BRIEFINGS IN BIOINFORMATICS. VOL 16. NO 1. 153–168
Advance Access published on 14 September 2013

doi:10.1093/bib/bbt059

On the validity of time-dependent AUC estimators

Matthias Schmid, Hans A. Kestler and Sergej Potapov

Submitted: 23rd May 2013; Received (in revised form): 5th July 2013

Abstract

Recent developments in molecular biology have led to the massive discovery of new marker candidates for the prediction of patient survival. To evaluate the predictive value of these markers, statistical tools for measuring the performance of survival models are needed. We consider estimators of discrimination measures, which are a popular approach to evaluate survival predictions in biomarker studies. Estimators of discrimination measures are usually based on regularity assumptions such as the proportional hazards assumption. Based on two sets of molecular data and a simulation study, we show that violations of the regularity assumptions may lead to over-optimistic estimates of prediction accuracy and may therefore result in biased conclusions regarding the clinical utility of new biomarkers. In particular, we demonstrate that biased medical decision making is possible even if statistical checks indicate that all regularity assumptions are satisfied.

Keywords: *molecular markers; survival analysis; time-dependent AUC*

INTRODUCTION

A key interest in biomedical research is the analysis of patient survival based on molecular data [1–3]. In cancer research, for example, gene expression signatures are used to predict the time to occurrence of metastases [4, 5], time to progression [6] and overall survival [7, 8]. To assess the prognostic performance of new biomarkers, statistical measures for evaluating the prediction accuracy of time-to-event models are needed.

Measuring the performance of time-to-event models is particularly important in situations where the predictive value of a newly discovered biomarker has to be compared with existing prediction rules based on clinical patient data [9, 10]. Because of the high-dimensional nature of gene expression data ($p \gg n$), combinations of genes that are 'significantly' associated with the outcome are often detected even if the true predictive power of the genes is small [11]. It is therefore questionable whether predictions based on molecular markers are more precise than those based on (readily available) clinical predictors. Consequently, the 'added predictive value' of gene signatures needs to be assessed [12].

Here, we consider discrimination measures for time-to-event data, which have become a widely used approach to measure the performance of biomarkers in survival studies [1, 7, 13]. The idea behind discrimination measures is to evaluate the discriminative ability of a marker, i.e. to use the marker to distinguish between (i) patients having an event at or before a particular time point and (ii) patients having the event afterwards. Consequently, time-dependent misclassification rates and receiver operating characteristic (ROC) curves can be derived for each threshold of a marker. Calculating the areas under the time-dependent ROC curves results in the time-dependent AUC curve, which is a measure of the discriminative ability of a marker at each time point under consideration [14]. By considering the area under the AUC curve, it is further possible to

Corresponding author. Matthias Schmid, Department of Statistics, University of Munich, 80539 Munich, Germany. Tel: +49 (0)89 2180 3165; Fax: +49 (0)89 2180 5308; E-mail: m.schmid@stat.uni-muenchen.de

Matthias Schmid is Professor of Computational Statistics at the University of Munich. He is former head of the Computational Biostatistics working group at the Department of Biometry at Erlangen University.

Hans A. Kestler is a senior lecturer at Ulm University. He heads the Research Group on Bioinformatics and Systems Biology and leads the research core SyStaR and the Genomics Facility.

Sergej Potapov is a postdoctoral researcher at the Department of Biometry at Erlangen University. His research interests include statistical learning and classification methods in bioinformatics.

calculate a time-independent summary measure of prediction accuracy (see the 'Methods' section). Summary measures can, for example, be computed for both a prediction model containing clinical data only and for a prediction model containing clinical data plus a newly discovered biomarker. By comparing the two summary measures, it is possible to investigate the added predictive value of the biomarker (cf. [15]).

Because time-dependent ROC and AUC curves are difficult to compute in the presence of censoring, a variety of estimators for discrimination measures has been developed ([14, 16, 17], see also the 'Methods' section for details and further references). Although these estimators are widely applied in biomarker studies, their use is not unproblematic. This is because many of the estimators rely on specific regularity assumptions (such as the Cox proportional hazards assumption) that have to be satisfied for the estimators to be valid. Consequently, if the working model assumed for the estimation of discrimination measures is not specified correctly, estimates of prediction accuracy may be biased and may therefore lead to wrong conclusions about the usefulness of new markers. This problem is often overseen in practice. Even worse, it may happen that regularity assumptions are impossible to hold in both the 'full' model (containing clinical predictors and biomarkers) and the 'reduced' model (containing the clinical predictors only) at the same time. For example, if a biomarker has a non-zero effect in a correctly specified Cox model, the proportional hazards assumption is unlikely to hold in the reduced model as well.

The main aim of the article is to compare the properties of some commonly used estimators of discrimination measures in situations where the added predictive value of a new marker needs to be evaluated. Specifically, we show that violations of regularity assumptions may result in over-optimistic estimates of prediction accuracy and may therefore lead to biased conclusions regarding the prognostic effect of new biomarkers. As will be demonstrated, this may even happen if biostatisticians carefully check all regularity assumptions.

The first part of our analysis ('Simulation Study' section) is based on a simulation study that evaluates biomarker combinations using Cox proportional hazards regression. Considering a proportional hazards working model is convenient not only because Cox regression is the predominant model in survival analysis [18, 19] but also because some of the aforementioned estimators explicitly rely on this model. The second part of the analysis ('Breast Cancer Data' section) is based on two biomarker studies that use molecular data for predicting the time to distant metastases in breast cancer patients. The first data set was collected by van de Vijver *et al.* [4] to validate a 70-gene expression signature reported by van't Veer *et al.* [20]. The second data set was collected by Desmedt *et al.* [5] to validate a 76-gene prognostic signature identified by Wang *et al.* [21]. Using these data, we investigate whether estimators of discrimination measures are able to capture the predictive value of the gene expression measurements in addition to traditional predictions based on clinical predictor variables only.

METHODS

Discrimination measures for time-to-event data

Let $T \in \mathbf{R}^+$ be a survival time and $X \in \mathbf{R}^p$ a vector of predictor variables. For example, X could be composed of a subvector representing clinical predictors (such as patient characteristics) and another subvector representing a set of molecular markers. Denote the conditional survival function of T given X by $S(t|X) = P(T > t|X = x)$. Let $f(t)$ and $S(t)$ be the unconditional probability density and survival functions of T, respectively. Let $C \in \mathbf{R}^+$ be a random censoring time and denote the observed survival time by $\tilde{T} := \min(T, C)$. The random variable $\Delta := \mathrm{I}(T \leq C)$ indicates whether $\tilde{T}$ is right-censored $(\Delta = 0)$ or not $(\Delta = 1)$. Throughout this article, we use the Cox proportional hazards model

$$S(t|X) = \exp\left(-\Lambda_0(t) \cdot \exp(X^T\beta)\right) \tag{1}$$

to estimate $S(t|x)$, where $\Lambda_0(t)$ is the cumulative baseline hazard function and β is a vector of coefficients relating the predictor variables to conditional survival. By definition, small values of the risk score $\eta := X^T\beta$ imply large expected survival times and vice versa. We further assume that β is estimated from an i.i.d. learning sample $\{(\tilde{T}_i^L, \Delta_i^L, X_i^L), i = 1, \ldots, n\}$ using the (possibly regularized) maximization of the partial log-likelihood. Denote the resulting estimate of $S(t|x)$ by $\hat{S}_n^L(t|X) = \exp\left(-\hat{\Lambda}_0^L(t) \cdot \exp(X^T\hat{\beta}^L)\right)$, where $-\hat{\Lambda}_0^L$ and $\hat{\beta}^L$ are the Breslow and maximum partial log-likelihood estimates of Λ_0 and β, respectively.

Throughout the article, we will use η as a marker for predicting patient survival. Prediction accuracy of $\hat{S}_n^L(t|X)$ (and thus of the marker $X^T\hat{\beta}^L$) will be evaluated by using an independent i.i.d. test sample $\{(\tilde{T}_j, \Delta_j, X_j), j = 1, \ldots, N\}$ that follows the same distribution as the learning data. Prediction of survival is based on the estimated marker values $\hat{\eta}_j := X_j^T\hat{\beta}^L, j = 1, \ldots, N$.

The discrimination measures considered in this article make use of the fact that T can be considered as a time-dependent binary variable with values 'event' and 'no event'. Consequently, at each time point $t > 0$, ROC analysis for the evaluation of binary outcomes can be used to distinguish between patients having an event ('cases') and those having no event ('controls', see Heagerty & Zheng [14]). Time-dependent 'cumulative and incident true positive rates' are defined as

$$\text{TPR}^C(c, t) := P(\eta > c|T \leq t), \tag{2}$$

$$\text{TPR}^I(c, t) := P(\eta > c|T = t), \tag{3}$$

respectively, where c is a threshold of interest. The two TPRs defined earlier in the text have a slightly different interpretation: When considering TPR^C, the aim is to distinguish between 'observations having an event at or before t' and 'observations having an event after t'. Thus, for fixed t, observations with $T_j \leq t, j = 1, \ldots, N$ are considered as cases, whereas observations with $T_j > t$ are considered as controls. Similarly, when considering TPR^I, the aim is to distinguish between 'observations having an event at t' and 'observations having an event after t'. In this case, observations with $T_j = t$ are considered as cases, whereas observations with $T_j > t$ are considered as controls. Although TPR^I has been adopted by most methodologists [1], we will analyze estimators of both versions of TPR.

The time-dependent 'dynamic false positive rate' is defined as

$$\text{FPR}^I(c, t) := P(\eta > c|T > t), \tag{4}$$

see [14]. It is called 'dynamic' because FPR^D depends on the time point t under consideration, whereas the 'static false positive rate' (which is not considered in this article) uses a single 'static' time point t_{stat} as reference point. Summarizing TPR^C, TPR^I and FPR^D results in 'cumulative/dynamic' and 'incident/dynamic ROC curves' defined as

$$\text{ROC}^{C/D}(c, t) := \left\{\text{FPR}^D(c, t), \text{TPR}^C(c, t)\right\}, \tag{5}$$

$$\text{ROC}^{I/D}(c, t) := \left\{\text{FPR}^D(c, t), \text{TPR}^I(c, t)\right\}, \tag{6}$$

respectively. Calculating the area under the cumulative/dynamic and incident/dynamic ROC curves results in time-dependent 'cumulative/dynamic' and 'incident/dynamic AUC curves' (denoted by $\text{AUC}^{C/D}(t)$ and $\text{AUC}^{I/D}(t)$, respectively). By definition, time-dependent AUC curves quantify the discriminative ability of a marker at each time point under consideration.

When comparing different survival models, it can be helpful to use a summary index that evaluates the 'overall' accuracy of a prediction rule. In this respect, the above-defined measures have the disadvantage that they are time-dependent and therefore need to be evaluated at each individual time point. To obtain a time-independent discrimination index, the area under the time-dependent AUC curve can be computed. For incident/dynamic AUC, Heagerty & Zheng [14] suggested to use the index

$$C^* := \int_t \text{AUC}^{I/D}(t) \cdot w(t)dt \tag{7}$$

with weights $w(t) := 2f(t)S(t)$. The authors showed that C^* equals the probability $P(\eta_{j1} > \eta_{j2}|T_{j1} < T_{j2})$, which is a concordance index measuring the probability that observations with large values of η have shorter survival times than observations with small values of η. (Here, η_{j1}, η_{j2}, T_{j1} and T_{j2} denote the markers and survival times of two randomly chosen observations in the test sample.) The concordance index C^* equals 0.5 in case a non-informative marker (independent of T) is used. Markers predicting better than chance should therefore result in values of C^* lying in the interval (0.5, 1], provided that there is a monotonic relationship between η and T.

In contrast to C^*, which is based on incident/dynamic AUC, no measures summarizing the cumulative/dynamic AUC have been derived yet [14]. We therefore propose to use the expected value

$$C^*_{\text{cum}} := \text{E}_T(\text{AUC}^{C/D}(T)) = \int_t \text{AUC}^{C/D}(t) \cdot f(t)dt \tag{8}$$

as a summary measure of $\text{AUC}^{C/D}$.

Estimators of discrimination measures

In the literature, various approaches to estimate cumulative/dynamic discrimination measures have been proposed (e.g. [14, 16, 17]). Some of these approaches rely on estimators of time-dependent true and false-positive rates (which can subsequently be used to calculate estimates of AUC, C^* and C^*_{cum}), whereas other approaches rely on estimating C^* and C^*_{cum} directly. Because not all of the approaches provide formulae for the estimation of TPR/FPR, we focus on the time-independent discrimination indices C^* and C^*_{cum}. The following estimators will be considered in our numerical studies:

Estimators based on incident/dynamic AUC

(i) Estimation approach by Heagerty & Zheng [14]. Heagerty & Zheng proposed two approaches to estimate $\left\{\mathrm{FPR}^D(c,t), \mathrm{TPR}^I(c,t)\right\}$. The first estimator, which is based on the assumptions of a Cox proportional hazards model, is given by

$$\widehat{\mathrm{TPR}}^{I}_{\mathrm{HZ}_{\mathrm{Cox}}}(c,t) := \sum_{j=1}^{N} \frac{\mathrm{I}(\hat{\eta}_j > c)\mathrm{I}(\tilde{T}_j \geq t)\exp(\hat{\eta}_j)}{W(t)}, \tag{9}$$

$$\widehat{\mathrm{FPR}}^{D}_{\mathrm{HZ}_{\mathrm{Cox}}}(c,t) := \sum_{j=1}^{N} \frac{\mathrm{I}(\hat{\eta}_j > c)\mathrm{I}(\tilde{T}_j > t)}{\sum_{k=1}^{N}\mathrm{I}(\tilde{T}_k > t)} \tag{10}$$

with $W(t) := \sum_k \mathrm{I}(\tilde{T}_k \geq t)\exp(\hat{\eta}_k)$. Estimates of $\mathrm{AUC}^{I/D}$ and C^* are based on numerical integration of Equations (9) and (10) using the empirical probability density and survival functions of T estimated from the learning data to calculate the weights $w(t)$ in Equation (7).

In cases where the proportional hazards assumption is critical, Heagerty & Zheng [14] suggested to use an alternative estimator of $\left\{\mathrm{FPR}^D(c,t), \mathrm{TPR}^I(c,t)\right\}$ that is based on a varying-coefficient Cox model (see Section 3.2 of [14]). Both estimators suggested by Heagerty & Zheng are consistent for $\left\{\mathrm{FPR}^D(c,t), \mathrm{TPR}^I(c,t)\right\}$ in case the assumptions of the respective working models are satisfied.

(ii) Estimation approach by Song & Zhou [16]. Similar to Heagerty & Zheng [14], Song & Zhou based their approach on the assumptions of a Cox proportional hazards model. The authors suggested to estimate TPR^I and FPR^D by

$$\widehat{\mathrm{TPR}}^{I}_{\mathrm{SZ}}(c,t) := \sum_{j=1}^{N} \frac{\exp(\hat{\eta}_j)\hat{S}^L_n(t|X_j)\mathrm{I}(\hat{\eta}_j > c)}{\sum_{k=1}^{N}\exp(\hat{\eta}_k)\hat{S}^L_n(t|X_k)}, \tag{11}$$

$$\widehat{\mathrm{FPR}}^{D}_{\mathrm{SZ}}(c,t) := \sum_{j=1}^{N} \frac{\hat{S}^L_n(t|X_j)\mathrm{I}(\hat{\eta}_j > c)}{\sum_{k=1}^{N}\hat{S}^L_n(t|X_k)}. \tag{12}$$

The estimators given in Equations (11) and (12) are consistent in case the assumptions of the Cox model are met. In contrast to the estimators suggested by Heagerty & Zheng [14], Song & Zhou's approach remains valid even if the censoring times depend on X [1]. Analogously to the approach of Heagerty & Zheng [14], estimates of $\mathrm{AUC}^{I/D}$ and C^* are obtained using numerical integration of Equations (11) and (12).

(iii) Concordance probability estimator by Gönen & Heller [22]. Gönen & Heller's approach is directly based on the assumptions of a Cox proportional hazards model. Let $\hat{\eta}_{ij} := (X_i - X_j)^T\hat{\beta}^L$. Gönen & Heller [22] showed that the concordance index C^* is consistently estimated by

$$C^*_{\mathrm{GH}} := \frac{2}{N(N-1)}\sum_{i<j}\frac{\mathrm{I}(\hat{\eta}_{ij} < 0)}{1+\exp(\hat{\eta}_{ij})} + \frac{\mathrm{I}(\hat{\eta}_{ji} < 0)}{1+\exp(\hat{\eta}_{ji})} \tag{13}$$

in case the assumptions of the Cox model are met. In this case, C^*_{GH} remains a valid estimator of C^* even if the censoring times depend on X.

(iv) Censoring-adjusted C-statistic by Uno *et al.* [17]. Uno *et al.* [17] suggested to estimate C^* by

$$C^*_{\mathrm{UnoC}} := \frac{\sum_{j,k}\left(\hat{S}^L_{C,n}(\tilde{T}_j)\right)^{-2}\mathrm{I}(\tilde{T}_j < \tilde{T}_k)\mathrm{I}(\hat{\eta}_j > \hat{\eta}_k)\Delta_j}{\sum_{j,k}\left(\hat{S}^L_{C,n}(\tilde{T}_j)\right)^{-2}\mathrm{I}(\tilde{T}_j < \tilde{T}_k)\Delta_j} \tag{14}$$

where $\hat{S}^L_{C,n}(t)$ denotes the Kaplan–Meier estimator of the unconditional survival function of C (estimated from the learning data). In contrast to the estimators in (i) to (iii), the estimator in Equation (14) does not rely on the assumptions of a Cox model. It is assumed, however, that C is independent of X ('random censoring assumption'). Under this assumption, C^*_{UnoC} is a consistent estimator of C^*. Consistency of C^*_{UnoC} is ensured by using the weights $\Delta_j/(\hat{S}^L_{C,n}(\tilde{T}_j))^2$, which account for the inverse probability that an observation in the test data is censored ['inverse probability of censoring weighted' (IPCW) estimation [23]].

Estimators based on cumulative/dynamic AUC

(i) Estimation approach by Song & Zhou [16]. In addition to the estimator of incident TPR given in Equation (11), Song & Zhou [16] proposed an estimator of the cumulative TPR. This estimator is defined as

$$\widehat{\mathrm{TPR}}^{C}_{SZ}(c,t) := \frac{\sum_{j=1}^{N}(1-\hat{S}^L_n(t|X_j))\mathrm{I}(\hat{\eta}_j \geq c)}{\sum_{j=1}^{N}(1-\hat{S}^L_n(t|X_j))}. \tag{15}$$

The estimator in Equation (15) is consistent if the assumptions of the Cox model hold. Estimates of C^*_{cum} are based on numerical integration of Equations (11) and (15) using the empirical probability density function of T estimated from the learning data to calculate $f(t)$ in Equation (8).

(ii) Estimation approach by Uno *et al.* [10]. Uno *et al.* [10] proposed to estimate TPR^C and FPR^D by

$$\widehat{\mathrm{TPR}}^C_{\mathrm{Uno}}(c,t) := \frac{\sum_j \Delta_j \mathrm{I}(\hat{\eta}_j > c \cap \tilde{T}_j \le t)/\hat{S}^L_{C,n}(\tilde{T}_j)}{\sum_j \Delta_j \mathrm{I}(\tilde{T}_j \le t)/\hat{S}^L_{C,n}(\tilde{T}_j)}, \tag{16}$$

$$\widehat{\mathrm{FPR}}^D_{\mathrm{Uno}}(c,t) := 1 - \frac{\sum_j \Delta_j \mathrm{I}(\hat{\eta}_j \le c \cap \tilde{T}_j > t)}{\sum_j \mathrm{I}(\tilde{T}_j > t)}. \tag{17}$$

Unlike the estimators proposed by Song & Zhou, Equations (16) and (17) do not rely on the assumptions of a Cox model. If the censoring times are independent of X, Equations (16) and (17) are consistent estimators of TPR^C and FPR^D, respectively. Analogously to the censoring-adjusted C-statistic proposed by Uno *et al.* [17], consistency of Equation (16) is ensured by using the IPC weights $\Delta_j/\hat{S}^L_{C,n}(\tilde{T}_j)$. Estimates of $\mathrm{AUC}^{C/D}$ and C^*_{cum} are obtained by using numerical integration of Equations (16) and (17).

Remark #1: Another popular estimator of the concordance probability C^* is 'Harrell's C for survival data' [24, 25]. Because Harrell's C is usually upward biased in the presence of censoring [26, 27], we will not consider this estimator in our numerical studies.

Remark #2: Apart from the estimators presented in the 'Estimators of Discrimination Measures' section, several other approaches for the estimation of discrimination measures exist (e.g. [28, 29]). Because the aim of this article is to demonstrate that violations of regularity assumptions may lead to biased medical decision making, we only focus on a selection of these estimators here.

SIMULATION STUDY

Data-generating model

The aim of the simulation study was to analyze the behavior of the estimators presented in the 'Estimators of Discrimination Measures' section in situations where the underlying regularity assumptions were violated. Here, we focus on the situation where the added predictive value of a newly discovered biomarker needs to be evaluated. We consider violations of both the proportional hazards assumption and the random censoring assumption, which are the key assumptions for guaranteeing the validity of the estimators presented in the 'Estimators of Discrimination Measures' section.

For the simulation study, we considered two standard normally distributed variables $X^{(1)}$ (representing a new biomarker) and $X^{(2)}$ (representing a combination of classical predictor variables). For the sake of simplicity, we assumed $X^{(1)}$ and $X^{(2)}$ to be independent. Survival times were generated by using the log normal regression model

$$\log(T) = U \cdot \left(X^{(1)}\gamma_1 + X^{(2)}\gamma_2\right) + \sigma \cdot Z, \tag{18}$$

where $\gamma_1 = 0.1$ and $\gamma_2 = 0.5$ were the regression coefficients of $X^{(1)}$ and $X^{(2)}$, respectively, and σ was the standard deviation of the noise that was added to the predictor $\gamma := U \cdot (X^{(1)}\gamma_1 + X^{(2)}\gamma_2)$. The noise variable Z was assumed to follow a normal distribution with zero mean and standard deviation $\sigma = 0.6$. The random variable U was a binary variable taking the values -1 and 1 with probability 0.5 each. Hence, U served as a grouping variable that resulted in different signs of the linear predictor $X^{(1)}\gamma_1 + X^{(2)}\gamma_2$. By definition, $T|X$ followed a log normal distribution in each of the two subgroups defined by U.

In the remainder of this section, we will assume that the existence of the grouping variable U is unknown to the biostatistician analyzing the data. Regarding biomedical practice, this is a realistic assumption because unobserved heterogeneity is a common issue in biomarker studies. With U being ignored, however, it is straightforward to show that any linear combination of the predictors $X^{(1)}$ and $X^{(2)}$ in model (18) results in markers with no discriminative ability at all. Specifically, any linear predictor $\hat{\eta}$ obtained from Cox regression results in a non-informative marker with $C^* = C^*_{\mathrm{cum}} = 0.5$, regardless of whether the full model containing both $X^{(1)}$ and $X^{(2)}$ or the reduced model containing $X^{(2)}$ only is used. Also, the proportional hazards assumption is violated, regardless of whether U is included in the Cox model.

To analyze whether the estimators of C^* and C^*_{cum} were able to reflect this situation, we applied Cox regression to 500 learning samples $\{(\tilde{T}^{\mathrm{L}}_{\mathrm{i}}, \Delta^{\mathrm{L}}_{\mathrm{i}}, X^{\mathrm{L}}_{\mathrm{i}}), i = 1, \ldots, n\}$ of size $n = 100$ each. Both the full model and the reduced model were fitted to each of the learning samples. Discrimination measures for the marker $\hat{\eta}$ (being a linear combination of

$X^{(1)}$ and $X^{(2)}$) were estimated from 500 independently generated test samples $\{(\tilde{T}_j, \Delta_j, X_j), j = 1, \ldots, N\}$. In the first step, we considered independent censoring times that were generated from an exponential distribution with rate 0.66. This strategy resulted in a moderately high censoring rate of ~50%.

All simulations were carried out with the R System for Statistical Computing (version 3.0.0, [30]) using the add-on packages **risksetROC** [31] and **survAUC** [32].

Results of the simulation study

Figure 1 shows the results obtained for the estimators of C^* proposed by Uno *et al.* (denoted by C^*_{UnoC}), Heagerty & Zheng (denoted $C^*_{\mathrm{HZ}_{\mathrm{Cox}}}$ and C^*_{HZ}), Song & Zhou (denoted by C^*_{SZ}) and Gönen & Heller (denoted by C^*_{GH}). In case of the estimators proposed by Heagerty & Zheng, $C^*_{\mathrm{HZ}_{\mathrm{Cox}}}$ refers to the Cox-based estimator, whereas C^*_{HZ} refers to the estimator based on a varying-coefficient Cox model.

As seen from Figure 1, the finite sample bias of the IPCW-based estimator C^*_{UnoC} is close to zero. Moreover, no systematic difference in estimated prediction accuracy between the full (shown in gray) and reduced Cox models (shown in white) is observable (see also Figure 2 for an additional illustration). This result, which reflects the true discriminatory power of the full and reduced models, is due to the fact that censoring times were generated independently of T, and that C^*_{UnoC} is consistent for C^* as long as the random censoring assumption holds. A similar result was obtained for the IPCW-based estimator of C^*_{cum} proposed by Uno *et al.* [10] (denoted by $C^*_{\mathrm{Uno}_{\mathrm{cum}}}$, see Figure 1).

In contrast to the IPCW-based estimators C^*_{UnoC} and $C^*_{\mathrm{Uno}_{\mathrm{cum}}}$, the Cox-based estimators in Figure 1 all show an upward bias. Although the bias is not large in absolute value, it is problematic because the Cox-based estimators indicate a positive effect of the non-informative marker $\hat{\eta}$ on prediction accuracy. Hence, by relying on the Cox-based estimators, we would wrongly conclude that $\hat{\eta}$ increases the discriminatory power of the prediction model. This result was not only observed for the Cox-based estimators of C^* but also for the Cox-based estimator of C^*_{cum} proposed by Song & Zhou [16] (denoted by $C^*_{\mathrm{SZ}_{\mathrm{cum}}}$).

On the other hand, the estimators $C^*_{\mathrm{HZ}_{\mathrm{Cox}}}$ and C^*_{HZ} are 'valid' in the sense that they do not indicate an additional effect of the biomarker $X^{(1)}$ on prediction accuracy. The values of $C^*_{\mathrm{HZ}_{\mathrm{Cox}}}$ and C^*_{HZ} obtained from the full and reduced models are almost identical. In contrast, the Cox-based estimators C^*_{SZ}, C^*_{GH} and $C^*_{\mathrm{SZ}_{\mathrm{cum}}}$ uniformly indicate superiority of the full model. Relying on the latter three estimators would therefore lead to the wrong conclusion that adding the biomarker $X^{(1)}$ to the prediction model increases discriminatory power (Figure 2).

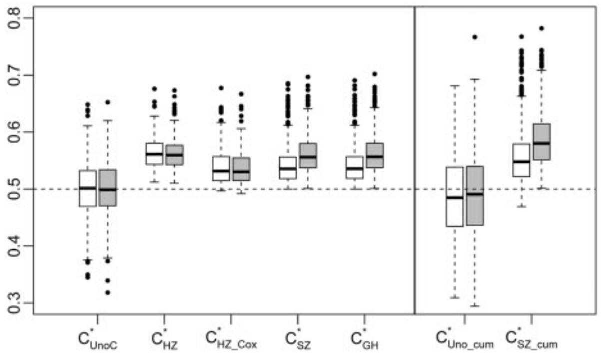

Figure 1: Results of the simulation study. Boxplots correspond to estimates of C^* and C^*_{cum} using the methods described in the 'Estimators of Discrimination Measures' section. Predictions were obtained by fitting Cox regression models to 500 learning samples (of size $n = 100$ each) generated from Equation (18) and by applying the resulting prediction rules to 500 independent test samples (of size $N = 100$ each). Censoring times were independent of T, and the censoring rate was approximately equal to 50%. White boxplots correspond to the reduced model (using $X^{(2)}$ only), whereas gray boxplots correspond to the full model (using both $X^{(1)}$ and $X^{(2)}$). The dashed horizontal line corresponds to the true value of C^* and C^*_{cum}.

The model selection bias of the Cox-based estimators can be attributed to the fact that these estimators explicitly rely on the proportional hazards assumption. Because this assumption is violated in the true data-generating model, Cox-based estimation of C^* and C^*_{cum} is no longer valid. Hence, the estimated baseline hazard and the coefficient estimates obtained from Cox regression are biased. This bias is implicitly incorporated in the Cox-based estimators of C^* and C^*_{cum} presented in Figures 1 and 2. The results presented in this subsection are largely insensitive to the choice of the residual standard deviation σ (see Appendix).

Checks of the proportional hazards assumption

As demonstrated in the previous subsection, generating simulated data from model (18) implies the

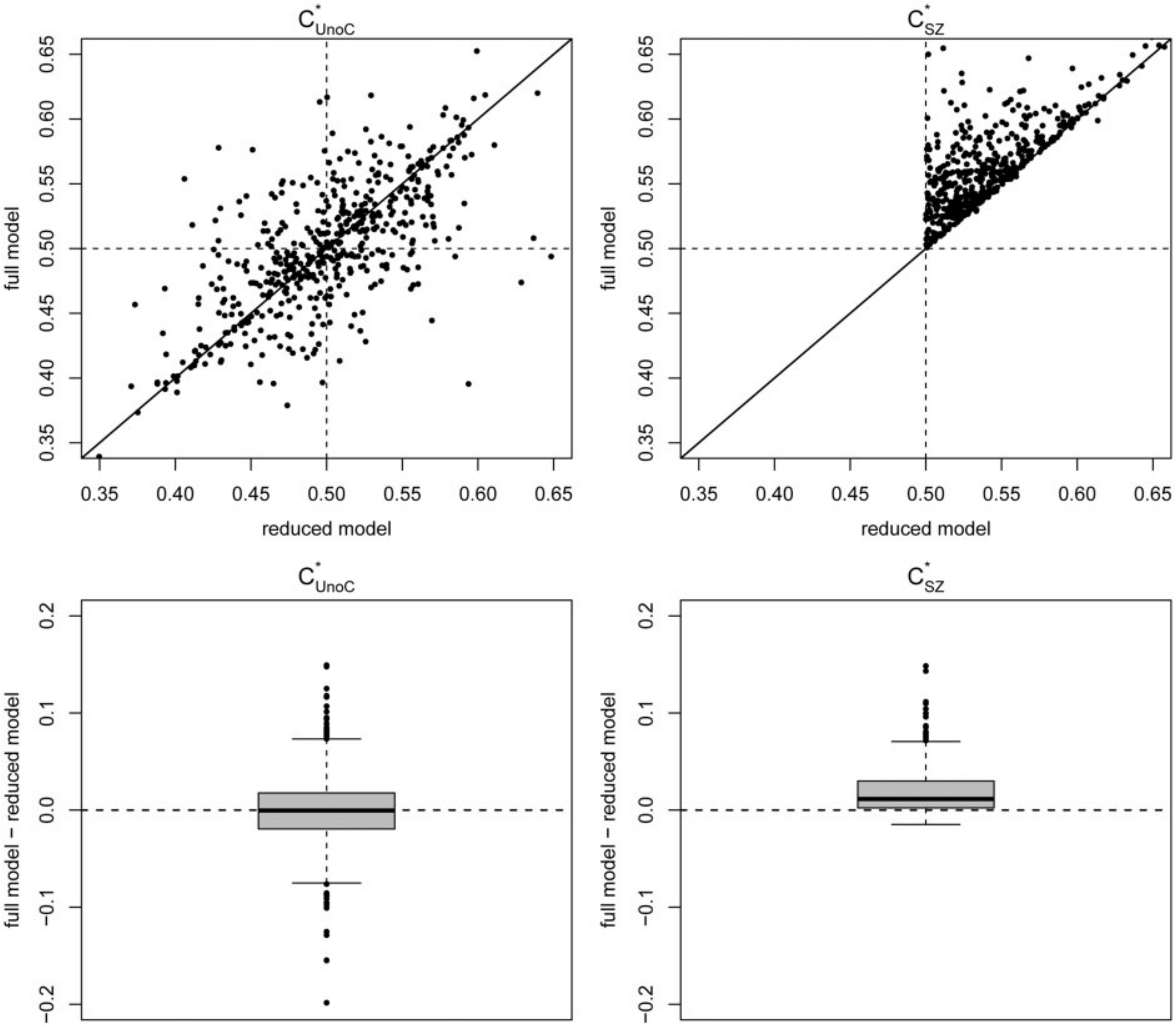

Figure 2: Results of the simulation study. The upper left panel shows the estimated values of the IPCW-based estimator C^*_{UnoC} for both the full and reduced models [as obtained from the 500 learning and test samples generated from Equation (18)]. The upper right panel shows the values of the Cox-based estimator C^*_{SZ} that were obtained from the same samples and models. Obviously, C^*_{UnoC} does not systematically prefer any of the two models, thereby indicating that the biomarker $X^{(1)}$ has no additional effect on prediction accuracy. In contrast, C^*_{SZ} wrongly indicates that $X^{(1)}$ increases the discriminatory power of the model. The relatively large variance of C^*_{UnoC} (seen in the upper two panels) is of minor interest when the focus is on the 'ranking' of the two models: In 88.2% of the samples, C^*_{SZ} wrongly indicates a positive effect of $X^{(1)}$ (despite its relatively small variance), whereas C^*_{UnoC} indicates this effect in only 48.8% of the samples (see the boxplots of the differences between full and reduced models presented in the lower two panels).

violation of the proportional hazards assumption and therefore a bias in the Cox-based estimators of C^* and C^*_{cum}. Consequently, it has to be investigated whether departures from the proportional hazards assumption could have been detected by analyzing the 500 sets of learning data.

To this purpose, we checked the proportional hazards assumption by applying $T(G)$-tests [33] to each of the 500 pairs of full and reduced Cox models. The tests are based on scaled Schoenfeld residuals and examine the hypothesis of non-proportionality for each of the covariates included in a Cox regression model. An additional 'global' $T(G)$-test allows for checking the proportional hazards assumption in cases where the Cox model contains more than one predictor variable.

The P-values obtained from the tests are presented in Figure 3. Obviously, the tests failed to detect significant departures from the proportional hazards assumption in 85.8% of the full and reduced model

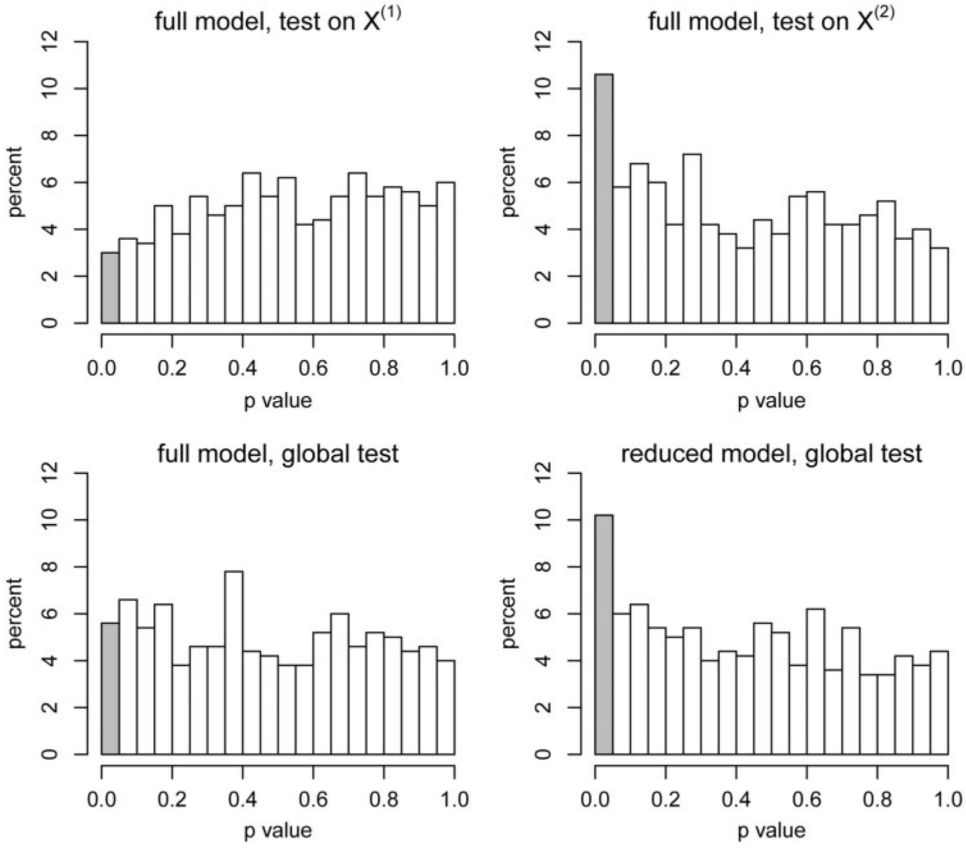

Figure 3: Checks of the proportional hazards assumption. The four panels show the distribution of *P*-values that were obtained by applying *T*(*G*)-tests to the Cox regression models discussed in the 'Results of the Simulation Study' section. The gray bars correspond to the fraction of tests with a *P*-value smaller than 0.05 (indicating significant departures from the proportional hazards assumption). In 71 of the 500 samples, at least one of the four tests resulted in $P < 0.05$. Consequently, the proportional hazards assumption was wrongly adopted in 85.8% of the samples.

fits (at significance level $\alpha = 0.05$). Consequently, in 429 of the 500 samples, we would have wrongly concluded that the proportional hazards assumption was satisfied—and hence would have trusted the results of the Cox-based estimators. Again, this result is largely insensitive to the choice of the standard deviation σ in Equation (18) (see Appendix).

In conclusion, Figure 3 demonstrates that biased medical decision making is possible even if a biostatistician carefully checks all underlying regularity assumptions.

Effects of sample size and censoring rate

The results of the simulation study presented in the 'Results of the Simulation Study' section are based on samples of size $n = N = 100$ and a censoring rate of ~50%. Although these numbers are typical for many biomarker studies, one might speculate that the results presented in Figures 1 and 2 are affected by sample size and censoring rate. To analyze these effects, we repeated the simulation study, this time using a larger sample size ($n = N = 1000$) and zero percent censoring.

The results are presented in Figure 4. Here, the upward bias of the Cox-based estimators is much smaller than the corresponding bias in Figure 1. However, the estimators C^*_{SZ}, C^*_{GH} and $C^*_{\mathrm{SZ_{cum}}}$ still show a tendency to prefer the full model over the null model. In contrast, the IPCW-based estimators C^*_{UnoC} and $C^*_{\mathrm{Uno_{cum}}}$ remain unbiased and do not indicate a positive effect of the non-informative biomarker $X^{(2)}$ on prediction accuracy. In summary, the results presented in Figure 4 suggest that biased medical decision making is possible even in favorable settings where the sample size is large and where the rate of censored observations is small.

Violations of the random censoring assumption

The IPCW-based estimators C^*_{UnoC} and $C^*_{\mathrm{Uno_{cum}}}$, which outperformed the Cox-based estimators of C^* and C^*_{cum} in the previous subsections, rely on

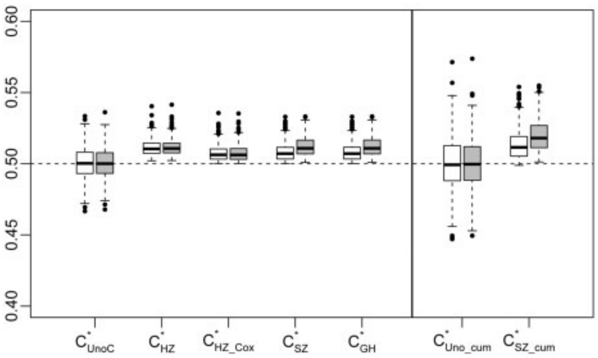

Figure 4: Results of the simulation study. Boxplots correspond to estimates of C^* and C^*_{cum} using the methods described in the 'Estimators of Discrimination Measures' section. Predictions were obtained from fitting Cox regression models to 500 learning samples generated from Equation (18), this time using a larger sample size ($n = N = 1000$) and 100% uncensored observations. Censoring times were independent of T. White boxplots correspond to the reduced model, whereas gray boxplots correspond to the full model. The dashed horizontal line corresponds to the true value of C^* and C^*_{cum}.

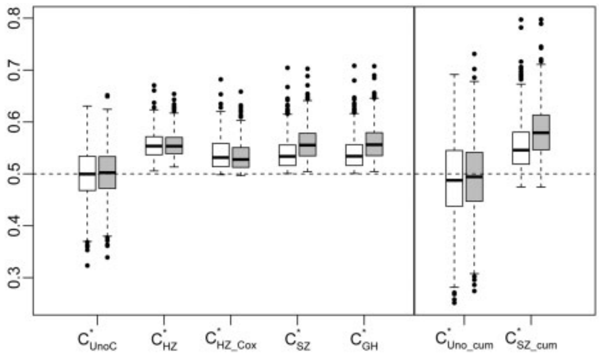

Figure 5: Results of the simulation study. Boxplots correspond to estimates of C^* and C^*_{cum} using the methods described in the 'Estimators of Discrimination Measures' section. Predictions were obtained from fitting Cox regression models to 500 learning samples generated from Equation (18). Censoring times were generated from Equation (19), which implied that censoring was no longer independent of T. White boxplots correspond to the reduced model, whereas gray boxplots correspond to the full model. The dashed horizontal line corresponds to the true value of C^* and C^*_{cum}.

the assumption that the censoring time C is independent of T. To analyze whether C^*_{UnoC} and $C^*_{\text{Uno}_{\text{cum}}}$ retain their favorable behavior in the presence of non-random censoring, we repeated the simulation study, this time generating censoring times from an exponential distribution with rate

$$\lambda_{\text{cens.viol}} := \lambda_{\text{I}} \cdot \exp(\bar{T}) / \exp(T), \qquad (19)$$

where $\bar{T}$ was the mean of T in the learning sample, and $\lambda_{\text{I}} = 0.5695$ was an additional parameter that resulted in ~50% censored observations. By choosing $\lambda_{\text{I}} = 0.5695$, the correlation between T and C^*_{HZ} was approximately equal to 0.45. Consequently, the random censoring assumption was violated, and observations with large survival times tended to have a small probability of being censored at early time points.

The results of the additional simulation study are presented in Figure 5. There is only little difference to the results presented in Figure 1. Specifically, the IPCW-based estimators C^*_{UnoC} and $C^*_{\text{Uno}_{\text{cum}}}$ are almost unaffected by the violation of the random censoring assumption (although they show a slight increase in variance).

BREAST CANCER DATA

Breast cancer data by van de Vijver *et al.*

We first analyzed a data set of 144 lymph node positive breast cancer patients that was collected by the Netherlands Cancer Institute [4]. The data set, which is publicly available as part of the R add-on package **penalized** [34], was used by van de Vijver *et al.* [4] to validate a 70-gene signature for metastasis-free survival after surgery [20]. In addition to the expression levels of the 70 genes, the data set contains five clinical predictor variables [tumor diameter, number of affected lymph nodes, estrogen receptor (ER) status, grade of the tumor and patient age]. Observed metastasis-free survival times ranged from 0.06 months to 17.66 months, with 67% of the survival times being censored.

The main aim of our analysis was to investigate whether the estimators presented in the 'Estimators of Discrimination Measures' section were able to discriminate between the full model (containing clinical predictors + 70 genes) and the reduced model (containing clinical predictors only). Using Cox regression, we first built a predictor for the full model. To avoid overfitting the data, a ridge penalty was imposed on the effects of the 70 genes (see [35]). Using a ridge penalty shrinks coefficient estimates toward zero, thereby reducing the variance of the coefficient estimates and avoiding multicollinearity problems. Ridge-penalized regression is also applicable in high-dimensional settings with several hundreds (or even thousands) or genes. In contrast to the gene expression measurements, no penalty was imposed on the clinical predictor variables. This strategy ensured that established predictors (such as the clinical variables) were favored over molecular predictors whose added predictive value had yet to be investigated [12]. The use of the ridge penalty automatically resulted in misspecified Cox models (because effect estimates were shrunken towards zero). In case of the reduced model, $T(G)$-tests indicated significant departures from the proportional hazards assumption ($P = 0.009$ for the global test).

To ensure that the combined predictor did not only work on the data it was derived from but also on 'external' validation data, we split the breast cancer data randomly into a learning sample containing two-thirds of the observations ($n = 96$) and a test sample containing one-third of the observations ($N = 48$). Cox regression was applied to the learning data using the R package **penalized**. Internal 5-fold cross-validation on the learning sample was used to determine the optimal tuning parameter of the ridge penalty. The resulting combined predictor was used as a marker to predict survival of the observations in the test data. To evaluate the performance of the combined predictor/marker, we used the estimators of C^* and C^*_{cum} presented in the 'Estimators of Discrimination Measures' section. The learning and validation procedure was repeated 100 times using different random splits of the data. In addition to fitting the full models, we used the same 100 sets of learning and test data and fitted reduced Cox models based on the five clinical variables only.

Figure 6 shows the estimates of C^* and C^*_{cum}, as obtained from the estimators presented in the 'Estimators of Discrimination Measures' section. It can be observed that all estimators took larger values on average when the 70 genes were added to the clinical predictor variables. The magnitude of the added predictive value was generally small, with the differences in the estimates of C^* and C^*_{cum} being smaller than 0.05 throughout. Such results are frequently observed in biomarker studies and do not seem to depend on the performance measure used for model evaluation (see, e.g. Steyerberg *et al.* [3]). On the other hand, Wilcoxon signed-rank tests on the differences between full and reduced models suggested that the 70 genes significantly improved prediction accuracy (Bonferroni–Holm-adjusted $P < 0.001$ for all methods). Consequently, although being different in magnitude, all estimators indicated a small benefit of adding the 70 genes to the clinical predictor variables. This result is confirmed by the cumulative/dynamic AUC curves presented in Figure 7.

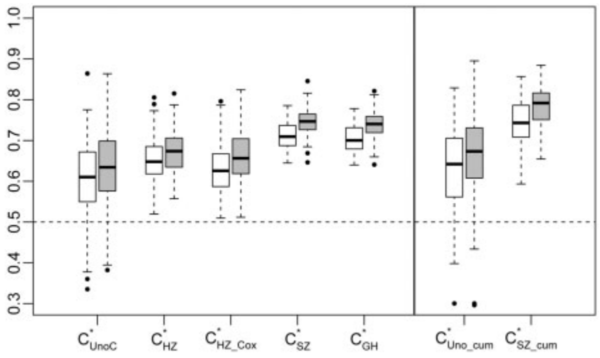

Figure 6: Breast cancer data by van de Vijver *et al.* [4]. Boxplots correspond to estimates of C^* and C^*_{cum}, as obtained from 100 random splits of the data ($n = 96$, $N = 48$). White boxplots correspond to the reduced Cox regression model (using the clinical predictors only), whereas gray boxplots correspond to the full model (ridge-penalized Cox regression based on the clinical predictors and the 70-gene signature by van't Veer *et al.* [20]).

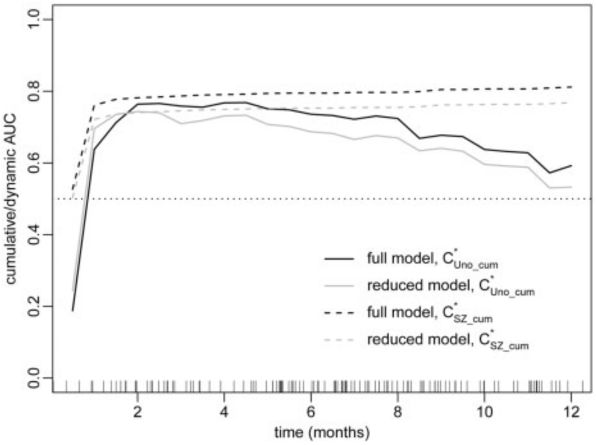

Figure 7: Breast cancer data by van de Vijver *et al.* [4]. Cumulative/dynamic AUC curves were obtained by applying the estimators proposed by Uno *et al.* [10] and Song & Zhou [16] to 100 test samples of size $N = 44$. The solid and dashed lines correspond to the average cumulative/dynamic AUC curves. It is seen that both estimators indicate improved prediction accuracy if the 70 genes are added to the clinical predictors. Interestingly, the differences between the two estimators (which are also seen in Figure 6) are largest at late time points. Similar results were obtained when applying estimators of incident/dynamic AUC to the 100 test samples.

Interestingly, the results presented in Figure 6 are similar to the results of the simulation study presented in Figure 1. For example, the IPCW-based estimates are smaller on average than their

Cox-based counterparts. Also, the differences between the full and reduced models are larger for the Cox-based estimators C^*_{SZ}, C^*_{GH}, $C^*_{\mathrm{SZ}_{\mathrm{cum}}}$ than for the IPCW-based estimators C^*_{UnoC}, $C^*_{\mathrm{Uno}_{\mathrm{cum}}}$ and for the estimators $C^*_{\mathrm{HZ}_{\mathrm{Cox}}}$, C^*_{HZ} proposed by Heagerty & Zheng. Based on the results of the simulation study, we may therefore speculate that the Cox-based estimators show a slight upward bias, and that the differences between the full and reduced models are slightly overestimated by the Cox-based estimators C^*_{SZ}, C^*_{GH} and $C^*_{\mathrm{SZ}_{\mathrm{cum}}}$.

In a final step, we assessed the specificity of the estimators by fitting non-informative null models to the breast cancer data. To this purpose, we repeated the analysis described earlier in the text, this time applying a random permutation to the observed survival times (and also to the values of the event indicator Δ). Consequently, the association between the predictor variables and the survival outcome was destroyed, and the 'true' prediction accuracy of both the full and the reduced model was equal to $C^* = C^*_{\mathrm{cum}} = 0.5$. Figure 8 shows that none of the estimators indicated a notable difference between the full and the reduced model in this case. This result is plausible because of the random permutation of the survival times and also because there was no violation of the Cox proportional hazards assumption in the non-informative null model. On the other hand, the average values of $C^*_{\mathrm{HZ}_{\mathrm{Cox}}}$, C^*_{HZ}, C^*_{SZ}, C^*_{GH} and $C^*_{\mathrm{SZ}_{\mathrm{cum}}}$were distinctly larger than 0.5, implying that the Cox-based estimators overestimated the true prediction accuracy of the models. This result suggests that all Cox-based estimators have a non-ignorable small-sample bias (whereas the IPCW-based estimators seem to have a relatively large small-sample variance). It is consequently not advisable to use the value 0.5 as a benchmark when Cox-based estimators are applied to evaluate the discriminatory power of a prediction model. Instead, we suggest that estimates of C^* and C^*_{cum} should additionally be compared with the estimates obtained from a null model that is based on a random permutation of the observed survival times.

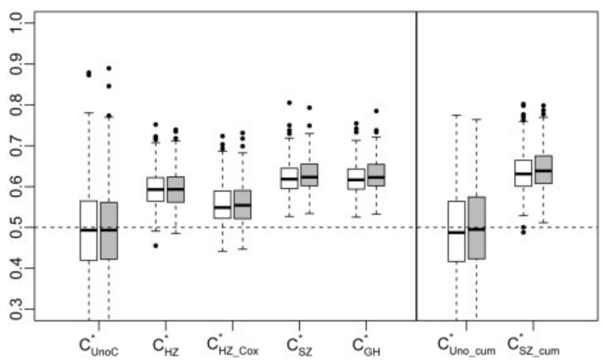

Figure 8: Breast cancer data by van de Vijver *et al.* [4]. Boxplots correspond to estimates of C^* and C^*_{cum}, as obtained from 100 random splits of the data ($n = 96$, $N = 48$). In contrast to the results shown in Figure 6, all estimates are based on random permutations of the observed survival times. Consequently, the true value of both C^* and C^*_{cum} was equal to 0.5. White boxplots correspond to the reduced Cox regression model (using the clinical predictors only), whereas gray boxplots correspond to the full model (ridge-penalized Cox regression based on the clinical predictors and the 70-gene signature by van't Veer *et al.* [20]).

Breast cancer data by Desmedt *et al.*

Desmedt *et al.* [5] collected a data set of 196 node-negative breast cancer patients to validate a 76-gene expression signature developed by Wang *et al.* [21]. The signature, which is based on Affymetrix microarrays, was developed separately for ER-positive patients (60 genes) and ER-negative patients (16 genes). In addition to the expression levels of the 76 genes, four clinical predictor variables were considered (tumor size, ER status, grade of the tumor and patient age). The data are publicly available on GEO (http://www.ncbi.nlm.nih.gov/geo, accession number GSE 7390).

Similar to Wang *et al.* [21], we used the time from diagnosis to distant metastases as primary outcome and investigated whether adding the 76 genes to the clinical variables increased prediction accuracy. Observed metastasis-free survival ranged from 125 days to 3652 days, with 79.08% of the survival times being censored.

To analyze the added predictive value of the 76 genes, we compared the clinical model with the combined model incorporating the clinical variables and the 76 genes. Analogous to the previous subsection, we used Cox regression to build the combined predictor and imposed a ridge penalty on the effects of the 76 genes. Cross-validation was performed as before by splitting the data randomly (100 samples, $n = 131$, $N = 65$). To account for the fact that the 76-gene signature was developed separately for ER-positive and ER-negative patients, we further included interaction terms between ER status and the 76 genes in the model equation. Additionally, we included a five-level factor variable that represented the hospitals in which the data were collected.

In case of the reduced model using the clinical variables only, $T(G)$ tests for age and ER status indicated significant departures from the proportional hazards assumption ($p = 0.007$ and $p = 0.041$, respectively).

Figure 9 shows the estimates of C^* and C^*_{cum}, as obtained from the estimators presented in the 'Estimators of Discrimination Measures' section. Similar to the breast cancer data by van de Vijver *et al.* [4], all estimators took larger values on average when the 76 genes were added to the clinical predictor variables. Wilcoxon signed-rank tests on the differences between combined and clinical estimates suggested that the 76 genes significantly improved prediction accuracy (Bonferroni–Holm-adjusted $P < 0.001$ for all methods). Again, the boxplots presented in Figure 9 are similar to those presented in Figures 1 and 6. Specifically, the IPCW-based estimates are smaller on average than their Cox-based counterparts. This result is confirmed by the cumulative/dynamic AUC curves presented in Figure 10.

In the same way as in the 'Breast cancer data by van de Vijver *et al.*' section, we additionally assessed the specificity of the estimators of C^* and C^*_{cum} by fitting non-informative null models to the breast cancer data. Again, this was accomplished by applying random permutations to the observed survival times and to the values of the event indicator Δ. Similar to the breast cancer data by van de Vijver *et al.* [4], none of the estimators indicated a notable difference between the full and the reduced models in this case (Figure 11). Also, the average values of $C^*_{\text{HZ}_{\text{Cox}}}$, C^*_{HZ}, C^*_{SZ}, C^*_{GH} and $C^*_{\text{SZ}_{\text{cum}}}$ were again larger than 0.5, implying that the Cox-based estimators overestimated the prediction accuracy of the models.

Summarizing the results obtained from the two breast cancer data sets, all estimators of discrimination measures indicated a small increase in prediction accuracy when the gene signatures were added to the clinical models. This result suggests that effects of a gene signature on prediction accuracy are likely to be detected by all estimators, even if the underlying regularity assumptions are violated to a certain degree.

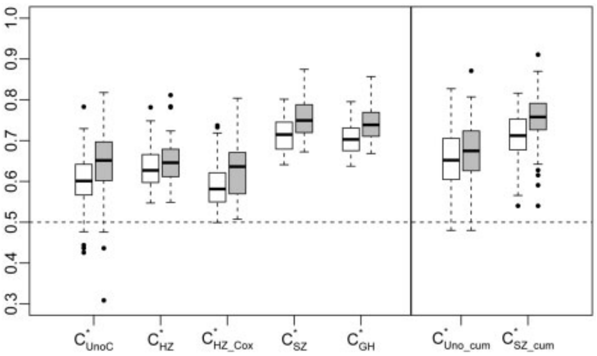

Figure 9: Breast cancer data by Desmedt *et al.* [5]. Boxplots correspond to estimates of C^* and C^*_{cum}, as obtained from 100 random splits of the data ($n = 131$, $N = 65$). White boxplots correspond to the reduced Cox regression model (using the clinical predictors only), whereas gray boxplots correspond to the full model (ridge-penalized Cox regression based on the clinical predictors and the 76-gene signature by Wang *et al.* [21]).

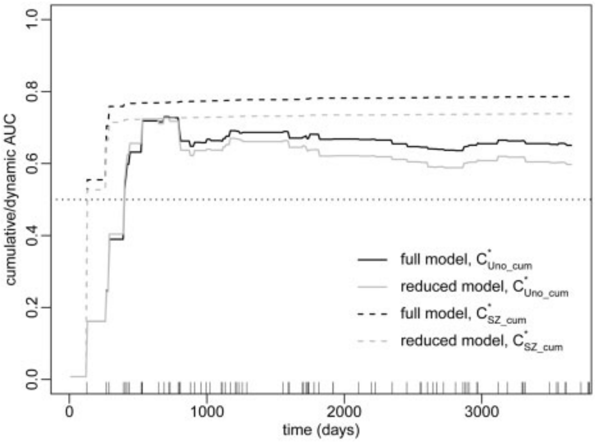

Figure 10: Breast cancer data by Desmedt *et al.* [5]. Cumulative/dynamic AUC curves were obtained by applying the estimators proposed by Uno *et al.* [10] and Song & Zhou [16] to 100 test samples of size $N = 65$. The solid and dashed lines correspond to the average cumulative/dynamic AUC curves. It is seen that both estimators indicate improved prediction accuracy if the 76 genes are added to the clinical predictors. Similar results were obtained by applying estimators of incident/dynamic AUC to the 100 test samples.

DISCUSSION

Deriving unbiased estimates of prediction accuracy is essential for the evaluation of molecular markers and gene signatures. In this respect, several authors have pointed out that the calculation of P-values is not sufficient for measuring the prognostic effect of new markers [36, 37]. This is not only because statistical hypothesis tests may be biased due to small sample sizes or the violation of regularity assumptions but also because effect sizes of significant biomarkers are often small and may therefore not lead to relevant improvements in the prediction of patient survival.

The time-dependent AUC curves and summary indices considered in this article are measures of the discriminative ability of biomarkers and hence constitute an easy-to-interpret approach for the quantification of the accuracy of survival predictions. During the past years, time-dependent ROC analysis has therefore become a popular tool for marker comparisons in biomedical studies [1, 7, 15]. Most important, AUC-based discrimination measures for survival analysis share the same characteristics as traditional AUC-based measures in classification tasks [14, 38].

Because discrimination measures cannot be calculated directly in the presence of censoring, estimators of time-dependent ROC and AUC curves have to be applied. In a previous study [27], we showed that the finite-sample behavior of these estimators is strongly affected by confounding factors (such as censoring rate, model misspecification and the degree of information contained in the predictor variables). In this article, we went one step further and focused directly on the effect of model misspecification on medical decision making. In other words, we did not analyze the magnitude of the finite-sample bias induced by the confounders but focused on the ranking of competing prediction models. By using the summary indices C^* and C^*_{cum} as ranking criteria (as previously proposed by Yu *et al.* [15] and Ma & Song [39]), we analyzed whether the violation of regularity assumptions affected the analysis of the 'added predictive value' of newly discovered biomarkers. Our simulation study showed that estimates of C^* and C^*_{cum} may lead to wrong conclusions about the utility of new biomarkers if they are based on misspecified Cox regression models.

Statistical practitioners might argue that the problems discussed in this article can easily be avoided by carrying out routine checks of the proportional hazards assumption. As demonstrated in the 'Simulation Study' section, however, this supposition is not true at all: our simulation study shows that departures from the proportional hazards assumption may be hard to detect even if the true data-generating model is simple and even if sample sizes are moderately large ($n = 100$). Consequently, biased medical decision making is possible even if biostatisticians carefully check all regularity assumptions.

Based on the results presented in this article, we recommend to use of the IPCW-based estimators C^*_{UnoC} and $C^*_{\mathrm{Uno_{cum}}}$ to evaluate the discriminatory power of survival models. Although the validity of these estimators depends on the assumption that survival and censoring times are independent (at least conditional on X), simulation results have suggested that C^*_{UnoC} and $C^*_{\mathrm{Uno_{cum}}}$ are barely affected by violations of the random censoring assumption ('Violations of the Random Censoring Assumption' section, see also [27]). Alternatively, we propose to apply the varying-coefficient estimator C^*_{HZ} by Heagerty & Zheng [14]. Although this estimator showed a slight upward bias in our simulation study, it did not result in biased rankings of competing prediction models–and hence did not lead to wrong conclusions about the utility of new biomarkers.

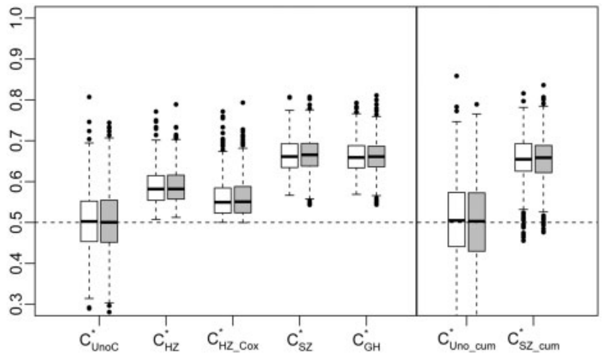

Figure 11: Breast cancer data by Desmedt *et al.* [5]. Boxplots correspond to estimates of C^* and C^*_{cum}, as obtained from 100 random splits of the data ($n = 131$, $N = 65$). In contrast to the results shown in Figure 9, all estimates are based on random permutations of the observed survival times. Consequently, the true value of both C^* and C^*_{cum} was equal to 0.5. White boxplots correspond to the reduced Cox regression model (using the clinical predictors only), whereas gray boxplots correspond to the full model (ridge-penalized Cox regression based on the clinical predictors and the 76-gene signature by Wang *et al.* [21]).

Key Points

- Recent developments in molecular biology have led to the massive discovery of new biomarkers for the prediction of survival outcomes.
- The estimation of AUC-based discrimination measures is a popular approach to evaluate the accuracy of these predictions.
- Estimators of discrimination measures depend on regularity assumptions that have to be satisfied to guarantee the validity of the estimators.
- Violations of the regularity assumptions may lead to a non-ignorable bias in estimators of discrimination measures and therefore to biased medical decision making.
- Our results suggest that inverse-probability-of-censoring-weighted estimators of discrimination measures are a robust approach to address the problem of biased medical decision making.

FUNDING

Deutsche Forschungsgemeinschaft (Project SCHM 2966/1–1 to M.S.); German Federal Ministry of Education and Research (BMBF) within the framework of the program of medical genome research (PaCa-Net; project ID PKB-01GS08) and within the framework Gerontosys II (Forschungskern SyStaR, project ID 0315894A) (to H.A.K.).

References

1. Pepe MS, Zheng Y, Jin Y, *et al*. Evaluating the ROC performance of markers for future events. *Lifetime Data Anal* 2008;**14**:86–113.
2. Simon RM, Subramanian J, Li MC, *et al*. Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Brief Bioinform* 2011; **12**:203–14.
3. Steyerberg EW, Vickers AJ, Cook NR, *et al*. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;**21**:128–38.
4. van de Vijver MJ, He YD, van't Veer LJ, *et al*. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002;**347**:1999–2009.
5. Desmedt C, Piette F, Loi S, *et al*. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independentvalidation series. *Clin Cancer Res* 2007;**13**:3207–14.
6. Kok M, Linn SC, Van Laar RK, *et al*. Comparison of gene expression profiles predicting progression in breast cancer patients treated with tamoxifen. *Breast Cancer Res Treat* 2009;**13**:275–83.
7. Li H, Gui J. Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics* 2004; **20**:i208–15.
8. Chang HY, Sneddon JB, Alizadeh AA, *et al*. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol* 2004;**2**:e7.
9. Cai T, Pepe MS, Zheng Y, *et al*. The sensitivity and specificity of markers for event times. *Biostatistics* 2006;**7**: 182–97.
10. Uno H, Cai T, Tian L, *et al*. Evaluating prediction rules for t-year survivors with censored regression models. *J Am Stat Assoc* 2007;**102**:527–37.
11. Eden P, Ritz C, Rose C, *et al*. ''Good old'' clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *Eur J Cancer* 2004;**40**: 1837–41.
12. Boulesteix A-L, Sauerbrei W. Added predictive value of high-throughput molecular data to clinical data and its validation. *Brief Bioinform* 2011;**12**:215–29.
13. Kaderali L, Zander T, Faigle U, *et al*. CASPAR: a hierarchical bayesian approach to predict survival times in cancer from gene expression data. *Bioinformatics* 2006;**22**: 1495–502.
14. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005;**61**:92–105.
15. Yu Y, Li J, Ma S. Adjusting confounders in ranking biomarkers: a model-based ROC approach. *Brief Bioinform* 2012;**13**:513–23.
16. Song X, Zhou XH. A semiparametric approach for the covariate specific ROC curve with survival outcome. *Stat Sin* 2008;**18**:947–65.
17. Uno H, Cai T, Pencina MJ, *et al*. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011;**30**:1105–117.
18. Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* 2005;**21**: 3001–8.
19. Subramanian J, Simon R. An evaluation of resampling methods for assessment of survival risk prediction in high-dimensional settings. *Stat Med* 2011;**30**:642–53.
20. van't Veer LJ, Dai HY, van de Vijver MJ, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;**415**:530–6.
21. Wang Y, Klijn JG, Zhang Y, *et al*. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005;**365**:671–9.
22. Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 2005;**92**:965–70.
23. van der Laan MJ, Robins JM. *Unified Methods for Censored Longitudinal Data and Causality*. Springer, New York. 2003.
24. Harrell FE, Califf RM, Pryor DB, *et al*. Evaluating the yield of medical tests. *JAMA*;**247**:2543–6.
25. Harrell FE, Lee KL, Califf RM, *et al*. Regression modeling strategies for improved prognostic prediction. *Stat Med* 1984;**3**:143–52.
26. Antolini L, Boracchi P, Biganzoli E. A time-dependent discrimination index for survival data. *Stat Med* 2005;**24**: 3927–44.
27. Schmid M, Potapov S. A comparison of estimators to evaluate the discriminatory power of time-to-event models. *Stat Med* 2012;**31**:2588–609.
28. Chambless LE, Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Stat Med* 2006;**25**:3474–86.
29. Hung H, Chiang CT. Estimation methods for time-dependent AUC models with survival data. *Can J Stat* 2010; **38**:8–26.
30. R Core Team *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. http://www.r-project.org.
31. Heagerty PJ. *RisksetROC: Riskset ROC Curve Estimation from Censored Survival Data* 2012. R package version 1.0.4. http://cran.r-project.org/web/packages/risksetROC/ index.html.
32. Potapov S, Adler W, Schmid M. *SurvAUC: Estimators of Prediction Accuracy forTime-to-Event Data* 2012. R package version 1.0–5. http://cran.r-project.org/web/packages/ survAUC/index.html.
33. Grambsch P, Therneau T. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994; **81**:515–26.
34. Goeman J. *Penalized: L1 (Lasso) and L2 (Ridge) Penalized Estimation in GLMs and in the Cox Model* 2012. R package version 0.9–42.

35. van Houwelingen HC, Bruinsma T, Hart AMM, *et al.* Cross-validated Cox regression on microarray gene expression data. *Stat Med* 2006;**25**:3201–16.

36. Rosthoj S, Keiding N. Explained variation and predictive accuracy in general parametric statistical models: the role of model misspecification. *Lifetime Data Anal* 2004;**10**:461–72.

37. Uno H, Tian L, Cai T, *et al.* A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data. *Stat Med* 2013;**32**: 2430–42.

38. Berrar D, Flach P. Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). *Brief Bioinform* 2012;**13**:83–97.

39. Ma S, Song X. Ranking prognosis markers in cancer genomic studies. *Brief Bioinform* 2011;**12**:33–40.

APPENDIX 1

EFFECT OF THE STANDARD DEVIATION σ ON THE RESULTS OF THE SIMULATION STUDY

To investigate whether changes in the value of the standard deviation σ in Equation (18) affect the results of the simulation study, we repeated the analysis of the 'Results of the Simulation Study' section using different values of σ. Figure 12 visualizes the estimated differences between the full and reduced models that were obtained from the estimators of C^* and C^*_{cum}. Obviously, different values of σ resulted in the same patterns as those presented in Figure 1: Although the mean differences between the full and reduced models were close to zero for the IPCW-based estimators estimators C^*_{UnoC} and $C^*_{\text{Uno}_{\text{cum}}}$ and for the estimators $C^*_{\text{HZ}_{\text{Cox}}}$and C^*_{HZ} by Heagerty & Zheng [14], the Cox-based estimators C^*_{SZ}, C^*_{GH} and $C^*_{\text{SZ}_{\text{cum}}}$ resulted in mean differences that were distinctly larger than zero. Consequently, the latter estimators wrongly indicated an additional effect of the biomarker $X^{(1)}$ on prediction accuracy, regardless of the value of σ.

Similarly, there were no qualitative changes regarding the results of the $T(G)$-tests to check the proportional hazards assumption. For example, estimates of the probabilities, that at least one of the four tests resulted in a P-value smaller than 0.05, are shown in Table 1. Because the predictors in the two subgroups of Equation (18) become less informative as σ becomes larger, there was a decrease in the fraction of P-values smaller than 0.05 when the value of σ was increased. However, even for very small values of σ, only ~20% of the $T(G)$-tests detected departures from the Cox proportional hazards assumption.

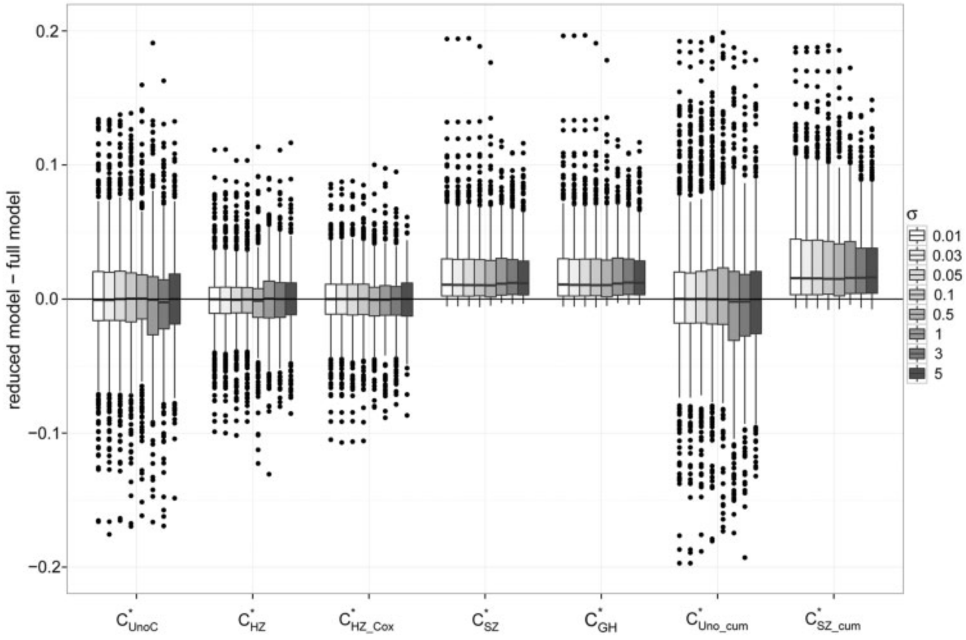

Figure 12: Results of the additional simulation study presented in the Appendix. The boxplots visualize the differences between the full and reduced models when estimating C^* and C^*_{cum} by the methods described in the 'Estimators of Discrimination Measures' section. Each boxplot corresponds to a specific value of the standard deviation σin Equation (18). Predictions were obtained by fitting Cox regression models to 500 learning samples generated from Equation (18). The horizontal line corresponds to the true mean difference between the full and reduced models.

Table 1: Results of the additional simulation study presented in the Appendix

σ	**0.01**	**0.03**	**0.05**	**0.10**	**0.50**	**1.00**	**3.00**	**5.00**
Fraction of $P < 0.05$ (%)	20.4	20.0	20.2	20.4	15.6	8.0	10.4	10.4

The table contains the estimated probabilities that at least one of the four *T(G)*-tests discussed in the 'Checks of the Proportional Hazards Assumption' section resulted in a *P*-value smaller than 0.05. All estimates are based on 500 learning samples generated from Equation (18).