Published in final edited form as:

Stat Med. 2013 September 10; 32(20): 3459–3471. doi:10.1002/sim.5776.

Landmark Risk Prediction of Residual Life for Breast Cancer Survival

Layla Parast^{1,*} and Tianxi Cai²

¹RAND Corporation, 1776 Main Street, Santa Monica, CA 90401

²Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115

Abstract

The importance of developing personalized risk prediction estimates has become increasingly evident in recent years. In general, patient populations may be heterogeneous and represent a mixture of different unknown subtypes of disease. When the source of this heterogeneity and resulting subtypes of disease are unknown, accurate prediction of survival may be difficult. However, in certain disease settings the onset time of an observable short term event may be highly associated with these unknown subtypes of disease and thus may be useful in predicting long term survival. One approach to incorporate short term event information along with baseline markers for the prediction of long term survival is through a landmark Cox model, which assumes a proportional hazards model for the residual life at a given landmark point. In this paper, we use this modeling framework to develop procedures to assess how a patient's long term survival trajectory may change over time given good short term outcome indications along with prognosis based on baseline markers. We first propose time-varying accuracy measures to quantify the predictive performance of landmark prediction rules for residual life and provide resampling-based procedures to make inference about such accuracy measures. Simulation studies show that the proposed procedures perform well in finite samples. Throughout, we illustrate our proposed procedures using a breast cancer dataset with information on time to metastasis and time to death. In addition to baseline clinical markers available for each patient, a chromosome instability genetic score, denoted by CIN25, is also available for each patient and has been shown to be predictive of survival for various types of cancer. We provide procedures to evaluate the incremental value of CIN25 for the prediction of residual life and examine how the residual life profile changes over time. This allows us to identify an informative landmark point, t_0 , such that accurate risk predictions of the residual life could be made for patients who survive past t_0 without metastasis.

Keywords

landmark prediction; biomarkers; disease prognosis; predictive accuracy; risk prediction; survival analysis

1. Introduction

The importance of developing personalized risk prediction estimates has become increasingly evident in recent years. It is no longer generally expected that patients with similar clinical symptoms will experience similar future clinical events. Rather, subjects

with similar clinical symptoms may have drastically different disease outcomes. For example, breast cancer patients sharing the same clinical features, such as lymph node status and histological grade, may have significantly different survival profiles [1, 2]. In general, patient populations may be heterogeneous and represent a mixture of different unknown subtypes of disease.

When the source of this heterogeneity and resulting subtypes of disease are unknown, accurate prediction of survival may be difficult. However, in certain disease settings the onset time of an observable short term event may be highly associated with these unknown subtypes of disease, and thus may be useful in predicting long term survival. For example, Cortese and Anderson [3] and Lee et al. [4] examined how the development of acute graft-versus-host disease affects survival of acute leukemia patients following bone marrow or stem cell transplantation. Additionally, Hirschtick et al. [5] demonstrated an association between occurrence of bacterial pneumonia and death among HIV-positive patients. Recently, the extent to which change in oncological status after surgery can predict long term survival for breast cancer patients was investigated in Hatteville et al. [6].

To optimally select prevention and treatment strategies in these types of settings, it would be of great interest to develop comprehensive prognostic systems for patients that could make predictions about long term survival, T_L , using the status of the short term event time, T_S^* . Methods developed specifically to incorporate T_S^* information along with other baseline predictors for the prediction of T_L has largely focused on semi-parametric models in a multi-state framework [7, 8, 9, 10, 6, 3]. For example, Klein et al. [9] use a standard Cox regression analysis to incorporate information on whether or not the intermediate event has occurred by a certain time. Hatteville et al. [6] extend this method to include non-proportional hazards and propose estimation. Other semi-parametric models for the cause-specific hazard functions [11, 12, 13] or the joint distribution of the short term and long term events [14, 15, 16, 17, 18, 19, 20, 21] can also be used to predict the long term survival given the short term event.

While these methods provide valuable tools to leverage information about the short term event status, the model assumptions may not always hold. Under model mis-specification, these model-based procedures may lead to unstable or inaccurate prediction rules and invalid inference about the predictive performance of the prediction rules. To make robust predictions of T_L without imposing restrictive assumptions on the relationship between T_L and T_S^* , a landmark prediction model was proposed by Van Houwelingen [22] and Van Houwelingen & Putter [23, 24]. Given some landmark time, t_0 , this model, which imposes a proportional hazards assumption for the residual life, $R_L^{t_0} = T_L - t_0$, among patients who survive past t_0 without experiencing the short term event, can be a convenient tool for constructing prediction rules for the residual life. Van Houwelingen [22] provides robust procedures for making inference about the hazard ratio parameter under possible model mis-specification. In this paper, we use this modeling framework to develop procedures for the assessment of how the long term survival trajectory may change over time given good short term outcome indications along with prognosis based on baseline markers. The incorporation of short term event time information has also been examined in Parast et al. [25] in which the actual time of the short term event is used in prediction. However, in Parast et al. [25], only a single binary marker measured at baseline is considered.

There are three goals in this paper. First, we derive landmark risk prediction rules for the residual life and evaluate their predictive performance by extending existing time-specific accuracy measures, such as the receiver operating characteristic curve to the setting with residual life landmark risk prediction. Robust inference procedures for such accuracy

measures are provided based on a resampling procedure. When a new marker is available to assist in risk prediction, it is important to evaluate its incremental value due to potential costs in measuring the marker. Van Houwelingen [22] and Van Houwelingen & Putter [24] propose to evaluate the usefulness of a new marker by examining the magnitude and significance of the corresponding hazard ratio parameter in the landmark model. However, it is well known that relative risk parameters are not adequate for assessing the predictive ability of the model [26, 27]. Our second goal here is to evaluate the incremental value of a new marker for landmark residual life prediction based on the improvement in prediction accuracy. In addition, the choice of the landmark time t_0 may be difficult in practice. We provide simultaneous inference procedures to evaluate the prediction performance and incremental value over a suitable range for t_0 . Such inference tools will enable us to identify an informative landmark point at which accurate prediction rules may be available for the residual life.

Throughout, we illustrate our proposed procedures using a dataset originally used in van de Vijver et al. [28] to evaluate a 70-gene risk score for breast cancer prognosis. This dataset consists of 260 women, with individual information on time to death, time to distant metastasis, age, tumor grade, size of tumor, baseline lymph node status, and estrogen receptor (ER) status. All patients had been treated by modified radical mastectomy or breast-conserving surgery and time is measured from time of surgery. Figure 1 displays the Kaplan Meier estimates for survival and metastasis-free survival. In addition to the above clinical markers, we also have a chromosome instability genetic score, denoted by CIN25, which has been previously shown as predictive of survival for various types of cancer [29]. We first obtain landmark prediction models for the residual life based on all predictors and then evaluate how the residual life profiles for patients with good or poor prognosis at baseline, change over time. Furthermore, we evaluate the incremental value of the genetic score for the prediction of $R_{\mathbb{L}}^{t_0}$ and conclude that the improvement due to the genetic score is limited when t_0 is small but could be substantial among patients who survive 2.5 years without metastasis.

2. The Landmark Prediction Model

Throughout, we let $T_{\mathbb{L}}$ denote the time of the long term event and $T_{\mathbb{S}}^*$ denote the time of the short term event. For example, $T_{\mathbb{L}}$ may represent time to death and $T_{\mathbb{S}}^*$ may represent time to distant metastasis. Due to censoring and semi-competing risks, one may not observe $T_{\mathbb{L}}$ or $T_{\mathbb{S}}^*$ directly. In the presence of competing risks, the short term event $T_{\mathbb{S}}^*$ such as distant metastasis may not be observable if $T_{\mathbb{L}}$ corresponds to a terminal event such as death. However, we assume that $T_{\mathbb{L}}$ is only subject to non-informative censoring, i.e. $T_{\mathbb{L}}$ will not be censored by $T_{\mathbb{S}}^*$. Under such assumptions, both $T_{\mathbb{L}}$ and $T_{\mathbb{S}} = \min(T_{\mathbb{S}}^*, T_{\mathbb{L}})$ are only subject to independent censoring. Thus for $T_{\mathbb{L}}$, we observe $\tilde{T}_{\mathbb{L}} = \min(T_{\mathbb{L}}, C)$, $\delta_{\mathbb{L}} = I(T_{\mathbb{L}} \leq C)$, for $\mathbb{L} = \mathbb{S}, \mathbb{L}$ where C is time of censoring. We assume that C is independent of $(T_{\mathbb{L}}, T_{\mathbb{S}})$ and $\mathbf{Z}$ with a common survival function $G(t) = P(C > t)$. Let $\mathbf{Z}$ denote the $p \times 1$ vector of predictors measured at baseline and suppose data for analysis consist of n independent and identically distributed random vectors $\{(\tilde{T}_{\mathbb{L}i}, \delta_{\mathbb{L}i}, \tilde{T}_{\mathbb{S}i}, \delta_{\mathbb{S}i}, \mathbf{Z}_i), i = 1, \dots, n\}$.

Here, we are particularly interested in the landmark prediction of the residual life $R_{\mathbb{L}}^{t_0} = T_{\mathbb{L}} - t_0$ among those who survive past t_0 without experiencing the short term event, defined as $\Omega_{t_0} = \{T_{\mathbb{S}} > t_0\}$, where $T_{\mathbb{S}} = \min(T_{\mathbb{S}}^*, T_{\mathbb{L}})$. We aim to develop conditional prognostic rules for classifying subjects as having a good or poor long term prognosis given they survive past t_0 without experiencing the short term event. Such a prediction rule will provide information on a patient's long term survival trajectory. To this end, we fit the landmark prediction model as given in Van Houwelingen [22],

$$P_{\Omega_{t_0}}(R_{\mathbb{L}}^{t_0} > \tau | \mathbf{Z}) = \exp \{ -\Lambda_0^{t_0}(\tau) \exp(\beta_{t_0}^\top \mathbf{Z}) \} \quad (1)$$

where $\Lambda_0^{t_0}(\cdot)$ is the unspecified baseline cumulative hazard function among Ω_{t_0} and β_{t_0} is the unknown covariate effect. When the model (1) holds, the same arguments as given in McIntosh & Pepe [30] can be used to show that the risk score $\mathcal{Z} = \beta_{t_0}^\top \mathbf{Z}$ has the optimal limiting receiver operating characteristic (ROC) curve (see Section 3.1) for classifying $I(R_{\mathbb{L}}^{t_0} \leq \tau)$ among Ω_{t_0} . To estimate β_{t_0} , one may use the subgroup with $\tilde{T}_{\mathbb{S}} > t_0$ as an unbiased random sample for the subpopulation Ω_{t_0} since $P(R_{\mathbb{L}}^{t_0} \leq \tau, \mathbf{Z} = \mathbf{z} | \tilde{T}_{\mathbb{S}} \geq t_0) = P_{\Omega_{t_0}}(R_{\mathbb{L}}^{t_0} \leq \tau, \mathbf{Z} = \mathbf{z})$. Let $\hat{\beta}_{t_0}$ denote the maximizer of the partial likelihood equation from this subpopulation as given in Van Houwelingen [22]. Note that when $t_0 = 0$ this is equivalent to no landmarking; we will refer to model 1 with $t_0 = 0$ as the global Cox model. Since Van Houwelingen [22] does not provide theoretical justifications for the asymptotic properties of $\hat{\beta}_{t_0}$, we provide details concerning the convergence of $\hat{\beta}_{t_0}$ as a process in t_0 in Appendix A of the Supplementary Materials.

3. Evaluating the Landmark Prediction Model

3.1. Estimation Procedures for Measures of Predictive Performance

To evaluate the accuracy of $\hat{\beta}_{t_0}^\top \mathbf{Z}$ for the landmark prediction of residual life, we next extend various time dependent accuracy measures proposed for the standard risk prediction setting [31, 32, 33, 34] to incorporate the conditional prognosis given $T_{\mathbb{S}} > t_0$. Specifically, for any given risk score $\mathcal{Z}$ and fixed t_0 , the discrimination accuracy of $\mathcal{Z}$ for classifying $I(R_{\mathbb{L}}^{t_0} \leq \tau)$ among Ω_{t_0} can be summarized by the time-specific sensitivity and specificity functions:

$$\text{Sens}_{t_0, \tau}(c) = P_{\Omega_{t_0}}(\mathcal{Z} > c | R_{\mathbb{L}}^{t_0} \leq \tau), \quad \text{and} \quad \text{Spec}_{t_0, \tau}(c) = P_{\Omega_{t_0}}(\mathcal{Z} < c | R_{\mathbb{L}}^{t_0} > \tau)$$

where $P_{\Omega_{t_0}}$ represents the probability taken over the sub-population in Ω_{t_0} and c is a threshold value. To adequately summarize the inherent discrimination accuracy for a score, one may summarize the trade-offs between the sensitivity and specificity using the receiver operating characteristic (ROC) curve [35, 36, 37]. The time-specific ROC curve in this setting can be defined as $\text{ROC}_{t_0, \tau}(u) = \text{Sens}_{t_0, \tau}\{\text{Spec}_{t_0, \tau}^{-1}(1-u)\}$ and the area under the ROC curve (AUC), $\text{AUC}_{t_0, \tau} = \int \text{ROC}_{t_0, \tau}(u) du$, can be used as an overall accuracy measure.

These classification accuracy measures are useful for describing the inherent capacity a score has in discriminating $I(R_{\mathbb{L}}^{t_0} \leq \tau)$ and thus for identifying optimal scoring systems and developing rules for assigning subjects into good or poor prognostic groups based on a selected threshold value c . After such a rule is identified, it would be important to evaluate the survival trajectory for patients assigned into the good or poor prognosis groups and examine how it varies over the landmark time t_0 . For such purposes, one may consider the positive predictive value (PPV) and negative predictive value (NPV) functions, defined as

$$\text{PPV}_{t_0, \tau}(c) = P_{\Omega_{t_0}}(R_{\mathbb{L}}^{t_0} \leq \tau | \mathcal{Z} > c), \quad \text{and} \quad \text{NPV}_{t_0, \tau}(c) = P_{\Omega_{t_0}}(R_{\mathbb{L}}^{t_0} > \tau | \mathcal{Z} \leq c).$$

To estimate these accuracy measures at any given t_0 and τ , we propose the following inverse probability weighted estimators with t_0 and τ chosen such that $P(\tilde{T}_{\mathbb{L}} > t_0 + \tau, \tilde{T}_{\mathbb{S}} > t_0) > 0$. Specifically, let

$$\begin{aligned}\widehat{\text{Sens}}_{t_0, \tau}(c) &= \frac{\sum_{i: \tilde{T}_{Si} > t_0} \widehat{W}_{t_0i} I(\widehat{\beta}_{t_0}^\top \mathbf{Z}_i \geq c) I(\tilde{R}_{Li}^{t_0} \leq \tau)}{\sum_{i: \tilde{T}_{Si} > t_0} \widehat{W}_{t_0i} I(\tilde{T}_{Li} \leq t_0 + \tau)}, & \widehat{\text{Spec}}_{t_0, \tau}(c) &= \frac{\sum_{i: \tilde{T}_{Si} > t_0} \widehat{W}_{t_0i} I(\widehat{\beta}_{t_0}^\top \mathbf{Z}_i < c) I(\tilde{R}_{Li}^{t_0} > \tau)}{\sum_{i: \tilde{T}_{Si} > t_0} \widehat{W}_{t_0i} I(\tilde{T}_{Li} > t_0 + \tau)} \\ \widehat{\text{PPV}}_{t_0, \tau}(c) &= \frac{\sum_{i: \tilde{T}_{Si} > t_0} \widehat{W}_{t_0i} I(\widehat{\beta}_{t_0}^\top \mathbf{Z}_i \geq c) I(\tilde{R}_{Li}^{t_0} \leq \tau)}{\sum_{i: \tilde{T}_{Si} > t_0} \widehat{W}_{t_0i} I(\widehat{\beta}_{t_0}^\top \mathbf{Z}_i \geq c)}, & \widehat{\text{NPV}}_{t_0, \tau}(c) &= \frac{\sum_{i: \tilde{T}_{Si} > t_0} \widehat{W}_{t_0i} I(\widehat{\beta}_{t_0}^\top \mathbf{Z}_i < c) I(\tilde{R}_{Li}^{t_0} > \tau)}{\sum_{i: \tilde{T}_{Si} > t_0} \widehat{W}_{t_0i} I(\widehat{\beta}_{t_0}^\top \mathbf{Z}_i < c)}.\end{aligned}$$

where $\widehat{\beta}_{t_0}$ is the maximizer of

$$\widehat{\ell}_{t_0}(\beta) = \sum_{i: \tilde{T}_{Si} > t_0} \delta_i [\beta^\top \mathbf{Z}_i - \log \{ \sum_{j: \tilde{T}_{Sj} > t_0} e^{\beta^\top \mathbf{Z}_j} I(\tilde{R}_{Lj}^{t_0} \geq \tilde{R}_{Li}^{t_0}) \}], \quad \tilde{R}_{Li}^{t_0} = \tilde{T}_{Li} - t_0, \quad \widehat{W}_{t_0i} = \frac{I(\tilde{T}_{Li} > t_0 + \tau)}{\widehat{G}(t_0 + \tau)} + \frac{I(\tilde{T}_{Li} \leq t_0 + \tau) \delta_{Li}}{\widehat{G}(\tilde{T}_{Li})}$$

and $\widehat{G}(\cdot)$ is the Kaplan-Meier estimator of $G(\cdot)$. Subsequently, $\text{ROC}_{t_0, \tau}(u)$ can be estimated as $\widehat{\text{ROC}}_{t_0, \tau}(u) = \widehat{\text{Sens}}_{t_0, \tau} \{ \widehat{\text{Spec}}_{t_0, \tau}^{-1}(1-u) \}$. Similarly, $\text{AUC}_{t_0, \tau}$ can be estimated by

$$\widehat{\text{AUC}}_{t_0, \tau} = \frac{\sum_{i: \tilde{T}_{Si} > t_0, \tilde{R}_{Li}^{t_0} \leq \tau} \sum_{j: \tilde{T}_{Sj} > t_0, \tilde{R}_{Lj}^{t_0} > \tau} I(\widehat{\beta}_{t_0}^\top \mathbf{Z}_i > \widehat{\beta}_{t_0}^\top \mathbf{Z}_j) \widehat{W}_{t_0i} \widehat{W}_{t_0j}}{\sum_{i: \tilde{T}_{Si} > t_0, \tilde{R}_{Li}^{t_0} \leq \tau} \sum_{j: \tilde{T}_{Sj} > t_0, \tilde{R}_{Lj}^{t_0} > \tau} \widehat{W}_{t_0i} \widehat{W}_{t_0j}}.$$

Note that though we use AUC as a measure of overall predictive performance at $t_0 + \tau$, many other measures of predictive performance are available. For example, if one is not interested in a specific time point but rather overall predictive performance across a range of time, overall measures such as the concordance C-index could be used [38, 39].

3.2. Inference Procedures for Measures of Predictive Performance

In Appendix B of the Supplementary Materials we show that the above accuracy functions are consistent for their theoretical counterparts, uniformly in (c, t_0, τ) , regardless of the adequacy of the landmark prediction model. Furthermore, we show that

$\sqrt{n}(\widehat{\text{Sens}}_{t_0, \tau}(c) - \text{Sens}_{t_0, \tau}(c))$ converges weakly to a zero-mean Gaussian process. Similarly, this holds for the other accuracy measures and it is not difficult to show that these convergences hold jointly. It then follows from the stochastic equi-continuity of these processes that $\sqrt{n}(\widehat{\text{ROC}}_{t_0, \tau}(u) - \text{ROC}_{t_0, \tau}(u))$ also converges weakly to a Gaussian process. Similar weak convergence results for other measures with cut-off values selected to achieve a sensitivity level of u , e.g. $\widehat{\text{PPV}}_{t_0, \tau} \{ \widehat{\text{Sens}}_{t_0, \tau}^{-1}(u) \}$ can be shown.

However, it is difficult to empirically estimate the variance associated with these processes as they involve derivative functions that are difficult to estimate especially under model misspecification. To overcome such difficulties, we propose using a perturbation-resampling method [40, 41, 33] to approximate the distributions of the proposed estimators. In

particular, let $\{V_i^{(b)}, i = 1, \dots, n, b = 1, \dots, B\}$ be nB independent realizations of a positive random variable V from a known distribution with unit mean and unit variance. Let $\widehat{\beta}_{t_0}^{(b)}$ be the minimizer of $\widehat{\ell}_{t_0}^{(b)}(\beta) = \sum_{i: \tilde{T}_{Si} > t_0} \delta_i V_i^{(b)} [\beta^\top \mathbf{Z}_i - \log \{ \sum_{j: \tilde{T}_{Sj} > t_0} V_j^{(b)} e^{\beta^\top \mathbf{Z}_j} I(\tilde{R}_{Lj}^{t_0} \geq \tilde{R}_{Li}^{t_0}) \}]$. Then one can estimate the variance of $\widehat{\beta}_{t_0}$ as the empirical variance of $\{ \widehat{\beta}_{t_0}^{(b)}, b = 1, \dots, B \}$. Note that robust estimation procedures for the variance of $\widehat{\beta}_{t_0}$ are proposed in Van Houwelingen [22], but we require the calculation of $\widehat{\beta}_{t_0}^{(b)}$ in order to derive variance estimates for the accuracy measures below. Furthermore, such resampling procedures will enable us to make joint

inference about these parameters over a range of t_0 . For accuracy measures such as $AUC_{t_0, \tau}$ and $Sens_{t_0, \tau}(c)$, let

$$\widehat{AUC}_{t_0, \tau}^{(b)} = \frac{\sum_{i: \tilde{T}_{Si} > t_0, \tilde{R}_{Li}^0 \leq \tau} \sum_{j: \tilde{T}_{Sj} > t_0, \tilde{R}_{Lj}^0 > \tau} I(\mathbf{Z}_i^T \widehat{\beta}_{t_0}^{(b)} > \mathbf{Z}_j^T \widehat{\beta}_{t_0}^{(b)}) \widehat{W}_{t_0i}^{(b)} \widehat{W}_{t_0j}^{(b)} V_i^{(b)} V_j^{(b)}}{\sum_{i: \tilde{T}_{Si} > t_0, \tilde{T}_{Li} \leq \tau} \sum_{j: \tilde{T}_{Sj} > t_0, \tilde{R}_{Lj}^0 > \tau} \widehat{W}_{t_0i}^{(b)} \widehat{W}_{t_0j}^{(b)} V_i^{(b)} V_j^{(b)}}$$

$$\widehat{Sens}_{t_0, \tau}^{(b)}(c) = \frac{\sum_{i: \tilde{T}_{Si} > t_0, \tilde{R}_{Li}^0 \leq \tau} \widehat{W}_{t_0i}^{(b)} I(\mathbf{Z}_i^T \widehat{\beta}_{t_0}^{(b)} \geq c) V_i^{(b)}}{\sum_{i: \tilde{T}_{Si} > t_0, \tilde{R}_{Li}^0 \leq \tau} \widehat{W}_{t_0i}^{(b)} V_i^{(b)}}$$

where $\widehat{W}_{t_0i}^{(b)} = I(\tilde{T}_{Li} > t_0 + \tau) / \widehat{G}^{(b)}(t_0 + \tau) + I(\tilde{T}_{Li} \leq t_0 + \tau) \delta_{Li} / \widehat{G}^{(b)}(\tilde{T}_{Li})$ and $\widehat{G}^{(b)}$ is the Kaplan-Meier estimator of $G(\cdot)$ with weights $V_i^{(b)}$. The empirical distributions of these realizations can be used to approximate the distribution of the corresponding estimators. For example, the variance of $\widehat{AUC}_{t_0, \tau}$ can be estimated using

$$\widehat{\sigma}_{t_0, \tau}^2 = B^{-1} \sum_{b=1}^B \{ \widehat{AUC}_{t_0, \tau}^{(b)} - \widehat{AUC}_{t_0, \tau} \}^2.$$

A $100(1 - 2\alpha)\%$ confidence interval (CI) for $AUC_{t_0, \tau}$ could be obtained either using the normal CI or the empirical quantiles of the perturbed samples. The validity of the resampling procedure can be justified based on the large sample theory given in Appendix B of the Supplementary Materials along with similar arguments as given in Jin et al. [40]. Variance estimates for all proposed accuracy measures can be estimated similarly.

The aforementioned accuracy estimators, commonly referred to as apparent accuracy estimates, tend to be overly optimistic, particularly when the dimension of $\mathbf{Z}$ is not small. To reduce this potential bias, we consider a general cross-validation where we randomly split the data into a training set of size n_t and a validation set of size $n_v = n - n_t$. For the k th split, we estimate β_{t_0} using the training set, denoted by $\widehat{\beta}_{t_0}^{(CV)k}$. Then for any accuracy measure of interest, say $AUC_{t_0, \tau}$ we estimate the AUC for the linear score $\widehat{\beta}_{t_0}^{(CV)kT} \mathbf{Z}$ based on the validation set, denoted by $\widehat{AUC}_{t_0, \tau}^{(CV)k}$. Let $\widehat{AUC}_{t_0, \tau}^{(CV)}$ be the average of all $\widehat{AUC}_{t_0, \tau}^{(CV)k}$ over K random splits. Using similar justification as given by Tian et al. [42] and Uno et al. [33], it can be shown that $\widehat{AUC}_{t_0, \tau}^{(CV)}$ and $\widehat{AUC}_{t_0, \tau}$ have the same limiting distribution at the first order. Therefore, one may construct a 95% CI for $AUC_{t_0, \tau}$ as $\widehat{AUC}_{t_0, \tau}^{(CV)} \pm 1.96 \widehat{\sigma}_{t_0, \tau}$.

3.3. Simulation Study

To examine finite-sample properties of the proposed accuracy measure estimation procedures, we conducted simulation studies and focused on the case with $T_L = T_S = T$ for simplicity. Two types of models were used to generate T : a mixture of log-normal in setting (i) and a Cox model in setting (ii). The covariate vector $\mathbf{Z}$ consists of 5 components: $(Z_2, Z_3, Z_4) \sim N(\mathbf{0}, \Sigma)$ with unit variance and a weak correlation ranging from -0.03 and 0.05 ; $Z_1 \sim$ a bernoulli distribution with success probability $\Phi(Z_2 + Z_3)$; and $Z_5 \sim \text{Uniform}[\Phi(Z_4), 1 + \Phi(Z_4)]$. We use setting (i) to illustrate that the robustness of the landmark model may lead to higher accuracy under model mis-specification, and setting (ii) to assess the efficiency loss due to landmarking when the global Cox model holds. In each setting, we generated 1000

realizations of size n for $n=200, 500$ and $1,000$ and obtained $\hat{\beta}_{t_0}$ as well as $\hat{\beta}$, the standard maximum partial likelihood estimator under a global Cox model. Accuracy measures were calculated for both β estimates. Here, in addition to assessing the overall accuracy based on the AUC, we consider the specificity, PPV and NPV at a cut-off value corresponding to a sensitivity level of 0.90. The standard error estimates were obtained based on 500 perturbations each.

In setting (i), the survival time was generated from $T = \exp\{(\beta_1^T Z + \varepsilon_1)/6\}$ with $\varepsilon_1 \sim N(-3, 1)$ if $\mathcal{B} = 1$ and $T = \exp\{(\beta_0^T Z + \varepsilon_0)/6\} + 1$ with $\varepsilon_0 \sim N(10, 4)$ if $\mathcal{B} = 0$, where $\mathcal{B} \sim \text{Bernoulli}(.4)$, $\beta_1 = (3, 6, 1.5, 0, 0)^T$ and $\beta_0 = \alpha(0, 0, 1.5, 3, 1.5)$. This mimics a clinical setting in which (Z_1, Z_2, Z_3) are predictive of the short term event, (Z_3, Z_4, Z_5) are predictive of long term survival. For illustration, we chose t_0 to be year 1 to reflect an early indication of disease and τ to be year 5. Therefore, among patients surviving 1 year, we wish to estimate their probability of having residual life greater than 5 years, given baseline covariates. Under these conditions, $P(T < t_0) = .23$ and $P(T < \tau + t_0 / T > t_0) = .49$. We first consider the case when C was generated from an Exponential distribution with rate .11, yielding a censoring rate of approximately 40%. The results are shown in Table 1. The resulting linear score has a higher overall accuracy with $\text{AUC}_{t_0, \tau} \sim .74$ for the landmark method compared to $\text{AUC}_{t_0, \tau} \sim .67$ using the global Cox model. All point estimators have negligible bias. The standard errors estimated from the resampling method are close to the empirical standard errors and the empirical coverage levels are close to their nominal level. Regression coefficient estimates are not shown here though coverage in this setting was also close to the empirical level of 0.95.

In setting (ii), we generated T from a Cox model with $1.25 \log(T) = 2 + \beta_0^T Z + \varepsilon$, where ε is from an extreme value distribution and $\beta_0 = (1, .5, .5, 1, .5)^T$. Under this configuration, $P(T < t_0) = .20$ and $P(T < \tau + t_0 / T > t_0) = .34$. Censoring time was generated from an Exponential distribution with rate .05, yielding approximately 40% of censoring. Since $\hat{\beta}$ is semi-parametric efficient and $\hat{\beta}_{t_0}$ is obtained based on the subset in Ω_{t_0} , which consists of 83% of the entire sample, there is a significant efficiency loss due to landmarking with efficiency loss ranging from 0.08 to 0.15. However, the efficiency loss in estimating the accuracy measures is negligible as shown in Table 2.

3.4. Predictive Performance in Breast Cancer Dataset at a fixed t_0

Using the breast cancer study mentioned previously, we first consider the landmark risk prediction of 10-year residual life among patients surviving year $t_0 = 3$ since surgery without metastasis. The 3-year rate of death was about 0.17. In Table 3(a), we present the regression coefficients estimates for the available clinical variables based on the landmark Cox model. Both the gene score CIN25 and the ER status are highly significant while the tumor size and the lymph node status do not appear to be predictive. Accuracy results are summarized in Table 4(a). The cross-validated estimate of the AUC is 0.694 with (s.e. = 0.067), suggesting that the linear score based on these markers has moderate accuracy in predicting 10-year residual life among Ω_3 i.e. patients who survive year 3 without metastasis. By choosing a threshold value to achieve sensitivity level of 0.90, one may classify patients as having a poor long term prognosis if the risk score is higher than the threshold value and good long term prognosis otherwise. Then for patients in Ω_3 , the chance of surviving another 10 years is about 95% (s.e. = 1.6%) for the good prognosis group but only 67% (s.e. = 10.5%) for the poor prognosis group.

4. Incremental Value of a New Marker

4.1. Inference Procedures for the Incremental Value

When a new marker such as a genetic or biomarker is available for risk prediction, determining the incremental value of the new marker is often of clinical interest, particularly if measuring the marker is expensive or invasive. Several procedures have been developed to quantify the overall incremental value from a new marker for the entire population [42, 33, 27, 43, 44]. Our proposed method could potentially shed light on how the incremental value may vary across sub-populations. If the new marker is only useful for predicting T_{L} among patients with $T_{\text{S}} > t_0$, these procedures may provide a useful tool for practitioners to decide when the new marker is needed in addition to conventional risk factors. Therefore, one should directly assess the increase in prediction performance resulting from the inclusion of the new marker.

Let (M_1) denote the model with routine markers only and (M_2) denote the model with both routine markers and the new marker. One may quantify the incremental value of the new marker based on $\Delta\text{AUC}_{t_0, \tau} = \text{AUC}_{t_0, \tau}^{(M_2)} - \text{AUC}_{t_0, \tau}^{(M_1)}$, where $\text{AUC}_{t_0, \tau}^{(M_k)}$ denotes the AUC of (M_k) , for $k = 1, 2$. To make inference about $\Delta\text{AUC}_{t_0, \tau}$, let $\widehat{\text{AUC}}_{t_0, \tau}^{(\text{CV})(M_k)}$ denote the cross-validated estimate of $\text{AUC}_{t_0, \tau}^{(M_k)}$ and $\{\widehat{\text{AUC}}_{t_0, \tau}^{(M_k)(b)}, b = 1, \dots, B\}$ be the perturbed estimates of $\text{AUC}_{t_0, \tau}^{(M_k)}$. Then one may construct confidence intervals for $\Delta\text{AUC}_{t_0, \tau}$ based on $\Delta\widehat{\text{AUC}}_{t_0, \tau}^{(\text{CV})} = \widehat{\text{AUC}}_{t_0, \tau}^{(\text{CV})(M_2)} - \widehat{\text{AUC}}_{t_0, \tau}^{(\text{CV})(M_1)}$ and the empirical variance of $\{B^{-1} \sum_{i=1}^n \{\widehat{\text{AUC}}_{t_0, \tau}^{(M_2)(b)} - \widehat{\text{AUC}}_{t_0, \tau}^{(M_1)(b)}, b = 1, \dots, B\}$. When interest lies in the accuracy of a binary classification rule, one may choose the cut-off value according to a desired sensitivity or specificity and then assess the incremental value with respect to other dimensions of the accuracy. For example, the incremental value of the new markers can be quantified based on $\Delta\text{Spec}_{t_0, \tau, u_0} = \text{Spec}_{t_0, \tau}^{(M_2)}\{\text{Sens}_{t_0, \tau}^{(M_2)-1}(u_0)\} - \text{Spec}_{t_0, \tau}^{(M_1)}\{\text{Sens}_{t_0, \tau}^{(M_1)-1}(u_0)\}$ where u_0 is the cut-off value that achieves the desired sensitivity level. Corresponding point and interval estimates may be calculated using the cross-validation and resampling as for $\Delta\text{AUC}_{t_0, \tau}$.

4.2. Incremental Value of CIN25 in the Breast Cancer Dataset

To evaluate the incremental value of CIN25 in predicting long term survival, we compared the estimated accuracy from Section 3.4 where we used all available predictors to that obtained by fitting the model without the gene score. In Table 3(b), we present the new regression coefficients estimates for fixed $t_0 = 3$ and $\tau = 10$. Without the gene score, the most informative markers appear to be ER status and tumor grade. Accuracy results are summarized in Table 4(b). The AUC estimate decreases from 0.694 to 0.621 when the gene score is removed from the model. A 95% confidence interval for the difference in the AUC is $(-0.04, 0.19)$ suggesting that the gene score slightly improves the accuracy for the landmark prediction though not significantly.

5. Examining Predictive Performance and Incremental Value Across a Range of t_0

5.1. Simultaneous Inference Across a Range of t_0

It is important to note that our proposed methods hold for any choice of t_0 and τ . In practice, the choice of t_0 may be difficult unless clinically meaningful suggestions exist. One may consider the proposed estimator across a range of t_0 and/or τ as a basis for selecting an appropriate time point such that the prediction model is most accurate. For illustration,

suppose that τ is fixed and one would like to examine how the accuracy of the landmark prediction of $I(R_{\mathbb{L}}^{t_0} \leq \tau)$ varies over t_0 to determine when $\mathbf{Z}$ information is most useful in prediction. Simultaneous inference about $\text{AUC}_{t_0, \tau}$ across a range of t_{0j} , denoted by $\mathcal{S}$, can be made by constructing point-wise and simultaneous CIs using the perturbation-resampling method as described above. A $100(1 - \alpha)\%$ simultaneous CI for $\{\widehat{\text{AUC}}_{t_0, \tau}^{(\text{CV})}, t_{0j} \in \mathcal{S}\}$ can be obtained as $\{\widehat{\text{AUC}}_{t_0, \tau}^{(\text{CV})} \pm \widehat{c}_\alpha \widehat{\sigma}_{t_0, \tau}, t_{0j} \in \mathcal{S}\}$, where $\widehat{c}_\alpha$ is the $(1 - \alpha)$ empirical quantile of $\{\sup_{t_{0j} \in \mathcal{S}} |\widehat{\text{AUC}}_{t_0, \tau}^{(b)} - \widehat{\text{AUC}}_{t_0, \tau}| / \widehat{\sigma}_{t_0, \tau}, b = 1, \dots, B\}$. The simultaneous CI will ensure the control of family-wise type I error when selecting a set of t_{0j} 's such that the overall accuracy is above a certain threshold value. To examine how the incremental value varies over t_0 , one may assess $\Delta \text{AUC}_{t_0, \tau}$ across a range of $t_0 \in \mathcal{S}$ with a fixed τ and identify the range of t_0 such that M_2 is better than M_1 . To this end, one may construct a plot of $\Delta \widehat{\text{AUC}}_{t_0, \tau}^{(\text{CV})} = \widehat{\text{AUC}}_{t_0, \tau}^{(\text{CV})(M_2)} - \widehat{\text{AUC}}_{t_0, \tau}^{(\text{CV})(M_1)}$. Simultaneous confidence intervals for $\{\Delta \text{AUC}_{t_0, \tau}, t_{0j} \in \mathcal{S}\}$ could be constructed using similar procedures as for $\{\widehat{\text{AUC}}_{t_0, \tau}^{(\text{CV})}, t_{0j} \in \mathcal{S}\}$.

When considering overall accuracy of a set of predictors, AUC is a valuable and meaningful measure. However, from a patient's perspective, estimates for their survival probabilities conditional on their prognosis status at baseline may be more useful for risk benefit decisions. For example, consider the breast cancer setting where all patients in Ω_{t_0} are classified as having either good or poor prognosis at baseline for their 10-year residual life from t_0 across $t_0 \in \mathcal{S}$. Now, for patients with poor prognosis at baseline, we have information on their probability of surviving 10 years from baseline, $\mathcal{P}_0$, and suppose that the 10-year residual life survival probability is highest at some landmark point t_0^* . With this information we can further tell the patient that despite their poor prognosis at baseline, if they survive to t_0^* without metastasis, then their probability of surviving an additional 10 years from t_0^* may be increased to $\mathcal{P}_0 + \varepsilon$. This could potentially be extremely valuable information to the patient depending on the magnitude of ε and the corresponding confidence intervals. In other words, we wish to provide the patient with some t_0^* which, provided they survive to that point without metastasis, their probability of survival is significantly higher than the probability provided to them without this information. To provide a more complete picture of the long term trajectory of disease progression, one may examine estimates of these residual survival probabilities over a range of landmark points by plotting $1 - \widehat{\text{PPV}}_{t_0, \tau}^{(\text{CV})}(c)$ and $\widehat{\text{NPV}}_{t_0, \tau}^{(\text{CV})}(c)$ over a set of $t_{0j} \in \mathcal{S}$, where c is the cut-off value corresponding to a given sensitivity or specificity level. Both point-wise and simultaneous CIs can be constructed again using the perturbation-resampling method as described above.

5.2. Residual Life Landmark Prediction Across a Range of t_0 for the Breast Cancer Study

For the breast cancer study, we are interested in examining how the accuracy of the landmark model for predicting $I(R_{\mathbb{L}}^{t_0} \leq 10)$ varies over $t_0 \in \mathcal{S} = [0, 4]$ years. Due to the small sample size and relatively low event rate, we choose $t_{0j} = 0.5j$ for $j = 1, \dots, 8$ and summarize the overall predictive accuracy at t_{0j} as $\overline{\text{AUC}}_{t_0, \tau} = 2 \int_{t_{0j}-0.5}^{t_{0j}} \text{AUC}_{t_0, \tau} dt_0$. Figure 2 displays the cross-validated estimates of $\overline{\text{AUC}}_{t_0, \tau}$, $\overline{\text{NPV}}_{t_0, \tau}(c) = 2 \int_{t_{0j}-0.5}^{t_{0j}} \text{NPV}_{t_0, \tau} dt_0$, and $1 - \overline{\text{PPV}}_{t_0, \tau}(c) = 1 - 2 \int_{t_{0j}-0.5}^{t_{0j}} \text{PPV}_{t_0, \tau} dt_0$ for $j = 1, \dots, 8$ along with their corresponding pointwise and simultaneous confidence intervals. Based on the AUC estimates, the overall accuracy of the landmark model for predicting 10-year residual life does not appear to change with t_{0j} initially with point estimates ranging from 0.70 to 0.73 for $t_{0j} = 2.5$, but

begins to decrease for larger t_{0j} 's and drops to 0.60 when $t_{0j} = 4$. Now, for patients classified as having good prognosis at baseline, the probability of surviving 10 *additional years after the landmark point* i.e. NPV stays near 90% across all t_{0j} 's. On the other hand, it is interesting to observe that for patients with poor prognosis at baseline, their probability of surviving 10 *additional years after the landmark point* appears to increase by landmarking at $t_{0j} = 3$ years. At baseline, patients in the poor prognosis group have a probability of 0.60 (95 % CI: 0.47, 0.71) of surviving to year 10. Based on the landmarking information, we could further inform the poor prognosis group at baseline that if they survive the first 3 years without metastasis, then their chance of surviving an additional 10 years will increase to 0.67 (95 % CI: 0.47, 0.80). Note that classification of a patient into the good or poor prognosis group is done at baseline with baseline predictors. If updated predictor information is available at t_0 , it would be possible for patients to be re-classified based on this updated information.

To evaluate the incremental value of the gene score across different values of t_0 , let M_2 and M_1 indicate models with and without the gene score, respectively. Figure 3(a) displays the cross-validated estimates of $\Delta\overline{AUC}_{t_{0j},\tau} = \overline{AUC}_{t_{0j},\tau}^{(M_2)} - \overline{AUC}_{t_{0j},\tau}^{(M_1)}$, across the aforementioned values of t_{0j} . At $t_0 = 0$, the gene score appears to have minimal incremental value in landmark residual life risk prediction with $\Delta\overline{AUC}_{0,\tau}$ estimated as 0.04 (95 % CI: -0.03, 0.12). However, at $t_0 = 2.5$ years, $\Delta\overline{AUC}_{2.5,\tau}$ increases to 0.13 (95% CI: 0.04, 0.21) indicating a gain in predictive accuracy. As previously mentioned, Van Houwelingen [22] evaluates the usefulness of a new marker by examining the estimated hazard ratio, though it has previously been shown that significance of relative risk parameters are not adequate for assessing the predictive ability of a model. Figure 3(b) displays the hazard ratio estimates of the gene score across different values of t_0 . Note that at $t_0 = 0$, the estimated hazard ratio of gene score is significant even though there is no incremental value as measured by the AUC.

In practice, this may be extremely useful information as it may indicate that the gene score need not be measured at baseline, but rather should be measured on patients that survive past a chosen landmark time. In this case, it would seem that gene score should perhaps only be obtained from patients that survive past the first 2.5 years after surgery without metastasis. This may greatly reduce the time and cost related to collecting this information. In settings where collecting the new marker is invasive or may cause harm to the patient, this would decrease the burden of harm to the patient.

6. Remarks

In this paper we propose robust procedures for evaluating the landmark prediction model proposed in Van Houwelingen [22] for the prediction of residual life. When a new marker is available for prediction, the incremental value of such a marker must be evaluated. We additionally develop procedures to examine the incremental value of such markers in residual life landmark risk prediction. When the landmark time t_0 is unknown, one may wish to examine the proposed prediction performance measures and incremental value of the new marker over a suitable range for t_0 . We provide simultaneous inference procedures to carry out this task. This may lead to information guiding the choice of which markers should be ascertained at baseline, and which should be ascertained on only a subset of the population that survives past the landmark point.

We have specifically focused on providing predictions to patients who have survived past t_0 without experiencing the short term event. While this is of interest in many practical settings, it may also be desirable to provide predictions to patients who have survived past t_0 and *have* experienced the short term event. Methods that incorporate short term event

information from all patients who survive past t_0 such as those proposed in [25, 45] may be useful in such settings.

For simplicity, we assumed independent censoring with $C \perp \mathbf{Z}$. However, our proposed procedures could easily accommodate marker-dependent censoring by replacing $\hat{G}$ in $\hat{W}_{t_0}$ with $\hat{G}_Z$ obtained as the Kaplan-Meier estimate of $P(C > t | Z = z)$ for the subgroup with $Z_i = Z$. To obtain a more robust estimate of β_{t_0} , one could consider administrative censoring at $t_0 + \tau$ for each combination of t_0 and τ when fitting the model [22, 23, 24, 46].

The robustness of the landmark prediction rule is particularly important when T_L and T_S represent two different outcomes as for such settings, it is difficult if not impossible to identify bivariate survival models that fully capture the complex relationship between the correlated outcomes and the predictors. It is thus desirable to have model-free procedures such as those proposed in this paper to evaluate prediction models and compare models.

7. Supplementary Material

The reader is referred to the on-line Supplementary Materials for technical appendices.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The majority of this work was completed while Parast was a Ph.D. candidate in biostatistics at Harvard University. Support for this research was provided by National Institutes of Health grants T32GM074897 (Parast), R01AI052817 (Parast), R01-GM079330 (Cai) and U54LM008748 (Cai).

References

1. Dent R, Trudeau M, Pritchard KI, Hanna WM, Kahn HK, Sawka CA, Lickley LA, Rawlinson E, Sun P, Narod SA. Triple-negative breast cancer: clinical features and patterns of recurrence. *Clinical Cancer Research*. 2007; 13(15):4429.10.1158/1078-0432.CCR-06-3045 [PubMed: 17671126]
2. Goldhirsch A, Wood WC, Senn HJ, Glick JH, Gelber RD. Meeting highlights: international consensus panel on the treatment of primary breast cancer. *Journal of the National Cancer Institute*. 1995; 87(19):1441.10.1093/jnci/87.19.1441 [PubMed: 7674331]
3. Cortese G, Andersen PK. Competing risks and time-dependent covariates. *Biometrical Journal*. 2010; 52(1):138–158.10.1002/bimj.200900076 [PubMed: 20029852]
4. Lee SJ, Klein JP, Barrett AJ, Ringden O, Antin JH, Cahn JY, Carabasi MH, Gale RP, Giralt S, Hale GA, et al. Severity of chronic graft-versus-host disease: association with treatment-related mortality and relapse. *Blood*. 2002; 100(2):406–414.10.1182/blood.V100.2.406 [PubMed: 12091329]
5. Hirschtick RE, Glassroth J, Jordan MC, Wilcosky TC, Wallace JM, Kvale PA, Markowitz N, Rosen MJ, Mangura BT, Hopewell PC, et al. Bacterial pneumonia in persons infected with the human immunodeficiency virus. *New England Journal of Medicine*. 1995; 333(13):845.10.1056/NEJM199509283331305 [PubMed: 7651475]
6. Hatteville L, Mahe C, Hill C. Prediction of the long-term survival in breast cancer patients according to the present oncological status. *Statistics in Medicine*. 2002; 21(16):2345–2354.10.1002/sim.1046 [PubMed: 12210619]
7. Putter H, van der Hage J, de Bock GH, Elgalta R, van de Velde CJH. Estimation and Prediction in a Multi-State Model for Breast Cancer. *Biometrical Journal*. 2006; 48(3):366–380.10.1002/bimj.200510218 [PubMed: 16845902]
8. Kay R. A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*. 1986; 42(4):855–865.10.2307/2530699 [PubMed: 2434150]

9. Klein JP, Keiding N, Copelan EA. Plotting summary predictions in multistate survival models: Probabilities of relapse and death in remission for bone marrow transplantation patients. *Statistics in Medicine*. 1994; 12(24):2315–2332.10.1002/sim.4780122408 [PubMed: 8134735]
10. Datta S, Satten GA, Datta S. Nonparametric Estimation for the Three-Stage Irreversible Illness–Death Model. *Biometrics*. 2000; 56(3):841–847.10.1111/j.0006-341X.2000.00841.x [PubMed: 10985224]
11. Fine J, Jiang H, Chappell R. On semi-competing risks data. *Biometrika*. 2001; 88(4):907–919.
12. Siannis F, Farewell V, Head J. A multi-state model for joint modeling of terminal and non-terminal events with application to whitehall ii. *Statistics in medicine*. 2007; 26(2):426–442.10.1002/sim.2342 [PubMed: 16220522]
13. Jiang H, Fine JP, Chappell R. Semiparametric analysis of survival data with left truncation and dependent right censoring. *Biometrics*. 2005; 61(2):567–575.10.1111/j.1541-0420.2005.00335.x [PubMed: 16011706]
14. Shih JH, Louis TA. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*. 1995; 51(4):1384–1399.10.2307/2533269 [PubMed: 8589230]
15. Klaassen CAJ, Wellner JA. Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*. 1997; 3(1):55–77.
16. Oakes D, Ritz J. Regression in a bivariate copula model. *Biometrika*. 2000; 87(2):345.10.1093/biomet/87.2.345
17. Pitt M, Chan D, Kohn R. Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*. 2006; 93(3):537.10.1093/biomet/93.3.537
18. Hougaard P. Frailty models for survival data. *Lifetime Data Analysis*. 1995; 1(3):255–273.10.1007/BF00985760 [PubMed: 9385105]
19. Ha ID, Lee Y, Song J. Hierarchical likelihood approach for frailty models. *Biometrika*. 2001; 88(1):233–243.10.1093/biomet/88.1.233
20. Cai T, Cheng S, Wei L. Semiparametric Mixed-Effects Models for Clustered Failure Time Data. *Journal of the American Statistical Association*. 2002; 97(458):514–523.10.1198/016214502760047041
21. Zeng D, Lin D. Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society, Series B Statistical Methodology*. 2007; 69(4):507.10.1111/j.1369-7412.2007.00606.x
22. Van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*. 2007; 34(1):70.
23. Van Houwelingen HC, Putter H. Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. *Lifetime Data Analysis*. 2008; 14(4):447–463.10.1007/s10985-008-9099-8 [PubMed: 18836831]
24. Van Houwelingen, HC.; Putter, H. *Dynamic Prediction in Clinical Survival Analysis*. CRC Press; 2012.
25. Parast L, Cheng SC, Cai T. Incorporating short-term outcome information to predict long-term survival with discrete markers. *Biometrical Journal*. 201110.1002/bimj.201000150
26. Ware J. The limitations of risk factors as prognostic tools. *New England Journal of Medicine*. 2006; 355(25):2615.10.1056/NEJMp068249 [PubMed: 17182986]
27. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology*. 2004; 159(9):882–90.10.1093/aje/kwh101 [PubMed: 15105181]
28. van de Vijver M, He Y, van't Veer L, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*. 2002; 345(25):1999–2009.10.1056/NEJMoa021967 [PubMed: 12490681]
29. Carter SL, Eklund AC, Kohane IS, Harris LN, Szallasi Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nature Genetics*. 2006; 38(9):1043–1048.10.1038/ng1861 [PubMed: 16921376]
30. McIntosh MW, Pepe MS. Combining several screening tests: Optimality of the risk score. *Biometrics*. 2002; 58(3):657–664.10.1111/j.0006-341X.2002.00657.x [PubMed: 12230001]

31. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. 2000; 56(2):337–344.10.1111/j.0006-341X.2000.00337.x [PubMed: 10877287]
32. Cai T, Pepe MS, Zheng Y, Lumley T, Jenney NS. The sensitivity and specificity of markers for event times. *Biostatistics*. 2006; 7(2):182–97.10.1093/biostatistics/kxi047 [PubMed: 16079162]
33. Uno H, Cai T, Tian L, Wei LJ. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*. 2007; 102:527–37.
34. Cai T, Cheng S. Robust combination of multiple diagnostic tests for classifying censored event times. *Biostatistics*. 2008; 9(2):216.10.1093/biostatistics/kxm037 [PubMed: 18056687]
35. Swets, JA.; Pickett, RM. *Evaluation of Diagnostic Systems: Methods from Signal Detection theory*. Academy press; New York: 1982.
36. Hanley HA. Receiver Operating Characteristic (ROC) methodology: the state of the art. *Clinical Reviews in Diagnostic Imaging*. 1989; 29:307–35.
37. Begg CB. Advances in statistical methodology for diagnostic medicine in the 1980's. *Statistics in Medicine*. 1991; 10:1887–1895. [PubMed: 1805316]
38. Harrell FE, Lee KL, Mark D. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*. 1996; 15:361–387. [PubMed: 8668867]
39. Uno H, Cai T, Pencina MJ, Agostino DRB, Wei LJ. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*. 2011; 30(10):1105–1117.10.1002/sim.4154 [PubMed: 21484848]
40. Jin Z, Ying Z, Wei LJ. A simple resampling method by perturbing the minimand. *Biometrika*. 2001; 88(2):381–390.10.1093/biomet/88.2.381
41. Park Y, Wei LJ. Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika*. 2003; 90:717–23.
42. Tian L, Cai T, Goetghebeur E, Wei L. Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika*. 2007; 94(2):297.10.1093/biomet/asm036
43. Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, Zheng Y. Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology*. 2008; 167(3):362.10.1093/aje/kwm305 [PubMed: 17982157]
44. Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine*. 2008; 27(2):157–172. [PubMed: 17569110]
45. Parast L, Cheng SC, Cai T. Landmark Prediction of long term survival incorporating short term event time information. *Journal of the American Statistical Association*. 2012; 107:1492–1501.10.1080/01621459.2012.721281 [PubMed: 23293405]
46. Xu R, Quigley OJ. Estimating average regression effect under non-proportional hazards. *Biostatistics*. 2000; 1(4):423–439.10.1093/biostatistics/1.4.423 [PubMed: 12933565]

Figure 1. Kaplan Meier estimates and pointwise confidence intervals for survival (a) and metastasis-free survival (b) from the breast cancer dataset

Figure 2.

Cross-validated estimates across $t_{0j} \in \{0, 0.5, 1, \dots, 4\}$ along with their corresponding pointwise (dotted) and simultaneous (dashed) confidence intervals of **(a)** $\overline{AUC}_{t_{0j}, \tau}(c)$ **(b)** $\overline{NPV}_{t_{0j}, \tau}(c)$ and **(c)** $1 - \overline{PPV}_{t_{0j}, \tau}(c)$ where c is the cut-off value corresponding to a sensitivity level of 0.90. Vertical lines appear at each t_{0j} as a visual aid. The gray line represents the Kaplan-Meier estimate of 10-year residual life survival without any predictor information.

Figure 3.

Cross-validated estimates across $t_{0j} \in \{0, 0.5, 1, \dots, 4\}$ with corresponding pointwise (dotted) and simultaneous (dashed) confidence intervals of (a) $\Delta \overline{AUC}_{t_{0j}, \tau}$ and (b) $\exp(\hat{\beta}_{t_0})$. Vertical lines appear at each t_{0j} as a visual aid. For comparison, there is dark gray line drawn where the vertical axis is 0 for (a) and 1 for (b).

Table 1

Proposed prediction performance measure estimates using the landmark method (Land) vs. the global Cox method (Cox) from setting(i) under a log-normal mixture model for T , with corresponding empirical standard errors (ESE), average of the standard error estimates from the perturbation-resampling method (ASE), and empirical coverage levels (Coverage). The cutoff value c was chosen such that the sensitivity level is 0.90.

n=200											
Truth		Average		ESE		ASE		Coverage			
Land	Cox	Land	Cox	Land	Cox	Land	Cox	Land	Cox	Land	Cox
AUC	0.741	0.668	0.756	0.675	0.051	0.067	0.048	0.064	0.064	0.891	0.920
Spec	0.319	0.207	0.360	0.243	0.119	0.106	0.129	0.117	0.117	0.952	0.960
NPV	0.771	0.686	0.790	0.707	0.074	0.114	0.078	0.120	0.120	0.885	0.892
PPV	0.557	0.519	0.577	0.535	0.067	0.059	0.072	0.065	0.065	0.952	0.962

n=500											
Land	Cox	Land	Cox	Land	Cox	Land	Cox	Land	Cox	Land	Cox
AUC	0.741	0.668	0.746	0.671	0.031	0.041	0.031	0.041	0.041	0.937	0.946
Spec	0.319	0.207	0.336	0.223	0.074	0.065	0.081	0.071	0.071	0.954	0.960
NPV	0.771	0.686	0.779	0.696	0.046	0.071	0.050	0.075	0.075	0.931	0.931
PPV	0.557	0.519	0.565	0.526	0.042	0.038	0.044	0.040	0.040	0.958	0.960

n=1000											
Land	Cox	Land	Cox	Land	Cox	Land	Cox	Land	Cox	Land	Cox
AUC	0.741	0.668	0.744	0.671	0.023	0.030	0.022	0.029	0.029	0.942	0.937
Spec	0.319	0.207	0.329	0.216	0.055	0.046	0.057	0.049	0.049	0.958	0.958
NPV	0.771	0.686	0.775	0.691	0.034	0.051	0.035	0.052	0.052	0.922	0.923
PPV	0.557	0.519	0.561	0.522	0.030	0.026	0.031	0.027	0.027	0.957	0.963

Proposed prediction performance measure estimates using the landmark method (Land) vs. the global Cox method (Cox) from setting(ii) when the global Cox model holds, with corresponding empirical standard errors (ESE), average of the standard error estimates from the perturbation-resampling method (ASE), and empirical coverage levels (Coverage). The cutoff value c was chosen such that the sensitivity level is 0.90.

Table 2

n=200										
	Truth		Average		ESE		ASE		Coverage	
	Land	Cox	Land	Cox	Land	Cox	Land	Cox	Land	Cox
AUC	0.865	0.865	0.870	0.869	0.030	0.030	0.029	0.030	0.910	0.915
Spec	0.588	0.588	0.601	0.599	0.102	0.102	0.112	0.112	0.931	0.935
NPV	0.907	0.907	0.914	0.914	0.021	0.021	0.024	0.025	0.921	0.927
PPV	0.568	0.568	0.584	0.582	0.074	0.074	0.080	0.080	0.951	0.949

n=500										
AUC	0.865	0.865	0.867	0.866	0.019	0.019	0.019	0.019	0.946	0.947
Spec	0.588	0.588	0.591	0.591	0.066	0.067	0.073	0.073	0.956	0.960
NPV	0.907	0.907	0.910	0.910	0.013	0.013	0.015	0.015	0.929	0.940
PPV	0.568	0.568	0.573	0.573	0.047	0.047	0.051	0.051	0.959	0.961

n=1000										
AUC	0.865	0.865	0.866	0.866	0.013	0.013	0.014	0.014	0.959	0.960
Spec	0.588	0.588	0.588	0.587	0.046	0.046	0.051	0.051	0.953	0.955
NPV	0.907	0.907	0.908	0.908	0.010	0.009	0.010	0.010	0.947	0.950
PPV	0.568	0.568	0.569	0.569	0.033	0.033	0.036	0.036	0.965	0.956

Table 3

Regression parameter estimates based on the landmark procedure where T_S represents time to metastasis and T_L represents time to death in the breast cancer dataset with $t_0 = 3$ and $\tau = 10$ with standard error (SE) estimates based on the proposed resampling methods.

	(a) with gene score		(b) without gene score	
	Estimate	SE	Estimate	SE
gene score	0.962	0.279	-	-
age	-0.052	0.037	-0.042	0.035
grade	0.041	0.277	0.445	0.244
size	0.017	0.024	0.015	0.022
ER status	-0.128	0.502	-0.705	0.468
posLN	-0.040	0.397	0.161	0.397

Table 4

Apparent (AP) and cross-validated (CV) accuracy measure estimates of the prognosis rules based on the landmark procedure in predicting $I(R_L^0 \leq \tau) | T_S > t_0$, where T_S represents time to metastasis and T_L represents time to death in the breast cancer dataset with $t_0 = 3$ and $\tau = 10$ with standard error (SE) estimates based on the proposed resampling methods. The cutoff value c was chosen such that the sensitivity level is 0.90.

	(a) with gene score		(b) without gene score	
	Estimate	SE	Estimate	SE
	AP	CV	AP	CV
AUC	0.805	0.694	0.736	0.621
Spec	0.607	0.434	0.464	0.374
NPV	0.969	0.945	0.954	0.951
PPV	0.392	0.333	0.318	0.294