Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring

Thomas A. Gerds\(^1\), Michael W Kattan\(^2\), Martin Schumacher\(^3\) and Changhong Yu\(^2\)

December 9, 2010

Abstract Given a continuous marker and a time-to-event response variable the proportion of concordant pairs in a data set is called \(C\) statistic (Harrell et al. 1982). A specifically useful marker is the risk predicted by a survival regression model. The \(C\) statistic has been widely used to assess and compare prediction models with respect to their ability to discriminate individual risks. This article extends the existing methodology for estimating the \(C\) index suitable for survival analysis where the length of the follow-up period may depend on the covariates. A new estimation method is proposed which is applicable to all kinds of prediction models and does not require a correctly specified survival regression model. Furthermore, a new performance measure is defined that captures time trends for models that allow crossing of individually predicted survival curves. For comparing rival modelling strategies on the same set of data, we discuss bootstrap confidence limits for differences in discrimination ability. For illustration we re-analyze a study on prediction of prostate cancer patients outcome (Kattan et al. 2000).

---

1 Department of Biostatistics, University of Copenhagen
2 Quantitative Health Sciences, Cleveland Clinic
3 Institute of Medical Biometry and Medical Informatics, University of Freiburg
1 Introduction

The concordance index $C$ (Harrell et al. 1982) is a well recognized measure of discrimination for models that predict a time to event. It has been reported more often than any other prediction model metric in the survival setting. The $C$ statistic plays an important role among the recommended validation tools for biomarker discovery and the development of prediction models (Taylor et al. 2008). In many applications it is desirable to compare the discriminative ability of several regression modelling strategies and the resulting prediction models. For example, in a recently published scientific statement on the evaluation of new markers (Hlatky et al. 2009, Table 1) it was recommended to calculate $C$ both for the model with established risk factors and for the model including the novel marker and the established markers. Another aim is to use $C$ to compare models derived in different statistical cultures (Breiman 2001), for example a Cox regression model and a random survival forest model.

For uncensored data, $C$ is very easy to interpret, as the relative frequency of concordant pairs among all pairs of individuals. A pair is called concordant if the individual with the shorter survival time received the higher predicted risk. However, from the data of a typical medical study with limited follow-up, where some event times are right censored, $C$ as defined cannot be estimated, since the order of the event times of two censored patients remains unknown. This limitation is already implicitly contained...
Time-dependent concordance index for survival

in previous work on $\mathcal{C}$ where pairs of subject where both event times are right censored are called unusable (Harrell et al. 1982). However, when unusable pairs are simply ignored, one estimates a population parameter which depends on the censoring distribution and the frequency of unusable pairs. Clearly, the longer the follow-up period of the study the fewer unusable pairs. Indeed, this dependence of $\mathcal{C}$ on the follow-up period has not been treated rigorously or explicitly by many authors.

For the case of a single continuous marker, Brown et al. (1974) introduced a simple estimate of the $\mathcal{C}$-index for right censored survival data, which was made popular by Harrell et al. (1982, 1996), and has been used widely and prominently in the medical literature. The simple estimate has a different interpretation, as the relative frequency of definitely concordant pairs. Brown et al. (1974) pointed out that the simple estimate does not use all the available information, and the authors proposed an alternative estimate based on imputed survival times. However, their imputation method is based on the marginal Kaplan-Meier estimator, and thus it is not well adapted to multivariable situations where the censoring distribution may depend on the covariates.

In this article, we show that the simple estimate of $\mathcal{C}$ asymptotically depends on the censoring distribution, and derive the explicit form of the asymptotic bias. The severeness of this bias is further investigated in a small simulation study. Gönen & Heller (2005) also investigated the bias of the simple estimate for estimation of the unrestricted $\mathcal{C}$ via simulation. They proposed a new estimate which requires that the prediction model is a Cox regression model and that this Cox regression model is correctly specified. Thus, their method cannot be used to compare two rival risk prediction models, since both must be Cox regression models and both cannot be correct. We discuss the problem of model based estimates of prediction performance in more detail.
Time-dependent concordance index for survival below. Note also that the methods of Pencina & D’Agostino (2004) are restricted to type-I censoring which rarely occurs in clinical settings.

To overcome the identifiability problem without assuming that the prediction model is correctly specified, we define a time truncated version $C(t)$ which is interpreted as the ability of the prediction model to discriminate event times that occurred before $t$. For varying values of $t$ this yields a discrimination curve. To consistently estimate $C(t)$, we adapt an inverse of the probability of censoring weighted (IPCW) estimate. The proposed estimate is based on general theory for censored data (Van der Laan & Robins 2003) and here shown to be unbiased under the assumption that the censoring times are conditionally independent of the survival times given the predictor variables. The method requires a regression model for the censoring times, but no further assumptions are needed regarding the dependence between the survival times and the predictor variables. Hence the proposed method can be used to compare rival prediction models of any kind. Recent work (Liu & Jin 2009, Uno et al. 2009) also used IPCW weighting to estimate $C(t)$, however, in both studies it is assumed that the censoring distribution is independent of the predictor variables. Our simulation study shows a bias of the IPCW method when this assumption fails.

Our approach extends previous work also in that we provide a framework for comparing modelling strategies with different degree of flexibility, like Cox regression with automated variable and functional form selection, or tools from machine learning. This is achieved by adapting a bootstrap cross-validation approach, also known as leave-one-out bootstrap (Efron & Tibshirani 1993), where the data are repeatedly split into training and test data. In all training data sets all steps of modelling are repeated. The concordance index as considered here uses survival time as response. It thus differs conceptually from the generalization of the area under the ROC curve for sur-
vival analysis (Heagerty et al., 2000, Chambless & Diao 2006) which uses the binary survival status at given time points as response. The time-dependent area under the ROC curve AUC(t) quantifies how well the prediction model can order the survival status at a given time point, whereas \( C(t) \) quantifies how well the prediction model can order the survival times up to time \( t \). Note further that the setup in Efron (1967) (see also Koziol & Jia 2009) is also conceptually different as it deals with the comparison of exactly two survival curves, that is a single binary marker, whereas we consider as marker the risk predicted by a model, which usually takes values on a continuous scale between 0 and 1.

The further developments of this article are motivated by the data and the aims of a study on prediction models for prostate cancer patients (Kattan et al., 2000). A particularity of the study was that the treatment dose increased with calendar time. Thus patients enrolled late received higher doses, and the censoring times depend heavily on the predictor variables. Consequently the assumption that the censoring is independent of the covariates is not justified for this study. The original study compared eight prediction models with Somers’ rank correlation coefficient which is a simple transformation of the \( C \)-index: \( D = 2 \times C - 1 \). Here we selected two of the previously considered prediction models and compare them based on the proposed IPCW estimate of \( C(t) \).

The discrimination curve analysis presented here can display the time course of predictive performance. This is particularly relevant for risk prediction models in which the order of risk predicted for a pair of subjects may change over time. For the purpose of illustration we also consider two prediction models based on Aalen’s additive hazard regression (Scheike 2001), and on a random survival forest approach (Ishwaran et al., 2008), both methods allow for time dependent risk ratios.
2 Data structure

Denote $T_i$ the time of death, or time to another event of interest, of subject $i$ and $X_i = (X_{i1}, \ldots, X_{iL})$ a vector of time independent predictor variables. Observed are only $\tilde{T}_i = \min(T_i, C_i)$, $\Delta_i = 1\{T_i \leq C_i\}$, and $X_i$, where $C_i$ is the individual time where the follow-up period for subject $i$ ends right censored.

We suppose a training set consisting of the data of $m$ independent subjects $D_n = \{\tilde{T}_i, \Delta_i, X_i\}, i \in \mathcal{I}_n\}$ and a separate validation set consisting of the data of $n$ independent subjects $D_m = \{\tilde{T}_i, \Delta_i, X_i\}, i \in \mathcal{I}_m\}$. It is further assumed that the sets $\mathcal{I}_n$ and $\mathcal{I}_m$ are disjunct. We will discuss how to proceed in practice when there is only one data set available in section 6.2.

3 Prediction models

A statistical prediction model is usually obtained by applying a regression modelling strategy $S$ to the training $D_n$. For example, one could apply a suitable variable selection algorithm and eventually fit a Cox regression model which includes the selected variables. Formally, a prediction model is a mapping from the range of the predictor variables to a survival probability for each time point $t$ in a predefined range of time points. For a subject $j \in \mathcal{I}_m$ with covariate vector $X_j$, we denote

$$M_n(t, X_j) = S(D_n)(t, X_j)$$

for the predicted probability that the person survives time $t$. Prediction models can differ with respect to the selection of predictor variables, and also by the number and type of parameters, link functions, modelling assumptions, and data dependent optimization steps (e.g. Harrell 2001, Hand 2001, Gerds et al. 2008).
For the definition and estimation of $C(t)$ in the following sections, the only requirement is that the model can rank the risk of any two subjects based on the information available at baseline. For a fixed time point $t$, if $M_n(t, X_i) < M_n(t, X_j)$ then the model predicts that person $i$ has a higher risk of dying until time $t$ than person $j$.

Note that often the prediction models for which $C$ has been reported do not allow that differences or ratios between the individual risks of two subjects change over time, in which case $M_n(t, X_i) < M_n(t, X_j)$ does not depend on $t$. However, in certain applications it may be important to allow for more flexible modelling strategies. The discrimination curve defined in section 4.2 may pick up time trends in the predictive performance of prediction models.

### 4 The risk discrimination index

#### 4.1 Unrestricted concordance index

Fix a time point $t$. The risk discrimination index $C$ is defined as

$$C = P \left\{ M_n(t, X_i) > M_n(t, X_j) \mid T_i < T_j \right\}.$$ (1)

Here $i, j \not\in \mathcal{I}_n$ represent two subjects from the population whose data is not included in the training data $D_n$. If the event time of subject $i$ turns out to be smaller than the event time of subject $j$ then a good model should have predicted a higher risk for subject $i$. Thus, the index $C$ is the ability of the model to rank the event times based on the risk predicted for a certain time point $t$ and given information $(X_i, X_j)$ available at the time origin. This makes most sense when the predicted ranking does not depend on time, i.e. when $M_n(t, X_i) > M_n(t, X_j)$ implies that $M_n(s, X_i) > M_n(s, X_j)$ for all time points $s$. 
The parameter $C$ defined in (1) has a nice and elegant interpretation, but it is not generally identifiable in a medical study with limited follow-up. The reason is that it is not possible to assess the order of the event times that occur beyond the maximal follow-up time of the study.

4.2 Truncated concordance and discrimination curve

To overcome the identifiability problem we define the following truncated version of $C$:

$$C(t) = P\left\{ M_n(t, X_i) > M_n(t, X_j) \mid T_i < T_j, T_i \leq t \right\},$$

(2)

where again $i, j \notin I_n$. The value of $C(t)$ can be interpreted as the ability of the model to rank the event times that occur before time $t$. By varying $t$ we obtain a discrimination curve which can be identified from right censored data until $t$ reaches a certain maximal time point which depends on the follow-up period of the study (see section 5). For proportional hazard models the discrimination curve will only vary due to different truncation times. For more flexible models, the discrimination curve will also vary due to changes of the ranking of subjects over time.

5 Estimation

We allow that the distribution of $C_i$ depends on $X_i$ but assume throughout that

$$C_i \text{ is conditionally independent of } T_i \text{ given } X_i.$$  \hfill (A1)

Denote $S(t \mid x) = P(T_i > t \mid x)$ the conditional survival function and $G(t \mid x) = P(C_i > t \mid x)$ the conditional censor survival function. Under (A1) we have

$$P(\tilde{T}_i > t \mid x) = S(t \mid x)G(t \mid x).$$

(3)
Note that \( G \) describes also the so-called inverse probability of censoring function (Van der Laan & Robins 2003)

\[
G(t^{-} \mid x) = P(\Delta_i = 1 \mid X_i = x, T_i = t) = P(C_i \geq t \mid X_i = x), \tag{4}
\]

where \( G(t^{-} \mid x) \) denotes the left hand limit of \( G(t \mid x) \).

To describe and analyze several estimators based on the observed data we introduce the following notation:

\[
\tilde{Y}_i(t) = 1\{T_i \leq t\} \quad \quad Q_n^{i,j}(t) = 1\{M_n(t, X_i) > M_n(t, X_j)\}
\]

\[
\tilde{N}_{ij} = 1\{\tilde{T}_i < \tilde{T}_j\} \quad \quad N_{ij} = 1\{T_i < T_j\}.
\]

Note that \( E(N_{ij}|X_i, X_j) = \int_{0}^{\infty} S(u \mid X_j) S(du \mid X_i) \). Conditional on the training data used to select the model \( M_n \), the truncated concordance (2) can be written as

\[
C(t) = \frac{E \left\{ Q_n^{i,j}(t) \int_{0}^{t} S(u \mid X_j) S(du \mid X_i)|D_n \right\}}{E \left\{ \int_{0}^{t} S(u \mid X_j) S(du \mid X_i) \right\}}, \tag{5}
\]

where the expectation is taken with respect to the marginal distributions of \( X_i \) and \( X_j \) and where \( i, j \notin I_n \) represent subjects whose data are not included in the training set.

5.1 Simple estimate

A frequently used estimator of \( C \) has been derived from a simple modification of Kendall’s rank correlation for censored data (Brown et al. 1974); it was made popular by Harrell et al. (1982, 1996) and in our notation based on the validation data reads

\[
\hat{C}_{simp}(t) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} \bar{N}_{ij} Q_n^{i,j}(t) \tilde{Y}_i(t) \Delta_i}{\sum_{i=1}^{m} \sum_{j=1}^{m} N_{ij} \tilde{Y}_i(t) \Delta_i}.
\]
The estimate can be interpreted as the frequency of definitely concordant pairs, because if $\bar{N}_{ij} = 1$ and $\Delta_i = 1$, then it can be deduced from the observable data that $T_j > T_i$.

The advantage of the simple estimate is that it does not require further modelling of the data. The main disadvantage of the simple estimate is that its asymptotic value depends on the censoring distribution, as shown below. Note further that, as it is commonly applied, the simple estimate is evaluated at the maximum follow-up time of the study.

To express the asymptotic bias of $\hat{C}_{simp}$, first note that under conditional independence one has (see Begun et al. 1983)

$$E(\Delta_i \bar{Y}_i(t) \mid X_i) = - \int_0^t G(u - \mid X_i) S(du \mid X_i).$$

Hence, by using the independence of subjects $i$ and $j$ and equation (3) one finds

$$E(\Delta_i \bar{Y}_i(t) \bar{N}_{ij} \mid X_i, X_j) = \int_0^t G(u - \mid X_i) G(u \mid X_j) S(u \mid X_j) S(du \mid X_i).$$

Define the weight function $W(u, x, y) = G(u - \mid x) G(u \mid y)$. By using empirical process theory (Van der Vaart 1998) it is straightforward to show that for $m \to \infty$ and fixed $n$ the simple estimate $\hat{C}_{simp}(t)$ converges in probability to

$$C_{simp}^*(t) = \frac{E \left\{ Q_{i,j}^n(t) \int_0^t W(u, X_i, X_j) S(u \mid X_j) S(du \mid X_i) \mid D_n \right\}}{E \int_0^t W(u, X_i, X_j) S(u \mid X_j) S(du \mid X_i)}. \quad (6)$$

Subtracting the true value (5) shows that the asymptotic bias $C_{simp}^*(t) - C(t)$ has the form

$$E \left\{ Q_{i,j}^n(t) \int_0^t \left[ P(T_i < T_j) W(u, X_i, X_j) - P(\bar{T}_i < \bar{T}_j) \right] S(u \mid X_j) S(du \mid X_i) \mid D_n \right\}$$

$$P(T_i < T_j) P(\bar{T}_i < \bar{T}_j)$$

This shows that also the truncated version of the simple estimate suffers a large-sample bias for estimating the truncated $C$-index. This theoretical result is confirmed by the simulation study presented in section 8.
5.2 IPCW estimate

The asymptotic value of the simple estimate in (6) suggests an inverse of the probability of censoring weighted (IPCW) estimate. Based on an estimate $\hat{G}_m$ of $G$ define weights

$$\hat{W}_{ij} = \hat{G}_m(T_i \mid X_j) \hat{G}_m(T_i^- \mid X_i)$$

and then the following IPCW estimate of the parameter (2)

$$\hat{C}_{ipcw}(t) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} \tilde{N}_{ij} \tilde{Q}_{ij}^{n}(t) \tilde{Y}_i(t) \Delta_i \hat{W}_{ij}^{-1}}{\sum_{i=1}^{m} \sum_{j=1}^{m} \tilde{N}_{ij} \tilde{Y}_i(t) \Delta_i \hat{W}_{ij}^{-1}}.$$  \hspace{1cm} (7)

The estimator is well defined at time $t$ if it holds uniformly in $x$ that $\hat{G}_m(t \mid x) > 0$. Thus, a sensible choice for $t$ should be made dependent on the method used to estimate the conditional censoring distribution. In practice, the times $t$ where the estimate is interpreted can be chosen in similar way as for the Kaplan-Meier estimator, i.e. $t$ should be such that the number at risk in all subgroups defined by the covariates is not too small.

Suppose $\hat{G}_m$ is a uniformly consistent estimate of $G$. Then $\hat{W}_{ij}$ consistently estimates the weight function $W$ and it follows immediately that the IPCW estimate is consistent. More precisely, for $m \to \infty$ and fixed $n$, the IPCW estimate $\hat{C}_{ipcw}(t)$ converges in probability to:

$$E \left\{ \frac{Q_n^{2i}(t) \int_0^t \{W(u, X_i, X_j)\}^{-1} W(u, X_i, X_j) S(u \mid X_j) S(du \mid X_i) | D_n \} \right\} = C(t).$$

The advantage of IPCW estimates is that they do not assume anything directly about the relation between survival times and predictor variables. They can thus be used to compare rival survival prediction models without bias towards one model. The disadvantage of the IPCW estimate is that one has to model the censoring distribution. If it is reasonable to assume that the $C_i$ are independent of $(T_i, X_i)$ then $G$ can be...
estimated by the Kaplan-Meier estimator for the censoring times (Graf et al. 1999, Liu & Jin 2009, Uno et al. 2009). However, this assumption is clearly violated in the application considered in this article (see section 7). In the prostate cancer study it is reasonable to assume conditional independence of the censoring times and the event times given the predictor variables. Thus, $G$ can be estimated with a regression model for the censoring times, as discussed below.

In practice, a bias is introduced if the model for $G$ is misspecified. Suppose $\hat{G}_m$ converges in probability to $G^* \neq G$. Then, as $m \to \infty$ the weights $W_{ij}$ converge to $W^*(u, x, y)$ and $\hat{C}_{ipcw}(t)$ converges to:

$$\hat{C}_{ipcw}(t) = \frac{\mathbb{E}\left\{ \int_0^t W^*(u, X_i, X_j) S(u | X_j) | D_n \right\}}{\mathbb{E}\left\{ \int_0^t W^*(u, X_i, X_j) M_n(u | X_j) | D_n \right\}}.$$ 

If the censoring model is misspecified, then $\hat{C}_{ipcw}(t) - C(t)$ is the asymptotic bias of $\hat{C}_{ipcw}(t)$. However, since this bias does not depend on the model $M_n$, the IPCW estimate can be recommended for comparing rival prediction models.

5.3 Model based estimate

The representation (5) suggests a model based estimate of $C(t)$, where one substitutes the model $M_n$ for $S$:

$$\hat{C}_{mb}(t) = \frac{\mathbb{E}\left\{ \int_0^t M_n(u | X_j) M_n(u | X_j) | D_n \right\}}{\mathbb{E}\left\{ \int_0^t M_n(u | X_j) M_n(u | X_j) | D_n \right\}}.$$ 

This type of estimate was discussed by Korn & Simon (1990) in a more general loss function setting, and by Gönen & Heller (2005) for the concordance index in the special case where $M_n$ is coming from a Cox regression model. The model based estimation of predictive model performance has been criticized because it relies on that the survival model is correctly specified (Graf et al. 1999). More realistically all models will be
misspecified to some extend. Suppose $M_n$ converges to $M^* \neq S$, then $\hat{C}_{mb}(t)$ converges to
\[
\hat{C}_{mb}^*(t) = \frac{E\left\{Q^{i,j}_n(t) \int_0^t M^*(u \mid X_j) M^*(du \mid X_i) \mid D_n\right\}}{E\left\{ \int_0^t M^*(u \mid X_j) M^*(du \mid X_i) \right\}}.
\]
Thus, the asymptotic bias of $\hat{C}_{mb}(t)$ depends on the model. We conclude that if the
model based estimate is used to compare rival prediction models, say $M^{(1)}_n$ and $M^{(2)}_n$, then a third model $M^{(3)}_n$ should be used as a plug-in for $S$ in (5). It remains unclear
how one would specify $M^{(3)}_n$ in practice.

6 Comparison of prediction models

In this section we develop bootstrap confidence limits based on the IPCW estimate for
the difference in $C(t)$ of rival prediction models.

6.1 Independent validation data

Suppose two rival prediction models, $M^{(1)}_n$ and $M^{(2)}_n$ have been built on the training
data $D_n$. Then, $D_m$ can be used without hesitation to estimate and compare $\hat{C}(t)$
of the two models. In this situation, the asymptotic distribution of $\hat{C}_{ipcw}(t)$ can be
derived using general theory on estimation of parameters in the presence of a nuisance
parameter (Bickel et al. 1993). The special case where the nuisance parameter is a
conditional censoring distribution has been extensively discussed in (Van der Laan &
Robins 2003, Example 1.12). They derive the efficient influence curve for estimation
of general statistics of $(T_i, X_i), i = 1, \ldots, n$, the $C$-index as defined in (5) being a
special case. However, they argue (see also Robins & Ritov 1997) that due to the
curse of dimensionality in the typical application with several continuous covariates
it is not possible to construct fully efficient estimates. One needs a “working” model
for the conditional censoring distribution and a corresponding estimate \( \hat{G}_m \) which is asymptotically Gaussian regular, that is

\[
\sqrt{n}(\hat{G}_m(t \mid x) - G(t \mid x)) = m^{-1/2} \sum_{i=1}^m IC^*_G(\hat{T}_i, \Delta_i, X_i; t, x) + \text{op}(1)
\]  

(A2)

where \( IC^*_G \) is the influence function of \( \hat{G}_m \) (Bickel et al. 1993). Then the functional delta method (Van der Vaart 1998) shows that \( \hat{C}_{ipcw}(t) \) is asymptotically Gaussian regular and

\[
\sqrt{m}(\hat{C}_{ipcw}(t) - C^{*}_{ipcw}(t)) = m^{-1/2} \sum_{i=1}^m IC^*_{ipcw}(\hat{T}_i, \Delta_i, X_i; t) + \text{op}(1).
\]

Here \( IC^*_{ipcw}(t) \) is the influence function of \( \hat{C}_{ipcw}(t) \), which is straightforward but tedious to derive the explicit form. However, for practical purposes under (A1) and (A2) the correspondence between the functional delta method and the bootstrap (Gill 1989) justifies the following construction of confidence limits for comparing two rival prediction models.

a) Draw bootstrap sets \( D^b_m \) for \( b = 1, \ldots, B \) from the validation set \( D_m \) either by sampling with replacement of size \( m \) or without replacement of size \( m_0 < m \).

b) Evaluate the predictions \( M_n^{(1)}(t; X_i) \) and \( M_n^{(2)}(t; X_i) \) of two rival models at all bootstrap samples \( i \in D^b_m \).

c) Compute \( \hat{G}_b \) and then \( \hat{W}_{bi,j} \) based on a model for the conditional censoring distribution and the data in \( D^b_m \).

d) For each of the two models compute

\[
\hat{C}^b_{ipcw}(t) = \frac{\sum_{i \in D^b_m} \sum_{i \in D^b_m} \bar{N}_{ij} Q_{ij}(t) \bar{Y}_i(t) \Delta_i \hat{W}^{-1}_{bi,j}}{\sum_{i \in D^b_m} \sum_{i \in D^b_m} \bar{N}_{ij} \bar{Y}_i(t) \Delta_i \hat{W}^{-1}_{bi,j}}.
\]

e) For each \( b = 1, \ldots, B \) compute the differences \( \xi_b = \hat{C}^b_{ipcw}(t; M_n^{(1)}) - \hat{C}^b_{ipcw}(t; M_n^{(2)}) \) and derive 95% bootstrap confidence limits as the 0.025 and 0.975 quantiles of \( \xi_1, \ldots, \xi_B \).
6.2 Bootstrap cross-validation

In many applications rival models have to be developed and compared based on the same data set. For the remaining of this section suppose that the training data \( D_n \) is all we have. The estimation of prediction performance in this situation is a complex task, since regression strategies may vary in their level of complexity, and in particular may overfit the current data set Efron & Tibshirani (1997). The aim is to estimate for each prediction model the discrimination ability which can be expected in new patients. The estimate of \( \hat{C}(t) \) obtained when testing the model with the same data where the model was developed is the so-called apparent or re-substitution estimate. It may severely overestimate the expected discrimination ability of the model. Cross-validation estimates of discrimination performance are obtained as the average over repeated splits of the data into training set and validation set. A golden rule is that all steps of modelling, including selection of candidate predictor variables, are repeated in each training set (Simon 2008). A suitable cross-validation algorithm which uses the bootstrap to split the data to compare two rival modelling strategies is as follows.

a) Draw bootstrap samples \( D_{nb}^b \) for \( b = 1, \ldots, B \), from \( D_n \) either with replacement of size \( n \) or without replacement of size \( m < n \).

b) Apply two modelling strategies \( S^{(1)} \) and \( S^{(2)} \) to \( D_{nb}^b \) to obtain prediction models \( M_{nl,b}^{(1)} \) and \( M_{nl,b}^{(2)} \) which solely depend on \( D_{nb}^b \).

c) Predict survival by evaluating predictions from both models \( M_{nl,b}^{(l)} \), \( l = 1, 2 \), for all subjects not in the \( b \)th bootstrap sample.

d) For each model use the subjects not sampled into \( D_{nb}^b \) to compute

\[
\hat{C}_{ipcw}^b(t) = \frac{\sum_{i \notin D_n} \sum_{j \notin D_n} \tilde{N}_{ij} \hat{Q}_{ij}(t) \tilde{Y}_i(t) \Delta_i \tilde{W}_{i,j}}{\sum_{i \notin D_n} \sum_{j \notin D_n} \tilde{N}_{ij} Y_i(t) \Delta_i \tilde{W}_{i,j}}.
\]
e) For each model compute the bootstrap cross-validation estimate

$$C_{ipcw,bcv}(t) = \frac{1}{B} \sum_{b=1}^{B} C_{ipcw}^{b}(t).$$  \hfill (8)

f) Compute the differences $\xi_b = C_{ipcw}^{b}(t; M_{n,b}^{(1)}) - C_{ipcw}^{b}(t; M_{n,b}^{(2)})$ for $b = 1, \ldots, B$ and construct confidence limits for the difference as the 0.025 and 0.975 quantiles of $\xi_1, \ldots, \xi_B$.

This algorithm seems well suited to compare the discriminative ability of rival prediction modelling strategies, as it directly compares the models built on the same bootstrap sample.

However, there are two issues. The first is that it is well known that the bootstrap-cross-validation estimate will systematically underestimate the expected discrimination of the prediction model. This bias is common to cross-validation estimates and due to the reduced information in the bootstrap training set. In various attempts to clear this problem Efron (1983), Efron & Tibshirani (1997) proposed several improvements, amongst others the “.632 bootstrap” and “.632+ bootstrap”. Both are linear combinations of the apparent and the bootstrap-cross-validation performance. However, the theoretical results regarding consistency of the “.632+ bootstrap” are still incomplete, and for a method like the random survival forest, for which the apparent performance is near perfect, its validity is unclear. On the other hand, for the direct comparison of two models the bootstrap cross-validation performance yields fair comparison under the mild assumption that the learning curves (increase in performance for increasing sample size) of the rival modelling strategies are comparable.

The second issue is that based on each bootstrap sample both models predict the same subjects not in sample. Hence by using the paired situation one could improve the power of the method (van de Wiel et al. 2009).
It is essential that the models are build and validated using the same bootstrap samples. The R library pec (Gerds 2009) will provide software for this task, for the simple and the IPCW estimates of the truncated \( C \)-index.

7 The prostate cancer study

The main objective of the study published by Kattan et al. (2000) was to develop a nomogram that accurately predicts the response to radiotherapy treatment of prostate cancer patients. Various statistical modelling approaches were tried and compared by means of Somers’ rank correlation coefficient. This is defined as \( D = 2 \cdot C - 1 \) where \( C \) is the unrestricted version of the concordance index (see section 4.1). Here we re-consider two of the models and compare them to two other not previously tried methods: an Aalen additive hazard regression model and a random survival forest model.

7.1 The data

The primary endpoint used in the study is time to tumor recurrence which was observed for 268 of the 1042 enrolled patients. The remaining 774 patients were right censored at the end of their follow-up time. There are 5 predictors considered: PSA, clinical stage (CS), Gleason score sum (GSS), radiation dose (RD) and hormone therapy (HT). The predictors PSA, Gleason score sum and radiation dose are on a continuous scale, hormone therapy is binary, either yes or no, and the clinical stage is categorical, taking one of the values (T1c, T2a, T2b ,T2c, T3ab ,T3c).
7.2 Prediction models

We consider the two rival Cox regression models considered previously in Kattan et al. (2000). They are defined and fitted in R using the following formulae:

Model 1 (Cox): 
\[ \text{cph}(\text{Surv}(\text{time}, \text{status}) \sim \log(\text{PSA}) + \text{CS} + \text{HT} + \text{GSS} + \text{RD}) \]

Model 2 (CoxSpline): 
\[ \text{cph}(\text{Surv}(\text{time}, \text{status}) \sim \text{rcs}(\log(\text{PSA}),3) + \text{CS} + \text{HT} + \text{rcs}(\text{GSS},3) + \text{rcs}(\text{RD},3)) \]

using the Cox regression routine \text{cph} of the R add-on library \text{rms} (Harrell 2009). The R notation reads as follows: \text{Surv}(\text{time}, \text{status}) is the survival response, \text{rcs}(x,3) denotes restricted cubic splines with 3 knots applied to predictor \( x \), and \( \log(x) \) the logarithmic transformation applied to predictor \( x \). We also consider a non-parametric Aalen additive regression model fitted with the functionality of the R add-on library \text{timereg}:

Model 3: 
\[ \text{Aalen}(\text{Surv}(\text{time}, \text{status}) \sim \log(\text{PSA}) + \text{CS} + \text{HT} + \text{GSS} + \text{RD}) \]

Model 3 is different from the other two models as it allows that the ranking of the predicted survival chances for two patients may change over time. Based on the data of the prostate cancer patients the Aalen regression model identifies a significant time-varying cumulative regression coefficient of hormone therapy. The effect is first positive and then turns sign showing a negative after 20 months. Finally we consider a random survival forest model with 1000 survival trees (Ishwaran et al. 2008) fitted with the R add-on library \text{randomSurvivalForest}:

Model 4: 
\[ \text{rsf}(\text{Survrsf}(\text{time}, \text{status}) \sim \log(\text{PSA}) + \text{CS} + \text{HT} + \text{GSS} + \text{RD}) \]

Model 4 differs from the other models as it belongs to the machine learning culture and thus is optimized for prediction making by internal cross-validation (Breiman 2001).
Fig. 1 The survival and the censoring probability in four radiation dose strata (RD1: [6000, 7000), RD2: [7000, 7500), RD3: [7500, 8000), RD4: [8000, 8640]). In the left panel the curves are estimated with the usual Kaplan-Meier estimator, in the right panel with the Kaplan-Meier estimator for the censoring times. The numbers at risk are given below the figure.

7.3 Estimating the weights

By study design the dose of radiation therapy depends on when the patient entered the study. This yields strong effects of almost all predictors on the censoring mechanism. To illustrate this, Figure 1 shows the dependence of the censoring distribution on the radiation dose. For estimation of (7) the weights $\hat{W}$ are derived from the following Cox regression model for $G$:

$$\text{cph(Surv(time,1-status) } \sim \text{ PSA + CS + HT + GSS + RD).}$$

(9)
Table 1  Results Cox regression (model 1) applied to the survival times (left panel) and of the model in display (9) applied to the censoring times (right panel). The radiation dose values have been centered at the mean value and scaled with root mean square value.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Survival distribution</th>
<th>Censoring distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Risk ratio</td>
<td>CI 95</td>
</tr>
<tr>
<td>Psa</td>
<td>1.003</td>
<td>[1.001; 1.004]</td>
</tr>
<tr>
<td>Radiation dose</td>
<td>0.630</td>
<td>[0.544; 0.731]</td>
</tr>
<tr>
<td>Gleason score sum</td>
<td>1.197</td>
<td>[1.095; 1.308]</td>
</tr>
<tr>
<td>Clinical stage=T1c</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Clinical stage=T2a</td>
<td>1.662</td>
<td>[0.984; 2.806]</td>
</tr>
<tr>
<td>Clinical stage=T2b</td>
<td>2.446</td>
<td>[1.580; 3.787]</td>
</tr>
<tr>
<td>Clinical stage=T2c</td>
<td>3.453</td>
<td>[2.235; 5.335]</td>
</tr>
<tr>
<td>Clinical stage=T3ab</td>
<td>3.039</td>
<td>[1.810; 5.102]</td>
</tr>
<tr>
<td>Clinical stage=T3c</td>
<td>4.183</td>
<td>[2.704; 6.468]</td>
</tr>
<tr>
<td>Hormone therapy=Yes</td>
<td>0.915</td>
<td>[0.689; 1.215]</td>
</tr>
</tbody>
</table>

The results of this Cox regression analysis are given in Table 1, together with the results of Model 1. The table shows a strong dependence of the censoring hazards on the predictor variables.

7.4 Discrimination results

We estimate the discrimination curves on a grid of time points between the time origin and 6 years of length 6 months. The estimation is based on equation (8) for which $B = 1000$ bootstrap samples are drawn with replacement from the original data. The weights $\hat{W}_{i,j}$ are derived from the model described in equation (9) and Table 1. Figure 2 and Table 2 show and compare the discrimination ability of the four models over time.
Table 2  Differences in $\hat{C}_{pcw,bcv}(t)$ between modelling strategies. In brackets 95% bootstrap confidence limits for differences as described in section 6.2.

<table>
<thead>
<tr>
<th>$t$ (months)</th>
<th>Cox vs RSF</th>
<th>Cox vs CoxSpline</th>
<th>Cox vs Aalen</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>$-1.16 [-6.99; 4.43]$</td>
<td>$-1.19 [-3.63; 1.16]$</td>
<td>$-2.42 [-7.28; 2.44]$</td>
</tr>
<tr>
<td>24</td>
<td>$0.09 [-4.13; 4.55]$</td>
<td>$-1.18 [-3.14; 0.75]$</td>
<td>$-0.32 [-3.1; 2.35]$</td>
</tr>
<tr>
<td>36</td>
<td>$1.27 [-2.71; 5.28]$</td>
<td>$-0.73 [-2.36; 1.09]$</td>
<td>$0.41 [-1.25; 2.32]$</td>
</tr>
<tr>
<td>48</td>
<td>$1.21 [-3.15; 5.73]$</td>
<td>$-0.6 [-2.24; 1.38]$</td>
<td>$0.26 [-1.65; 2.26]$</td>
</tr>
<tr>
<td>60</td>
<td>$0.91 [-3.59; 5.35]$</td>
<td>$-0.51 [-2.26; 1.83]$</td>
<td>$0.52 [-1.93; 3.72]$</td>
</tr>
</tbody>
</table>

Fig. 2 Discrimination curves comparing the four models. The direct comparison of model 4 (RSF) to model 3 (Aalen) and to model 2 (CoxSpline) are not shown. The curves are estimated using the bootstrap cross-validation procedure described in section 6 with 1000 bootstrap samples and weights derived from the model in (9) for the censoring distribution.
For example, $\hat{C}_{ipcw,bcv}(12)$ includes only pairs of patients where at least one died within the first 12 months after inclusion into the study. Here the Aalen model (Model 3) performs best showing $\hat{C}_{ipcw,bcv}(12) = 77.47\%$. The Cox model (Model 1) achieves $\hat{C}_{ipcw,bcv}(12) = 74.67\%$ and the spline version (Model 2) $\hat{C}_{ipcw,bcv}(12) = 75.6\%$ and the random survival forest (Model 4) $\hat{C}_{ipcw,bcv}(12) = 76.2\%$.

At five years the picture has changed: the spline Cox model wins with $\hat{C}_{ipcw,bcv}(60) = 72.8\%$ whereas the Cox model (Model 1) achieves $\hat{C}_{ipcw,bcv}(60) = 72.3\%$, the Aalen model (Model 3) $\hat{C}_{ipcw,bcv}(60) = 71.8\%$, and the random survival forest (Model 4) $\hat{C}_{ipcw,bcv}(60) = 71.4\%$.

The corresponding bootstrap cross-validation results, obtained with the simple estimator and the same bootstrap samples, at 60 months, also yielded $\hat{C}_{simp,bcv}(60) = 71.24\%$ for Model 1, $\hat{C}_{simp,bcv}(60) = 72.16\%$ for Model 2 and $\hat{C}_{simp,bcv}(60) = 70.64\%$ for Model 3. Anyhow, all differences are small.

8 Simulation studies

We first consider only a single binary and simulate data to compare the empirical properties of different IPCW estimators and the simple estimate. We then simulate data based on the prostate cancer study of section 7 to illustrate the effect of a hypothetical new predictive marker on $C(t)$.

8.1 A single binary marker

The following steps were repeated for different sample sizes: N=100, 200, 500, 1000. To generate a single binary marker, we drew independent random numbers with Bernoulli distribution with success probability 60\%. These values were then used to generate
failure times following a Weibull proportional hazard model with shape parameter equal to 2, scale parameter equal to 0.01 (Bender et al. 2006), and repeated for different regression coefficients $\beta = (0.7, 1.3, 2)$. Similarly the binary marker values were used to generate censoring times with a second Weibull proportional hazard model with shape parameter equal to 1, scale parameter equal to 0.1, and repeated for different regression coefficients $\gamma = (0.7, 1.3, 2)$. Right censored survival times were computed and their distribution estimated in both groups defined by the binary marker values. This yielded on between 40% and 80% censored. For each $(\beta, \gamma)$ constellation a single evaluation time point $t_0$ was chosen such that the expected number at risk at $t_0$ was roughly 10% of the data. This time point was hold fix for all sample sizes. Nevertheless, the results of few simulation runs, were the absolute number at risk at time $t_0$ was below 5 were removed.

A Cox regression model was fitted to a single simulated training data set of size $N$. To compute $\hat{C}(t_0)$ for this model the predicted survival order of two patients was defined by the survival probabilities predicted by this model at $t$. The true value of $\hat{C}(t_0)$ was approximated by the mean of 100 repetitions, were each time $\hat{C}(t_0)$ was estimated based on a large uncensored data set ($N=30,000$) simulated independently of the training data. For each $(\beta, \gamma, N)$ constellation, $\hat{C}(t_0)$ was estimated in each of 10,000 different test data sets simulated independently of the training data. We then computed for $M = 10,000$, $\text{MSE} = 1/M \times \sum_{i=1}^{M} (\hat{C}(t_0) - \hat{C}(t_0))^2$ and absolute bias=$| \hat{C}(t_0) - 1/M \sum_{i=1}^{M} \hat{C}(t_0) |$, for the simple estimate ($\hat{C}_{simp}$), and also for three different IPCW estimates, where the weights were estimated with the marginal Kaplan-Meier estimate for the censored times ($\hat{C}_{ipcw.1}$), with a Cox estimate for the censored times ($\hat{C}_{ipcw.2}$), and with the stratified Kaplan-Meier estimate for the censored times ($\hat{C}_{ipcw.3}$), respectively. Note that $\hat{C}_{ipcw.1}$ has a large-sample bias if $\gamma > 0$. 
Commonly in all settings where the binary marker has an effect on survival ($\beta > 0$), the simulation results show smaller MSE for the IPCW estimates $\hat{C}_{ipcw,2}$ and $\hat{C}_{ipcw,3}$ as compared to the simple estimate. Table 3 shows the results for sample size 200. The table also shows that the bias of $\hat{C}_{simp}$ and $\hat{C}_{ipcw,1}$ increase with increasing $\gamma$. This was the case for all other sample sizes (results not shown). For sample sizes 100 (results not shown) and 200, if the effect on the censoring is stronger than on survival ($\gamma > \beta$), the bias of $\hat{C}_{ipcw,2}$ and of $\hat{C}_{ipcw,3}$ was larger than the bias of the simple estimate. However, for sample sizes 500 and 1000 this phenomenon was only seen for $\beta = 0$ (results not shown).

For $\beta = 1.3$ and $\gamma = 1.3$, Figure 3 shows boxplots over 10,000 simulations for all sample sizes. The figure clearly illustrates the large-sample bias of the estimates $\hat{C}_{simp}$ and of $\hat{C}_{ipcw,1}$, consistency of $\hat{C}_{ipcw,2}$ and of $\hat{C}_{ipcw,3}$, and uniformly less variability of the latter.

8.2 Effect of a new marker

To simulate the situation of the prostate cancer study, in a first step we fitted parametric Cox-Weibull regression models to the event times and two the censoring times using the same variables as in Section 7. In a second step, the joint distribution of the predictor variables was approximated by assuming independence of the predictor variables and by normal distributions for logPSA, ggtot and dose using observed means and variances, and by binomial distributions for hormones and stage using observed frequencies. The result of both steps yielded a fully parametric model for the joint distribution of $(T_i, C_i, X_i)$. We simulated data from the parametric model which yielded
Table 3 Results of 10,000 simulations with a single binary predictor for sample size 200 and varying effects of the predictor on the survival distribution ($\beta$) and the censoring distribution ($\gamma$). The column Cens shows average percentage of censored over 10,000 runs. The table shows the mean squared error (MSE) and the absolute bias. For the IPCW estimates the weights were estimated with the marginal Kaplan-Meier estimate for the censored times ($\hat{C}_{ipcw.1}$), with a Cox estimate for the censored times ($\hat{C}_{ipcw.2}$), and with the stratified Kaplan-Meier estimate for the censored times ($\hat{C}_{ipcw.3}$), respectively.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$\gamma$</th>
<th>$C(t_0)$</th>
<th>Cens</th>
<th>MSE</th>
<th>Absolute</th>
<th>bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\hat{C}_{simp}$</td>
<td>$\hat{C}_{ipcw.1}$</td>
<td>$\hat{C}_{ipcw.2}$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>50</td>
<td>54.6</td>
<td>10.2</td>
<td>8.6</td>
<td>8.6</td>
</tr>
<tr>
<td>0</td>
<td>0.7</td>
<td>50</td>
<td>67.4</td>
<td>16.5</td>
<td>14.2</td>
<td>13.6</td>
</tr>
<tr>
<td>0</td>
<td>1.3</td>
<td>50</td>
<td>75.3</td>
<td>31.3</td>
<td>28.0</td>
<td>32.9</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>49.9</td>
<td>79.8</td>
<td>94.0</td>
<td>84.3</td>
<td>147.3</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>58</td>
<td>48.1</td>
<td>8.0</td>
<td>7.2</td>
<td>7.2</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7</td>
<td>58</td>
<td>61.2</td>
<td>12.4</td>
<td>11.6</td>
<td>9.8</td>
</tr>
<tr>
<td>0.7</td>
<td>1.3</td>
<td>57.8</td>
<td>71.2</td>
<td>24.2</td>
<td>23.7</td>
<td>18.0</td>
</tr>
<tr>
<td>0.7</td>
<td>2</td>
<td>57.7</td>
<td>78.1</td>
<td>73.2</td>
<td>73.2</td>
<td>67.4</td>
</tr>
<tr>
<td>1.3</td>
<td>0</td>
<td>63.9</td>
<td>43.0</td>
<td>6.6</td>
<td>6.2</td>
<td>6.1</td>
</tr>
<tr>
<td>1.3</td>
<td>0.7</td>
<td>63.7</td>
<td>55.5</td>
<td>8.9</td>
<td>9.1</td>
<td>7.2</td>
</tr>
<tr>
<td>1.3</td>
<td>1.3</td>
<td>63.3</td>
<td>66.6</td>
<td>17.4</td>
<td>18.9</td>
<td>10.4</td>
</tr>
<tr>
<td>1.3</td>
<td>2</td>
<td>62.7</td>
<td>75.8</td>
<td>49.7</td>
<td>53.9</td>
<td>26.0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>69.1</td>
<td>38.0</td>
<td>5.3</td>
<td>5.0</td>
<td>4.9</td>
</tr>
<tr>
<td>2</td>
<td>0.7</td>
<td>68.8</td>
<td>49.0</td>
<td>6.6</td>
<td>7.8</td>
<td>5.5</td>
</tr>
<tr>
<td>2</td>
<td>1.3</td>
<td>68.2</td>
<td>60.3</td>
<td>12.9</td>
<td>16.9</td>
<td>6.6</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>67.1</td>
<td>71.8</td>
<td>36.8</td>
<td>45.6</td>
<td>10.7</td>
</tr>
</tbody>
</table>
Fig. 3 Results of the simulation study with a single binary predictor. Boxplots of results of 10,000 simulation runs are shown for $\beta = \gamma = 1.3$ and varying sample sizes.

Kaplan-Meier and Cox regression results that corresponded nicely to the corresponding results with the observed data of the prostate cancer study.

Then a hypothetical new binary biomarker was simulated, independent from the other predictor variables, based on the Bernoulli distribution with 50% chance of being positive. For different hypothetical regression effects ($\beta$) on survival of the hypothetical new predictor, we added the marker to the parametric Cox-Weibull model just described. We further assumed zero effect on censoring for the new marker simulated.
Table 4  Simulation results showing the differences in estimated $C(t)$ between a Cox regression model with and without adding a hypothetical new binary marker. The estimates are based on 1000 bootstrap-cross-validation steps and the IPCW estimate using a Cox regression model for estimating the weights.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>Difference</th>
<th>CI-95 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.74</td>
<td>[−0.74; 1.76]</td>
</tr>
<tr>
<td>1</td>
<td>4.85</td>
<td>[1.89; 7.33]</td>
</tr>
<tr>
<td>1.5</td>
<td>9.32</td>
<td>[5.65; 12.56]</td>
</tr>
<tr>
<td>2</td>
<td>15.81</td>
<td>[11.29; 19.89]</td>
</tr>
<tr>
<td>2.5</td>
<td>17.7</td>
<td>[13.4; 22.08]</td>
</tr>
<tr>
<td>3</td>
<td>24.81</td>
<td>[15.98; 35.49]</td>
</tr>
</tbody>
</table>

data sets with sample size $N = 1042$. For each $\beta$, $C(t)$ was estimated at different time points ($t = 6, 12, 18, 24, 36, 42$) months for two rival models: a Cox regression model (Model 1 of Section 7) and a new Cox regression model where we added $\beta$ times the hypothetical new marker to the linear predictor of the first model. Discrimination was estimated with weights based on a Cox regression model for the censoring times as described in section 7.3, and for each $\beta$, 1000 bootstrap-cross-validation steps were performed as described in Section 6.1.

Table 4 shows results at $t = 36$ months. Results at the other time points were similar (not shown). The results indicate that first for large effects $\beta \geq 1$, and corresponding risk ratio greater than 2.7, the analysis shows a significant improvement of discrimination ability. This finding confirms earlier results on binary response variables, which show that large odds ratios are needed to significantly improve predictive performance (see e.g. Pepe et al. 2004).
9 Discussion

We addressed a limitation of a popular approach for assessing the ability of a prediction model to discriminate individual risks: Various studies reported discrimination by a simple estimate of \( C(t_{\text{max}}) \), where \( t_{\text{max}} \) is the time point were the last patient either had the event or was lost to follow-up, and without explicitly stating this time point.

We have defined the discrimination curve whose values are truncated \( C \) statistics and introduced a new IPCW estimate which is consistent even if the censoring distribution depends on the predictor variables. In a small simulation study we compared the bias and mean squared error of a truncated version of the simple estimate, the here proposed IPCW estimate, and the IPCW estimate proposed by Liu & Jin (2009), Uno et al. (2009). We could demonstrate a large-sample bias of the simple estimate, and under dependent censoring, also for the latter IPCW estimate. Note that Gönen & Heller (2005) investigated the bias of the unrestricted version of the simple estimate: \( \hat{C}_{\text{simp}}(\infty) \); thus, their results do not cover the presented simulation results.

The methods presented here require an estimate of the censoring distribution. Generally if this estimate is not consistent then also the estimate of the truncated concordance index is not consistent. But the bias would be the same for two rival prediction models and thus the direct comparison should not be affected. In many applications it will be reasonable to assume that the censoring mechanism is independent of the predictor variables and then the marginal Kaplan-Meier estimate (ignoring the predictor variables) for the censoring times can be used to derive a consistent IPCW estimate. However, according to theory (Bickel et al. 1993, Van der Laan & Robins 2003) weighting based on the marginal Kaplan-Meier estimate of \( G \) (Liu & Jin 2009, Uno et al. 2009) yields inefficient estimates for \( C(t) \), even if the censoring is independent of the survival
times and the predictors. However, within the limitations of our simulation study we saw only a minor improvement of efficiency when the censoring times were independent of the predictors.

The proposed IPCW estimate does not assume that the prediction model is correctly specified, and it can thus be used to assess and compare all kinds of prediction models. The only requirement is that the models rank the risk or the survival chances of all subjects. The ranking can be obtained according to the predicted cumulative event probability at time $t$, or by predicted mean restricted lifetime, or a similar measure. By using the same estimating procedure for each one of several rival prediction models the results can be compared directly. This is a clear advantage over existing estimates.

When several rival modelling strategies are compared with a single performance measure, in many applications one would simply pick the model with the highest score. In some cases, however, it is desirable to infer significant differences between two prediction modelling strategies. Unfortunately, even though the IPCW estimate and its asymptotic properties are well studied (Van der Laan & Robins 2003) it is not straightforward to assess the variability of the IPCW estimates of $C(t)$, because the variability depends further on the variability of the prediction model. Uno et al. (2009) derive the asymptotic distribution when the prediction model is a regular, fully specified regression model and not subject to model selection. In many applications, however, the prediction model is developed using the data at hand and typically the strategies include data dependent model selection steps which should not be ignored for assessing the model (Harrell 2001, Simon 2005). It is thus difficult to theoretically assess the variability of the estimated concordance index unless the prediction model is given externally and the analysis based on independent validation data.
We presented algorithms using bootstrap and cross-validation to derive confidence limits also for the situation where only a single data set is available. Recently van de Wiel et al. (2009) introduced a general testing procedure which combines rank tests based on patient individual residuals with cross-validation. The concordance index is a correlation coefficient and does not define individual residuals. However, it seems possible to implement a version of the van de Wiel test by testing differences based on the concordance indicator of all pairs of patients: $Q_{ij}(t)N_{ij}$.

The main disadvantage of the simple estimate is that it estimates a quantity which depends on the follow-up period. This potentially gives the study designer or the data analyst an added opportunity to artificially terminate the follow-up period at a time point where the results are suitable. On the other hand, the results of the prostate cancer study presented here show that the differences are rather small between the truncated version of the simple estimates and the corresponding IPCW estimates.

References


Uno, H., Cai, T., Pencina, M. J., D’Agostino, R. B. & Wei, L. J. (2009). On the c-
statistics for evaluating overall adequacy of risk prediction procedures with censored 

van de Wiel, M., Berkhof, J. & van Wieringen, W. N. (2009). Testing the pre-
diction error difference between 2 predictors. Biostatistics, Advance access-

data and causality. Springer.