Measuring the temporal prognostic utility of a baseline risk score

Sean M Devlin, Mithat Gönen, and Glenn Heller

Department of Epidemiology and Biostatistics

Memorial Sloan Kettering Cancer Center

New York, NY

April 6, 2020

Abstract

In the time-to-event setting, the concordance probability assesses the relative level of agreement between a model-based risk score and the survival time of a patient. While it provides a measure of discrimination over the entire follow-up period of a study, the probability does not provide information on the longitudinal durability of a baseline risk score. It is possible that a baseline risk model is able to segregate short-term from long-term survivors but unable to maintain its discriminatory strength later in the follow-up period. As a consequence, this would motivate clinicians to re-evaluate the risk score longitudinally. This longitudinal re-evaluation may not, however, be feasible in many scenarios since a single baseline evaluation may be the only data collectible due to treatment or other clinical or ethical reasons. In these scenarios, an attenuation of the discriminatory power of the patient risk score over time would indicate decreased clinical utility and call into question whether this score should remain a prognostic tool at later time points. Working within the concordance probability paradigm, we propose a method to address this clinical scenario and evaluate the discriminatory power of a baseline derived risk score over time. The methodology is illustrated with two examples: a baseline risk score in colorectal cancer defined at the time of tumor resection, and for circulating tumor cells in metastatic prostate cancer.

1 Introduction

Biomarker discovery has become an integral part of clinical research and the number of potential biomarkers has grown exponentially with the advent of molecular technologies. These developments have led to the need for specialized statistical methods focusing on the evaluation of novel markers and their utility within statistical models for the prediction of clinical outcomes.

One common tool to assess the performance of a statistical model is the concordance probability. For survival models, the concordance probability assesses the relative level of agreement between the model-based risk scores and the survival times of patients. An estimate of the concordance probability, however, does not provide information on the durability of a baseline risk score over time. This is a critical component when integrating new biomarkers into clinical care and patient surveillance.

If a model is able to segregate short-term from long-term survivors but unable to maintain its discriminatory strength later in the follow-up period, this would motivate re-evaluation of the risk score longitudinally. This re-evaluation may not, however, be feasible for all biomarkers contained in the risk score; a single baseline measurement may be the only possibility. For example, in patients whose tumors are completely resected, it is not possible to obtain longitudinal measurements of tumor biomarkers, which require pathological evaluation. Even when the tumor is not resected it may be unethical to obtain post-baseline biopsies. Alternatively, cost considerations of novel technologies, such as new imaging biomarkers, may render repeated measurements impractical. In these scenarios, an attenuation of the discriminatory power of the patient risk score over time would indicate decreased model clinical utility and call into question whether it should remain part of the diagnostician's armamentarium. The methodology proposed in this manuscript addresses this clinical scenario. Working within the concordance probability paradigm, we propose a method to evaluate the discriminatory power over time of a baseline or peri-treatment derived risk score.

The proposed methodology outlined in this manuscript differs from the dynamic landmarking approach (for example, see Van Houwelingen 2007) as we focus on how the discriminatory ability of a baseline risk score changes over time instead of assessing how hazard ratio estimates of a biomarker change through repeated landmarks. The considered clinical setting is analogous to Heagerty and Zheng (2005) and Parast and Cai (2013) as both evaluate discrimination of a baseline risk score either at a point in time or over a time period. We review both approaches in the following section. The contribution of this manuscript is how the discrimination over time can be derived as a model-based concordance probability estimate, and we highlight scenarios when such an estimation procedure is advantageous.

The methodology is illustrated in two examples. The first example estimates the discrimination of a baseline risk score evaluated at the time of tumor resection in colorectal cancer. The risk score incorporates information of the surgically removed tumor, and therefore cannot be longitudinally reassessed. The second example evaluates the clinical utility of a relatively new diagnostic biomarker, circulating tumor cells (CTC), in metastatic prostate cancer research. The introduction of CTC to assess prognosis for patients with metastatic prostate cancer has improved the magnitude of the estimated concordance probability (Scher et al. 2009; Heller et al. 2017). However, due to cost considerations, CTC are not used throughout patient follow-up to monitor disease progression. While baseline CTC has strong overall discriminatory power, it is unclear whether the baseline CTC maintains its strength uniformly throughout the follow-up period, and hence whether the early evaluation is sufficient for clinical decisions made throughout the follow-up period.

The article continues with Section 2 providing concordance probability definitions, modeling approaches, and the extension of the methodology to the clinical scenario under consideration. Section 3 presents various simulation scenarios to assess the operating characteristics of the proposed approach. Section 4 illustrates the use of the methodology by applying it to the evaluation of a surgical risk score in colorectal cancer and CTCs for metastatic prostate cancer. The article concludes in Section 5 with a discussion of the key results and implications for future use.

2 Methods

For a survival outcome, the concordance probability is defined as

$$\Pr[\boldsymbol{\beta}^T \boldsymbol{X}_2 > \boldsymbol{\beta}^T \boldsymbol{X}_1 | T_1 > T_2, T_2 < \tau], \tag{1}$$

where $\boldsymbol{\beta}^T \boldsymbol{X}$ represents a model-based risk score composed of a linear combination of risk factors $\boldsymbol{X}$, patient survival time T, and τ denotes the maximum follow-up time under consideration. The concordance probability ranges between 0.5 and 1.0, where 1.0 represents perfect concordance between the risk score and the survival time and 0.5 indicates no relationship between them. In this paper, it is assumed that the risk score is derived using the proportional hazards model

$$h(t|\boldsymbol{x}) = h_0(t) \exp[\boldsymbol{\beta}^T \boldsymbol{x}],$$

although other models may be utilized, such as the proportional odds model (see Zhang and Shao, 2018).

The most frequently applied estimate of the concordance probability is the c-index

$$C_n^*(\hat{\boldsymbol{\beta}}; \tau) = \frac{\sum_i \sum_{j \neq i} \delta_j I(y_j < y_i, y_j < \tau) I(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i < \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_j)}{\sum_i \sum_{j \neq i} \delta_j I(y_j < y_i, y_j < \tau)},$$

where for each individual, y represents the minimum of the survival time and censoring time, δ is an indicator function denoting whether the observed time is the survival time, and $\hat{\boldsymbol{\beta}}$ is the estimated regression coefficient from the proportional hazards model (Harrell et al. 1996; Pencina and D'Agostino 2004) . This concordance measure is robust to model misspecification in that it does not require a properly specified model, used to create the risk score, for its interpretation. The c-index, however, does not consistently estimate the concordance probability (1). As a result, the weighted c-index was developed

$$C_n(\hat{\boldsymbol{\beta}};\tau) = \frac{\sum_i \sum_{j\neq i} \delta_j I(y_j < y_i, y_j < \tau) I(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i < \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_j)\{\hat{G}(y_j)\}^{-2}}{\sum_i \sum_{j\neq i} \delta_j I(y_j < y_i, y_j < \tau)\{\hat{G}(y_j)\}^{-2},}.$$

where G is the survival function of the underlying censoring times. (Uno et al., 2011) The weighted c-index is consistent and asymptotically normal, but utilizes inverse probability censoring weights (IPCW), and as a result, may be sensitive to large failure times. In addition, if the censoring distribution is a function of the covariates, then IPCW may require a model based conditional survival function to specify this relationship (Gerds et al. 2013).

An alternative measure of concordance with survival data, under proportional hazards, is the concordance probability estimate (Gönen and Heller 2005). The concordance probability estimate (CPE) defined under a maximum follow-up time (τ) is

$$K_n(\hat{\boldsymbol{\beta}};\tau) =$$

$$\frac{\sum_i \sum_{j\neq i} I(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i < \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_j) \left[1 + \exp\{\hat{\boldsymbol{\beta}}^T (\boldsymbol{x}_i - \boldsymbol{x}_j)\}\right]^{-1} \left[1 - \hat{S}(\tau|\boldsymbol{x}_i)\hat{S}(\tau|\boldsymbol{x}_j)\right]}{0.5 \ \times \ \sum_i \sum_{j\neq i} \left[1 - \hat{S}(\tau|\boldsymbol{x}_i)\hat{S}(\tau|\boldsymbol{x}_j)\right]},$$

where $\hat{S}(t|\boldsymbol{x})$ represents the estimated survival function from the Cox model.

When the proportional hazards assumption is correct and the risk score is continuous, the CPE consistently estimates the concordance probability (1). Estimation of the concordance probability when the risk score is discrete has additionally been developed (Heller and Mo, 2016). The consistency of the CPE is unaffected by the conditionally independent censoring distribution and does not require an external inverse probability weight to derive desirable asymptotic properties. The CPE, however, does require a properly specified proportional hazards model for its consistency and interpretation.

The concordance probability parameter (1) is an evaluation of the baseline risk score ordering relative to the ordered survival times. To assess the prognostic utility of the risk scores for patients who survive beyond time s, when the risk score has not been updated, the concordance probability is modified so that the evaluation occurs in a bounded interval

$$\Pr[\boldsymbol{\beta}^T \boldsymbol{X}_2 > \boldsymbol{\beta}^T \boldsymbol{X}_1 | T_1 > T_2, s < T_2 < \tau].$$

The weighted c-index estimate of the concordance probability, evaluated within the interval (s, τ), is

$$C_n(\hat{\boldsymbol{\beta}}; s, \tau) = \frac{\sum_i \sum_{j \neq i} \delta_j I(y_j < y_i, s < y_j < \tau) I(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i < \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_j) \{\hat{G}(y_j)\}^{-2}}{\sum_i \sum_{j \neq i} \delta_j I(y_j < y_i, s < y_j < \tau) \{\hat{G}(y_j)\}^{-2}}. \tag{2}$$

This derivation is analogous to Parast and Cai (2013) where the performance of a risk score is evaluated at a specific time point instead of a time interval.

The corresponding CPE is $K_n(\hat{\boldsymbol{\beta}}; s, \tau) =$

$$\frac{\sum_i \sum_{j \neq i} I(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i < \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_j) \left[1 + \exp\{\hat{\boldsymbol{\beta}}^T (\boldsymbol{x}_i - \boldsymbol{x}_j)\}\right]^{-1} \left[\hat{S}(s|\boldsymbol{x}_i)\hat{S}(s|\boldsymbol{x}_j) - \hat{S}(\tau|\boldsymbol{x}_i)\hat{S}(\tau|\boldsymbol{x}_j)\right]}{0.5 \ \times \ \sum_i \sum_{j \neq i} \left[\hat{S}(s|\boldsymbol{x}_i)\hat{S}(s|\boldsymbol{x}_j) - \hat{S}(\tau|\boldsymbol{x}_i)\hat{S}(\tau|\boldsymbol{x}_j)\right]} \tag{3}$$

A derivation of this estimate and its asymptotic normal distribution is provided in Supplemental Section A.1.

An alternative approach to the measurement of post-baseline discrimination is the area under the curve (AUC) estimate developed by Heagerty and Zheng (2005). The authors estimated the local AUC under proportional hazards using the model-derived estimates of sensitivity and specificity proposed by Xu and O'Quigley (2005). This estimate provides a measure of discrimination at a given point during follow-up, in contrast to our approach, which is based on a follow-up time interval. However, using integration of these estimates over time, the concordance probability can be estimated in the interval (s, τ) using

$$A_n(\hat{\boldsymbol{\beta}}; s, \tau) = \int_s^\tau \widehat{\mathrm{AUC}}(t)\ \hat{w}(t)\ dt, \tag{4}$$

where the weight function $w(t)$ is estimated using a discrete approximation to the survival time density function, via the jumps in the Kaplan-Meier estimate. A necessary condition for this estimation procedure is that the censoring time is independent of the survival time and the covariate.

Throughout the manuscript, equations (2), (3), and (4) are referred to as the weighted c-index, CPE, and the integrated AUC, respectively.

3 Simulations

3.1 Independent Censoring Under Proportional Hazards

Two simulation scenarios evaluated the performance of the weighted c-index, integrated AUC, and CPE under proportional hazards when the censoring distribution is independent of the survival time and the biomarker. Under the first scenario, the true concordance remained high and near constant as the distance away from baseline s increased and the evaluation interval moved away from baseline. Under the second scenario, the true concordance decreased from high to a level that indicates that while the marker is associated with survival, its usefulness for risk stratification is unclear. The value of τ was fixed in each scenario based on the underlying censoring distribution, as described below.

Data under both scenarios were generated from a proportional hazards relationship using the Weibull regression model $t_i = \exp\{1 - 0.2x_i\} \times \epsilon_i$, and ϵ_i were generated from independent and identically distributed Weibull random variables with scale parameter 1 and shape parameter 4, which produced a true concordance probability from baseline risk scores equal to approximately 0.70. The distribution of biomarker x_i differed across the two simulation scenarios: the first scenario used a standard normal distribution, N(0,1), while the second used a log-normal distribution, LN(μ=0, σ=1). The latter distribution is evaluated as its longer tail enables greater variability in the risk scores, producing a greater range in the true concordance as s increases.

Censoring times were generated from a uniform distribution Un(0, b), and b varied to correspond to censoring proportions of 25%, 50% and 75% in both scenarios. The values of b were 3.18, 5.05, and 10.1 for the normal marker distribution. In this scenario, $\tau = 3.18$ and s varied across four equally spaced values between 0 and $\frac{3}{5} \times \tau$. When x_i was generated from a log-normal distribution, b=2.36, 3.75, and 7.5, and $\tau = 2.36$. If τ exceeded the maximum of the simulated survival times within an iteration, it was truncated to the observed maximal event time.

The relative performance of the weighted c-index, integrated AUC, and CPE were compared based on the average bias and the relative efficiency. In addition, the estimated standard error of these estimates were evaluated relative to their simulation standard errors. The CPE and the integrated AUC utilized bootstrap resampling to estimate the standard error, while the weighted c-index used a perturbation-resampling method as previously described (Uno et al., 2011).

The sample size for all simulations was 300, and the simulation results were averaged across 2,000 iterations. The values for the true underlying concordance probability for various values of s were approximated by the average of 2,000 iterations with 3,000 uncensored observations.

The first row of Figure 1 provides a graphical summary of the results for the simulation using a standard normal marker distribution. The results are also presented in Supplemental Table S1. All three approaches were minimally biased. The average estimated standard error aligned closely with the simulation standard error for all estimates in Table S1, except for the weighted c-index with 75% censoring. In any given scenario, the standard error for the model-based CPE and integrated AUC were lower than the weighted c-index; this translated into improved efficiency of CPE over the weighted c-index. There were marginal gains in efficiency of CPE over the integrated AUC.

The results for the log-normal marker distribution are provided in the second row of Figure 1 and Table S2. All approaches were again minimally biased across the various s values; the bias increased slightly for the higher censoring proportions. There were larger gains in relative efficiency of CPE over the weighted c-index due to the smaller standard error for the CPE. This was most notable for higher degrees of censoring and larger values of s. While smaller than the gains for CPE compared to the weighted c-index, the relative efficacy improvement of CPE over the integrated AUC ranged from 21% to 61% across the values of s.

3.2 Conditionally Independent Censoring Under Proportional Hazards

Data were generated under two scenarios where the censoring distribution depends on the value of the biomarker, inducing survival times conditionally independent of censoring times. Data were again generated using the Weibull regression model $t_i = \exp\{1 - 0.2x_i\} \times \epsilon_i$ and ϵ_i were generated from Weibull random variables with scale parameter 1 and shape parameter 4. The distribution of the biomarker x_i was the same as the previous scenarios: standard normal for the first scenario and log-normal for the second scenario. However, the censoring times were now generated as $\exp\{a - 0.2x_i\} \times \epsilon_i$, where a was selected to achieve censoring proportions of 25%, 50% and 75% and ϵ_i was similarly from a Weibull distribution with scale parameter 1 and shape parameter 4. In the first scenario, under a normal marker distribution, the values of a were 1.274, 1.0, and 0.726. Under the log-normal marker distribution, the values were 1.275, 1.0, and 0.726.

Results are shown in Figure 2 and tables S3 and S4. As expected, the average bias for the weighted c-index and integrated AUC were substantial, particularly under higher degrees of censoring and under the lognormal marker distribution. CPE remains unbiased under conditionally independent censoring. The large average bias difference across the methods translated to considerable gains in efficiency for CPE compared to either the weighted c-index or integrated AUC.

3.3 Non-proportional Hazards

Two simulation scenarios compared the relative performance of the three methods under non-proportional hazards. The data generation process for both simulations used a standard normal distribution for x_i and a Weibull regression model $t_i = \exp\{1 - 0.6x_i\} \times \epsilon_i$. However, now ϵ_i were generated from Weibull random variables with scale parameter of 1 and shape parameter that depends on x_i to induce non-proportional hazards. In the first simulation scenario, designed to represent a minor deviation from proportional hazards, the shape was equal to $1 - 0.1x_i$. In the second scenario, representing a larger deviation, the shape was equal to $1 - 0.175x_i$.

For each of the 2,000 simulated data sets with 300 observations, the deviation from the proportional hazards assumption was evaluated using a score test of a time-varying interaction with x_i (Grambsch and Therneau 1994). In the first scenario, the test was rejected in 18%, 14%, and 9% of the simulations when the censoring proportion was 25%, 50%, and 75%, respectively. Without censoring, the test is rejected in 22% of the simulations. In the second scenario, the test was rejected in 52%, 39%, and 22%, respectively, of the simulations; without censoring, the rejection rate was 62%.

Censoring times were similarly generated from a uniform distribution Un(0, b). The values of b were 1.504, 4.5, and 12.04 in the first scenario and 1.51, 4.6, and 12.25 in the second scenario.

Figure 3 along with tables S5 and S6 provide the average bias, standard error, and relative efficiency for the two scenarios. As anticipated the biases for CPE and integrated AUC were larger for the simulation with a larger deviation from proportional hazards. For both scenarios, the bias decreased as the censoring proportion increased. The weighted c-index maintained a minimal bias in both scenarios.

While having a large bias, the estimated standard error for CPE was notably smaller than the weighted c-index across all scenarios. This translated to a gain in relative efficiency when either s moved away from 0, where the weighted c-index has less information, or the censoring rate increased. There were minimal gains in efficiency for CPE compared to the integrated AUC.

3.4 Summary

Under the evaluated scenarios, CPE performed as well or better than the integrated AUC. The largest advantage to CPE was observed when the marker followed a long tailed distribution and when the survival times were conditionally independent of the censoring times. Therefore, CPE is advantageous to the integrated AUC as it is valid under both independent censoring and conditionally independent censoring.

The improvement in efficiency for the model based CPE relative to the nonparametric weighted c-index is expected, but requires that the proportional hazards specification is correct. The gain in precision when using CPE translates in practice into tighter 95% confidence bounds, so CPE will provide clearer evidence of the strength or weakness of a model.

Finally, CPE may be considerably biased under non-proportional hazards. However, the largest biases were observed under the scenarios when the score test had moderate-to-high power to reject the hypothesis that there is no time-varying interaction with x_i. This highlights the importance of investigating the proportional hazards assumption prior to any data analysis.

Therefore, if the analyst deems a data set to be well approximated by proportional hazards, CPE would be an efficient and unbiased approach to estimate the temporal discrimination of a baseline risk score.

4 Data Analysis

4.1 Analysis of a Surgical Risk-score for Colorectal Cancer

Primary treatment for localized colorectal cancer is surgical resection of the tumor. In this example we develop a risk score for 1,364 colorectal cancer patients at the time of the surgical resection of the tumor. There are a number of important applications of this risk score. Many of these patients recur and there is some evidence that post-operative (adjuvant) chemotherapy is beneficial in reducing recurrence rates. This needs to be balanced with the morbidity and costs of treatment. An accurate risk score helps physicians select patients for adjuvant treatment. Some patients, however, are unable to start chemotherapy shortly after resection, and hence it is important to determine whether the baseline risk score remains informative 6-12 months post surgery. Another application for the baseline risk score is to determine the intensity of surveillance. Patients at high risk are evaluated with imaging, lab tests and clinical exams more often after the resection when compared with low risk patients. Whether a particular surveillance schema, formulated at baseline, is appropriate, is again a function of the accuracy of the risk score over time.

The risk score was a weighted combination of the characteristics of the resected tumor in addition to the patient's age and comorbidity score. The tumor characteristics included the tumor size (T-stage), whether the cancer cells were well or poorly differentiated, and the number of positive lymph nodes removed during surgery. The weighted combination for the risk score was generated under a proportional hazards model. The discriminatory strength of the baseline risk score was evaluated over different time intervals. The lower bound s of the time interval varied from 0 to 36 months. The upper bound τ was fixed at 50 months; the estimated survival probability at this time was 0.66 (95% CI: 0.63-0.68).

A total of 139 patients died by 50 months following surgery.

To examine the appropriateness of the proposed estimation process to these data, the proportional hazards assumption was examined. In Supplemental Figure S1, the loess curve of the scaled Schoenfeld residuals + $\hat{\beta}$ for each of the covariates was approximately constant with respect to time. In addition, the global test of the proportional hazards assumption had a corresponding p-value of 0.945; the p-value for each individual covariate score test is shown in Supplemental Figure S1. Collectively, this suggests the data may be suitably modeled by proportional hazards, leading to the use of the CPE to evaluate the temporal prognostic utility of the surgical risk score.

The CPEs for the different values of s are provided in Figure 4. When the lower bound started from the baseline at $s = 0$, indicating that the entire follow-up time up until τ is used to compute the CPE, the estimated concordance was relatively high at 0.671 (95% CI: 0.650-0.692). Further, the estimated concordance only minimally decreased as s moved away from 0. The estimated concordance was 0.666 (95% CI:0.646-0.686) when estimated between 12 and 50 months, and 0.653 (0.632-0.674) between 36 and 50 months.

These results suggest that the risk score defined at the time of surgical resection of the tumor is able to retain its discriminatory ability following the baseline evaluation. At 12 months following resection, there has been only a minimal decrease of 0.005 in the estimated concordance probability; therefore, the risk score continues to segregate those at highest risk of death at this post-surgery landmark time. This may be useful for patient surveillance in the year following surgery even though the risk score cannot be updated over time.

4.2 Analysis of CTC for Metastatic Prostate Cancer

The biomarker circulating tumor cells (CTC) was evaluated on 332 patients with metastatic castration-resistant prostate cancer at the time of treatment. While the discriminatory power of the baseline and peritreatment CTC have been previously described, the duration of time the baseline marker remains discriminatory for survival is unknown. This information will inform the duration of time the CTC evaluation can be used as part of clinical prognostication; currently the cost associated with the CTC assay precludes long-term follow-up evaluation in many settings.

The concordance probabilities for baseline CTC were evaluated over the same intervals as the previous example. The lower bound s of the time interval varied from 0 to 36 months and the upper bound τ was fixed at 50 months. The estimated survival probability at 50 months was 0.03 (95% CI: 0.01-0.09). A total of 243 patients died during the follow-up period. Baseline CTC was log-transformed in the analysis.

The graphical examination of the proportional hazards assumption (Supplemental Figure S2), and the score test for a time-varying interaction with CTC (p-value = 0.81), suggested that the data may be suitably modeled by proportional hazards.

Figure 5(A) provides the CPEs for CTC. When the lower bound started from the baseline at $s = 0$, indicating that the entire follow-up time up until τ is used to compute the CPE, CTC has a high discriminatory ability: the estimated concordance for CTC was 0.693 (95% CI: 0.659-0.726). However, as the survival time conditioning set moved away from baseline, the estimated concordance for baseline CTC dropped. For patients who survived at least 36 months, the estimated concordance for CTC was 0.633 (0.544-0.722).

These results illustrate that while baseline CTC has strong overall discrimination, it loses its discriminatory power for patients surviving later in the follow up period. The implication is that the removal of the earliest failures attenuates the CPE for baseline CTC. To further demonstrate this finding, the CPE interval measure is reoriented to assess the CPE in the follow-up interval $(0, t)$, where $0 < t < 24$ months. This conjecture is corroborated in Figure 5(B), where by retaining the earliest failures in the conditioning set $(0, t)$, the CPE remains stable at approximately 0.70 for a follow-up period out to 24 months.

Therefore, while baseline CTC may have a high degree of prognostic utility when evaluated, the drop in discrimination over time casts doubt on whether the baseline marker should be used for disease management 12-24 months following evaluation. Due to cost considerations, however, early evaluation is its primary usage.

5 Discussion

Commonly applied predictive accuracy methods for survival models include measures of calibration and discrimination. Calibration computes how close the model predictions are to the true survival times while discrimination calculates the models ability to distinguish between long-term and short-term survivors. In this article we focus on discrimination and present a novel method to evaluate the durability in the concordance probability over time for a baseline risk score in the setting of a censored time-to-event outcome. A decrease in the discriminatory power may indicate that the baseline biomarker has diminishing clinical utility over time. This is an apposite consideration in oncology, as there are many new prognostic biomarkers being integrated into cancer clinical care.

We illustrated the proposed methodology with a risk score developed at the time of surgical resection in colorectal cancer that retains its discrimination over a year following surgery, and for a new marker in metastatic prostate cancer with high yet decreasing discrimination.

In addition to the three methods evaluated in this manuscript, we additionally considered a landmark-based CPE approach, analogous to the weighted c-index in Equation (2). In this approach, the concordance was estimated just among individuals who remained at risk at the landmark time (denoted s in Section 2). While this landmark-based estimate remained unbiased across the proportional hazards scenarios, higher variance was observed as the landmark time s moved away from baseline, where there were fewer patients available for estimation. In contrast, the proposed CPE measure in Equation (3) was able to more efficiently borrow information to estimate the concordance probability for larger values of s in the bounded interval $T_1 > T_2, s < T_2 < \tau$. As the landmark-based CPE ~~had minimal gains~~ over the main three methods, it was omitted from the manuscript.

As mentioned throughout the manuscript, the estimation framework proposed here relies on the assumption that the data are well approximated by a proportional hazards model. Blanche, Kattan, and Gerds (2019) note that ~~in general,~~ concordance estimates are not proper scores and hence their population values are not maximized at the true regression parameter values (β). The implication is that one can choose an unorthodox estimator for β that would create a more optimistic value for the concordance probability estimate. The proper scoring issue aligns with our warning that suitable diagnostics need to be employed to provide assurance that the proportional hazards model specification is appropriate, leading to the use of conventional maximum partial likelihood estimates

$(\hat{beta})$ to estimate the concordance probability.

Under the proportional hazards specification, there can be large efficiency gains, as shown in the simulation scenarios in Section 3, from using a concordance measure derived directly from the properly specified model. However, it is important to investigate the proportional hazards assumption prior to using the proposed methods. In the event this assumption does not hold, the estimated CPE may be biased. When either proportional hazards is violated or the clinical interest is in t-year survival, alternative modeling strategies, such as a nonparametric estimate of $\widehat{\mathrm{AUC}}(t)$ in Equation (4), should be considered and further evaluated.

While the current methodology does not accommodate non-proportional hazards, extensions of the model may be able to accommodate certain deviations. A direct extension occurs for the proportional odds model, where the work of Zhang and Shao (2018) can be extended to compute the concordance probability estimate in the interval (s, τ). In addition, future work will investigate local CPE estimation approaches under time-varying coefficient models, where the concordance probability changes over time due to a post-baseline change in the covariate effect. Alternatively, when there are multiple biomarkers and a subset do not satisfy proportional hazards, a stratified proportional hazards model can be implemented; the strata could be defined based on a categorization of the biomarkers not satisfying the proportional hazards assumption. Additional research is warranted in these settings for methodology development and evaluation of the corresponding operating characteristics.

Acknowledgements

This work was supported by NIH Grants P30CA008748 and R01CA207220.

References

[1] Blanche P., Kattan M.W., and Gerds T.A. (2019). The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics* **20(2)**, 347–357.

[2] Gerds T.A., Kattan M.W., Schumacher M., and Yu C. (2013). Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine* **32(13)**, 2173–2184.

[3] Gönen M., and Heller G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* **92(4)**, 965–970.

[4] Grambsch P.M., and Therneau T.M. (1994).Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika* **81(3)**, 515–526.

[5] Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Multi-variable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361–387.

[6] Heagerty P.J., and Zheng Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61(1)**, 92-105.

[7] Heller G, and Mo Q. (2016). Estimating the concordance probability in a survival analysis with a discrete number of risk groups. *Lifetime Data Analysis* **22(2)**, 263-279.

[8] Heller G, Fizazi K, McCormack RT, Molina A, MacLean D, Webb IJ, Saad F, de Bono JS, and Scher HI. (2017). The added value of circulating tumor cell enumeration to standard markers in assessing prognosis in a metastatic castration-resistant prostate cancer population.*Clinical Cancer Research* **23(8)**, 1967-1973.

[9] Parast L., and Cai T. (2013). Landmark risk prediction of residual life for breast cancer survival. *Stat Med.* **10(32)**,3459-3471.

[10] Pencina M.J.,and D'Agostino R.B. (2004).Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation.*Statistics in Medicine* **23(13)**, 2109-2123.

[11] Scher HI, Jia X, de Bono JS, Fleisher M, Pienta KJ, Raghavan D, and Heller G. (2009). Circulating tumor cells as prognostic markers in progressive, castration-resistant prostate cancer: a reanalysis of IMMC38 trial data. *Lancet Oncology* **10 (3)**, 233-239.

[12] Uno H., Cai T., Pencina M.J., D'Agostino R.B., and Wei L.J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data.*Statistics in Medicine* **30(10)**, 1105–1117.

[13] Van Houwelingen HC.(2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics* **34(1)**, 70-85

[14] Xu, R. and O'Quigley, J. (2000). Proportional hazards estimate of the conditional survival function. *Journal of the Royal Statistical Society, Series B, Methodological* **62**, 667-680.

[15] Zhang Y., and Shao Y. (2018). Concordance measure and discriminatory accuracy in transformation cure models. *Biostatistics* **19(1)**, 14-26.

Figure 1: Independent censoring: estimated bias and relative efficiency for the four values of s under the normal and log-normal marker simulations. Data were generated under proportional hazards.

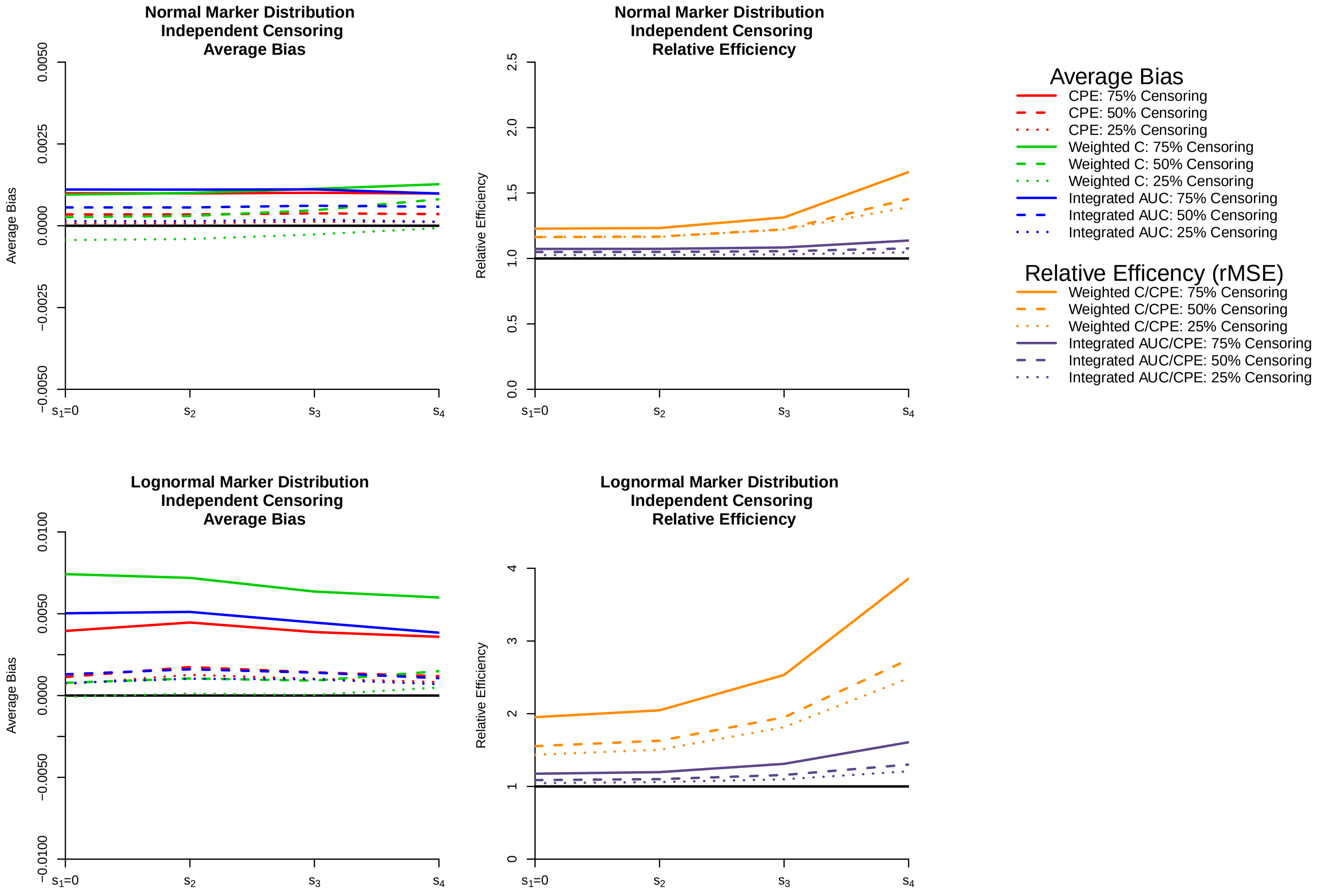


Figure 2: Independent censoring conditional on marker: estimated bias and relative efficiency for the four values of s under the normal and log-normal marker simulations. Data were generated under proportional hazards.

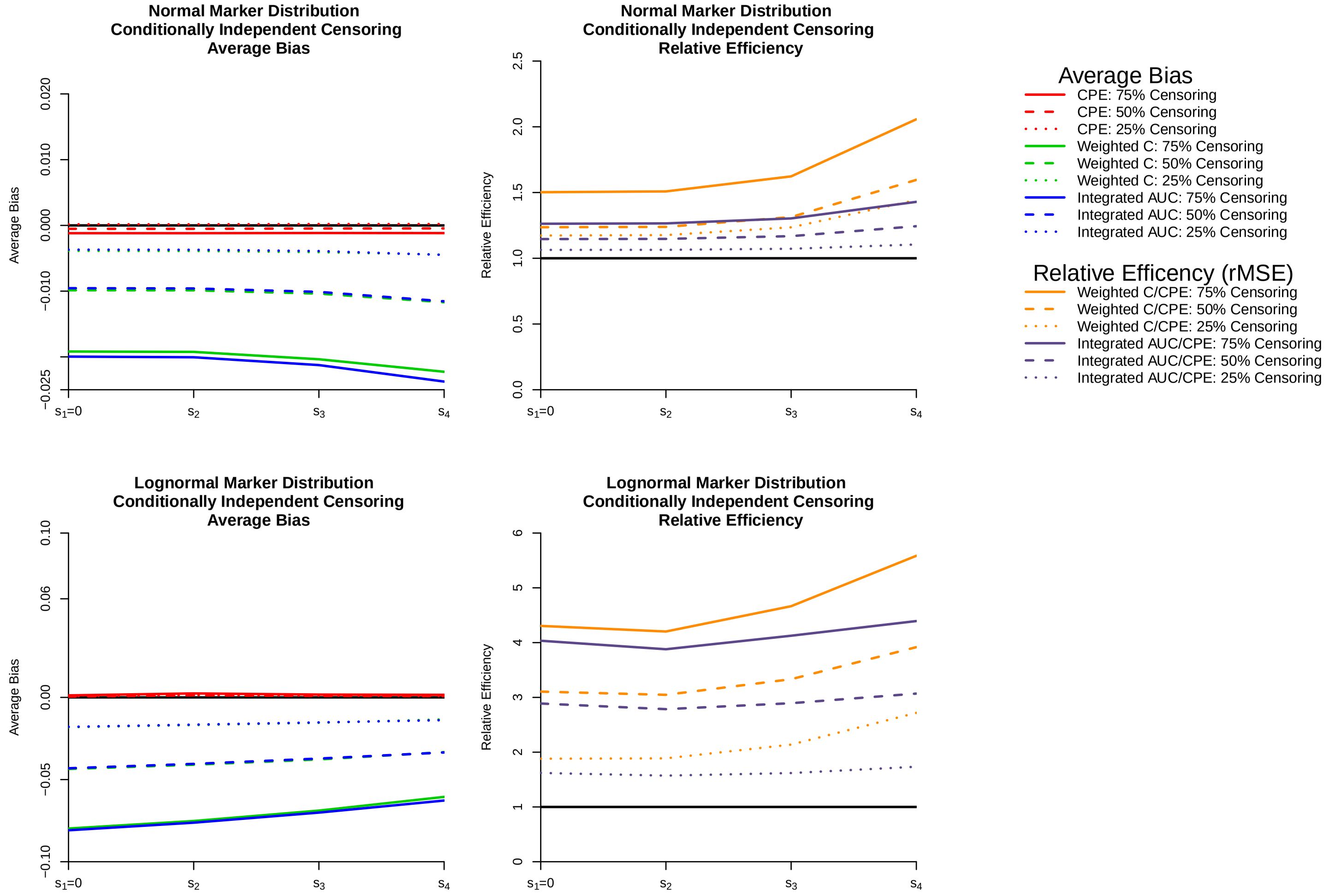


Figure 3: Non-proportional hazards: estimated bias and relative efficiency for the four values of s when censoring times were generated from a Weibull random variable with shape parameter equal to either $1 - 0.1X$ or $1 - 0.175X$, corresponding to a small and large deviation, respectively, from proportional hazards.

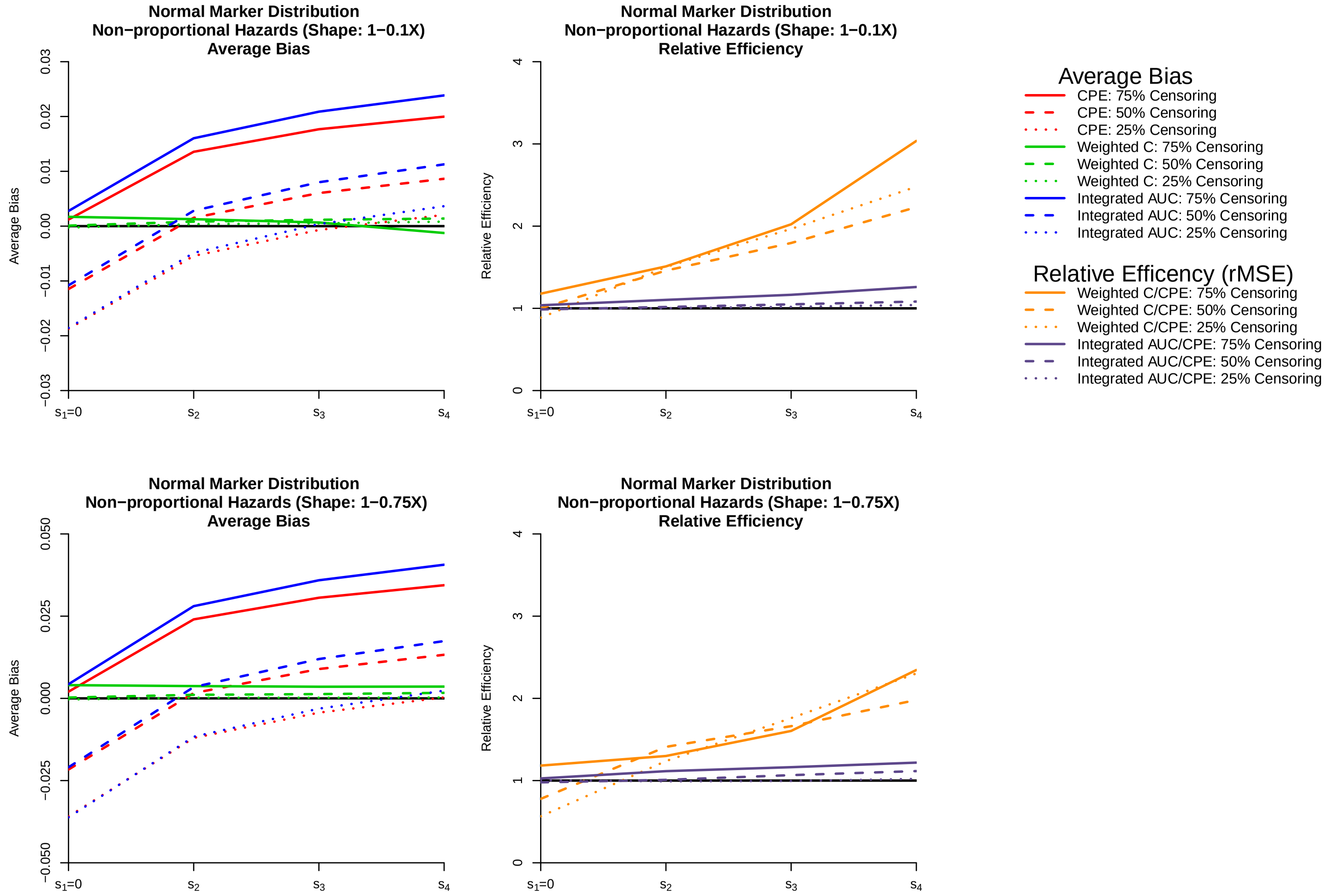


Figure 4: The estimated CPE in the time interval (s, τ) with a 95% confidence band for surgical-based risk score among 1,364 patients with colorectal cancer.

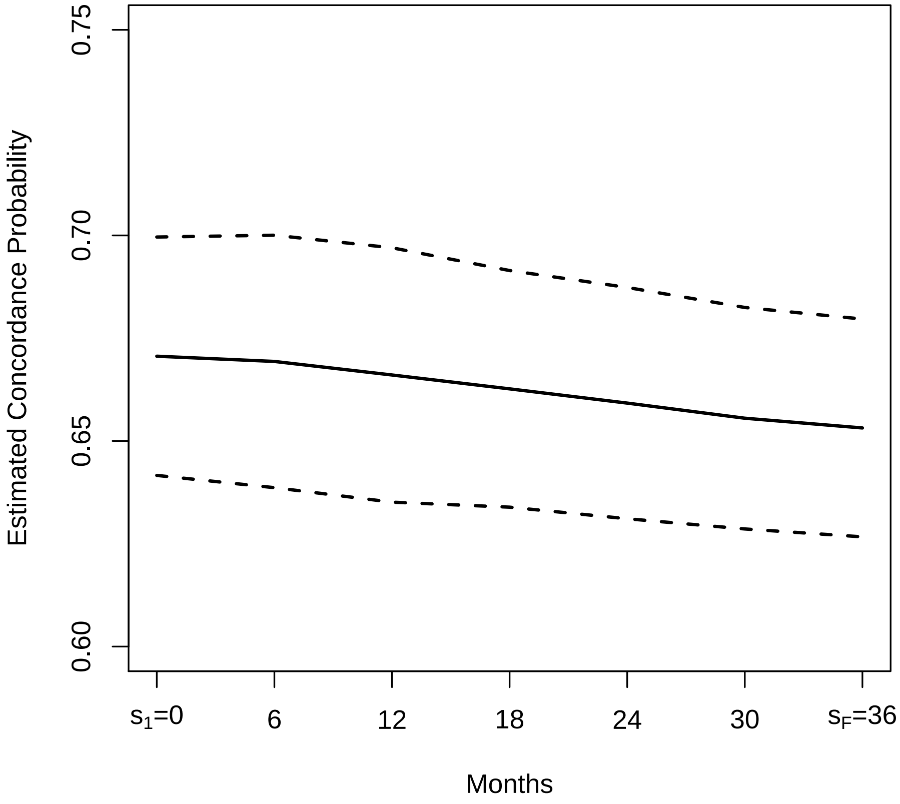
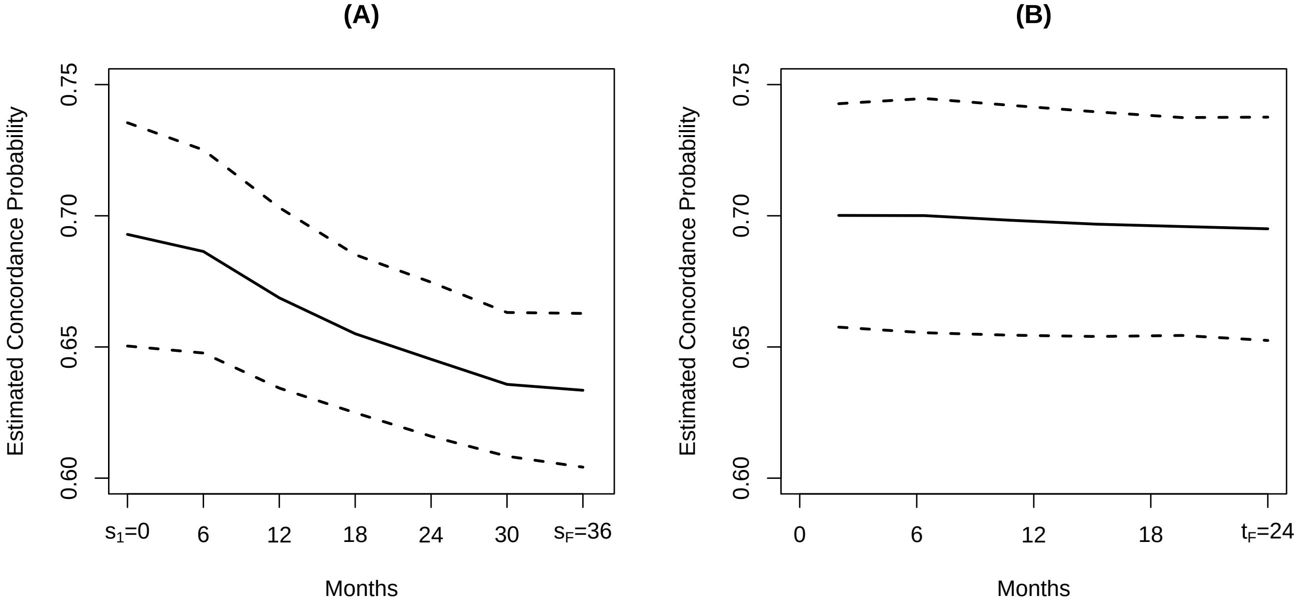


Figure 5: (A) The estimated CPE with a 95% confidence band in the time interval (s, τ) for CTC among 332 patients with metastatic castration-resistant prostate cancer. (B) The CPE values when estimated in the time interval $(0, t)$.



A Supplementary Web Material

A.1 *Asymptotic distribution of CPE(s,τ)*

Assumptions and notation

For each subject, denote the observed time Y as the minimum of the failure time (T) and censoring time (C), $\delta = I(T \leq C)$ is the censoring indicator, and $\boldsymbol{X}$ the covariate vector. It is assumed that the individual copies of the random vector $(T, C, \boldsymbol{X})$ are independent and identically distributed. In addition, define $N(t) = I(T \leq t, \delta = 1)$ as the counting process, $\psi(t) = I(Y \geq t)$ as the at risk process.

It is assumed that the failure times are generated from a proportional hazards model

$$h(t|\boldsymbol{x}) = h_0(t) \exp[\boldsymbol{\beta}^T \boldsymbol{x}].$$

The estimated conditional survival function from the proportional hazards model $\hat{S}(t|\boldsymbol{x})$, may be written in terms of the estimated baseline cumulative hazard function $\hat{H}_0(t)$ and the estimated regression coefficients $\hat{\boldsymbol{\beta}}$

$$\hat{S}(t|\boldsymbol{x}) = \exp\left[-\hat{H}_0(t) e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{x}}\right] \qquad \text{and} \qquad \hat{H}_0(t) = \int_{u<t} \frac{\sum_i dN_i(u)}{\sum_i \psi_i(u) \exp[\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i]}.$$

The CPE evaluated in the interval (s, τ) is written as

$$K_n(\hat{\boldsymbol{\beta}}, \hat{H}_0; s, \tau) = \frac{[n(n-1)]^{-1} \sum_i \sum_{j \neq i} w_{ij}(\hat{\boldsymbol{\beta}}, \hat{H}_0; s, \tau) I(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i < \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_j) \left[1 + \exp\{\hat{\boldsymbol{\beta}}^T (\boldsymbol{x}_i - \boldsymbol{x}_j)\}\right]^{-1}}{0.5 \times [n(n-1)]^{-1} \sum_i \sum_{j \neq i} w_{ij}(\hat{\boldsymbol{\beta}}, \hat{H}_0; s, \tau)}$$

where $w_{ij}(\hat{\boldsymbol{\beta}}, \hat{H}_0; s, \tau) = \hat{S}(s|\boldsymbol{x}_i)\hat{S}(s|\boldsymbol{x}_j) - \hat{S}(\tau|\boldsymbol{x}_i)\hat{S}(\tau|\boldsymbol{x}_j)$.

To show that the concordance probability $\Pr[\boldsymbol{\beta}^T\boldsymbol{X}_1 < \boldsymbol{\beta}^T\boldsymbol{X}_2 | T_1 > T_2, s < T_2 < \tau]$ may be consistently estimated by $K_n(\hat{\boldsymbol{\beta}}, \hat{H}_0; s, \tau)$, apply Bayes theorem to rewrite the concordance probability as

$\Pr[\boldsymbol{\beta}^T\boldsymbol{X}_1 < \boldsymbol{\beta}^T\boldsymbol{X}_2 | T_1 > T_2, s < T_2 < \tau] =$

$$\Pr[T_1 > T_2, s < T_2 < \tau | \boldsymbol{\beta}^T\boldsymbol{X}_1 < \boldsymbol{\beta}^T\boldsymbol{X}_2] \times \frac{\Pr[\boldsymbol{\beta}^T\boldsymbol{X}_1 < \boldsymbol{\beta}^T\boldsymbol{X}_2]}{\Pr[T_1 > T_2, s < T_2 < \tau]}$$

Evaluation of the three terms on the right hand side proceeds as follows.

Under the proportional hazards model,

$$\Pr[T_1 > T_2, s < T_2 < \tau | \boldsymbol{\beta}^T\boldsymbol{X}_1 < \boldsymbol{\beta}^T\boldsymbol{X}_2] =$$

$$2 \int \int_{\theta_1 < \theta_2} \tfrac{\theta_2}{\theta_1 + \theta_2} \left[\exp\left\{-\alpha_s(\theta_1 + \theta_2)\right\} - \exp\left\{-\alpha_\tau(\theta_1 + \theta_2)\right\}\right] dG(\theta_1) dG(\theta_2),$$

where $\theta_j = \exp[\boldsymbol{\beta}^T\boldsymbol{X}_j]$ and $\alpha_r = \int_{u=0}^{r} h_0(u) du$.

Under the assumption that the risk score is continuous, $\Pr[\boldsymbol{\beta}^T\boldsymbol{X}_1 < \boldsymbol{\beta}^T\boldsymbol{X}_2] = \frac{1}{2}$, and a straightforward calculation provides

$$\Pr[T_1 > T_2, s < T_2 < \tau] = \frac{S^2(s) - S^2(\tau)}{2},$$

where $S(\cdot)$ are the marginal survival functions.

It follows that substituting the consistent estimates $\hat{\boldsymbol{\beta}}$ and $\hat{S}(t|\boldsymbol{x})$ for the proportional hazards regression coefficient and conditional survival function produces the proposed estimate.

To compute the asymptotic distribution of the CPE, an asymptotically equivalent smooth version of K_n is used

$$\tilde{K}_n(\hat{\boldsymbol{\beta}}, \hat{H}_0; s, \tau) = \frac{[n(n-1)]^{-1} \sum_i \sum_{j \neq i} w_{ij}(\hat{\boldsymbol{\beta}}, \hat{H}_0; s, \tau) \Phi\left(\frac{\hat{\boldsymbol{\beta}}^T(\boldsymbol{x}_j - \boldsymbol{x}_i)}{h}\right) \left[1 + \exp\{\hat{\boldsymbol{\beta}}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\}\right]^{-1}}{\xi(s, \tau)}$$

where $\xi(s,\tau) = \lim_{n\to\infty} 0.5 \times [n(n-1)]^{-1} \sum_i \sum_{j\neq i} w_{ij}(\hat{\boldsymbol{\beta}}, \hat{H}_0; s, \tau)$ and h is the scale parameter of the local distribution function $\Phi(\cdot)$, and is chosen so that $nh^4 \to 0$ as $n \to \infty$ (Heller, 2004).

Thus, to derive the asymptotic distribution of the interval constrained CPE, we evaluate

$$n^{1/2} \left[\tilde{K}_n(\hat{\boldsymbol{\beta}}, \hat{H}_0; s, \tau) - \kappa(\boldsymbol{\beta}_0, H_0; s, \tau)\right] \tag{1}$$

where $\kappa(\boldsymbol{\beta}_0, H_0; s, \tau) = \lim_{n\to\infty} \tilde{K}_n(\hat{\boldsymbol{\beta}}, \hat{H}_0; s, \tau)$.

To simplify the notation for the derivation, denote the baseline cumulative hazard function at times $\{s, \tau\}$ as $\boldsymbol{\eta}_0^T = \{H_0(s), H_0(\tau)\}$.

To demonstrate the asymptotic distribution of (1), we decompose it into three terms

$$n^{1/2}[\tilde{K}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}) - \tilde{K}_n(\boldsymbol{\beta}_0, \hat{\boldsymbol{\eta}})] + n^{1/2}[\tilde{K}_n(\boldsymbol{\beta}_0, \hat{\boldsymbol{\eta}}) - \tilde{K}_n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)] + n^{1/2}[\tilde{K}_n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0) - \kappa(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)].$$

For the first two terms, Taylor expand around $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}_0$, respectively,

$$\left[\frac{\partial}{\partial \boldsymbol{\beta}} \tilde{K}_n(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]^T \left[n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right] + \left[\frac{\partial}{\partial \eta} \tilde{K}_n(\boldsymbol{\beta}_0, \boldsymbol{\eta})\bigg|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0}\right]^T \left[n^{1/2}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)\right]$$

and note that $\left[\frac{\partial}{\partial \boldsymbol{\beta}} \tilde{K}_n(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]$ and $\left[\frac{\partial}{\partial \eta} \tilde{K}_n(\boldsymbol{\beta}_0, \boldsymbol{\eta})\bigg|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0}\right]$ converge in probability.

In addition,

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = [n^{-1}I(\boldsymbol{\beta}_0)]^{-1}n^{-1/2}\sum_i u_i(\boldsymbol{\beta}_0) + o_p(1)$$

$$n^{1/2}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) = n^{-1/2}\sum_i \int_{u<t} g(u;\boldsymbol{\beta}_0)dM_i(u) + o_p(1) \qquad t = \{s, \tau\},$$

where I is the information matrix and $u_i(\boldsymbol{\beta})$ is the score vector for subject i from the partial likelihood,

$$M(t) = N(t) - \int_{u<t} \psi(u)h_0(u)\exp[\boldsymbol{\beta}_0^T\boldsymbol{x}]du \quad \text{is a martingale, and}$$

$$g(u;\boldsymbol{\beta}) = \lim_{n\to\infty} n^{-1}\sum_i \psi(u)\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i).$$

Therefore, the first two terms in the expansion can be asymptotically represented as the sum of independent identically distributed (iid) mean zero random variables.

The third term

$$n^{1/2}\left[\tilde{K}_n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0) - \kappa(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)\right]$$

is a degree 2 U-statistic, and using Hajek's projection lemma, it too may be asymptotically represented as the sum of iid mean zero random variables.

Therefore, combining the three terms

$$n^{1/2}\left[\tilde{K}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}) - \kappa(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)\right]$$

is asymptotically equivalent to the sum of iid mean zero random variables and application of the central limit theorem demonstrates that the interval constrained CPE is asymptotically normal with mean zero and an asymptotic variance that can be estimated using the empirical bootstrap.

Reference

Heller, G. (2004) Incorporating Follow-up Time in M-Estimation for Survival Data. *Lifetime Data Analysis* **10**, 51-64.

A.2 *Supplemental Tables for Simulations*

Table S1: Independent censoring: estimated bias, standard error, and relative efficiency for the normal marker distribution. The rows represent the three estimation approaches under varying degrees of censoring in terms of bias, standard error, and relative efficiency. The columns represent the four increasing values of s from 0 to $\frac{3}{5} \times \tau$, which represent the post-baseline follow-up times used to evaluate the estimates.

	Censoring	s_1	s_2	s_3	s_4
Concordance Probability		0.6941	0.694	0.6928	0.6884
Bias					
CPE	25%	0.0001	0.0001	0.0001	0.0001
Integrated AUC	25%	0.0001	0.0001	0.0002	0.0001
Weighted C	25%	-0.0004	-0.0004	-0.0003	-0.0001
CPE	50%	0.0003	0.0003	0.0004	0.0004
Integrated AUC	50%	0.0006	0.0006	0.0006	0.0006
Weighted C	50%	0.0003	0.0003	0.0005	0.0008
CPE	75%	0.001	0.001	0.001	0.001
Integrated AUC	75%	0.0011	0.0011	0.0011	0.001
Weighted C	75%	0.001	0.001	0.0011	0.0013
Estimated Standard Error/Simulation Standard Error					
CPE	25%	0.016/0.0166	0.016/0.0165	0.0157/0.0163	0.0149/0.0155
Integrated AUC	25%	0.017/0.017	0.0169/0.017	0.0168/0.0168	0.0163/0.0162
Weighted C	25%	0.0195/0.0193	0.0195/0.0193	0.02/0.0199	0.022/0.0215
CPE	50%	0.0188/0.0192	0.0187/0.0192	0.0184/0.0188	0.0175/0.0179
Integrated AUC	50%	0.0202/0.0201	0.0201/0.0201	0.02/0.0199	0.0195/0.0193
Weighted C	50%	0.0225/0.0223	0.0226/0.0224	0.0233/0.023	0.0262/0.0261
CPE	75%	0.0251/0.0256	0.0251/0.0255	0.0247/0.0251	0.0234/0.0239
Integrated AUC	75%	0.028/0.0274	0.028/0.0274	0.0279/0.0272	0.028/0.0271
Weighted C	75%	0.0342/0.0314	0.0342/0.0315	0.0357/0.033	0.0427/0.0396
Relative Efficiency (rMSE)					
Integrated AUC /CPE	25%	1.026	1.0263	1.0307	1.0469
Weighted C/CPE	25%	1.1643	1.1657	1.221	1.3912
Integrated AUC/CPE	50%	1.0494	1.0499	1.0555	1.0769
Weighted C/CPE	50%	1.1635	1.167	1.2223	1.4547
Integrated AUC/CPE	75%	1.0728	1.0734	1.0839	1.1368
Weighted C/CPE	75%	1.2275	1.2325	1.3136	1.6603

Table S2: Independent censoring: Estimated bias, standard error, and relative efficiency for the log-normal marker distribution.

	Censoring	s_1	s_2	s_3	s_4
Concordance Probability		0.7209	0.709	0.6864	0.6582
Bias					
CPE	25%	0.0007	0.0013	0.001	0.0008
Integrated AUC	25%	0.0008	0.001	0.001	0.0007
Weighted C	25%	-0.0001	0.0001	<0.0001	0.0005
CPE	50%	0.0011	0.0017	0.0014	0.0012
Integrated AUC	50%	0.0013	0.0016	0.0014	0.0011
Weighted C	50%	0.0008	0.001	0.0009	0.0015
CPE	75%	0.0039	0.0045	0.0039	0.0036
Integrated AUC	75%	0.005	0.0051	0.0045	0.0038
Weighted C	75%	0.0074	0.0072	0.0064	0.006
Estimated Standard Error/Simulation Standard Error					
CPE	25%	0.0138/0.0141	0.014/0.0138	0.0128/0.0124	0.0111/0.0106
Integrated AUC	25%	0.0155/0.0148	0.0153/0.0146	0.0145/0.0136	0.0134/0.0129
Weighted C	25%	0.0205/0.0203	0.0208/0.0208	0.0223/0.0225	0.0262/0.0265
CPE	50%	0.0148/0.015	0.0149/0.0148	0.0138/0.0134	0.012/0.0116
Integrated AUC	50%	0.0173/0.0163	0.0172/0.0163	0.0165/0.0156	0.0158/0.0151
Weighted C	50%	0.0237/0.0234	0.0242/0.0242	0.0264/0.0263	0.0319/0.032
CPE	75%	0.0174/0.0176	0.0176/0.0172	0.0163/0.0157	0.0143/0.0139
Integrated AUC	75%	0.0224/0.0205	0.0227/0.0207	0.0229/0.0208	0.0249/0.0228
Weighted C	75%	0.0369/0.0344	0.0382/0.0357	0.0429/0.0406	0.0567/0.0551
Relative Efficiency					
Integrated AUC/CPE	25%	1.0452	1.0592	1.0984	1.2077
Weighted C/CPE	25%	1.4356	1.5048	1.815	2.4869
Integrated AUC/CPE	50%	1.0873	1.1005	1.1573	1.3014
Weighted C/CPE	50%	1.554	1.6284	1.9509	2.7448
Integrated AUC/CPE	75%	1.175	1.1968	1.3105	1.6077
Weighted C/CPE	75%	1.9527	2.0468	2.5336	3.8586

Table S3: Independent censoring conditional on marker: estimated bias, standard error, and relative efficiency for the normal marker distribution.

	Censoring	s_1	s_2	s_3	s_4
Concordance Probability		0.6941	0.694	0.6928	0.6884
Bias					
CPE	25%	0.0002	0.0002	0.0002	0.0002
Integrated AUC	25%	-0.0037	-0.0037	-0.0039	-0.0045
Weighted C	25%	-0.0039	-0.0039	-0.0041	-0.0045
CPE	50%	-0.0005	-0.0005	-0.0005	-0.0005
Integrated AUC	50%	-0.0096	-0.0096	-0.0101	-0.0115
Weighted C	50%	-0.0099	-0.0099	-0.0104	-0.0117
CPE	75%	-0.0012	-0.0012	-0.0012	-0.0012
Integrated AUC	75%	-0.02	-0.0201	-0.0212	-0.0238
Weighted C	75%	-0.0192	-0.0192	-0.0204	-0.0223
Estimated Standard Error/Simulation Standard Error					
CPE	25%	0.0162/0.0164	0.0162/0.0163	0.016/0.0161	0.0151/0.0153
Integrated AUC	25%	0.017/0.017	0.017/0.017	0.0168/0.0168	0.0163/0.0162
Weighted C	25%	0.019/0.0188	0.019/0.0188	0.0195/0.0194	0.0216/0.0215
CPE	50%	0.0194/0.0197	0.0194/0.0196	0.0191/0.0193	0.0181/0.0183
Integrated AUC	50%	0.0205/0.0204	0.0206/0.0204	0.0203/0.0202	0.0198/0.0196
Weighted C	50%	0.0224/0.0222	0.0224/0.0222	0.0233/0.0231	0.0267/0.0267
CPE	75%	0.027/0.0277	0.027/0.0277	0.0266/0.0272	0.025/0.0257
Integrated AUC	75%	0.0291/0.0287	0.0291/0.0287	0.0289/0.0284	0.0289/0.028
Weighted C	75%	0.0394/0.037	0.0396/0.0371	0.0417/0.0392	0.0508/0.0479
Relative Efficiency (rMSE)					
Integrated AUC/CPE	25%	1.0638	1.0643	1.0723	1.1048
Weighted C/CPE	25%	1.1728	1.1763	1.2353	1.4428
Integrated AUC/CPE	50%	1.146	1.1474	1.1682	1.2439
Weighted C/CPE	50%	1.236	1.2386	1.3134	1.5961
Integrated AUC/CPE	75%	1.2625	1.2651	1.3032	1.4292
Weighted C/CPE	75%	1.5025	1.5085	1.6227	2.0572

Table S4: Independent censoring conditional on marker: estimated bias, standard error, and relative efficiency for the log-normal marker distribution.

	Censoring	s_1	s_2	s_3	s_4
Concordance Probability		0.7209	0.709	0.6864	0.6582
Bias					
CPE	25%	0.0007	0.0013	0.001	0.0008
Integrated AUC	25%	-0.0179	-0.0166	-0.0152	-0.0139
Weighted C	25%	-0.0182	-0.0169	-0.0155	-0.0134
CPE	50%	0.0005	0.0013	0.0009	0.0008
Integrated AUC	50%	-0.0431	-0.0405	-0.0372	-0.0335
Weighted C	50%	-0.0436	-0.041	-0.0378	-0.0333
CPE	75%	0.0012	0.0024	0.0017	0.0016
Integrated AUC	75%	-0.0809	-0.0762	-0.0701	-0.0628
Weighted C	75%	-0.0798	-0.0752	-0.0689	-0.0606
Estimated Standard Error/Simulation Standard Error					
CPE	25%	0.0142/0.0144	0.0143/0.0141	0.013/0.0126	0.0111/0.0107
Integrated AUC	25%	0.0158/0.0151	0.0154/0.0148	0.0143/0.0136	0.0129/0.0125
Weighted C	25%	0.0205/0.0202	0.0207/0.0206	0.022/0.0221	0.0256/0.026
CPE	50%	0.0159/0.0161	0.0159/0.0157	0.0143/0.0139	0.012/0.0117
Integrated AUC	50%	0.0179/0.0172	0.0175/0.0167	0.0159/0.0151	0.0141/0.0134
Weighted C	50%	0.0245/0.0243	0.0248/0.0248	0.0265/0.0267	0.0317/0.0318
CPE	75%	0.0208/0.0207	0.0206/0.0202	0.018/0.0175	0.0148/0.0147
Integrated AUC	75%	0.0228/0.021	0.0222/0.0203	0.0205/0.0181	0.0193/0.0162
Weighted C	75%	0.0439/0.0398	0.0445/0.0406	0.0484/0.0442	0.06/0.0559
Relative Efficiency					
Integrated AUC/CPE	25%	1.6208	1.572	1.6187	1.7341
Weighted C/CPE	25%	1.8805	1.8873	2.1383	2.7177
Integrated AUC/CPE	50%	2.8882	2.7853	2.8938	3.0694
Weighted C/CPE	50%	3.1054	3.0468	3.3319	3.9172
Integrated AUC/CPE	75%	4.0343	3.879	4.1257	4.3934
Weighted C/CPE	75%	4.3051	4.2025	4.6645	5.5848

Table S5: Non-proportional hazards: estimated bias, standard error, and relative efficiency when censoring times were generated from a Weibull random variable with shape parameter equal to $1 - 0.1X$, dependent on the underlying marker X.

	Censoring	s_1	s_2	s_3	s_4
Concordance Probability		0.674	0.6585	0.6522	0.6481
Bias					
CPE	25%	-0.0188	-0.0054	-0.0007	0.0021
Integrated AUC	25%	-0.0186	-0.0049	0.0004	0.0036
Weighted C	25%	-0.0003	0.0004	0.0005	0.0008
CPE	50%	-0.0115	0.0015	0.006	0.0086
Integrated AUC	50%	-0.0108	0.0028	0.008	0.0113
Weighted C	50%	0.0001	0.0008	0.0012	0.0014
CPE	75%	0.0012	0.0136	0.0177	0.02
Integrated AUC	75%	0.0028	0.016	0.0209	0.0239
Weighted C	75%	0.0017	0.0013	0.0007	-0.0013
Estimated Standard Error/Simulation Standard Error					
CPE	25%	0.0184/0.0193	0.0177/0.0185	0.0172/0.018	0.0168/0.0176
Integrated AUC	25%	0.0188/0.0193	0.0182/0.0187	0.0179/0.0183	0.0178/0.0181
Weighted C	25%	0.0238/0.0238	0.0286/0.029	0.0344/0.0354	0.043/0.044
CPE	50%	0.0211/0.0217	0.0201/0.0208	0.0196/0.0202	0.0191/0.0198
Integrated AUC	50%	0.0215/0.0217	0.0208/0.021	0.0206/0.0206	0.0205/0.0205
Weighted C	50%	0.0249/0.0246	0.0305/0.0304	0.0374/0.0379	0.0474/0.0481
CPE	75%	0.0273/0.0274	0.026/0.0262	0.0252/0.0254	0.0246/0.0249
Integrated AUC	75%	0.029/0.0284	0.0291/0.0283	0.0303/0.0294	0.038/0.0323
Weighted C	75%	0.0335/0.0323	0.0456/0.0445	0.063/0.0626	0.0947/0.0968
Relative Efficiency (rMSE)					
Integrated AUC /CPE	25%	0.9961	1.0022	1.0187	1.0406
Weighted C/CPE	25%	0.8862	1.5036	1.9667	2.4798
Integrated AUC/CPE	50%	0.9869	1.0165	1.0491	1.0825
Weighted C/CPE	50%	1.0007	1.4588	1.7949	2.2306
Integrated AUC/CPE	75%	1.0377	1.103	1.1646	1.2596
Weighted C/CPE	75%	1.1785	1.5101	2.0223	3.0366

Table S6: Non-proportional hazards: estimated bias, standard error, and relative efficiency when censoring times were generated from a Weibull random variable with shape parameter equal to $1 - 0.175X$, dependent on the underlying marker X.

	Censoring	s_1	s_2	s_3	s_4
Concordance Probability		0.6846	0.6587	0.6496	0.6439
Bias					
CPE	25%	-0.0359	-0.012	-0.0043	0.0003
Integrated AUC	25%	-0.0362	-0.0117	-0.0031	0.0023
Weighted C	25%	-0.0003	0.0002	0.0002	0.0003
CPE	50%	-0.0216	0.0016	0.0089	0.0133
Integrated AUC	50%	-0.021	0.0034	0.012	0.0174
Weighted C	50%	0.0002	0.0011	0.0013	0.0017
CPE	75%	0.002	0.024	0.0306	0.0344
Integrated AUC	75%	0.0043	0.028	0.0359	0.0406
Weighted C	75%	0.004	0.0037	0.0035	0.0036
Estimated Standard Error/Simulation Standard Error					
CPE	25%	0.0196/0.0211	0.0188/0.0203	0.0183/0.0198	0.018/0.0194
Integrated AUC	25%	0.0198/0.0209	0.0192/0.0203	0.0189/0.0199	0.0188/0.0197
Weighted C	25%	0.0236/0.0236	0.0289/0.0292	0.0347/0.0357	0.0434/0.0448
CPE	50%	0.0218/0.0228	0.0207/0.0218	0.0202/0.0212	0.0197/0.0208
Integrated AUC	50%	0.022/0.0225	0.0212/0.0217	0.021/0.0214	0.0209/0.0212
Weighted C	50%	0.0247/0.0244	0.0308/0.0307	0.0377/0.0382	0.0477/0.0488
CPE	75%	0.027/0.0275	0.0255/0.0261	0.0247/0.0253	0.0242/0.0248
Integrated AUC	75%	0.0285/0.028	0.0285/0.0278	0.03/0.0291	0.0389/0.0319
Weighted C	75%	0.0328/0.0324	0.0456/0.0459	0.0631/0.0636	0.0935/0.0994
Relative Efficiency					
Integrated AUC/CPE	25%	1.0032	0.9912	0.9951	1.0204
Weighted C/CPE	25%	0.5654	1.2372	1.76	2.3029
Integrated AUC/CPE	50%	0.9783	1.0088	1.0662	1.1154
Weighted C/CPE	50%	0.7776	1.41	1.6619	1.9805
Integrated AUC/CPE	75%	1.0262	1.1146	1.1628	1.2183
Weighted C/CPE	75%	1.1822	1.2989	1.6038	2.345

A.3 *Scaled Schoenfeld residuals for the prostate cancer and colorectal surgery data examples.*

The baseline risk scores for the colorectal and prostate cancer examples were both estimated in a proportional hazards regression model. To assess the proportional hazards assumption for both models, loess curves were fit to the scaled Schoenfeld residuals. As shown in Figures S1 and S2, the curves are approximately constant over time for all covariates, indicating no observed deviation from the proportion hazards assumption.

Figure S1: Scaled Schoenfeld residuals for colorectal surgery risk score when estimated from a proportional hazard regression model.

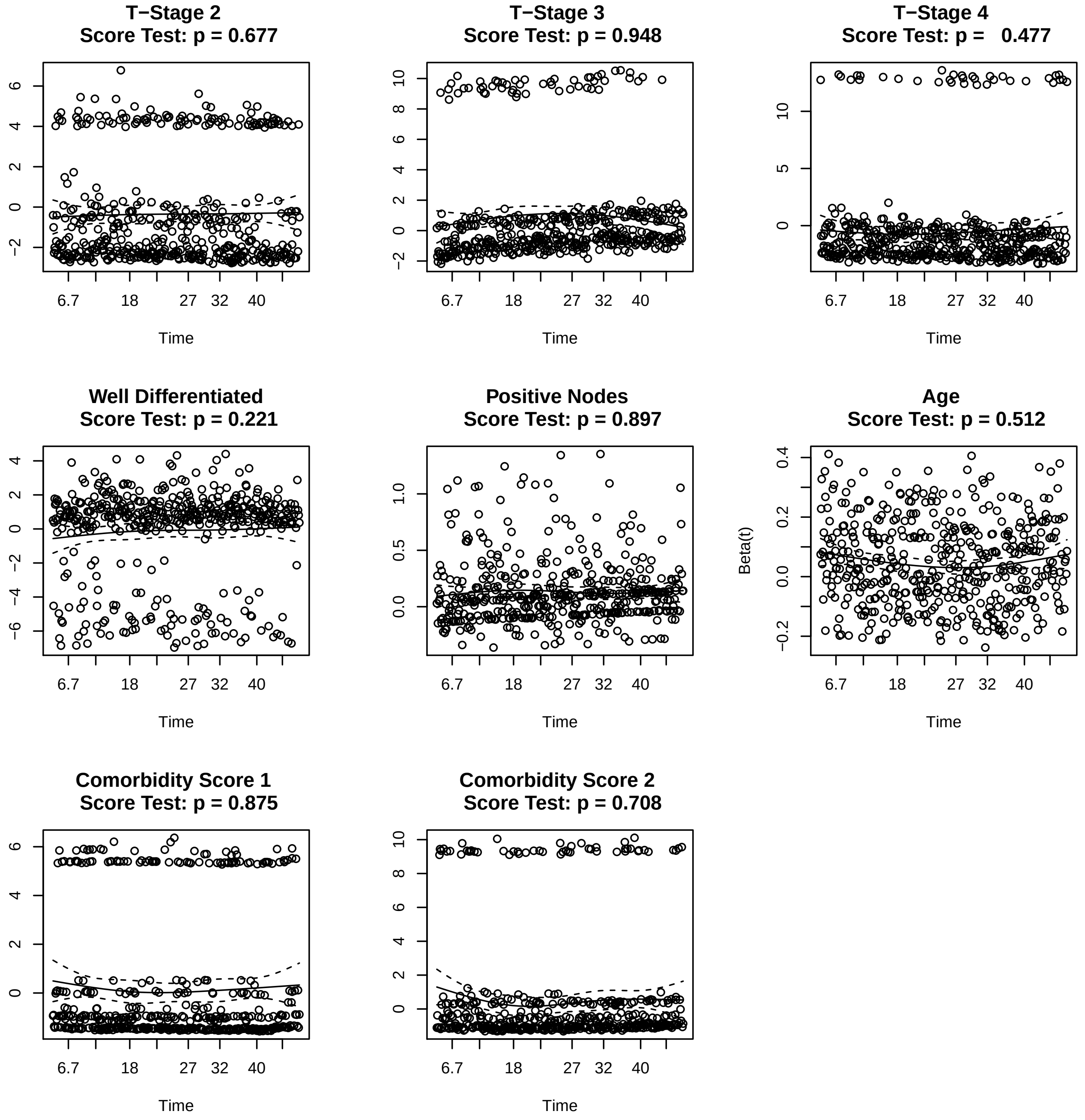
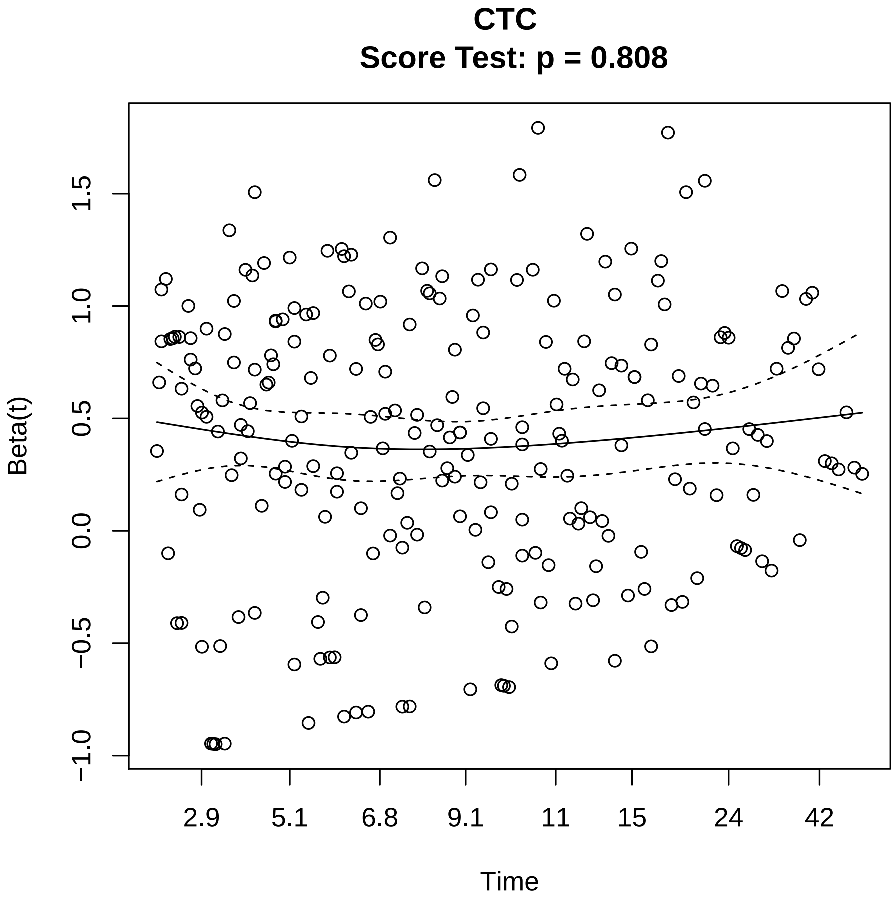


Figure S2: Scaled Schoenfeld residuals for CTC when estimated from a proportional hazard regression model.


CTC
Score Test: p = 0.808
Beta(t)
Time