing methods have not been designed to handle competing risks. In this study, we developed a landmark proportional subdistribution hazard (PSH) model and a comprehensive supermodel for dynamic risk prediction with competing risks. Simulations showed that our proposed models perform satisfactorily (assessed by the time-dependent relative difference, Brier score and area under the receiver operating characteristics curve) under PSH or non-PSH settings. The models were used to predict the probabilities of developing a distant metastasis among breast cancer patients where death was treated as a competing risk. Prediction can be estimated by using standard statistical packages.

**Keywords:** Competing risks; Landmark method; Risk prediction; Time-dependent variables; Time-varying effects

1. **Introduction**

An individualized risk prediction model that estimates the probability of a clinical event dynamically using time-dependent information can enable physicians to optimize personalized treatment strategies in realtime. For example, for an early stage breast cancer patient who received surgery, physicians can adjust the approach to the subsequent chemotherapies if they know the risk of a distant metastasis within the next few years given the patient’s loco-regional recurrence (LRR) measured over the follow-up. Note that LRR is a time-dependent intermediate event. Other types of time-dependent information include longitudinal measurements of biomarkers and time-varying covariate effects, which are also crucial to the success of a risk prediction model.
Two popular approaches in dynamic prediction that have received significant attention are joint modelling (Proust-Lima and Taylor, 2009; Rizopoulos, 2011; Mauguen et al., 2013; Rizopoulos et al., 2014) and landmarking (van Houwelingen, 2007; van Houwelingen and Putter, 2008, 2012; Parast et al., 2012). To incorporate time-dependent information, joint modelling captures the processes of longitudinal and time-to-event data simultaneously whereas the landmark method applies time-to-event modelling on patients who are still at risk at the updated prediction baseline, i.e. the landmark time, using the information that has been collected up to the landmark time. Although the joint models provide additional information from the joint distributions, the implementation is complex and the correlation structure is difficult to specify correctly and thus often results in inaccurate prediction. One could use multistate modelling to incorporate intermediate clinical events but would have to work with a number of models, up to the number of possible transitions between states, and related issues. One issue is that the computation will be complex when the number of states is large; another issue is overfitting when the sample size is not sufficiently large. In addition, the lack of goodness of fit of the transition models could greatly impact the predictive performance; van Houwelingen and Putter (2008) also pointed out that multistate modelling will cause computational complexity when the Markov assumption is not satisfied. By contrast, the landmark Cox model and supermodels that were proposed by van Houwelingen (2007) have shown significant advantages in simplicity, flexibility and accuracy in estimating the risk. Moreover, they become robust against model misspecifications by avoiding the full specification of the joint distribution. These models can easily incorporate various types of time-dependent information without using computationally intensive estimation procedures, and the estimation can be implemented in everyday clinical practice by using standard statistical software. Recently, landmark modelling has been used in many clinical applications (Zamboni et al., 2010; Fontein et al., 2015).

To account for data with competing events, Cortese and Andersen (2009) and Nicolaie et al. (2013a) applied the landmark method to the cause-specific hazards model. The disadvantage of this approach is that to estimate the risk from one cause we need to construct landmark models or supermodels on the basis of the cause-specific hazards for all causes. Nicolaie et al. (2013b) developed a landmark supermodel based on the pseudo-observations that are created at each landmark time point; however, this method will be computationally intensive in the calculation of risk when many landmark time points are included. Cortese et al. (2013) applied the landmark method to the Fine–Gray subdistribution hazards model (Fine and Gray, 1999). Although their approach can be used to estimate the risks directly at a small set of fixed landmark time points, it has the following limitations:

(a) there was a lack of theoretical proof that the target risk estimate can be calculated from the model proposed;
(b) they did not assess robustness of their proposed model’s prediction against violations of the proportional subdistribution hazard (PSH) assumption;
(c) they did not construct the comprehensive landmark supermodel to estimate the risks simultaneously for a sequence of prediction time points.

The major challenge that is encountered in developing a landmark model or supermodel based on the Fine–Gray model was in constructing the landmark subset at each landmark time to account properly for competing events in the setting of subdistribution hazards (Nicolaie et al., 2013a).

In this study, we overcame the aforementioned challenges, extended the landmark method to the Fine–Gray model and proposed a landmark PSH model and landmark PSH supermodel for dynamic prediction in the presence of competing risks. Our proposed models are more
straightforward in making the prediction of conditional cumulative incidence functions because we developed a direct link between the cause-specific covariate effects and the dynamic risk of the specific cause. The proposed landmark PSH model and supermodel can also provide accurate prediction even when the PSH assumption fails to hold. The landmark PSH supermodel uses smoothing techniques to aggregate a set of simple landmark PSH models. Therefore, users can make dynamic predictions for a set of landmark points by fitting only one model. This model also provides a simple and explicit estimation form to incorporate time-dependent covariates and/or time-varying covariate effects. These two proposed models can be effortlessly implemented with standard statistical software without computational burden.

This paper is organized as follows. In Section 2, we first introduce notation and dynamic predictive quantities for data with competing risks, and then review the Fine–Gray PSH model and present our proposed landmark PSH model and supermodel. In Section 3, we assess the performance in prediction for our proposed models by using simulations. In Section 4, we apply the proposed models to predict the dynamic cumulative incidences of distant metastasis based on the LRR status and a set of prognostic factors of breast cancer patients in a multicentre clinical trial. Discussion is provided in Section 5.

2. Dynamic prediction models

2.1. Notation and conditional cumulative incidence function

Let $T$ and $C$ be the failure and censoring time respectively, $\varepsilon \in \{1, \ldots, k\}$ be the cause of failure and $Z(\cdot)$ be a $p$-dimensional vector of covariates, which could be time-fixed covariates measured at the baseline or be time-dependent covariates. Here, for each subject $i$ ($i = 1, \ldots, n$), we assume that $C_i$ is independent of $T_i$ and $Z_i(\cdot)$ and refer to it as random censoring. For right-censored data, we observe an independently and identically distributed quadruplet of $\{X_i = \min(T_i, C_i), \Delta_i = I(T_i \leq C_i), \Delta_i \varepsilon_i, Z_i(\cdot)\}$. Suppose that cause 1 is the primary event of interest; the subdistribution function or cumulative incidence function (CIF) of cause 1 is defined as $F_{1}(t) = \Pr(T \leq t, \varepsilon = 1)$.

In dynamic prediction, the aim is to predict the conditional CIF, i.e. the cumulative incidence of occurrence of the primary event within the next $w$ time units (also named a prediction window with a fixed width of $w$) for a subject who has not failed from any cause at a landmark time (denoted by $s$) by incorporating the time-dependent information that is available up to the landmark time $s$. The width of prediction window $w$ depends on the clinical relevance that defines the time interval within which the risk of having the primary event is of interest. Then, the conditional CIF is defined as

$$F_{1,LM}\{s + w|Z(s), s\} = \Pr\{T \leq s + w, \varepsilon = 1|T > s, Z(s)\},$$

where $Z(s)$ are the covariates $Z(\cdot)$ whose values are measured or available at $s$. Using the definition of conditional probabilities, we can rewrite the conditional CIF as

$$F_{1,LM}\{s + w|Z(s), s\} = \frac{F_{1}\{s + w|Z(s)\} - F_{1}\{s|Z(s)\}}{S\{s|Z(s)\}} = \frac{F_{1}\{s + w|Z(s)\} - F_{1}\{s|Z(s)\}}{1 - \sum_{j=1}^{k} F_{j}\{s|Z(s)\}},$$

where $F_{j}\{s|Z(s)\}$ is the CIF for cause $j$ and $S\{s|Z(s)\}$ is the overall survival at time $s$.

One can estimate the conditional CIF $F_{1,LM}\{s + w|Z(s), s\}$ by using standard competing risks methods (e.g. a cause-specific hazards model and the Fine–Gray model) through estimating all cause-specific CIFs and overall survival, although it is often computationally challenging. We can estimate the conditional CIF directly with much simpler computation if there is a one-to-one relationship between the cause-specific covariate effects and the conditional CIF. In this study,
we extend the landmark method that was developed by van Houwelingen (2007) to the competing risks setting and propose a landmark PSH model that can be used to estimate the conditional CIF directly. Furthermore, our proposed model is also robust against misspecification of the PSHs.

2.2. Fine–Gray model
Fine and Gray (1999) proposed a regression model with respect to the subdistribution hazard

$$\lambda_1(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr\{t \leq T \leq t + \Delta t, \varepsilon = 1|T \geq t \cup (T \leq t \cap \varepsilon \neq 1)\}.$$  

With time-dependent covariates $Z(t)$, the Fine–Gray model takes the form

$$\lambda_1\{t|Z(t)\} = \lambda_{10}(t) \exp\{Z(t)^T \beta\},$$

and

$$F_1\{t|Z(t)\} = 1 - \exp\left\{- \int_0^t \lambda_{10}(u) \exp\{Z(u)^T \beta\} \, du\right\}.$$  

The regression coefficients $\beta$ are estimated through a partial likelihood approach with modified risk sets which are defined as $R(T_i) = \{j : (T_j \geq T_i) \cup (T_j \leq T_i \cap \varepsilon_j \neq 1)\}$ for the $i$th individual. $R(T_i)$ includes all subjects who have not failed from the event of interest by time $T_i$.

2.3. Landmark proportional subdistribution hazards model
To link the cause-specific covariate effect and cause-specific conditional CIF $F_{1,LM}\{s + w|Z(s), s\}$ directly at a given landmark time $s$, we define a conditional subdistribution hazard

$$\lambda_1\{t|Z(s), s\} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr\{t \leq T \leq t + \Delta t, \varepsilon = 1|T \geq t \cup (s \leq T \leq t \cap \varepsilon \neq 1), Z(s)\}$$  

for all $t \geq s$, which is the subdistribution hazard $\lambda_1\{t|Z(s)\}$ conditional on assuming that no event occurred from any cause before $s$. It can be shown that equation (1) is the hazard function of the conditional CIF:

$$F_{1,LM}\{s + w|Z(s), s\} = \Pr\{s < T \leq s + w, \varepsilon = 1|T > s, Z(s)\} = 1 - \Pr\{(T > s + w) \cup (s < T \leq s + w \cap \varepsilon \neq 1)|T > s, Z(s)\} = 1 - \exp\left[-\int_s^{s+w} \lambda_1\{t|Z(s), s\} \, dt\right],$$

where $\Lambda_1\{\cdot|Z(s), s\}$ is a cumulative conditional subdistribution hazard at landmark time $s$. Therefore, we propose a landmark PSH model with the form

$$\lambda_1\{t|Z(s), s\} = \lambda_{10}(t|s) \exp\{Z(s)^T \beta_{LM}\}, \quad (2)$$

for $s \leq t \leq s + w$.

To estimate the regression parameters $\beta_{LM}$ and the baseline conditional subdistribution hazard $\lambda_{10}(t|s)$, we can simply apply the Fine–Gray model to subjects who have not yet failed from any cause at $s$ and ignoring the events after $s + w$ by adding administrative censoring at $s + w$. As discussed in Liu et al. (2016), under a non-PSH assumption, the Fine–Gray model with administrative censoring at the horizon time provides an accurate prediction of the CIF at this
Interval time (but not other time points) without constructing complex procedures to estimate time-varying effects; and the partial likelihood estimator $\hat{\beta}_{LM}$ of $\beta_{LM}$ in model (2) is consistent for a weighted average of possibly time-varying effects over the interval $[s, s + w]$. Thus, the proposed landmark PSH model with truncation at the landmark time $s$ and administrative censoring at the horizon time $s + w$ is robust to violations of the PSH assumption. The conditional CIF can be easily estimated as

$$\hat{F}_{1,LM}\{s + w|Z(s), s\} = 1 - \exp[-\exp(Z(s)^T\hat{\beta}_{LM})\{\hat{\Lambda}_{10}(s + w) - \hat{\Lambda}_{10}(s-))],$$

in which $\hat{\beta}_{LM}$ and $\hat{\Lambda}_{10}(\cdot)$ are estimates calculated from the Fine–Gray model. Therefore, the landmark PSH model provides a simple and convenient way to predict directly the conditional CIF in dynamic prediction.

### 2.4. Landmark proportional subdistribution hazards supermodel

For dynamic prediction, it is required to estimate conditional CIFs at several landmark time points because the intermediate events may occur at any time point and covariate values may also be updated during the follow-up. Therefore, it is of interest to predict the risk probabilities $F_{1,LM}\{s + w|Z(s), s\}$ dynamically with fixed width $w$ for varying landmark times $s$ within an interval $[s_0, s_L]$. As discussed in van Houwelingen and Putter (2012), this provides a dynamic prediction framework with a sliding prediction window over the specified interval. Theoretically we can fit $L + 1$ landmark PSH models (2) to obtain $L + 1$ conditional CIFs. When the number of landmark points increases dramatically, it is less practical and is computationally inefficient to fit many models separately.

Instead, we adopt the smoothing strategy of van Houwelingen (2007) and propose a landmark PSH supermodel with the form

$$\lambda_1\{t|Z(s), s\} = \lambda_{10}(t) \exp(Z(s)^T\beta_{LM}(s) + \gamma(s))$$

for $s_0 \leq t \leq s_L + w$, which can estimate conditional CIFs at a set of landmark time points from $[s_0, s_L]$ in one step. As shown in Liu et al. (2016), $\hat{\beta}_{LM}$ of $\beta_{LM}$ in model (2) is consistent for a weighted average of $\beta(t)$ over the interval $[s, s + w]$, so we can expect the effect of $s$ on $\beta_{LM}(s)$ to be a continuous function and model $\beta_{LM}(s)$ as continuous functions of $s$. For $\lambda_{10}(t|s)$, as the Breslow estimator showed that the dependence on $s$ is through $\beta_{LM}(s)$, we can also expect that $\lambda_{10}(t|s)$ varies continuously with $s$ and rewrite $\lambda_{10}(t|s)$ as $\lambda_{10}(t|s) = \lambda_{10}(t) \exp\{\gamma(s)\}$. Hence, the sequence of landmark PSH models at a set of landmark time points is combined into a supermodel via $\beta_{LM}(s)$ and $\gamma(s)$ which are both continuous functions of $s$. For simplicity, we can fit linear models for $\beta_{LM}(s)$ and $\gamma(s)$: $\beta_{LM}(s) = \beta_{LM}(s|\theta) = \sum_{j=1}^{m\beta} \theta_j f_j(s)$ and $\gamma(s) = \gamma(s|\eta) = \sum_{j=1}^{m\gamma} \eta_j g_j(s)$, where $f(s)$ and $g(s)$ are two sets of parametric functions of $s$, $\theta$ and $\eta$ are vectors of parameters, and $m_{\beta}$ and $m_{\gamma}$ are the number of parameters in $\theta$ and $\eta$. Note that $g(s)$ does not need to contain the constant term and have the restriction of $g_j(s_0) = 0$ for all $j$ ($j = 1, \ldots, m_{\gamma}$) because of identifiability of the baseline subdistribution hazard, as discussed in van Houwelingen (2007).

To fit the landmark PSH supermodel, which is similar to the landmark supermodel for survival data without competing risks (van Houwelingen, 2007), we need to create an augmented data set which is constructed as follows: first select a set of landmark points $s$ from the interval $[s_0, s_L]$; for each $s$, create a landmark subset by selecting the subjects who have not yet failed from any cause at $s$ and adding administrative censoring at the prediction horizon $s + w$; then stack all the individual landmarking subsets into a superprediction data set. Then, model (3) can be fitted.
by applying a Fine–Gray model including landmark–covariate interactions \( Z(s) \ast f_j(s) \) to the stacked data set.

The parameters \((\theta, \eta)\) can be consistently estimated by maximizing a Breslow pseudo partial log-likelihood for tied events because, in the stacked data set, one subject with event time \( T_i \) has \( n_i = \# \{ s : s < T_i \leq s + w, s \in [s_0, s_L] \} \) repeated observations, and \( \Lambda_{10}(t) \) can also be estimated by a Breslow estimator for the cumulative subdistribution hazard (see the web appendix A for details). To obtain the standard errors for the estimated parameters, a robust sandwich estimator is required to adjust for the correlation between the risk sets which exists because the same subject is repeatedly used when we estimate the parameters on the basis of the stacked data set. Thus, the target dynamic prediction probabilities \( F_{1,LM}(s + w|Z(s), s) \) have a simple and explicit estimation form, which is given by

\[
\hat{F}_{1,LM}(s + w|Z(s), s) = 1 - \exp\left[ - \exp\{ Z(s)^T \hat{\beta}_{LM}(s) + \hat{\gamma}(s) \} \{ \hat{\Lambda}_{10}(s + w) - \hat{\Lambda}_{10}(s) \} \right]
\]

for all \( s \in [s_0, s_L] \). Therefore, the landmark PSH supermodel can provide a prediction of \( F_{1,LM}(s + w|Z(s), s) \) in any period of length \( w \) starting anywhere in the interval \([s_0, s_L] \).

For implementation, before stacking all landmarking subsets into a super data set, we need to transform each subset into counting process style and include time-varying weights by using the inverse probability of censoring weighting approach for the subjects who experienced competing events to adjust for the random right censoring (Geskus, 2011). In practice, fitting the landmark PSH supermodel in the stacked data set requires software that allows for delayed entry or left truncation at \( s \). The \texttt{R} function \texttt{coxph()} (R Core Team, 2019) can be used to fit model (3) and it can also provide a robust sandwich covariance matrix for \((\hat{\theta}, \hat{\eta})\) which can be used in the significance test of the estimated regression coefficients. As discussed, the landmark effect on \( \lambda_{10}(t|s) \) is through \( \beta_{LM}(s) \), so there is a correlation between \( \hat{\theta} \) and \( \hat{\eta} \). It is recommended to centre the covariates before fitting the model (van Houwelingen and Putter, 2012).

For simplicity, we assumed that \( C \) is independent of \( T \) and \( Z \), but it can be generalized to conditional independence between \( C \) and \( T \) given \( Z \), as discussed in the PSH model (Fine and Gray, 1999). The dependence between \( C \) and \( Z \) can be handled by modelling \( C \) as a function of \( Z \) in inverse probability of censoring weighting. Following the same strategy, the properties of the landmark PSH model and landmark PSH supermodel will be retained under conditionally independent censoring.

### 2.5. Measure of predictive performance

To evaluate the dynamic predictive performance of the procedures proposed, we adapted the time-dependent Brier score, relative difference (RD) between the observed number of main events and the expected number of events, and the area under the receiver operating characteristics curve, AUC, in terms of predictive accuracy, calibration and discrimination respectively.

The time-dependent Brier score is an estimate of the mean-squared prediction error, i.e. the average squared difference between the predicted event probabilities at \( s + w \) and the observed event status for subjects who are event free at landmark time \( s \) (Schoop et al., 2011; Cortese et al., 2013). For data with competing risks, let \( \pi_i(s, w) = F_{1,LM}(s + w|Z_i(s), s) \), the Brier score is defined as

\[
BS_{LM}(s, w) = \frac{1}{n_s} \sum_{i \in R_s} \left\{ D_i(s, w) - \hat{\pi}_i(s, w) \right\}^2,
\]

where \( R_s = \{ i : T_i > s \} \) is the set of subjects who are event free at landmark time \( s \), \( n_s \) is the number of subjects in \( R_s \), \( D_i(s, w) = I(s < T_i \leq s + w, \varepsilon_i = 1) \) and the estimated conditional CIF
\[ \hat{\pi}_i(s, w) = \hat{F}_{1,LM}(s + w | Z_i(s), s) \text{ is obtained from the predictive model. When random right censoring exists, we replaced } BS_{LM}(s, w) \text{ by a pseudovalue-based consistent estimator given by} \]

\[ \hat{BS}_{LM}(s, w) = \frac{1}{\tilde{n}_s} \sum_{i \in \tilde{R}_s} [\hat{Q}_{1,LM}^{(i)}(s + w | s) \{1 - 2\hat{\pi}_i(s, w)\} + \hat{\pi}_i(s, w)^2], \]

which was proposed by Cortese et al. (2013), where \( \tilde{R}_s = \{i : X_i > s\} \) and \( \tilde{n}_s \) is the number of subjects in \( \tilde{R}_s \).

\[ \hat{Q}_{1,LM}^{(i)}(s + w | s) = \hat{n}_s \hat{F}_{1,LM}(s + w | s) - (\hat{n}_s - 1) \hat{F}_{1,LM}^{(i)}(s + w | s) \]

is a jackknife pseudovalue for the \( i \)th subject who is still at risk at time \( s \), where \( \hat{F}_{1,LM}(s + w | s) \) is the non-parametric estimate of the marginal cumulative incidence \( \Pr(T \leq s + w | T > s) \), and \( \hat{F}_{1,LM}^{(i)}(s + w | s) \) is the same estimate but is based on the data where the \( i \)th subject has been removed.

To assess calibration, we adapted the time-dependent RD which is defined as the RD between the observed number of main events from the landmark time point to the prediction horizon time and the expected number of events estimated from the predictive model:

\[ RD_{LM}(s, w) = \frac{\sum_{i \in \tilde{R}_s} D_i(s, w)}{\sum_{i \in \tilde{R}_s} \hat{\pi}_i(s, w)} - 1. \]

As discussed in Pfeiffer and Gail (2017), to address the right censoring, the jackknife pseudovalue \( \hat{Q}_{1,LM}^{(i)}(s + w | s) \) was applied, and

\[ \hat{RD}_{LM}(s, w) = \frac{\sum_{i \in \tilde{R}_s} \hat{Q}_{1,LM}^{(i)}(s + w | s)}{\sum_{i \in \tilde{R}_s} \hat{\pi}_i(s, w)} - 1. \]

To measure the discriminatory ability of the landmark PSH models, we followed the work of Blanche et al. (2015) and Huang et al. (2016) to calculate the time-dependent AUC, which is given by

\[ AUCLM_{s, w} = \Pr\{\hat{\pi}_i(s, w) > \hat{\pi}_j(s, w) | D_i(s, w) = 1, D_j(s, w) = 0, T_i > s, T_j > s\} \]

for a pair of subjects \( \{i, j\} \) neither of whom has experienced any event at the landmark time \( s \). In the presence of right censoring, an IPCW estimator proposed by Blanche et al. (2015) was utilized.

3. Simulation studies

We evaluated the performance of the proposed dynamic prediction models by using simulated data under two settings of the non-PSH. Because the non-parametric method is model free, we used it as the reference to compare with other model-based methods in scenarios where only categorical variables are included. For simplicity, only two failure types were considered: type 1 failure is the primary event of interest; type 2 failure indicates competing events. We also simulated competing risks data under the non-PSH setting with a time-dependent covariate.

In the first non-PSH setting, we generated the type 1 failure times from a two-parameter Weibull mixture distribution with the subdistribution

\[ F_{1,1}(t; Z_i) = p(1 - \exp[-\{\lambda_1 \exp(Z_i/\beta_1)t\}^{\alpha_1}] \]), \]
where the rate parameter depends on a single binary covariate $Z_i$ from a Bernoulli(0.5) distribution, and the coefficient of $Z_i$ is constant. In the second non-PSH setting, we let the coefficient of $Z_i$ be a function of time; and the subdistribution of the primary event became

$$F_{1,2}(t; Z_i) = 1 - (1 - p\{1 - \exp\{-\langle\lambda_2\rangle^{\alpha_2}\}\})\exp\{Z_i\beta_{21} + Z_i\beta_{22}\ln(t+1)\}.$$

Parameter $p$ is used to control the proportion of main events. In both settings, we let $p = 0.3$, which produced about 30% main events at $Z_i = 0$ when there was no censoring. We generated the type 2 failure times from an exponential distribution $\Pr(T_i \leq t|\epsilon_i = 2, Z_i) = 1 - \exp\{-\exp(Z_i\beta_c t)\}$ by taking $\Pr(\epsilon_i = 2|Z_i) = 1 - \Pr(\epsilon_i = 1|Z_i)$, where $\beta_c = 0.5$. A sample size of $n = 1000$ was chosen.
and the data were simulated repeatedly for $N = 1000$ times. The censoring times were generated independently from a uniform distribution, which resulted in about 20% censoring.

The performances of the proposed landmark PSH model and the landmark PSH supermodel in dynamic prediction of the conditional CIFs using a fixed width of $w$ at a set of landmark points were compared with the prediction performance of the non-parametric method and the standard PSH model. The width of the prediction window $w$ was chosen on the basis of the distributions of the primary event of interest (web Fig. 1).

For fitting the landmark PSH supermodel, we set up a fine grid of landmark points with an equidistant step of 0.1 from 0 to 5 for the first non-PSH setting and from 0 to 4 for the second non-PSH setting. In both settings, we took ordinary polynomials for the basis functions as $f(s) = \{1, s, s^2\}$ and $g(s) = \{s, s^2\}$. Note that the landmark PSH supermodel in the absence of censoring cannot be fitted by using the `coxph()` function in R, because the competing risks data cannot be transformed into the counting process format if there is no censoring; and note that the model cannot be fitted by the `crr()` function in R either since the `crr()` function does not allow delayed entries. Therefore, in the absence of censoring, only the landmark PSH model is fitted to compare with other standard approaches.

Fig. 1 depicts the true and estimated conditional CIFs obtained from various approaches. We also provided the estimates of the conditional failure probability from the landmark Cox supermodel (van Houwelingen, 2007) by treating the competing risks as random censoring. We found that, in two different non-PSH scenarios, the performances of the landmark PSH model and the landmark PSH supermodel are as good as that of the non-parametric methods. But the landmark Cox supermodel overestimated the risk of failure for the main event if ignoring the competing events.

For each approach we evaluated the prediction errors in the dynamic conditional CIFs by estimating time-dependent Brier scores, and we used a threefold cross-validation to correct for possible overfitting. Web Table 1 presents averaged estimates of the cross-validated time-dependent Brier score and its empirical standard deviation. To quantify the improvement in predictive accuracy for the proposed landmark PSH model and the supermodel to the standard PSH model under non-proportional hazards, we utilized a relative increment (or reduction) of prediction errors by treating the non-parametric estimates as the reference. The relative increments of prediction errors are presented in Fig. 2. As expected, in both non-PSH settings the predictive accuracy of the landmark PSH model was almost the same as that obtained from the non-parametric method. Compared with the landmark PSH model or the non-parametric method, the landmark PSH supermodel has slightly lower accuracy, yet the differences in prediction errors are negligible.

To evaluate the predictive performance of the proposed landmark PSH model and the supermodel further for data with time-dependent covariates, we added $Z_i(t)$ to the second non-PSH setting by using the simulation strategy that was introduced in Huang et al. (2016) (see web appendix B for details). We chose a prediction window of $w = 0.4$ and sample size of $n = 5000$ to calculate the threefold cross-validated time-dependent RD, Brier score and AUC for the landmark PSH model and the supermodel, and compared them with the landmark Cox supermodel. To fit the supermodels, we selected landmark time points from 0 to 4 with a step of 0.1, and we used the same quadratic basis function $f(s)$ and $g(s)$ for $Z_i(t)$ because only the effect of $Z_i(t)$ depends on landmark $s$ in the Wald test. Table 1 provides averaged estimates of the RD, Brier score and AUC with the corresponding empirical standard errors from 1000 simulated data sets. The landmark PSH model and the supermodel showed good predictive performance in terms of calibration, predictive accuracy and discrimination.

In addition, we also assessed the predictive performance of the proposed models for data with smaller sample sizes under all the non-PSH settings. We evaluated the performance of sample
Fig. 2. Relative increment of prediction errors (and their standard deviation) at landmark points ($Z$ is the Bernoulli(0.5) variant, and $p = 0.3$, which produced about 30% main events at $Z_i = 0$ when there was no censoring) (▲, PSH versus non-parametric; ●, landmark PSH versus non-parametric; ■, landmark PSH supermodel versus non-parametric) (the prediction errors were cross-validated (three-fold) estimates for the time-dependent Brier scores, where $w = 3$ for (a), (b) the first non-PSH setting with $\alpha_1, \lambda_1, \beta_1 = 3.2, 0.18, -0.81$ and $w = 2$ for (c), (d) the second non-PSH setting with $\alpha_2, \lambda_2, \beta_{21}, \beta_{22} = 3.2, 0.12, 0.8, 0.3$): (a), (c) 0% censoring; (b), (d) 20% censoring.

sizes $n = 400$ and $n = 600$ under the first and second non-PSH settings, and $n = 1000$ under the non-PSH setting with time-dependent covariates. The results show that, in smaller sample sizes, the landmark PSH models have good predictive performance (web Table 2). When examining the time-dependent Brier score, AUC and RD, we found that the Brier score and AUC from data sets of smaller sample sizes are as good as that of larger sample sizes. For smaller sample sizes, the RD is slightly higher under the non-PSH setting with time-dependent covariates for the landmark PSH model, because threefold cross-validation further reduced the sample sizes in fitting the model (web Table 3). We also observed that the landmark supermodel is more sensitive to the sample size under all non-PSH settings; for example, the supermodel lost accuracy particularly at $n = 400$ after cross-validation was used (web Tables 2 and 3).
Table 1. Cross-validated (threefold) estimates of the time-dependent RD, Brier score and AUC with the corresponding standard errors in parentheses for the landmark PSH model, landmark PSH supermodel and landmark Cox supermodel in the second non-PSH setting with time-fixed and time-dependent covariates

| Landmark | RD (× 100) | Brier score (× 100) | AUC (× 100) |
|----------|------------|---------------------|-------------|
| **Landmark PSH model** | | | |
| 2.4 | 4.940 (5.377) | 1.255 (0.219) | 66.840 (5.913) |
| 2.6 | 2.561 (2.660) | 2.329 (0.294) | 68.292 (4.006) |
| 2.8 | 1.354 (1.593) | 4.038 (0.363) | 70.915 (2.770) |
| 3.0 | 0.925 (0.976) | 6.468 (0.426) | 75.479 (1.909) |
| 3.2 | 0.539 (0.735) | 9.698 (0.482) | 75.376 (1.601) |
| **Landmark PSH supermodel** | | | |
| 2.4 | 3.001 (15.356) | 4.237 (11.421) | 67.285 (5.805) |
| 2.6 | 1.166 (14.050) | 5.511 (11.836) | 68.113 (4.545) |
| 2.8 | −0.918 (12.702) | 6.906 (11.135) | 70.396 (3.749) |
| 3.0 | −6.604 (12.686) | 9.799 (11.981) | 74.628 (3.966) |
| 3.2 | −4.102 (14.165) | 13.664 (12.298) | 74.143 (4.253) |
| **Landmark Cox supermodel** | | | |
| 2.4 | −29.313 (44.813) | 37.405 (36.391) | 60.853 (8.206) |
| 2.6 | −34.983 (45.379) | 41.573 (37.761) | 60.820 (8.270) |
| 2.8 | −38.224 (44.256) | 43.415 (38.241) | 61.910 (9.160) |
| 3.0 | −43.607 (41.745) | 47.426 (37.717) | 63.376 (11.370) |
| 3.2 | −40.574 (42.258) | 48.259 (34.891) | 62.858 (11.656) |

Furthermore, we computed the predictive accuracy of each approach under the PSH setting by removing ln(t + 1) from the second non-PSH setting. As the results show in web Table 4, the prediction errors are similar among the models proposed, the non-parametric method and the PSH model with various sample sizes. Note that, although the landmark PSH supermodel had slightly larger prediction errors compared with the other methods, the differences are negligible.

4. Case-study—a randomized clinical trial for breast cancer

We applied our proposed landmark PSH models in dynamic prediction of the risk of distant metastasis for breast cancer patients on the basis of their prognostic information measured during follow-up. We used the data from a multicentre phase III clinical trial for breast cancer patients with oestrogen receptor positive and historically nodes negative (Fisher et al., 1997). In this trial, 2363 patients were randomly assigned to receive one of the following three regimens: tamoxifen 10 mg daily for 5 years, tamoxifen 10 mg daily for 5 years plus metrotrexate M and fluorouracil F, and tamoxifen 10 mg daily for 5 years plus M, F and cyclophosphamide C, which are denoted TAM, TAM+MF and TAM+CMF respectively for simplicity. The median follow-up time was 11.2 years.

Among the 2272 clinically eligible patients who were followed, 241 developed distant metastasis, 127 died from other causes before distant metastasis could occur (for simplicity, we refer to it as death) and the remaining 1904 were censored because of withdrawal from the study, loss to follow-up or event free on or before the analysis cut-off date. In analyses, the censoring cases were treated as random non-informative censoring where the justification is described in web Appendix C. For early stage breast cancer patients after surgery, development of LRR is...
Table 2. Estimated regression parameters of the landmark PSH supermodel for distant metastasis and death before distant metastasis†

| Covariate                        | Results for distant metastasis | Results for death before distant metastasis |
|----------------------------------|--------------------------------|---------------------------------------------|
|                                  | \( \hat{\beta} \) | Robust se(\( \hat{\beta} \)) | Wald test p-value | \( \hat{\beta} \) | Robust se(\( \hat{\beta} \)) | Wald test p-value |
| Treatment                        | Constant | −1.091 | 0.331 | 0.001 | −0.312 | 0.238 | 0.190 |
| TAM + MF versus TAM              | \( s \) | 0.482 | 0.160 | 0.010 |
|                                 | \( s^2 \) | −0.050 | 0.017 |
| TAM + CMF versus TAM             | Constant | −0.470 | 0.170 | 0.006 | 0.160 | 0.219 | 0.463 |
| Age \( \geq 50 \) versus \(< 50 \) years | \( s \) | −0.540 | 0.257 | 0.035 |
|                                 | \( s^2 \) | 0.407 | 0.135 | 0.006 |
| Surgery type                     | Mastectomy versus lumpectomy plus radiation therapy | Constant | 0.424 | 0.144 | 0.003 | −0.115 | 0.189 | 0.545 |
| Clinical tumour size             | \( > 2 \) versus \( \leq 2 \) cm | Constant | 0.284 | 0.141 | 0.044 | 0.334 | 0.196 | 0.089 |
| Tumour grade                     | Moderate versus well | Constant | 0.221 | 0.179 | 0.218 | 0.104 | 0.221 | 0.638 |
| Poor versus well                 | Constant | 0.752 | 0.189 | < 0.001 | 0.124 | 0.258 | 0.633 |
| LRR status                       | Constant | 2.315 | 0.237 | < 0.001 | 3.575 | 1.230 | 0.004 |
|                                 | \( s \) | 0.492 | 0.582 | 0.038 |
|                                 | \( s^2 \) | 0.008 | 0.056 |
| Baseline parameters              | \( \eta_1 \) | \( s \) | −0.400 | 0.088 | < 0.001 | −0.001 | 0.014 | < 0.001 |
|                                 | \( \eta_2 \) | \( s^2 \) | 0.035 | 0.009 | 0.002 | 0.001 |

†\( s \) is landmark time; \( \eta_1 \) and \( \eta_2 \) are baseline parameters defined in Section 2.4.

an important prognostic clinical event affecting the risk of distant metastasis. In these data, 15.1% of the patients developed LRR before progressing to distant metastasis, but only 6.6% of patients experienced LRR before death. Our main interest in this application is to predict dynamically the risk of distant metastasis within the subsequent 3 years for a breast cancer patient, based on her LRR status measured during follow-up and other prognostic covariates measured at baseline, including the treatment (TAM, TAM + MF or TAM + CMF), surgery type (lumpectomy plus radiation therapy versus mastectomy), age at the study entry (under 50 versus 50 years old or older), clinical tumour size (2 cm or smaller versus bigger than 2 cm) and tumour grade (well, moderate and poor). We also compared the dynamic 3-year fixed width probabilities of distant metastasis and death based on a patient’s LRR history.

Web Fig. 2(a) shows the estimated CIFs for both distant metastasis and death in which the maximum event time is around 13 years. Web Fig. 2(b) depicts the estimated CIF of LRR. Only a few random censoring events occurred during the first 10 years of follow-up (the figure is not shown). Thus, we chose the range of landmark points from 0 to 10 years and prediction window with a fixed width of 3 years. To fit a landmark PSH supermodel to this data set, we took 51 equally spaced landmark points \( s (0 \leq s \leq 10) \) and set the basis functions for \( \beta_{LM}(s) \) and \( \gamma(s) \) as \( \beta_{LM}(s) = \theta_1 + \theta_2 s + \theta_3 s^2 \) and \( \gamma(s) = \eta_1 s + \eta_2 s^2 \). The frequencies of distant metastasis and death in each of the landmark sub-data-sets for \( s = 0, 1, \ldots, 10 \) years are shown in web Fig. 2(c).
Fig. 3. Predicted 3-year fixed width conditional cumulative incidences of distant metastasis (——) and death before distant metastasis (---) with the associated bootstrap 95% confidence intervals (----) for different landmark time points, for patients younger than 50 years old with poor tumour grade, tumour size larger than 2 cm and treated with lumpectomy in each of the treatment groups and with the first LRR at none, 3 years, 5 years and 7 years: (a) treatment TAM; (b) treatment TAM+MF; (c) treatment TAM+CMF

We began our analysis using the backward selection procedure to select those covariates of which the effects were dependent on the landmark points. We tested the landmark–covariate interactions for each covariate via a Wald test based on the robust covariance matrix of the estimated coefficients from the landmark PSH supermodel. The main effects of all covariates and the significant landmark–covariate interactions were included in the supermodel. We found that,
for distant metastasis, the effects of the TAM + MF treatment and age were significantly dependent on the landmark points, whereas, for death, only the effect of LRR status was dependent on the landmark points. In the model, all covariates were included but only significant landmark–covariate interactions were kept. The estimated coefficient and the corresponding robust standard error for a given prognostic factor are summarized in Table 2. Multivariate Wald tests for the baseline parameters $\eta_1$, $\eta_2$ were significant for both distant metastasis and death, indicating that the baseline subdistribution hazard also depended on the choice of landmark points.

Given different LRR status (including no LRR developed over the course of study, with the first LRR at 3, 5 and 7 years) and various treatment regimens (TAM, TAM + MF or TAM + CMF), Fig. 3 depicts the predicted dynamic 3-year fixed width cumulative incidences of distant metastasis and death with the associated bootstrap 95% confidence intervals for patients who were younger than 50 years old with poorly differentiated tumours, tumour size larger than 2 cm and treated by lumpectomy plus radiation therapy. If a patient did not develop LRR, the risk of having distant metastasis within the subsequent 3 years is very close to the risk of death in any treatment group. However, if the patient had an LRR, she would have a much higher risk of developing distant metastasis compared with the risk of death, especially for the TAM+MF

Table 3. Estimates of the time-dependent RD, Brier score and AUC (with standard errors from the bootstrap with $B = 500$ in parentheses) at a selected set of landmark time points for distant metastasis and death before distant metastasis

| Outcome                        | Landmark (years) | RD ($\times$ 100) | Brier score ($\times$ 100) | AUC ($\times$ 100) |
|--------------------------------|------------------|-------------------|---------------------------|--------------------|
| **Landmark PSH model**         |                  |                   |                           |                    |
| Distant metastasis             | 1                | −0.188 (9.783)    | 4.195 (0.389)             | 68.102 (2.831)     |
|                                | 3                | −0.082 (10.032)   | 3.851 (0.370)             | 69.453 (2.992)     |
|                                | 5                | 0.074 (12.992)    | 2.829 (0.359)             | 74.907 (3.295)     |
|                                | 7                | 0.514 (15.304)    | 2.043 (0.303)             | 69.164 (5.055)     |
| Death before distant metastasis| 1                | 0.217 (18.887)    | 1.196 (0.223)             | 65.175 (5.303)     |
|                                | 3                | −0.447 (16.296)   | 1.420 (0.224)             | 77.240 (3.898)     |
|                                | 5                | −0.173 (16.553)   | 1.907 (0.306)             | 68.779 (4.043)     |
|                                | 7                | 0.115 (14.388)    | 2.546 (0.355)             | 66.395 (4.187)     |
| **Landmark PSH supermodel**    |                  |                   |                           |                    |
| Distant metastasis             | 1                | 2.055 (10.042)    | 4.237 (0.395)             | 66.174 (2.943)     |
|                                | 3                | −1.391 (9.938)    | 3.854 (0.370)             | 68.260 (3.156)     |
|                                | 5                | −0.866 (12.911)   | 2.865 (0.364)             | 74.688 (3.266)     |
|                                | 7                | −0.229 (15.288)   | 2.070 (0.308)             | 67.821 (4.718)     |
| Death before distant metastasis| 1                | 0.754 (19.187)    | 1.206 (0.224)             | 62.136 (6.003)     |
|                                | 3                | −2.035 (16.038)   | 1.439 (0.228)             | 76.705 (4.398)     |
|                                | 5                | 0.202 (16.506)    | 1.913 (0.308)             | 64.165 (4.530)     |
|                                | 7                | 0.208 (14.454)    | 2.553 (0.357)             | 63.816 (4.282)     |
| **Landmark Cox supermodel**    |                  |                   |                           |                    |
| Distant metastasis             | 1                | 1.099 (9.949)     | 4.265 (0.397)             | 66.175 (2.953)     |
|                                | 3                | 24.565 (12.581)   | 3.881 (0.382)             | 68.373 (3.151)     |
|                                | 5                | 1.843 (14.550)    | 2.885 (0.371)             | 74.639 (3.264)     |
|                                | 7                | 20.795 (18.517)   | 2.097 (0.316)             | 67.917 (4.692)     |
| Death before distant metastasis| 1                | 3.829 (19.861)    | 1.241 (0.228)             | 62.291 (6.009)     |
|                                | 3                | 10.630 (18.211)   | 1.478 (0.235)             | 76.497 (4.391)     |
|                                | 5                | 6.203 (17.515)    | 1.947 (0.314)             | 64.295 (4.539)     |
|                                | 7                | 9.095 (15.754)    | 2.585 (0.363)             | 64.924 (4.217)     |
In addition to the landmark PSH supermodel, we also applied the landmark PSH model at each landmark time point. The estimated coefficients are reported in web Table 5. We compared the risk predictive performance of our proposed landmark PSH model and supermodel against the landmark Cox supermodel in terms of time-dependent RD, Brier score and AUC at a selected set of landmark time points with bootstrap standard errors with \( B = 500 \). Because of the limited number of events in these application data, cross-validation was not used. The fitted landmark Cox supermodel is summarized in web Table 6. As results show in Table 3, both the landmark PSH model and supermodel are well calibrated, and the landmark PSH model has a slightly superior performance as demonstrated in simulation studies. The landmark Cox supermodel tends to underestimate the risk of event with larger RDs. Because the two supermodels have a similar model form (Table 2 and web Table 6), the differences in AUCs are limited in this application example.

To demonstrate an additional value that dynamic prediction provides, we compared the performance of the landmark PSH supermodel that dynamically incorporated LRR status at each landmark time with the PSH model that simply used LRR status as a predictor to predict the CIFs at the horizon time. The results are summarized in web Table 7. Apparently, the landmark PSH supermodel outperformed the PSH model in terms of both calibration and predictive accuracy.

5. Discussion

In this study, we developed dynamic prediction models for data containing competing risks by extending the landmark approach to the Fine–Gray model. The resulting landmark PSH model and the supermodel can be used to predict the dynamic cumulative incidences directly for the occurrence of a specific event within a given prediction window of a fixed width through incorporating all available information up to the landmark time under the condition that the patient did not fail at the landmark time.

Our proposed models have several advantageous features over the currently available methods in predicting dynamic cumulative inferences. Our landmark PSH model and its supermodel can provide accurate predictions under both PSH and non-PSH settings, whereas the model that was developed by Cortese et al. (2013) might lead to biased estimates of conditional CIFs after incorporating the landmark method in the Fine–Gray model where the PSH assumption is potentially violated. Therefore, when applying our landmark PSH models in practice, there is no need to test the PSH assumption. In addition, unlike the Fine–Gray model which does not allow the use of internal time-dependent covariates in prediction of the CIF (Kalbfleisch and Prentice, 2002; Beyersmann and Schumacher, 2008), the landmark PSH models can incorporate both internal and external time-dependent information. The landmark method provides a simpler and explicit form of estimates by bypassing modelling the covariate change process and the time-to-event outcome process. Thus, it is much easier to incorporate intermediate events and/or time-dependent covariates compared with those multistate and joint models which are more complicated computationally and prone to overfitting. Although joint models give better prediction by incorporating the whole trajectory of a time-dependent covariate, the landmark method provides a sufficiently good approximation when the covariate is sparsely measured or when many time-dependent covariates are included. In contrast with the landmark supermodel based on cause-specific hazards (Nicolaie et al., 2013a), our landmark PSH supermodel predicts the conditional CIFs simply in one step and provides a direct interpretation of the landmark-
specific effects on the predictive probabilities. Compared with the landmark supermodel based on pseudo-observations (Nicolaie et al., 2013b), our supermodel also shows simplicity in computation whereas the generalized estimating equation based method using pseudo-observations would have convergence issues for large sample sizes especially when dealing with many landmark points of interest.

We evaluated the prediction performance of our proposed models through simulations. We determined how closely the estimated conditional CIFs would approximate the true probabilities for our models. We also evaluated the performances of our landmark PSH model and its supermodel at different sample sizes from the perspectives of calibration, predictive accuracy and discriminatory abilities by utilizing time-dependent RDs, Brier scores and AUCs. As the simulation results showed, our landmark PSH model performed well with small and larger data sets after incorporating time-fixed and time-dependent covariates even when the PSH assumption was violated, whereas the landmark PSH supermodel lost prediction accuracy with small sample sizes, particularly when there is a time-dependent covariate.

We also compared the performances between our landmark PSH model and its supermodel for breast cancer data. The landmark PSH model showed a slightly superior performance over the supermodel and more efficient in computation for dynamic prediction at four landmark time points, because it does not need to construct the super stacked data set as the supermodel does. When there are more landmark time points of interest (e.g. more frequent monitoring the event risk), the supermodel will demonstrate its advantage as the stacked data set needs to be developed only once and the model needs to be fitted only once. Therefore, we expect that the landmark PSH model will be more favourable than the supermodel in practice when the data set is small, and the amount of landmark time points is small.

In the landmark PSH supermodel, we assume that the effect of $s$ on baseline hazards $\lambda_{10}(t|s)$ is an additive effect. As discussed in van Houwelingen (2007), this assumption will hold if the follow-up is not too long or the effect of covariates is not too large. If we choose an optimal width $w$ for the prediction window and a suitable range $[s_0, s_L]$ for the landmark time points, the landmark PSH supermodel can provide a correct approximation for the conditional CIF at time $s + w$ for any $s \in [s_0, s_L]$. An alternative method is to fit a stratified landmark PSH supermodel by applying a Fine–Gray model with the landmark–covariate interactions $Z(s) \times f_j(s)$ to the stacked data set with stratification on $s$, which estimates the baseline conditional subdistribution hazard for each $s$.

In the presence of time-dependent covariates, if the covariate values vary too often and too much over time, the estimated covariate effects could be attenuated in the models proposed and subsequently lead to biased estimation of the conditional CIF. This is also shown in our simulation result for data with a time-dependent covariate. To adjust for this issue, which was also discussed by van Houwelingen and Putter (2012), we shall explore a set of suitable landmark time points and use an additional monotonically decreasing function to model the attenuation process of the covariate effects in future work. These studies to handle time-dependent covariates in dynamic prediction have attracted attention from other researchers. Wu et al. (2020) have extended our proposed landmark PSH model framework to analysing biomarker information measured at irregularly spaced time points by using a kernel smoothing method.

6. Software

The R code with example data are available from https://github.com/Qliu428/risk-prediction-model.git and also from

https://rss.onlinelibrary.wiley.com/hub/journal/14679876/series-c-datasets.
Acknowledgements

We thank the Joint Editors, Associate Editor and the reviewers for their helpful comments and suggestions, which have led to substantial improvements of this paper.

References

Beyersmann, J. and Schumacher, M. (2008) Time dependent covariates in the proportional subdistribution hazards model for competing risks. *Biommetrics*, 9, 765–776.

Blanche, P., Proust-Lima, C., Loubere, L., Bertrand, F., Dartigues, J. and Jacqmin-Gadda, H. (2015) Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*, 71, 102–113.

Cortese, G. and Andersen, P. K. (2009) Competing risks and time-dependent covariates. *Biometr. J.*, 51, 138–158.

Cortese, G., Gerds, T. A. and Andersen, P. K. (2013) Comparing predictions among competing risks models with time-dependent covariates. *Statist. Med.*, 32, 3089–3101.

Fine, J. P. and Gray, R. J. (1999) A proportional hazards model for the subdistribution of a competing risk. *J. Am. Statist. Ass.*, 94, 496–509.

Fisher, B., Dignam, J., Wolmark, N., DeCillis, A., Emir, B., Wickerham, D. L., Bryant, J., Dimitrov, N., Abrams, N., Atkins, J. N., Shibata, H., Deschenes, L. and Morgenstern, R. G. (1997) Tamoxifen and chemotherapy for lymph node-negative, estrogen receptor-positive breast cancer. *J. Natn. Cancer Inst.*, 89, 1673–1682.

Fontein, D. B. Y., Grand, M. K., Nortier, J. W., Seynaeve, C., Krakenbarg, R., Dirkx, L. Y., Meershoek-Klein, E., van de Velde, C. J. H. and Putter, H. (2015) Dynamic prediction in breast cancer: proving feasibility in clinical practice using the team trial. *Ann. Oncol.*, 26, 1254–1262.

Geskus, R. B. (2011) Cause-specific cumulative incidence estimation and the Fine and Gray model under both left truncation and right censoring. *Biometrics*, 67, 39–49.

van Houwelingen, H. C. (2007) Dynamic prediction by landmarking in event history analysis. *Scand. J. Statist.*, 34, 70–85.

van Houwelingen, H. C. and Putter, H. (2008) Dynamic prediction by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. *Lifetim. Data Anal.*, 14, 447–463.

van Houwelingen, H. C. and Putter, H. (2012) *Dynamic Prediction in Clinical Survival Analysis*. Boca Raton: CRC Press.

Huang, X., Yan, F., Ning, J., Feng, Z., Choi, S. and Cortes, J. (2016) A two-stage approach for dynamic prediction of time-to-event distributions. *Statist. Med.*, 35, 2167–2182.

Kalbfleisch, J. D. and Prentice, R. L. (2002) *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Liu, Q., Tang, G., Costantino, J. P. and Chang, C. C. H. (2016). Robust prediction of the cumulative incidence function under non-proportional subdistribution hazards. *Can. J. Statist.*, 44, 127–141.

Mauguen, A., Rachet, B., Mathoulin-Pelissier, S., MacGrogan, G., Laurent, A. and Rondeau, V. (2013) Dynamic prediction of risk of death using history of cancer recurrences in joint frailty models. *Statist. Med.*, 32, 5366–5380.

Nicolaia, M. A., van Houwelingen, J. C., de Witte, T. M. and Putter, H. (2013a) Dynamic prediction in competing risks by landmarking. *Statist. Med.*, 32, 2031–2047.

Nicolaia, M. A., van Houwelingen, J. C., de Witte, T. M. and Putter, H. (2013b) Dynamic pseudo-observations: a robust approach to dynamic prediction in competing risks. *Biometrics*, 69, 1043–1052.

Parast, L., Cheng, S. C. and Cai, T. X. (2012) Landmark prediction of long-term survival incorporating short-term event time information. *J. Am. Statist. Ass.*, 107, 1492–1501.

Pfeiffer, R. M. and Gail, M. H. (2017) Absolute risk. In *Methods and Applications in Clinical Management and Public Health*. Boca Raton: CRC Press.

Proust-Lima, C. and Taylor, J. M. G. (2009) Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of post treatment PSA: a joint modeling approach. *Biometrics*, 10, 535–549.

R Core Team (2019) R: *a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

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

Rizopoulos, D., Hafitfield, L. A., Carlin, B. P. and Takkenberg, J. J. M. (2014) Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging. *J. Am. Statist. Ass.*, 109, 1385–1397.

Schoop, R., Beyersmann, J., Schumacher, M. and Binder, H. (2011) Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometr. J.*, 53, 88–112.

Wu, C., Li, L. and Li, R. (2020) Dynamic prediction of competing risk events using landmark sub-distribution
hazard model with multiple longitudinal biomarkers. *Statist. Meth. Med. Res.*, to be published, doi 10.1177/0962280220921553.

Zamboni, B., Yothers, G., Choi, M., Fuller, C., Dignam, J. J., Raich, P. C., Thomas, Jr, C. R., O’Connell, M. J., Wolmark, N. and Wang, S. J. (2010) Conditional survival and the choice of conditional set for patients with colon cancer: an analysis of NSABP trials C-03 through C-07. *J. Clin. Oncol.*, 28, 2544–2548.

*Supporting information*

Additional 'supporting information' may be found in the on-line version of this article:

‘Landmark proportional subdistribution hazards models for dynamic prediction of cumulative incidence functions’.