Concordance probability as a meaningful contrast across disparate survival times.

Sean M Devlin* and Glenn Heller†

Department of Epidemiology and Biostatistics

Memorial Sloan Kettering Cancer Center

New York, NY

October 19, 2020

Abstract

The performance of time-to-event models is frequently assessed in part by estimating the concordance probability, which evaluates the probabilistic pairwise ordering of the model-based risk scores and survival times. The standard definition of this probability conditions on any survival time pair ordering, irrespective of whether the times are meaningfully separated. Inclusion of survival times that would be deemed clinically similar attenuates the concordance and moves the estimate away from the contrast-of-interest: comparing the risk scores between individuals with disparate survival times. In this manuscript, we propose a concordance

*devlins@mskcc.org

†hellerg@mskcc.org

definition and corresponding method to estimate the probability conditional on survival times being separated by at least a minimum difference. The proposed estimate requires direct input from the analyst to identify a separable survival region, and in doing so, is analogous to the clinically-defined subgroups used for binary outcome area under the curve estimates. The method is illustrated in two cancer examples: a prognostic score in clear cell renal cell carcinoma and two biomarkers in metastatic prostate cancer.

1 Introduction

Discrimination is a common metric for time-to-event risk models to evaluate how separable the risk scores are relative to good and poor survival outcomes. A commonly applied parameter of discrimination is the concordance probability, 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 $\boldsymbol{X}$, T denotes survival time, and τ indicates the maximum follow-up time of the study.

Initially, an estimate of the concordance probability was developed by Harrell et al.,[1,2] using a ratio statistic based on pairwise inequalities of the survival times and risk scores. Over time, this estimate has been refined by incorporating inverse probability censoring weights,[3,4] assuming a proportional hazards risk model,[5] and integrating estimates of the true and false positive rates.[6,7]

An attractive feature of the survival-based concordance probability is that it is identical to the area under the receiver operating characteristic curve (AUC) when the outcome

is binary, and thus provides an extension to the well-known binary outcome AUC literature.[8,9] In the binary endpoint setting, risk scores are used to discriminate between individuals with and without a disease, or those who do and do not achieve a complete response to therapy. These binary endpoints are defined by clinically meaningful subgroups, and the assessment of discrimination through the area under the curve corresponds to how well the risk scores segregate across these well-defined subgroups.

For the concordance probability with a survival endpoint defined in (1), the probabilistic pairwise ordering of the risk scores are evaluated over the upper wedge of the survival time pair space defined in Figure 1A. Importantly, survival pairs in the space near but above the diagonal $T_1 = T_2$ with $T_2 < \tau$ are included in the conditioning argument, but are similar from a clinical viewpoint and thus their inclusion is not constructive for the concordance probability interpretation. A more natural extension of the binary endpoint AUC methodology is to restrict the survival pair space to elements where the survival times are at least Δ months apart (Figure 1B). This, however, will require input/expertise from the analyst as defining clinically meaningful survival time subgroups depends on the scenario-at-hand. In cancer, for example, a survival difference of 6 months may not be meaningful for certain localized diseases but may be significant for advanced metastatic disease.

In this paper, we propose a method to evaluate the concordance probability conditional on a meaningful difference in survival times. This probability is estimated within the framework of a properly specified survival regression model. The model choice is at the discretion of the analyst, but time-invariant risk scores are required. This paper continues with Section 2, which provides the probability definition and proposed methodology in this setting. Section 3 illustrates the use of the concordance probability

by applying it to two biomarkers in metastatic prostate cancer and to a prognostic score in clear cell renal cell carcinoma. Section 4 evaluates the operating characteristics of the methodology. Lastly, the article concludes in Section 5 with a discussion of the key results and implications for future use.

2 Methods

The clinical utility of a discrimination metric, applied to a survival model, is the ascertainment of whether the survival experience is reflected in the model-based risk score. Strong model discrimination indicates that clinical decisions, such as the aggressiveness of treatment, may in part be guided by the patient risk score. The concordance probability discrimination metric, as defined in (1), however, does not precisely address this concept of risk score separation for clinically distinct survival times. The limitation of the concordance probability is illustrated with two clinical examples that are further evaluated in Section 3. The first consists of 271 patients with metastatic prostate cancer who had a biomarker of tumor burden, circulating tumor cells (CTC), evaluated prior to treatment. The second consists of 421 patients with clear cell renal cell carcinoma (ccRCC) from The Cancer Genome Atlas; the risk score for two clinical features, SSIGN and age, was estimated using the proportional odds model.

A scatter plot of survival times is depicted in Figure 2(A) for CTC in prostate cancer and in Figure 2(B) for the risk score in ccRCC. There is evidence in both that high risk scores indicate short survival (region 1) and low risk scores predict longer survival (region 2). However, an estimate of the concordance probability will be attenuated due to the

evaluation of all data pairs within region 1, even though all patients in this region can be classified as short-term survivors.

Our proposal is to extend the concordance probability definition by focusing on survival times T that are separated by at least Δ units of time. The probability of interest is defined by

$$\Pr[\boldsymbol{\beta}^T \boldsymbol{X}_2 > \boldsymbol{\beta}^T \boldsymbol{X}_1 | T_1 > T_2 + \Delta, T_2 < \tau] \qquad \Delta > 0, \tag{2}$$

which we call dCP, for a delta concordance probability based on a separation of Δ units of time.

It is assumed that the risk score $\boldsymbol{\beta}^T \boldsymbol{x}$ is derived from a semiparametric regression model

$$m(t) = \boldsymbol{\beta}^T \boldsymbol{x} + \epsilon,$$

where $m(t)$ is a monotone transformation of the survival time and ϵ is the random error. Examples of semiparametric models include when m is unknown and the error distribution of ϵ is extreme value (proportional hazards) or logistic (proportional odds). Alternatively, an accelerated failure time model is produced when the transformation is specified as $m(t) = \log(t)$ and the error distribution is unknown.

Equation 2 for dCP may be rewritten as

$$K(\boldsymbol{\beta}, S; \Delta, \tau) = \frac{\int_{\boldsymbol{\beta}^T \boldsymbol{x}_2 > \boldsymbol{\beta}^T \boldsymbol{x}_1} \int_{t=0}^{\tau} S(t + \Delta | \boldsymbol{x}_1) dS(t | \boldsymbol{x}_2) dG(\boldsymbol{\beta}^T \boldsymbol{x}_1) dG(\boldsymbol{\beta}^T \boldsymbol{x}_2)}{\int_{t=0}^{\tau} S(t + \Delta) dS(t)},$$

where G represents the distribution function of $\boldsymbol{\beta}^T \boldsymbol{x}$ and S denotes a survival function. This probability can be estimated by

$$K_n(\hat{\boldsymbol{\beta}}, \hat{S}; \Delta, \tau) = \frac{\sum_i \sum_j \sum_k I\{\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i > \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_j\} I\{t_{(k)} < \tau\} \hat{S}(t_{(k)} + \Delta | \boldsymbol{x}_j) \left(\hat{S}(t_{(k)} | \boldsymbol{x}_i) - \hat{S}(t_{(k-1)} | \boldsymbol{x}_i) \right)}{\sum_i \sum_j \sum_k I\{t_{(k)} < \tau\} \hat{S}(t_{(k)} + \Delta | \boldsymbol{x}_j) \left(\hat{S}(t_{(k)} | \boldsymbol{x}_i) - \hat{S}(t_{(k-1)} | \boldsymbol{x}_i) \right)},$$

where $\hat{S}(t_{(k)}|\boldsymbol{x}_i)$ represents the model based estimated survival function for the ordered survival time $t_{(k)}$, where $\hat{S}(t_{(0)}|\boldsymbol{x}) = 1$.

The estimated concordance probability $K_n(\hat{\boldsymbol{\beta}}, \hat{S}; \Delta, \tau)$ is a ratio statistic with its denominator bounded away from zero. To develop its asymptotic distribution, the indicator function $I(\boldsymbol{\beta}^T\boldsymbol{x}_1 > \boldsymbol{\beta}^T\boldsymbol{x}_2)$ is approximated by a smooth local Gaussian distribution function $\Phi(\frac{\boldsymbol{\beta}^T\boldsymbol{x}_1 - \boldsymbol{\beta}^T\boldsymbol{x}_2}{h})$, where the bandwidth $h \to 0$ as $n \to \infty$.[15] To see this approximation, note that if $\boldsymbol{\beta}^T\boldsymbol{x}_1 > \boldsymbol{\beta}^T\boldsymbol{x}_2$, then $\Phi(\frac{\boldsymbol{\beta}^T\boldsymbol{x}_1 - \boldsymbol{\beta}^T\boldsymbol{x}_2}{h}) \to 1$ as $h \to 0$, and if $\boldsymbol{\beta}^T\boldsymbol{x}_1 < \boldsymbol{\beta}^T\boldsymbol{x}_2$, then $\Phi(\frac{\boldsymbol{\beta}^T\boldsymbol{x}_1 - \boldsymbol{\beta}^T\boldsymbol{x}_2}{h}) \to 0$, as $h \to 0$. It follows that as n gets large and $h \to 0$, $\Phi(\frac{\boldsymbol{\beta}^T\boldsymbol{x}_1 - \boldsymbol{\beta}^T\boldsymbol{x}_2}{h}) \to I(\boldsymbol{\beta}^T\boldsymbol{x}_1 > \boldsymbol{\beta}^T\boldsymbol{x}_2)$.

A first order Taylor series expansion on the smoothed version of the concordance probability statistic produces three sources of variation: the estimated regression coefficient $\hat{\boldsymbol{\beta}}$, the survival estimate $\hat{S}$, and the covariates $\boldsymbol{x}$. The asymptotic normality of the estimated concordance probability stems from the convergence of $n^{1/2}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}]$ to normality using either maximum partial likelihood or estimating equation theory, the convergence of $n^{1/2}[\hat{S}(t) - S(t)]$ to a Gaussian process using martingale arguments, and the convergence of $n^{1/2}[K_n(\boldsymbol{\beta}, S; \Delta, \tau) - \kappa(\boldsymbol{\beta}, S; \Delta, \tau)]$ to a Gaussian process by applying the functional central limit theorem for U processes, where $\kappa(\boldsymbol{\beta}, S; \Delta, \tau) = \lim_{n\to\infty} K_n(\boldsymbol{\beta}, S; \Delta, \tau)$. The asymptotic normality of the concordance probability statistic follows from the three components of the expansion. The specific arguments were established in a comparable statistic.[3] The asymptotic variance $K_n(\hat{\boldsymbol{\beta}}, \hat{S}; \Delta, \tau)$ and corresponding asymptotic confidence interval are computed using the bootstrap. The validity of the bootstrap in semiparametric models was established in Kosorok et al.[16]

3 Data Analysis

The concordance probability as defined in Equation (2) is illustrated in the two previously mentioned examples for varying degrees of separation in survival times. The first evaluates dCP for two biomarkers in metastatic castration-resistant prostate cancer. The second estimates dCP for a clinical risk score in clear cell renal cell carcinoma (ccRCC). While the risk scores in both settings have been shown to have moderate discrimination, the evaluations to date have conditioned on any survival time region ($\Delta = 0$) when estimating concordance. The analyses in this section aim to estimate dCP when the survival times are separated by a clinically meaningful difference in time.

3.1 CTC and ALK in metastatic prostate cancer

Two biomarkers were evaluated on 271 patients with metastatic castration-resistant prostate cancer at the time of treatment. The first biomarker was circulating tumor cells (CTC), a blood based assay, used in metastatic prostate cancer to determine tumor burden; the second biomarker, alkaline phosphatase (ALK), is also a measure of tumor burden. Both markers have improved model discrimination in metastatic prostate cancer studies.[19,20] In terms of the direction of effect, higher values of CTC and ALK are associated with worse prognosis.

For this cohort of patients, the estimated survival curve is shown in Figure 3(A). A total of 202 patients died during the follow-up period. Baseline CTC and ALK were log-transformed for the analysis. Assessment of the separate proportional hazards models for CTC and ALK indicated no apparent deviation from proportional hazards: see Supplemental Section A.1.

The time scale for a meaningful improvement in survival for individuals with metastatic prostate cancer is relatively small. In a randomized clinical trial designed by one of us, a median survival improvement of 1.33 years has been conferred on an experimental treatment relative to the conventional treatment. For baseline CTC and ALK, the dCP in Equation (2) was evaluated over different values of Δ, ranging from 0 to 2 years. The value of τ, used to define the concordance probability, was selected to be 2 years.

The results are shown in Figure 3(B). When $\Delta = 0$, CTC had a dCP of 0.64 (95% CI: 0.60-0.67), which was marginally better than ALK, with a dCP of 0.60 (0.56-0.63). Neither biomarker is distinguished in terms of prognostic utility. However, when $\Delta = 2$, the dCP for CTC increased dramatically to 0.78 (0.73-0.82), indicating that CTC has strong discriminatory power when comparing distinct survival time regions. In contrast, the dCP for ALK was only 0.70 (0.64-0.76) at $\Delta = 2$. The difference in dCP between markers was 0.08 (0.01-0.14).

3.2 SSIGN in clear cell renal cell carcinoma

This analysis includes 421 patients with clear cell renal cell carcinoma (ccRCC) from The Cancer Genome Atlas cohort with complete data on overall survival and the clinical components to calculate the Mayo Clinical Stage, Size, Grade, and Necrosis (SSIGN) score, as previously described.[10,11] A total of 144 (34%) patients died, with a Kaplan-Meier median survival estimate of 6.4 years from diagnosis. The estimated survival curve is shown in Figure 4(A).

The risk score was estimated using proportional odds regression for the covariates of SSIGN and age. SSIGN, with observed values between 0 to 15, was modeled using a

natural cubic spline due to observed non-linearity. Assessment of the proportional odds assumption is provided in the Supplemental Section. The value of τ was set to be 4; at this time point, the estimated survival was 0.67(95%: 0.63-0.73). dCP was estimated over the range $\Delta = 0$ to 3.5 years.

The estimated dCP are shown in Figure 4(B). The estimate, based on $\Delta = 0$, produces a dCP of 0.79 (0.76-0.83). While overall strong, this estimate, as noted earlier, penalizes the risk score for a lack of separation between patients with minimally different survival times. Importantly, dCP increased to 0.83 (0.79-0.87) and 0.85(0.81-0.89) when conditioning on at least a 2-year and 3-year survival difference, respectively. A separation of $\Delta = 3$ is similar to the projected 2.6 year difference in median overall survival used in the original design of a phase III study in non-metastatic RCC.[18]

While the risk score provides good discrimination when viewed within the traditional definition of concordance probability with $\Delta = 0$, the analysis illustrates the true potential of the risk score as a standalone prognostic marker in ccRCC when dCP is reevaluated for survival times meaningfully separated.

4 Operating Characteristics

Simulation scenarios evaluated the operating characteristics of the dCP estimation procedure. The first scenario mirrors the SSIGN risk score in ccRCC and evaluates the procedure under a proportional odds model. The second scenario is consistent with the CTC biomarker in metastatic prostate cancer under the proportional hazards model. The final scenario evaluates dCP estimation under model misspecification. The operating characteristics were evaluated by the average bias and the square root of the mean

square error (rMSE) calculated over 2,000 iterations. In addition, the estimated standard errors for the different values of Δ were compared to their simulation standard errors. Bootstrap resampling estimated the standard error for dCP, with 200 bootstrap resamples used for each simulation iteration. The true underlying dCP values for the different Δ values were approximated by the average of 2,000 iterations of 3,000 uncensored observations.

4.1 Proportional Odds

The first scenario generated data under a proportional odds model motivated by the ccRCC risk score, where dCP increased from good to excellent concordance as Δ increased from 0 to 3.75 years. Data were generated under the model $t_i = \exp\{7.4 - 1.15x_i + \epsilon_i\}$, with independent and identically distributed ϵ_i from a logistic distribution. The distribution of x_i was N($\mu = 4.24$, $\sigma = 1.452$), following the estimated mean and standard deviation of the risk score developed in Section 3. Censoring times were generated from a uniform distribution Un(0, 12) to approximate the censoring proportion of 66% observed in the data analysis. The value of τ was fixed at 4. Sample sizes of 225, 450, and 900 were considered.

Simulation results are provided in Table 1 and Figure 5(A). Overall the average bias across the simulation iterations was minimal; the average bias increased slightly as Δ increased. rMSE similarly increased slightly as Δ increased. As the sample size increased, the average bias and rMSE decreased across all values of Δ. The average estimated standard error aligned with the simulation standard error for all estimates.

4.2 Proportional Hazards

This scenario was designed to reflect the CTC data analysis in Section 3. Data were generated from a proportional hazards relationship using the regression model $t_i = \exp\{2 - 0.4x_i\} \times \epsilon_i$, and ϵ_i were generated from Weibull random variables with scale parameter 0.668 and shape parameter 1. The distribution of marker x_i followed a normal distribution, N($\mu = 2.33$, $\sigma = 1.76$), corresponding to the mean and standard deviation of the log-transformed CTC values in the data analysis. Similarly, the censoring times were generated from a uniform distribution Un(0, 8.5) to approximate a censoring proportion of 25% observed in Section 3. The value of τ was fixed at 2. The true concordance was 0.68 for $\Delta = 0$, which increased to 0.81 as Δ increased. Sample sizes of 150, 300, and 600 were considered.

Simulation results are provided in Table 2. The performance of the dCP method is again, excellent. A slightly higher bias was observed for larger values of Δ, which attenuated for the larger sample size, and the estimated standard errors were approximately equal to the simulation standard errors.

4.3 Non-proportional Hazards

Two simulation scenarios evaluated the performance of the method when a proportional hazards model was fit to data that violated the proportional hazards assumption. Both scenarios mirrored the simulations in Section 4.2. However, now ϵ_i were generated from Weibull random variables with scale parameter 0.668 and shape parameter dependent on the marker x_i to induce non-proportional hazards. In the first scenario, representing a minor deviation from proportional hazards, the shape parameter was equal to $1 - 0.04x_i$.

The second scenario, representing a larger deviation, the shape equaled $1 - 0.09x_i$. All other values remained the same as the previous section.

In terms of the assessment of the proportional hazards assumption, the goodness-of-fit test was rejected in 8.2%, 12%, and 19% of the 2,000 simulation iterations in the first scenario for the sample sizes of 150, 300, and 600, respectively, and 36%, 63%, and 91% of the simulations in the second scenario for the same sample sizes.

Simulation results are provided in Figure 6 and Tables 3 and 4. As expected, the bias of the dCP estimates depended on the degree of proportional hazards violation. For the first scenario, the average estimated bias ranged from -0.005 to -0.010 across the different sample sizes. For second scenario, the average bias increased substantially, ranging from -0.033 to -0.055.

4.4 Summary

These results illustrate that the proposed dCP methodology performed well under simulation scenarios motivated by the SSIGN risk score and the CTC biomarker data analyses in Section 3 under correct specification of the regression model. The proposed bootstrapping procedure for standard error estimation closely mirrored the simulation standard error.

When the selected regression model is misspecified, the dCP and its estimated standard error may be biased. These results reiterate the importance of investigating all modeling assumptions prior to data analysis. We note that the largest biases were observed in Section 4.3 when the goodness-of-fit test had the highest power to reject the null hypothesis of no time-varying effect.

5 Discussion

In this paper, we propose a novel definition of concordance probability that focuses on individuals with a clinically meaningful difference in survival. Conditioning on a separable survival region orients the concordance probability, termed dCP, around the quantity we care most about: how well the risk score segregates across individuals with disparate survival times. Framed in this way, the probability aligns with the area under the curve for binary outcomes, where outcomes are similarly based on clinically distinct groups.

The proposed dCP definition requires direct input the analyst; the minimum difference in survival (Δ) depends on the clinical scenario and data under consideration. While a single value of Δ may frequently be selected for reporting the discrimination of a model, the probability can be estimated and visually displayed over a range of Δ values. As shown in Section 3, this visual representation is particularly useful when evaluating the relative discrimination of two or more markers. The two markers, CTC and ALK, have moderate concordance when contiguous survival regions are considered; however, focusing on survival regions with a consequential separation highlights the prognostic utility of each marker, particularly in the case of CTC.

The dCP definition requires an estimate of the conditional survival function. As outlined in Section 2, this can be estimated within various time-to-event models, including the proportional hazards, the proportional odds, and the accelerated failure time models. It is left to the analyst to select the model most appropriate for the data under consideration and first evaluate any modeling assumptions.

Discrimination is one measure frequently used to assess model performance. When assessing the prognostic utility of a set of factors within a survival model, the evaluation

of discrimination should be done in conjunction with calibration and explained variation.

The formulation of dCP in this manuscript is for a continuous risk score. Future work will extend the methodology to a discrete risk score, analogous to the work developed for the c-index and CPE.[21] In addition, methodology will be developed to evaluate the added value of new biomarkers to an existing risk model within the context of the dCP estimate.

Acknowledgements

This work was supported by NIH Grants P30CA008748 and R01CA207220.

Data Accessibility

The authors are unable to publicly provide the data used in the metastatic prostate cancer example. The ccRCC data are available at cBioPortal (http://www.cbioportal.org). The R code for dCP along with the code to generate all simulation scenarios can be obtained from www.github.com/sedevlin.

References

[1] Harrell FE, Califf RM, Pryor DB, Lee KL., and Rosati R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association* **247**, 2543–2546.

[2] Harrell FE, Lee KL, and Mark DB. (1996). Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine* **15**, 361–387.

[3] 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.

[4] Gerds TA, Kattan MW, 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.

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

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

[7] Song X, and Zhou XH.(2008). A semiparametric approach for the covariate specific ROC curve with survival outcome. *Statistica Sinica* **18**, 947–965.

[8] Hanley JA, and McNeil BJ. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. **143**, 29—36.

[9] Pepe MS. (2003). The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press.

[10] Hakimi AA, Mano R, Ciriello G, et al. (2014). Impact of recurrent copy number alterations and cancer gene mutations on the predictive accuracy of prognostic models in clear cell renal cell carcinoma. *J Urol*. **192**, 24—29.

[11] The Cancer Genome Atlas Research Network. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499(7456)**: 43–49.

[12] Tsiatis AA. (1981). A large sample study of Cox's regression model. *Annals of Statistics* **9**, 93-108.

[13] Murphy SA, Rossini AJ, va der Vaart AW. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association* **92**, 968-976.

[14] Park Y, and Wei LJ. (2003). Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika* **90**, 717-723.

[15] Horowitz J.L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrika* **60**, 505-531.

[16] Kosorok MR, Lee BL, Fine JP. (2004). Robust inference for univariate proportional hazards frailty regression models. *Annals of Statistics* **32**, 1448-1491.

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

[18] Haas NB, Manola J, Uzzo RG, et al. (2016). Adjuvant sunitinib or sorafenib for high-risk, non-metastatic renal-cell carcinoma (ECOG-ACRIN E2805): a double-blind, placebo-controlled, randomised, phase 3 trial. **387(10032)** 2008 – 2016.

[19] 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.

[20] 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.

[21] 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.

Figure 1: (A) The survival space $T_1 > T_2$ with $T_2 < \tau$ used as the conditioning argument in a standard concordance probability definition. (B) The space reoriented around a survival difference of Δ units of time.

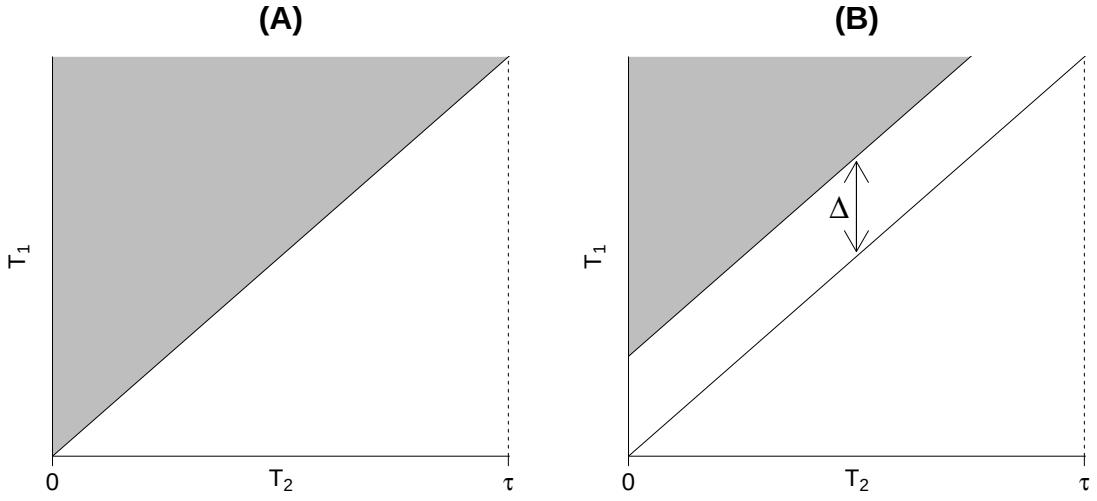


Figure 2: (A) Values of log CTC for the survival times in the metastatic prostate cancer example. (B) The estimated risk score in the clear cell renal cell carcinoma example.

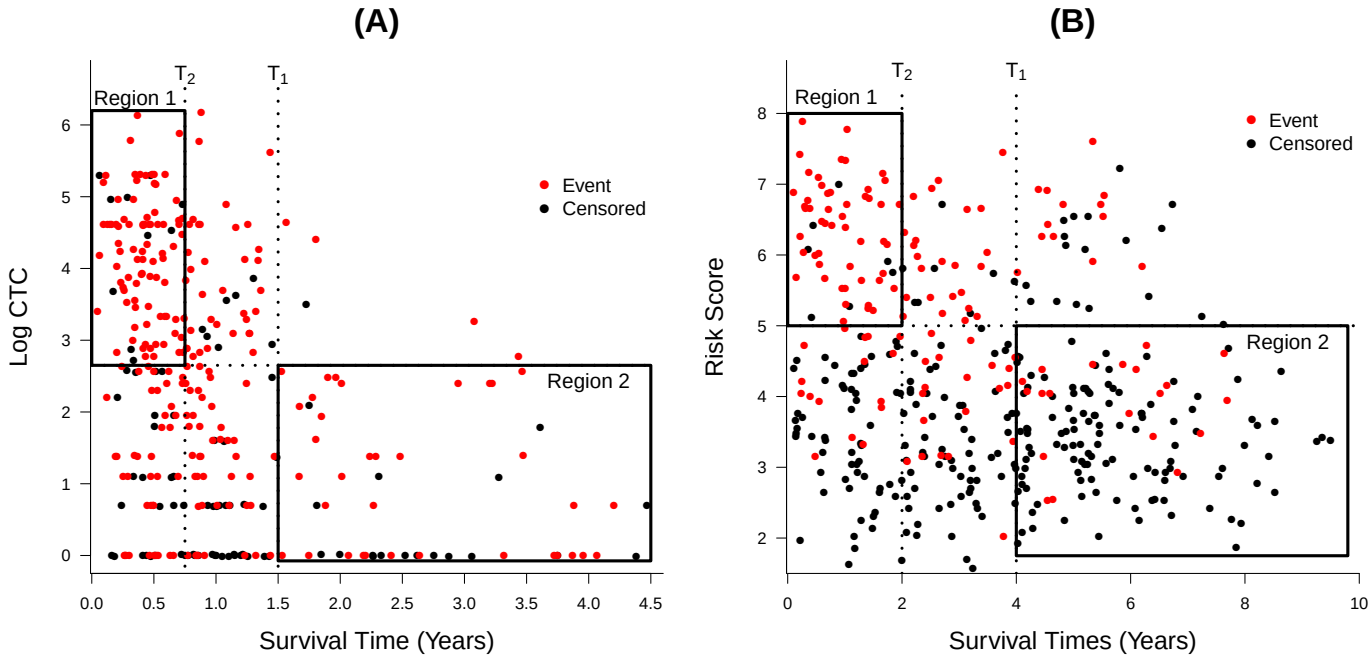


Figure 3: (A) Estimated overall survival for the 271 patients with metastatic prostate cancer starting from the time of treatment. (B) Estimated dCP for ALK and CTC for various degrees of separation in survival times (Δ).

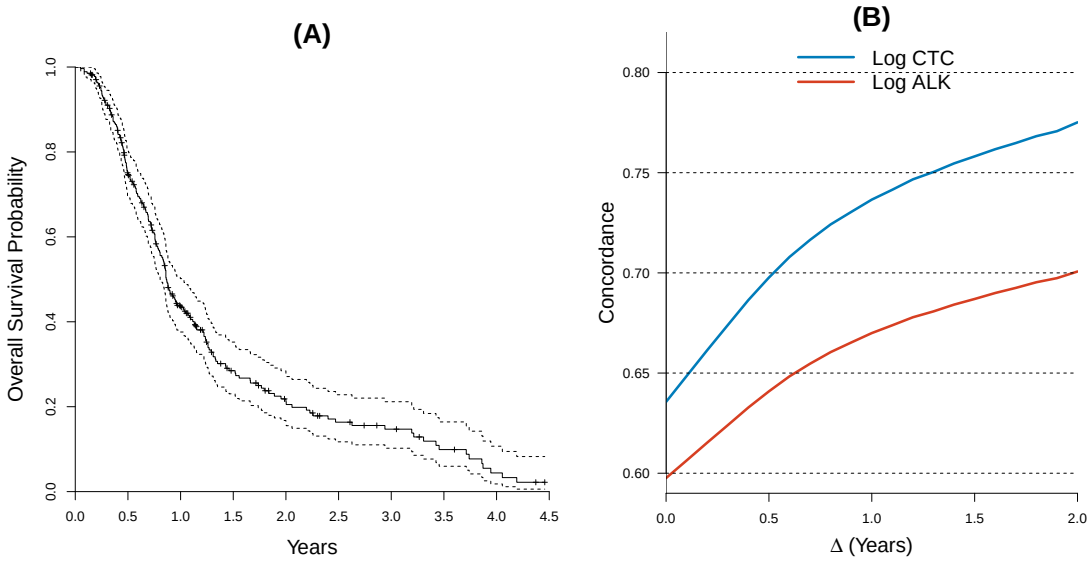


Figure 4: (A) Estimated overall survival for the 421 patients with clear cell renal cell carcinoma starting from the date of diagnosis. (B) Estimated dCP for the risk score based on SSIGN and age for various degrees of separation in survival times (Δ).

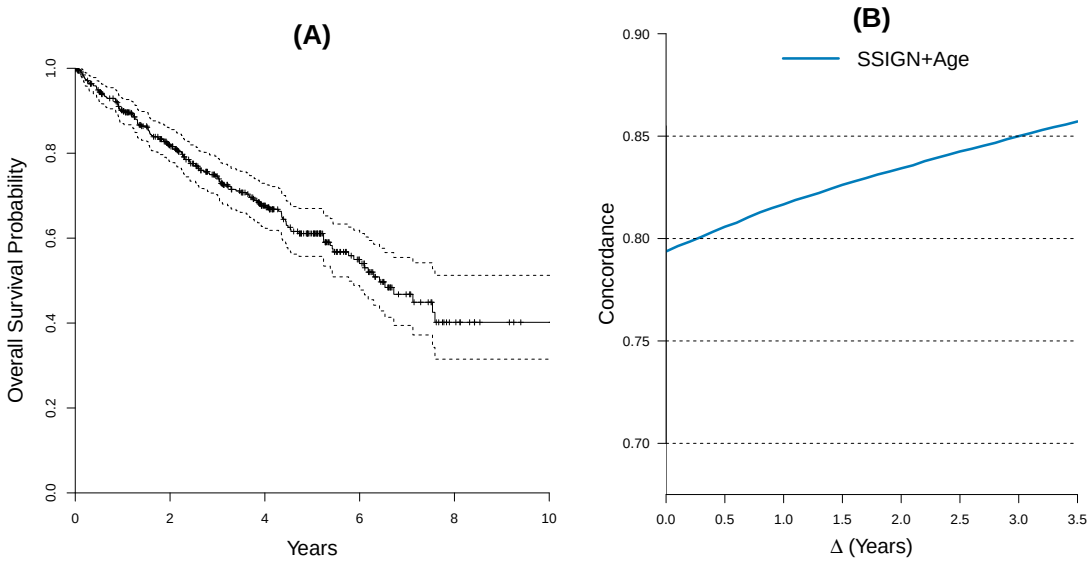


Figure 5: (A) The simulation results for 225, 450 and 900 observations in the proportional odds simulation. The value of τ was fixed at 4. (B) The average estimated dCP for 150, 300 and 600 observations across the values of Δ in the proportional hazards simulation. The value of τ was fixed at 2.

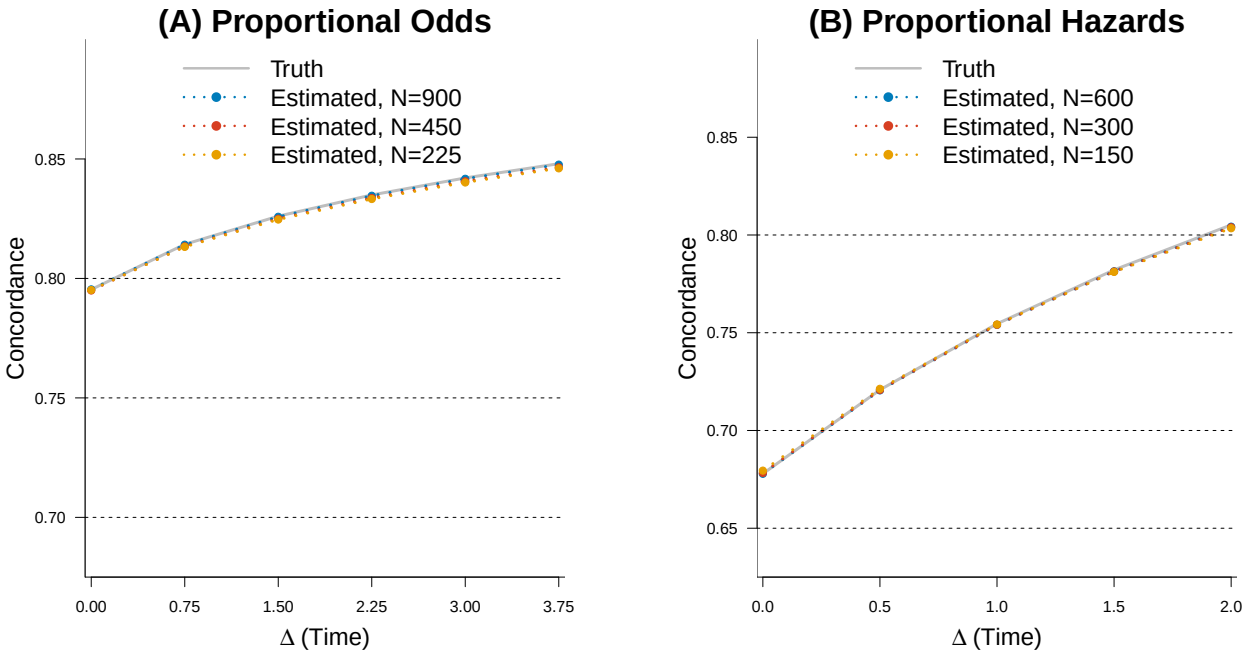


Figure 6: Simulation results for two scenarios evaluating different degrees of deviation from proportional hazards. (A) The average estimated dCP when survival times were generated using a Weibull random variable with shape equal to $1 - 0.04X$, dependent on the underlying marker X. (B) The average estimated dCP when survival times were generated using a Weibull random variable with shape equal to $1 - 0.09X$.

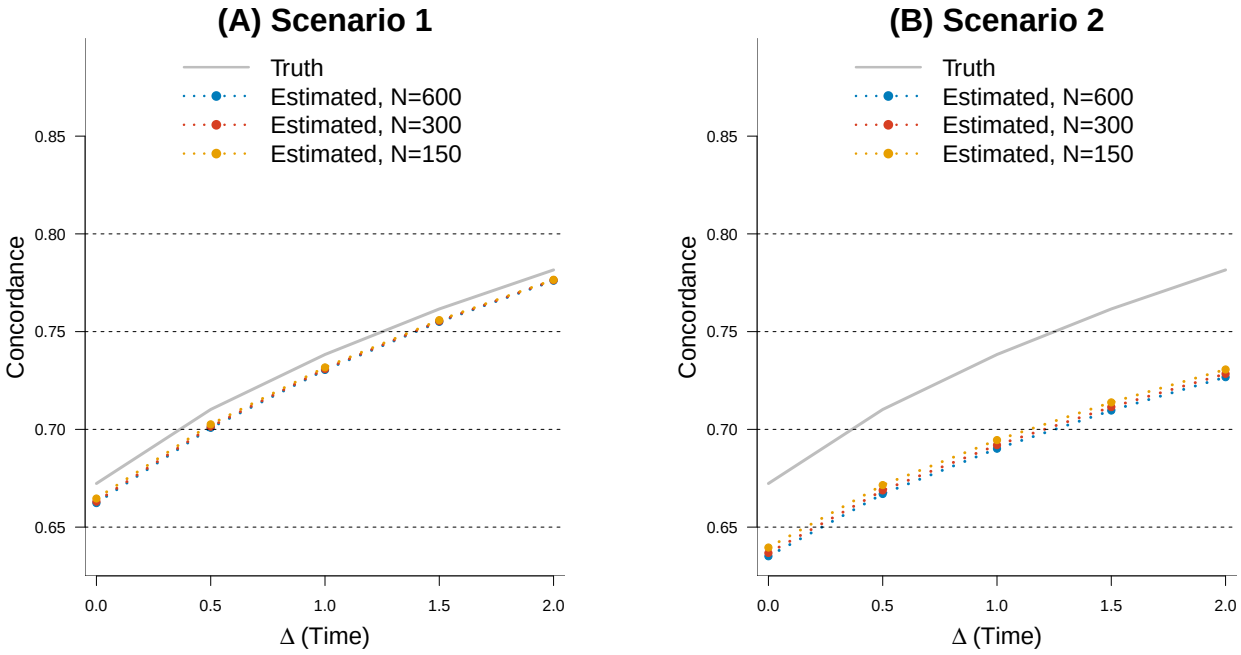


Table 1: Proportional Odds: estimated bias, root mean square error, and the estimated standard error compared to the simulation standard error averaged over 2,000 simulation iterations. The columns represent the increasing values of Δ from 0 to 3.75. The value of τ was set to 4.

N	$\Delta = 0$	0.75	1.5	2.25	3.0	3.75
True Probability						
	0.7954	0.8143	0.8261	0.8349	0.8421	0.8480
Average Bias						
225	-0.0002	-0.0011	-0.0014	-0.0017	-0.0019	-0.0019
450	-0.0004	-0.0008	-0.0011	-0.0012	-0.0014	-0.0015
900	0.0001	-0.0001	-0.0002	-0.0003	-0.0003	-0.0004
Square-root Mean Square Error (rMSE)						
225	0.0230	0.0238	0.0241	0.0242	0.0243	0.0244
450	0.0163	0.0168	0.0169	0.0170	0.0171	0.0172
900	0.0114	0.0118	0.0119	0.0119	0.0119	0.0119
Estimated Standard Error/Simulation Standard Error						
225	0.9962	0.9916	0.9895	0.9905	0.9936	1.0008
450	0.9931	0.9955	0.9953	0.9943	0.9948	0.9963
900	0.9981	1.0002	1.0003	1.0019	1.0043	1.0059

Table 2: Proportional Hazards: estimated bias, root mean square error, and the estimated standard error compared to the simulation standard error averaged over 2,000 simulation iterations. The columns represent the increasing values of Δ from 0 to 2. The value of τ was set to 2.

N	$\Delta = 0$	0.5	1	1.5	2
True Probability					
	0.6779	0.7208	0.7546	0.7821	0.8052
Average Bias					
150	0.0016	0.0006	-0.0003	-0.001	-0.0017
300	0.0004	-0.0001	-0.0005	-0.0008	-0.0012
600	<0.0001	-0.0003	-0.0005	-0.0007	-0.0009
Square-root Mean Square Error (rMSE)					
150	0.025	0.0291	0.0313	0.0324	0.033
300	0.0174	0.0203	0.0218	0.0226	0.0231
600	0.0119	0.0138	0.0149	0.0154	0.0157
Estimated Standard Error/Simulation Standard Error					
150	0.9555	0.9486	0.9477	0.9479	0.9496
300	0.9633	0.9606	0.9588	0.9598	0.9572
600	0.9946	0.9954	0.9960	0.9955	0.9942

Table 3: Non-proportional Hazards Scenario 1: estimated bias, root mean square error, and the estimated standard error compared to the simulation standard error when survival times were generated using a Weibull random variable with shape equal to $1-0.04X$, dependent on the underlying marker X. The columns represent the increasing values of Δ from 0 to 2. The value of τ was set to 2.

N	$\Delta = 0$	0.5	1	1.5	2
True Probability					
	0.6723	0.7102	0.7383	0.7616	0.7816
Average Bias					
150	-0.0077	-0.0076	-0.0066	-0.0057	-0.005
300	-0.0093	-0.0088	-0.0074	-0.0062	-0.0052
600	-0.0100	-0.0094	-0.0079	-0.0066	-0.0055
Square-root Mean Square Error (rMSE)					
150	0.0274	0.0321	0.0346	0.0361	0.0371
300	0.0205	0.0234	0.0248	0.0255	0.0261
600	0.0161	0.0176	0.0180	0.0183	0.0185
Estimated Standard Error/Simulation Standard Error					
150	0.9401	0.9316	0.9289	0.9283	0.9279
300	0.9562	0.9532	0.9511	0.9515	0.9491
600	0.9870	0.9872	0.9866	0.9856	0.9833

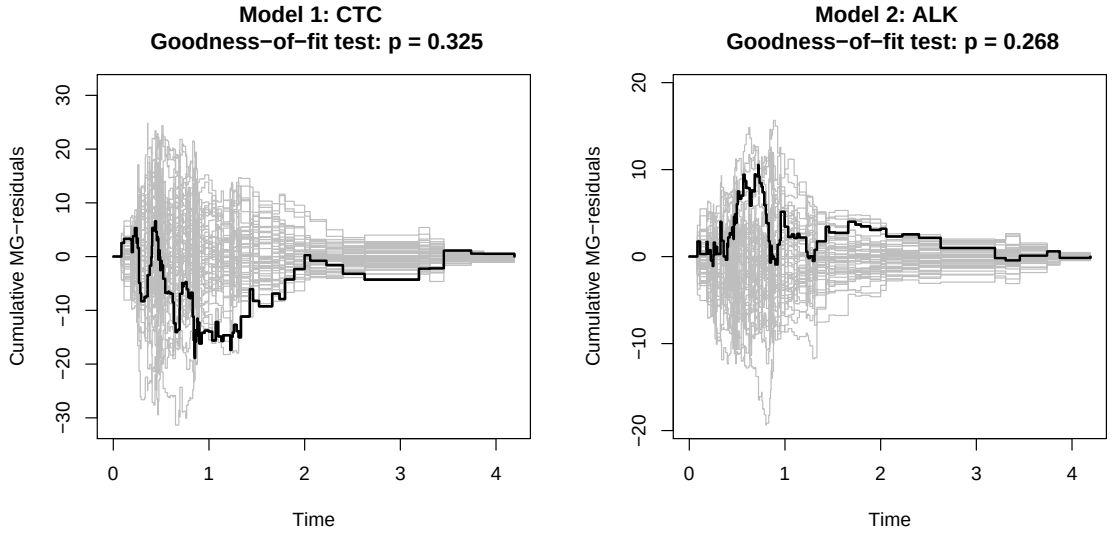
Table 4: Non-proportional Hazards Scenario 2: estimated bias, root mean square error, and the estimated standard error compared to the simulation standard error when survival times were generated using a Weibull random variable with shape equal to $1-0.09X$, dependent on the underlying marker X. The columns represent the increasing values of Δ from 0 to 2. The value of τ was set to 2.

N	$\Delta = 0$	0.5	1	1.5	2
True Probability					
	0.6737	0.713	0.7427	0.7672	0.7882
Average Bias					
150	-0.0327	-0.0387	-0.0438	-0.0478	-0.0509
300	-0.0355	-0.0413	-0.0462	-0.05	-0.0529
600	-0.0372	-0.0432	-0.0481	-0.0518	-0.0547
Square-root Mean Square Error (rMSE)					
150	0.0444	0.0531	0.0593	0.0641	0.0678
300	0.0414	0.0488	0.0544	0.0587	0.0619
600	0.0401	0.0469	0.0522	0.0562	0.0593
Estimated Standard Error/Simulation Standard Error					
150	0.9139	0.9053	0.901	0.8982	0.8953
300	0.9303	0.9259	0.921	0.9197	0.918
600	0.9504	0.9483	0.9464	0.9443	0.9414

A Supplementary Web Material

A.1 *Assessment of the proportional hazards assumption in Section 3.1*

Figure S1: Assessment of model fit for the two proportional hazards regression models in metastatic prostate cancer in Section 3.1. Both the goodness-of-fit tests and the estimated martingale residuals indicate no apparent deviation from proportional hazards for either model. Model estimates, the goodness-of-fit statistics, and residuals were estimated using the *timereg* package in R.



A.2 *Assessment of the proportional odds assumption in Section 3.2*

Figure S2: Assessment of model fit for the proportional odds regression model of SSIGN and age in clear cell renal cell carcinoma in Section 3.2. SSIGN was modeled using a natural cubic spline. Both the goodness-of-fit test and the estimated martingale residuals indicate no apparent deviation from proportional odds. Model estimates, the goodness-of-fit statistics, and residuals were estimated using the *timereg* package in R.