Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring

Paul Blanche*[,1,2], **Jean-François Dartigues**[1,2], and **Hélène Jacqmin-Gadda**[1,2]

[1] Université Bordeaux, ISPED, Centre INSERM U897-Epidemiologie-Biostatistique, F-33000 Bordeaux, France
[2] INSERM, ISPED, Centre INSERM U897-Epidemiologie-Biostatistique, F-33000 Bordeaux, France

Received 24 February 2012; revised 18 February 2013; accepted 17 April 2013

To quantify the ability of a marker to predict the onset of a clinical outcome in the future, time-dependent estimators of sensitivity, specificity, and ROC curve have been proposed accounting for censoring of the outcome. In this paper, we review these estimators, recall their assumptions about the censoring mechanism and highlight their relationships and properties. A simulation study shows that marker-dependent censoring can lead to important biases for the ROC estimators not adapted to this case. A slight modification of the inverse probability of censoring weighting estimators proposed by Uno et al. (2007) and Hung and Chiang (2010a) performs as well as the nearest neighbor estimator of Heagerty et al. (2000) in the simulation study and has interesting practical properties. Finally, the estimators were used to evaluate abilities of a marker combining age and a cognitive test to predict dementia in the elderly. Data were obtained from the French PAQUID cohort. The censoring appears clearly marker-dependent leading to appreciable differences between ROC curves estimated with the different methods.

Keywords: AUC; IPCW; Prediction; ROC curve; Survival analysis.

1 Introduction

For many diseases, it is useful to identify a marker or a combination of markers that enables the identification of subjects at high and low risk of the disease in the future. In prostate cancer, the value or the change in prostate-specific antigen level is frequently used to predict cancer recurrence after the initial therapy and then to decide whether or not to undergo a secondary therapy (Proust-Lima and Taylor, 2009). In Alzheimer's disease (AD), the decline in cognitive functions begins a long time before all the criteria for the clinical diagnosis are reached. To ensure safety and care of these declining patients, it would be useful to identify them as early as possible. In particular, AD treatments given after the clinical diagnosis have been shown to have modest effects and research is currently focussing on preventive treatment given in the prediagnosis phase (Vellas et al., 2006). To ensure sufficient power for this preventive trials and then to apply the preventive treatment if its efficacy is demonstrated, validated markers for detecting subjects at high risk of AD in next years are required.

The diagnostic accuracy of a quantitative marker is often evaluated by the ROC curve that displays the sensitivity (probability that the marker X be above the cutpoint c for a diseased subject) versus 1-specificity (where the specificity is the probability that X be below c for a healthy subject) for all the possible cutpoints c. The diagnostic accuracy is often summarized by the area under the ROC curve

*Corresponding author: e-mail: Paul.Blanche@isped.u-bordeaux2.fr, Phone: +33-5-57-57-95-76

(AUC) that may be interpreted as the probability that the marker value of a randomly chosen case is above the marker value of a randomly chosen healthy subject (Pepe, 2003). In a diagnostic study, the marker and disease are measured at the same time and are known for all participants. In prognostic studies, the marker is measured at a given time (considered as time 0 in the following) while the disease may occur at any time thereafter. Thus, sensitivity, specificity, and ROC curve are time-dependent and may be computed for different time durations t (window of prediction).

Heagerty and Zheng (2005) proposed several definitions of the ROC curve for survival time that are also discussed in Pepe et al. (2008) and Cai et al. (2006). In this paper, we focus on the *cumulative/dynamic* definition (Heagerty and Zheng, 2005) where a case is a subject diagnosed before time t and a healthy subject (a control) is a subject free of the disease at time t. From our point of view, this definition is the most relevant as clinicians often want to predict disease onset in a period of time rather than at a specific time t (as in *incident* sensitivity) and want to distinguish healthy subjects at the end of the same period rather than at a later prespecified time τ (as in *static* specificity).

Without censoring, one could easily estimate such quantities by empirical proportions. Sensitivity could be estimated by the observed true-positive fraction and specificity could be estimated by the true-negative fraction (Pepe, 2003). However, in practice, there is often loss to follow-up before the time point t. Therefore, for some subjects, it is impossible to know if the outcome occurred before the time point t. To deal with this censoring, several approaches have been proposed to estimate the cumulative sensitivity, the dynamic specificity, and the associated ROC curve. First, Heagerty et al. (2000) proposed estimators based on Bayes' theorem and the Kaplan–Meier estimator (denoted by $\mathrm{KM}_{\mathrm{HLP}}$ in the following). As this approach does not guarantee the monotonicity of estimators, they also proposed estimators based on the nearest neighbor estimator of the bivariate distribution of the marker and the time-to-event (denoted by NNE). Then, Chambless and Diao (2006) proposed two alternative methods. The first one deals with censoring by conditioning on observed event times as in the Kaplan–Meier estimator (denoted by $\mathrm{KM}_{\mathrm{CD}}$). The second method, studied in detail by Song and Zhou (2008), uses a model for $S(t|X)$, the conditional survival probability of the outcome at time t given the marker X. Finally, Uno et al. (2007) and Hung and Chiang (2010b) have independently proposed an inverse probability of censoring weighting method (denoted by IPCW).

Among these estimators, only the NNE (Heagerty et al., 2000) and the model-based one (Chambless and Diao, 2006) allow the censoring to depend on the marker, whereas this is often the case in epidemiology. For instance, in cohorts of elderly people, it is known that poor cognitive level at entry is associated with both AD risk and study dropout (Jacqmin-Gadda et al., 1997). More generally, the marginal independence assumption between censoring time and event time required by most of these estimators is much less tenable than the conditional independence assumption given the marker. However, the weighting approach of the IPCW estimator may be slightly modified to obtain a nonparametric estimator robust to marker-dependent censoring. The resulting estimator is denoted CIPCW in the following (for conditional IPCW). On the other hand, the model-based approach has two drawbacks: it does not preserve invariance to increasing transformation of the marker and can lead to biases when the model is misspecified. To our knowledge, all of these estimators have never been compared together; the most exhaustive comparisons by simulation were performed under the independence assumption (Viallon and Latouche, 2011) or under weak dependence structure (Chiang and Hung, 2010).

The goal of this paper is to review the estimators proposed in the literature, to point out their similarities and differences and to compare their behaviors when the censoring time depends on the marker.

The "Naive" estimator, the estimators assuming independent censoring and the estimators robust to marker-dependent censoring are reviewed in Sections 2, 3, and 4, respectively, and then compared for different censoring scenarios in a simulation study in Section 5. Section 6 presents an application to the PAQUID cohort of elderly people (Jacqmin-Gadda et al., 1997; Amieva et al., 2005) to evaluate a predictive marker for dementia based on a cognitive test. Section 7 discusses limitations and possible extensions.

2 Naive estimator of time-dependent ROC curve

2.1 Notations and definitions

Let X_i denote a quantitative marker, T_i a time-to-event, C_i a censoring time, $\delta_i = I(T_i \leq C_i)$ the indicator of event, and $T_i^* = \min(T_i, C_i)$ the observed time for a subject i. We observe a sample of n independent subjects i: $\{(T_i^*, \delta_i, X_i), i = 1, \ldots, n\}$. To simplify the presentation of the estimators, we assume there are no ties in either the observed times $\{T_i^*, i = 1, \ldots, n\}$ or in the quantitative marker values $\{X_i, i = 1, \ldots, n\}$. We discuss adaptation of estimators for ties at the end of Section 4. Hereafter, we denote $S(t)$ the survival probability $\mathrm{P}(T > t)$ and $S(t|X)$ the conditional survival probability $\mathrm{P}(T > t|X)$. $I(\cdot)$ denotes the indicator function.

For a threshold $c \in \mathbb{R}$ and a given time t, Heagerty and Zheng (2005) defined the *cumulative* sensitivity $Se(c, t)$ and the *dynamic* specificity $Sp(c, t)$ by

$$Se(c, t) = \mathrm{P}(X > c|T \leq t) \quad \text{and} \quad Sp(c, t) = \mathrm{P}(X \leq c|T > t).$$

The corresponding time-dependent ROC curve, denoted by $ROC(t)$, is defined as the plot of $Se(c, t)$ versus $1 - Sp(c, t)$ for all possible values of c, i.e.,

$$ROC(t) = \{(1 - Sp(c, t), Se(c, t)), c \in \mathbb{R}\}.$$

The area under the $ROC(t)$ curve, denoted by $AUC(t)$, is therefore equal to $\mathrm{P}(X_i > X_j|T_i \leq t, T_j > t)$ with i and j the indexes of two independent subjects. In the following, given any estimators of sensitivity and specificity $\widehat{Se}$ and $\widehat{Sp}$, we denote $\widehat{ROC}(t) = \{(1 - \widehat{Sp}(c, t), \widehat{Se}(c, t)), c \in \mathbb{R}\}$, and $\widehat{AUC}(t)$ the area under the $\widehat{ROC}(t)$ curve.

2.2 "Naive" estimator

When there is no censoring, $Se(c, t)$ can be estimated as the proportion of subjects with $X_i > c$ among subjects diagnosed before t and $Sp(c, t)$ can be estimated as the proportion of subjects with $X_i \leq c$ among subjects free of disease at time t.

With censored data, the "Naive" estimator is computed by removing all subjects censored before time point t. Thus, for each subject kept in this subsample, we know if the event occurred before time t or not. "Naive" estimators of sensitivity and specificity can then be defined by observed true-positive and true-negative fractions in this subsample,

$$\widehat{Se}(c, t) = \frac{\sum_{i=1}^{n} \delta_i I(X_i > c, T_i^* \leq t)}{\sum_{i=1}^{n} \delta_i I(T_i^* \leq t)}, \qquad \widehat{Sp}(c, t) = \frac{\sum_{i=1}^{n} I(X_i \leq c, T_i^* > t)}{\sum_{i=1}^{n} I(T_i^* > t)}. \tag{1}$$

As a consequence, the area under the resulting step $\widehat{ROC}(t)$ curve is equal to

$$\widehat{AUC}(t) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i I(T_i^* \leq t, T_j^* > t) I(X_i > X_j)}{\sum_{i=1}^{n} \delta_i I(T_i^* \leq t) \sum_{j=1}^{n} I(T_j^* > t)}.$$

These estimators induce a loss of information and are often biased depending on the censoring distribution. We show in the Appendix that as $n \to \infty$, we obtain:

$$\widehat{Se}(c,t) \xrightarrow{a.s} Se(c,t) \times \frac{\mathrm{P}\big(T \leq C | X > c, T \leq t\big)}{\mathrm{P}\big(T \leq C | T \leq t\big)},$$

$$\widehat{Sp}(c,t) \xrightarrow{a.s} Sp(c,t) \times \frac{\mathrm{P}\big(C > t | X \leq c, T > t\big)}{\mathrm{P}\big(C > t | T > t\big)},$$

$$\text{and} \quad \widehat{AUC}(t) \xrightarrow{a.s} AUC(t) \times \frac{\mathrm{P}(T_i \leq C_i, C_j > t | T_i \leq t, T_j > t, X_i > X_j)}{\mathrm{P}(T_i \leq C_i, C_j > t | T_i \leq t, T_j > t)},$$

with i and j the indexes of two independent subjects. These results show that the "Naive" specificity estimator is consistent if censoring is independent of X and T while the "Naive" sensitivity and AUC estimators may be biased in this case since T depends on X. For instance, with independent uniform censoring over $[0, t]$, $\mathrm{P}(T \leq C | X > c, T \leq t) \neq \mathrm{P}(T \leq C | T \leq t)$ for some c as soon as T depends on X. However, we will see in Section 5 that this bias can be small in practice due to conditioning on $T \leq t$.

3 Estimators assuming independent censoring

These estimators were proposed to account for the information brought by censored subjects with unknown status at time t.

3.1 Kaplan–Meier estimator of Heagerty et al. (2000) ($\mathrm{KM}_{\mathrm{HLP}}$)

To deal with censored data, Heagerty et al. (2000) first proposed to use Bayes' theorem to rewrite sensitivity and specificity as functions of easy computable terms in presence of censoring. They proposed the estimators

$$\widehat{Se}(c,t) = \frac{[1 - \hat{S}(t | X > c)](1 - \hat{F}_X(c))}{1 - \hat{S}(t)} \quad \text{and} \quad \widehat{Sp}(c,t) = \frac{\hat{S}(t | X \leq c)\hat{F}_X(c)}{\hat{S}(t)} \tag{2}$$

with $\hat{F}_X(\cdot)$ the empirical distribution function of the marker X, $\hat{S}(\cdot)$ the Kaplan–Meier estimator of the marginal survival function of the time-to-event T (Kaplan and Meier, 1958), and $\hat{S}(\cdot | X > c)$ and $\hat{S}(\cdot | X \leq c)$ the Kaplan–Meier estimators computed on the subset of subjects with marker values such as $X > c$ and $X \leq c$, respectively.

As the Kaplan–Meier estimator requires independence between the censoring time and the event time, these estimators are not robust to marker-dependent censoring. Another problem is that $\widehat{Se}(c,t)$ and $\widehat{Sp}(c,t)$ are not necessarily monotone in c nor bounded in $[0, 1]$. Therefore, the corresponding $\widehat{ROC}(t)$ curve is neither monotone nor necessarily included in the square $[0, 1] \times [0, 1]$. This is due to the fact that the conditional survival functions are estimated on different subsamples when c varies.

Heagerty et al. (2000) proposed a Bootstrap approach for computing the variances and the confidence intervals of these estimates.

3.2 Kaplan–Meier-like estimator of Chambless and Diao (2006) ($\mathrm{KM}_{\mathrm{CD}}$)

Chambless and Diao (2006) proposed a Kaplan–Meier-like estimator that consists in a recursive computation using the risk sets at each event time. Let us consider the ordered observed event times

$s_0 = 0 < s_1 < s_2 < \cdots < s_{m(t)}$, with s_k the k-th observed event time and $s_{m(t)}$ the last observed event time before time point t. Chambless and Diao (2006) proposed the following estimators, which we present with a slightly different formulation in order to make the comparisons with the other estimators easier:

$$\widehat{Se}(c,t) = \frac{\sum_{k=1}^{m(t)} I(X_{d(k)} > c)(\hat{S}(s_{k-1}) - \hat{S}(s_k))}{1 - \hat{S}(s_{m(t)})}$$

$$\text{and} \quad \widehat{Sp}(c,t) = \frac{\hat{F}_X(c) - \sum_{k=1}^{m(t)} I(X_{d(k)} \leq c)(\hat{S}(s_{k-1}) - \hat{S}(s_k))}{\hat{S}(s_{m(t)})}, \tag{3}$$

where $d(k)$ is the index of the subject who experiences the event at time t_k. Here again, $\hat{S}(\cdot)$ denotes the Kaplan–Meier estimator of the survival function of the time-to-event T. The indicator $I(X_{d(k)} > c)$ estimates $\mathrm{P}(X > c | s_{k-1} < T \leq s_k)$, the indicator $I(X_{d(k)} \leq c)$ estimates $\mathrm{P}(X \leq c | s_{k-1} < T \leq s_k)$, and the difference $\hat{S}(s_{k-1}) - \hat{S}(s_k)$ estimates $\mathrm{P}(s_{k-1} < T \leq s_k)$.

By contrast to $\mathrm{KM}_{\mathrm{HLP}}$, this sensitivity estimator is monotone from 0 to 1. However, the specificity estimator is not monotone and is not necessarily bounded in [0, 1]. Therefore, the corresponding $\widehat{ROC}(t)$ curve is neither monotone nor necessarily included in the square [0, 1] × [0, 1]. Indeed, if we order subjects according to X, the change of specificity between two thresholds corresponding to two successive observed values $X_{(i)} < X_{(i+1)}$ of the marker X is negative when $\delta_{(i+1)} = 1$ and $T^*_{(i+1)} < t$. This is due to the fact that the change of the Kaplan–Meier estimator between two successive observed times s_{k-1} and s_k is always greater than or equal to $1/n$.

Chambless and Diao (2006) suggested to use bootstrapping for computing variances of these estimators and their confidence intervals.

3.3 Inverse probability of censoring weighting

Uno et al. (2007) and Hung and Chiang (2010b) separately proposed to correct the "Naive" estimator by weighting the observations kept in the subsample of uncensored subjects before time t by their probability of being kept in the subsample, that is their probability of being uncensored; i.e., they proposed:

$$\widehat{Se}(c,t) = \frac{\sum_{i=1}^{n} I(T_i^* \leq t, X_i > c) \frac{\delta_i}{n\hat{S}_C(T_i^*)}}{\sum_{i=1}^{n} I(T_i^* \leq t) \frac{\delta_i}{n\hat{S}_C(T_i^*)}} \quad \text{and} \quad \widehat{Sp}(c,t) = \frac{\sum_{i=1}^{n} I(T_i^* > t, X_i \leq c)}{\sum_{i=1}^{n} I(T_i^* > t)}, \tag{4}$$

where $\hat{S}_C(\cdot)$ is the Kaplan–Meier estimator of the survival function of the censoring time C. $\hat{S}_C(T_i^*)$ estimates the probability of being uncensored at the observed time T_i^*.

This specificity estimator is the same as the "Naive" one because weights are all equal to $1/(n\hat{S}_C(t))$, which allows simplification of the formula. Although not mentioned by the authors, this sensitivity estimator is identical to the $\mathrm{KM}_{\mathrm{CD}}$ estimator given at formula (3). Indeed, Satten and Datta (2001) showed that, if we order subjects according to T^*, the change of the Kaplan–Meier estimator between two successive observed times $T^*_{(i-1)}$ and $T^*_{(i)}$ is $\hat{S}(T^*_{(i-1)}) - \hat{S}(T^*_{(i)}) = \delta_{(i)}/(n\hat{S}_C(T^*_{(i)}))$ and, as a consequence, we also have $\sum_{i=1}^{n} I(T_i^* \leq t) \frac{\delta_i}{n\hat{S}_C(T_i^*)} = 1 - \hat{S}(t)$.

Interestingly, IPCW sensitivity and specificity estimators and the corresponding $\widehat{ROC}(t)$ curve are monotone and bounded in [0, 1]. Given that $\frac{1}{n}\sum_{i=1}^{n} I(T_i^* > t) = \hat{S}_C(t)\hat{S}(t)$, some simple algebra lead to the following formula for the area under the resulting step $\widehat{ROC}(t)$ curve (Hung and Chiang, 2010a):

$$\widehat{AUC}(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I(T_i^* \le t, T_j^* > t) I(X_i > X_j) \dfrac{\delta_i}{n^2 \hat{S}_C(T_i^*)\hat{S}_C(t)}}{\hat{S}(t)(1-\hat{S}(t))}.$$

The usual $\sqrt{n}$-consistency and asymptotic normality of these estimators have been established by Uno et al. (2007) and Hung and Chiang (2010a, b), and resampling techniques such as bootstrapping can be used to estimate the variances of the estimators.

4 Estimators for marker-dependent censoring

Using the Kaplan–Meier estimator, previous approaches assume independence between censoring time and time-to-event. However, in epidemiology, censoring often depends on the marker. Thus, time-to-event and censoring cannot be assumed independent. They are more likely independent conditionally on the marker. The three following approaches allow censoring to depend on the marker and only assume the conditional independence assumption between censoring time and time-to-event given the marker.

4.1 Model-based approach

Chambless and Diao (2006) and Song and Zhou (2008) proposed to use a model-based estimator for the conditional survival probability $S(t|X)$. Modeling the probability of each subject i to be a case by $1 - S(t|X_i)$ or a control by $S(t|X_i)$ to deal with censoring, they proposed the estimators:

$$\widehat{Se}(c,t) = \frac{\sum_{i=1}^{n}(1-\hat{S}(t|X_i))I(X_i > c)}{\sum_{i=1}^{n} 1-\hat{S}(t|X_i)} \quad \text{and} \quad \widehat{Sp}(c,t) = \frac{\sum_{i=1}^{n}\hat{S}(t|X_i)I(X_i \le c)}{\sum_{i=1}^{n}\hat{S}(t|X_i)}. \tag{5}$$

Consequently, the area under the resulting $\widehat{ROC}(t)$ curve is

$$\widehat{AUC}(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{S}(t|X_j)(1-\hat{S}(t|X_i))I(X_i > X_j)}{\sum_{j=1}^{n}\sum_{i=1}^{n}\hat{S}(t|X_i)(1-\hat{S}(t|X_j))}.$$

Chambless and Diao (2006) suggested a Cox proportional hazards model $\hat{S}(t|X) = \hat{S}_0(t)^{\exp\hat{\beta}X}$; Song and Zhou (2008) have studied asymptotic properties in this case and have extended these estimators to include covariate adjustment. Usual $\sqrt{n}$-consistency and asymptotic normality have been established and resampling techniques such as Bootstrap can be used to estimate the variances. Recently, Foucher et al. (2010) proposed a similar model-based approach to deal with competing risks. This

method is appealing because sensitivity and specificity are monotone, bounded in [0, 1], and allow censoring to depend on the marker. Moreover, some simulations have shown that this approach is more efficient than the nonparametric ones (Chambless and Diao, 2006; Song and Zhou, 2008). However, this model-based approach can easily lead to biases if the survival model is misspecified as it was shown in a simulation study by Viallon and Latouche (2011). Moreover, this estimator does not preserve the invariance to an increasing transformation of the marker X, which is a desirable property of ROC curve estimator (Pepe, 2003, p. 125). For these reasons, we only focused on nonparametric approaches and chose not to compare this method to the others in our simulation study in Section 5.

4.2 NNE of Heagerty et al. (2000)

Heagerty et al. (2000) suggested to use the NNE of the bivariate distribution of (X, T) introduced by Akritas (1994). Although the presentation of their estimators is sligthly different in the original paper, we can define their estimators by formula (5) but replacing the model-based estimator of $S(t|X)$ by the nonparametric estimator proposed by Akritas (1994):

$$\hat{S}(t|X_i) = \prod_{j:T_j^* \le t} \left(1 - \frac{W_{\lambda_n}(X_i, X_j)}{\sum_l W_{\lambda_n}(X_i, X_l) I(T_l^* \ge T_j^*)}\right)^{\delta_j} \tag{6}$$

with $W_{\lambda_n}(X_i, X_j) = I(|\hat{F}_X(X_i) - \hat{F}_X(X_j)| < \lambda_n)$, and $\lambda_n \in (0, 1]$ a bandwidth. Thus, $\hat{S}(t|X_i)$ is the standard Kaplan–Meier estimator computed with the subset of the $2\lambda_n$ percent of subjects who have the nearest values of the marker X_i (except at the tails). This method was recently extended to the competing risks setting (Saha and Heagerty, 2010). One important feature of the NNE $\hat{S}(t|X_i)$ of Akritas (1994) compared to the more standard Beran (1981) estimator is that the resulting $\widehat{ROC}(t)$ curve is invariant to any monotone transformation of the marker. With this method, sensitivity and specificity estimators are monotone and bounded in [0, 1]. Moreover, this method allows the censoring to depend on the marker X and is nonparametric. Based on results of Akritas (1994), asymptotic properties of sensitivity, specificity, AUC, and partial AUC estimators have been studied by Cai et al. (2011), Hung and Chiang (2011), and Hung and Chiang (2010b). Usual $\sqrt{n}$-consistency and asymptotic normality have been established and resampling techniques such as Bootstrap can be used to estimate the variances. However, to our knowledge, no optimal rule has been proposed to choose the value of λ_n in practice, although results on real data set with finite sample size are sensitive to this choice. The theory only tells us that choosing $\lambda_n = \mathcal{O}(n^{-1/3})$ works when n is large enough (Heagerty et al., 2000).

Finally, let us note that due to smoothing, when applied to uncensored data, the NNE ROC curve estimator remains different from the usual empirical estimator based on empirical true- and false-positive fractions. By contrast, when applied to uncensored data, $\mathrm{KM}_{\mathrm{HLP}}$, $\mathrm{KM}_{\mathrm{CD}}$, and IPCW, all lead to the usual empirical estimator.

4.3 Conditional inverse probability of censoring weighting

Following the idea used by Gerds and Schumacher (2006) to extend the Brier score estimator of Graf et al. (1999), the IPCW estimator may be modified to be robust to marker-dependent censoring. The change consists in weighting the observations of uncensored subjects at time t by the conditional probability of being uncensored given the marker, instead of weighting by the marginal probability of

being uncensored. We suggest the following estimators, denoted CIPCW:

$$\widehat{Se}(c,t) = \frac{\sum_{i=1}^{n} I(X_i > c, T_i^* \le t)\frac{\delta_i}{n\widehat{S}_C(T_i^*|X_i)}}{\sum_{i=1}^{n} I(T_i^* \le t)\frac{\delta_i}{n\widehat{S}_C(T_i^*|X_i)}},$$

$$\text{and} \quad \widehat{Sp}(c,t) = \frac{\sum_{i=1}^{n} I(X_i \le c, T_i^* > t)\frac{1}{n\widehat{S}_C(t|X_i)}}{\sum_{i=1}^{n} I(T_i^* > t)\frac{1}{n\widehat{S}_C(t|X_i)}}. \tag{7}$$

The censoring survival probability $S_C(t|X) = \mathrm{P}(C > t|X)$ may be estimated using a Cox model or any other model. However, to avoid any parametric assumption, we propose to use the nonparametric estimator of Akritas (1994). Therefore, $\widehat{S}_C(t|X_i)$ corresponds to the right term of formula (6) where δ_j is replaced by $1 - \delta_j$. Let us remark that, by definition, $\widehat{S}_C(T_i^*|X_i)$ cannot be equal to zero when $\delta_i = 1$ and $\widehat{S}_C(t|X_i)$ cannot be equal to zero when $T_i^* > t$. Then, with the convention $0/0 = 0$, the estimators are always correctly defined.

The area under the resulting step $\widehat{ROC}(t)$ curve is

$$\widehat{AUC}(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I(X_i > X_j) I(T_i^* \le t, T_j^* > t)\frac{\delta_i}{n^2\widehat{S}_C(T_i^*|X_i)\widehat{S}_C(t|X_j)}}{\left(\sum_{i=1}^{n} I(T_i^* \le t)\frac{\delta_i}{n\widehat{S}_C(T_i^*|X_i)}\right)\left(\sum_{j=1}^{n} I(T_j^* > t)\frac{1}{n\widehat{S}_C(t|X_j)}\right)}$$

and corresponds to an estimator already suggested by Hung and Chiang (2010a).

When $\lambda_n < 1/n$, the CIPCW estimators are equal to the "Naive" ones; when $\lambda_n = 1$, for all subject i and all time t, $\widehat{S}_C(t|X_i)$ is equal to the marginal Kaplan–Meier estimator $\widehat{S}_C(t)$, and then CIPCW estimators are equal to the IPCWs. Thus, whatever the value of λ_n, CIPCW has a sensible interpretation. By contrast, choosing $\lambda_n = 1$ yields NNE estimators to $\widehat{Se}(c,t) = 1 - \widehat{Sp}(c,t)$ and so the $\widehat{ROC}(t)$ curve is equal to the first bisector, whereas choosing $\lambda_n < 1/n$ leads to $\widehat{S}(t|X_i) = 1 - I(T_i^* < t)\delta_i$ and so the NNE sensitivity estimator is equal to the "Naive" one but NNE specificity has no meaning in this case.

The CIPCW estimators are nonparametric, monotone, bounded in [0, 1], and robust to marker-dependent censoring. As for NNE, when using the Akritas' conditional Kaplan–Meier estimator, the $\widehat{ROC}(t)$ curve is invariant to any monotone increasing transformation of the marker X, and the bandwidth λ_n does not depend on the scale of X.

By contrast to NNE, when applied to uncensored data, CIPCW estimators are equal to usual empirical estimators, since $\widehat{S}_C(t|X_i)$ is equal to 1 for all time t and all subjects i in this case.

4.4 Adaptation of estimators with ties

Although the marker is assumed quantitative, ties among marker values are frequent in real data sets. In this case, $I(X_i > X_j)$ has to be replaced by $I(X_i > X_j) + \frac{1}{2}I(X_i = X_j)$ in all formulae for $\widehat{AUC}(t)$.

Some issues regarding the interpretation and the validity of ROC curves for ordinal data are discussed in Section 4.5 of Pepe (2003).

If there are ties among the observed times $\{T_i^*, i = 1, \ldots, n\}$, we need to adapt the definition of the conditional Kaplan–Meier estimator, and formula (6) becomes

$$\widehat{S}(t|X_i) = \prod_{s \in \mathcal{T}_n, s \leq t} \left(1 - \frac{\sum_{j=1}^{n} W_{\lambda_n}(X_i, X_j) I(T_j^* = s) \delta_j}{\sum_{l} W_{\lambda_n}(X_i, X_l) I(T_l^* \geq s)} \right),$$

with $\mathcal{T}_n$ the set of unique values of T_i^*. Moreover, in $\widehat{Se}(t, c)$ and $\widehat{Sp}(t, c)$ in Subsection 3.2, we need to replace $I(X_{d(k)} > c)$ by $\frac{\sum_{i=1}^{n} \delta_i I(T_i^* = t_k) I(X_i > c)}{\sum_{i=1}^{n} \delta_i I(T_i^* = t_k)}$ and $I(X_{d(k)} \leq c)$ by $\frac{\sum_{i=1}^{n} \delta_i I(T_i^* = t_k) I(X_i \leq c)}{\sum_{i=1}^{n} \delta_i I(T_i^* = t_k)}$.

5 Simulation study

The aim of the simulation study was to compare the behavior of the nonparametric estimators of the area under the $ROC(t)$ curve with both independent and marker-dependent censoring.

5.1 Simulation scenarios

Several scenarios were generated. For each one, we generated $N = 500$ samples $\{(T_i^*, X_i, \delta_i), i = 1, \ldots, n\}$ including $n = 300$ subjects. The marker X was generated from the standard normal distribution $\mathcal{N}(0, 1)$. The time-to-event T and the censoring time C were generated from proportional hazards models with a Weibull baseline hazard function: $\lambda_T(t; X) = \frac{\beta t^{\beta-1}}{\eta^\beta} \exp(\alpha X)$ and $\lambda_C(t; X) = \frac{\nu t^{\nu-1}}{\theta^\nu} \exp(\gamma X)$.

In each scenario, we set the proportion of censored data before time point $t = 1$ to be $\mathrm{P}(T_i^* \leq 1, \delta_i = 0) = 50\%$. We simulated 12 scenarios corresponding to two different true AUCs at time point $t = 1$ ($AUC(1) = 0.75$ and $AUC(1) = 0.85$, generated from the corresponding hazard ratios for one standard deviation increase of the marker $\exp(\alpha) \approx 2.30$ and $\exp(\alpha) \approx 4.15$, respectively), two different risks of event ($\mathrm{P}(T > 1) \approx 0.55$ or $\mathrm{P}(T > 1) \approx 0.73$), and three different hazard ratios for one standard deviation increase of the marker $HR_C = \exp(\gamma)$, for the association between the censoring time C and the marker X: independent censoring ($HR_C = 1$), moderate dependence ($HR_C = 1.35$), and strong dependence ($HR_C = 2.40$). The risk of censoring and the risk of event both increased with time in all scenarios ($\nu = 2$ and $\beta > 1$). Values of parameters η, β, and θ were chosen to control the previous values.

The proportion of censoring before $t = 1$ was equal to 50%, and the value $\mathrm{P}(T > 1) \approx 0.73$ was chosen to mimic the real data set of the PAQUID cohort at time $t = 10$ years. The hazard ratios HR_C have the same magnitude than the hazard ratio fitted on the PAQUID data presented in Section 6.

5.2 Results

For each scenario, we computed the bias and the root mean-squared error (RMSE) of the different estimators of $AUC(t)$ at time point $t = 1$. For NNE and CIPCW, we present results for $\lambda_n = 2.5\%$, 5%, and 10%. The true AUC was estimated as the mean over the 500 data sets of the AUC estimates computed by the usual empirical estimate with uncensored data $\{(T_i, X_i), i = 1, \ldots, n\}$. The biases

Table 1 Results of the simulation study : Average bias ($\widehat{AUC}(1) - AUC(1)$) multiplied by 100. Results from NNE and CIPCW estimates are given for $\lambda_n = 2.5\%$, 5%, and 10%.

						NNE			CIPCW		
AUC(1)	P($T > 1$)	"Naive"	KM_{HLP}	KM_{CD}	IPCW	2.5%	5%	10%	2.5%	5%	10%
$HR_C = 1$											
0.75	0.73	0.50	0.03	0.04	−0.06	−0.21	−0.43	−1.21	0.01	−0.01	−0.04
	0.53	0.94	0.02	0.10	−0.07	−0.32	−0.34	−0.99	0.34	0.12	−0.01
0.85	0.72	0.83	0.02	0.04	−0.03	−0.24	−0.38	−1.24	0.15	0.04	−0.03
	0.55	1.70	0.02	0.06	0.04	−0.70	−0.47	−0.97	0.58	0.23	0.04
$HR_C = 1.35$											
0.75	0.73	3.02	2.59	−4.26	2.20	−0.48	−0.61	−1.44	0.26	0.04	−0.07
	0.53	3.05	3.15	−7.96	1.81	−1.00	−0.60	−1.12	0.79	0.35	0.12
0.85	0.72	2.74	4.05	−3.39	1.84	−0.47	−0.48	−1.35	0.44	0.23	0.02
	0.55	3.53	3.90	−5.24	1.98	−1.25	−0.74	−1.12	1.01	0.49	0.33
$HR_C = 2.40$											
0.75	0.73	5.24	5.14	−13.82	3.92	−3.31	−2.44	−2.72	1.49	0.59	−0.06
	0.53	3.40	5.23	−23.11	1.85	−6.13	−3.87	−3.13	1.41	0.80	0.06
0.85	0.72	4.31	9.76	−11.67	3.21	−2.79	−2.00	−2.53	1.17	0.63	0.05
	0.55	4.52	8.27	−15.37	2.90	−4.11	−2.80	−2.67	1.60	1.02	0.66

and RMSE computed with reference to this ideal AUC estimate are displayed in Tables 1 and 2, respectively.

When censoring is independent of the marker ($HR_C = 1$), all estimators are consistent with similar RMSE. It is interesting to note that the bias for the "Naive" estimator is very small in these scenarios. As expected, simulations show that "Naive," KM_{HLP}, KM_{CD}, and IPCW are biased when censoring depends on the marker, and bias and RMSE increase with an increasing association between the marker and the censoring time. Whereas "Naive," KM_{HLP}, and IPCW overestimate the AUC, KM_{CD} underestimates it in these simulations. Even if IPCW assumes independent censoring like KM_{HLP} and KM_{CD}, it appears to be more robust to dependent censoring than KM_{HLP} and KM_{CD}. Interestingly, one can note that KM_{HLP} and KM_{CD} may be more biased than the "Naive" estimator for strongly dependent censoring.

NNE and CIPCW are both suitable when censoring depends on the marker. With respect to both bias and RMSE, these estimators perform as well when censoring is independent ($HR_C = 1$) or moderately dependent on the marker ($HR_C = 1.35$). However, when censoring strongly depends on the marker ($HR_C = 2.4$), NNE tends to be more biased than CIPCW. In these scenarios with $HR_C = 2.4$, there are neighborhoods of some X_i that include only subjects that either met the event before time t or are censored before time t. For these neighborhoods, the conditional Kaplan–Meier estimator of the time-to-event survival function $S(t|X_i)$ performs poorly. This is probably the main reason of the biases in NNE with highly dependent censoring, and this could explain why the bias increases when survival probability of time-to-event is lower (≈ 0.53). This also explains that NNE behavior depends more strongly on the size of the neighborhood λ_n when $HR_C = 2.4$. By contrast, CIPCW only requires the computation of the censoring survival function $S_C(t|X_i)$ for X_i such that $T_i^* \geq t$. Thus, there is always at least one subject at risk in each of these neighborhoods. Finally, the behavior of NNE appears more dependent on the choice of the bandwidth λ_n than the behavior of CIPCW. This is confirmed using other bandwidths $\lambda_n = 7.5\%$, 12.5%, and 15% (results not shown).

Table 2 Results of the simulation study: Root mean-squared error of $\widehat{AUC}(1)$ multiplied by 100. Results from NNE and CIPCW estimates are given for $\lambda_n = 2.5\%$, 5%, and 10%.

						NNE			CIPCW		
AUC(1)	P($T > 1$)	"Naive"	KM_{HLP}	KM_{CD}	IPCW	2.5%	5%	10%	2.5%	5%	10%
$HR_C = 1$											
0.75	0.73	4.12	4.27	4.27	4.25	4.20	4.18	4.31	4.27	4.13	4.08
	0.53	4.09	4.25	4.69	4.19	4.00	4.02	4.07	4.24	4.03	3.93
0.85	0.72	3.10	3.42	3.38	3.24	3.08	3.08	3.34	3.16	3.03	3.00
	0.55	3.34	3.44	3.57	3.31	3.03	3.02	3.16	3.27	3.09	3.01
$HR_C = 1.35$											
0.75	0.73	4.88	5.60	6.13	4.59	4.33	4.29	4.45	4.37	4.23	4.15
	0.53	4.82	5.85	9.26	4.35	4.29	4.14	4.27	4.32	4.21	4.11
0.85	0.72	3.91	5.78	4.93	3.57	3.15	3.17	3.48	3.16	3.10	3.13
	0.55	4.32	5.49	6.33	3.48	3.17	3.04	3.23	3.22	3.06	3.06
$HR_C = 2.40$											
0.75	0.73	6.57	9.02	14.62	5.76	6.22	6.01	6.06	5.30	5.46	5.54
	0.53	4.95	8.86	23.60	4.24	8.21	6.92	6.59	5.03	5.22	5.45
0.85	0.72	5.22	12.03	12.41	4.62	4.89	4.63	5.00	3.90	4.00	4.05
	0.55	5.08	9.76	15.86	3.95	5.41	4.51	4.47	3.49	3.33	3.42

6 Illustration on prediction of dementia

6.1 Objective

The objective of this analysis was to evaluate the predictive performance of a cognitive marker to predict dementia onset over a period of time of 5, 10, 15, or 17 years. The marker was previously defined as a linear combination of age at baseline (the main risk factor for dementia) and Digit Symbol Substitution Test (DSST) score (Wechsler, 1981) at baseline using a Cox model. The DSST explores attention and psychomotor speed and is highly associated with cognitive ageing and risk of dementia. It consists in filling in, in a 90-s interval, blank squares with the symbol that is paired to the digit displayed above the square, according to a code table of associated digits and symbols. Such a combined marker could be useful for preventive clinical trials to select population at high risk of dementia over the duration of the trial.

6.2 The PAQUID sample

The PAQUID cohort is a French prospective study on cognitive ageing including 3777 subjects aged 65 years and older and living at home at baseline. Subjects were initially interviewed at home in 1988 and 1, 3, 5, 8, 10, 13, 15, 17, and 20 years later. Their cognition was evaluated at each visit with several psychometric tests such as the DSST. In addition, dementia diagnosis was assessed at each visit by the investigating psychologist and then confirmed through a clinical examination by a neurologist if the subjects were screened positive by the psychologist.

The sample for this analysis included 2516 subjects nondemented at the initial visit, without missing value for the DSST at baseline and who were visited at least once thereafter. The baseline time for the analysis is the follow-up time from the baseline visit. Time to dementia onset is computed as the mean between the time at the visit of diagnosis and the time at the previous visit. Censoring time is the time at the last follow-up visit for the subject.

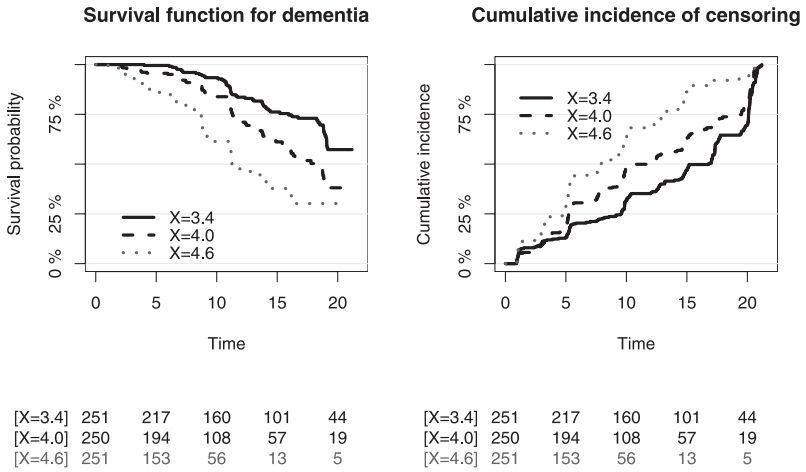

Figure 1 Survival function for dementia and cumulative incidence function for censoring. These curves are computed by conditional Kapan–Meier (formula (6) with $\lambda_n = 0.05$) for the first quartile ($X = 3.4$), the median ($X = 4$), and the third quartile ($X = 4.6$) of the distribution of the marker X. Numbers of subjects at risk at time $t = 0, 5, 10, 15, 20$ years are indicated below the graph.

Table 3 Proportion of censored, demented, and nondemented subjects for each time t ($n = 2516$ subjects).

Time t	Censored before t	Dementia before t	Free of dementia at t
5	566 (22.5%)	128 (5.1%)	1822 (72.4%)
10	1132 (45.0%)	326 (13.0%)	1058 (42.1%)
15	1389 (55.2%)	533 (21.2%)	594 (23.6%)
17	1444 (57.4%)	591 (23.5%)	481 (19.1%)

The mean age at enrollment for this sample was 74.1 (standard deviation = 6.2) and the mean DSST score was 27.5 (standard deviation = 11.5). The marker was defined by $X = 0.073 \times \text{AGE} - 0.052 \times \text{DSST}$ with AGE and DSST measured at enrollment and the coefficients 0.073 and 0.052 are the estimates from a Cox model for time to dementia. The mean of X was 4.00 (standard deviation = 0.89) with a minimum of 1.18 and a maximum of 6.79.

6.3 Results

The association between the marker and the risk of dementia and between the marker and the risk of censoring are illustrated in Fig. 1. A high value of the marker is associated with high risks of both dementia and censoring. For instance, the probability of being free of dementia is estimated at 0.93 and the probability of censoring at 0.31 at time $t = 10$ for the first quartile of the marker distribution, whereas they were, respectively, estimated at 0.63 and 0.61 for the third quartile. To illustrate the association between the censoring time and the marker, a Cox regression model was also fitted. The hazard ratio for one standard deviation increase of the marker value was estimated to $\widehat{HR}_C = 1.65$ (95% confidence interval [1.56, 1.74]). Thus, the censoring time clearly depends on the marker. We computed the ROC curves and the AUCs at time points $t = 5$, 10, 15, and 17 years. Table 3 displays the proportions of censored, demented, and healthy subjects before each time point.

Table 4 $\widehat{AUC}(t)$ of the $\widehat{ROC}(t)$ curve evaluating predictive ability of the cognitive marker ($X = 0.073 \times \text{AGE} - 0.052 \times \text{DSST}$) for dementia (PAQUID, $n = 2561$).

Time t	"Naive"	KM_{HLP}	KM_{CD}	IPCW	NNE	CIPCW
5	0.828	0.832	0.788	0.827	0.802	0.803
10	0.828	0.843	0.724	0.819	0.782	0.798
15	0.807	0.806	0.622	0.784	0.762	0.775
17	0.816	0.813	0.593	0.790	0.759	0.773

We present the $\widehat{ROC}(t)$ curve in Fig. 2 and the corresponding $\widehat{AUC}(t)$ in Table 4. For NNE and CIPCW estimators, due to the large sample size, we chose $\lambda_n = 0.01$ and therefore there were about $2 \times \lambda_n \times n \approx 50$ subjects included in each neighborhood (except for the boundaries) to compute the conditional Kaplan–Meier estimators. Comparison of AUC estimates leads to similar results to those of the simulation study where censoring depended moderately on the marker ($HR_C = 1.35$). For the two estimates that deal with marker-dependent censoring, NNE and CIPCW, results are close (the largest difference is 0.016) although the NNE is always slightly smaller than CIPCW. The "Naive," KM_{HLP}, and IPCW estimates are higher than those of NNE and CIPCW, while the KM_{CD} estimate is smaller than the NNE and CIPCW ones. As expected, the differences between estimates that handle independent censoring and the others increase with increasing time point t and so the percentage of censoring (Tables 3 and 4).

Finally, NNE and CIPCW estimates show that this cognitive marker exhibits good predictive power for the identification of dementia cases. As expected, $\widehat{AUC}(t)$ tends to decrease with increasing window of prediction t, reflecting poorer long-term predictive ability. However, this decline with time appears small, and the predictive ability remains quite good for a window of prediction as long as 17 years. This marker could probably be improved by including several cognitive tests and should be evaluated on a validation sample.

7 Discussion

In this paper, we have reviewed the estimators of *cumulative/dynamic* sensitivity, specificity, and ROC curve previously proposed for censored event-time, and detailed their properties and relationships. We then focused on the behavior of the nonparametric estimators when the censoring time depends on the marker, and thus is only conditionally independent from the event time given the marker. When the censoring depends on the marker, the simulation study has shown that some of these AUC estimators handling censoring may have poorer behavior than the "Naive" one that removes censored subjects. Only the NNE (Heagerty et al., 2000) and the CIPCW are robust in this case. As in Chambless and Diao (2006), our simulations have shown slight biases for the "Naive" estimator when censoring does not depend on the marker, and as in Hung and Chiang (2010a) they have confirmed that IPCW is quite robust to censoring that depends moderately on the marker. Based on the expression of the biases presented in Section 2.2, it is possible to simulate scenarios with independent censoring and larger biases for the "Naive" AUC. However, we failed to find simple realistic scenarios with independent censoring leading to large biases for the "Naive" estimators.

NNE and CIPCW both rely on the nearest neighbor Kaplan–Meier estimator proposed by Akritas (1994) to estimate the conditional survival function either for the event (NNE) or for the censoring (CIPCW). This partly drives their properties. First, NNE may have trouble in circumstances where

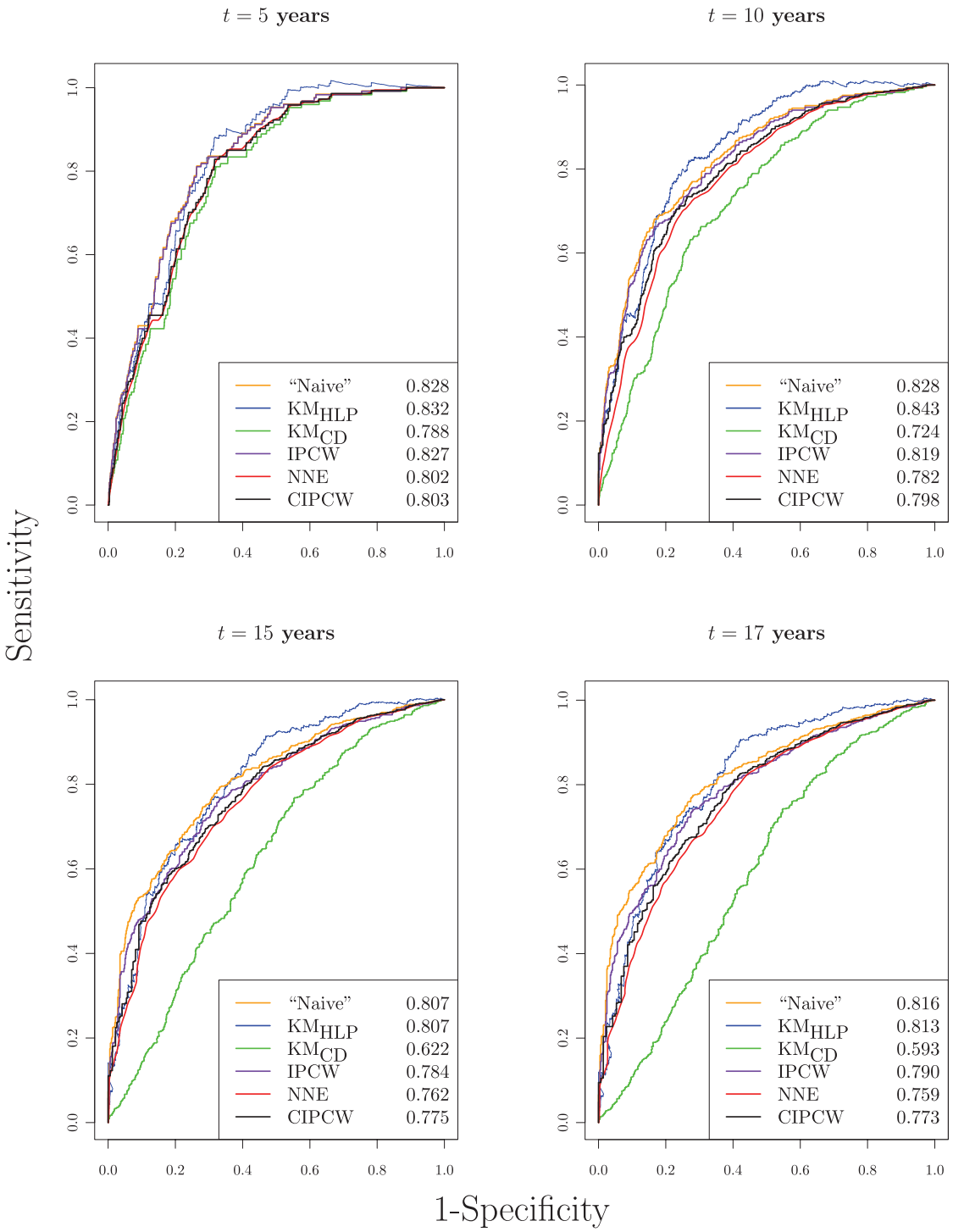

Figure 2 $\widehat{ROC}(t)$ curve and $\widehat{AUC}(t)$ for windows of prediction $t = 5$, 10, 15 and 17 years, estimated with the nonparametric estimators presented in Sections 2, 3, and 4. PAQUID, $n = 2516$ subjects.

there is no subject at risk at the time t of interest in some neighborhoods. However, this problem may only occur with small sample sizes, heavy censoring, and high association between marker and censoring. Performances of NNE and CIPCW should be equivalent with bigger sample sizes. Second, both estimators require the choice of a smoothing parameter λ_n (neighborhood size) but the two extreme values of λ_n correspond to sensible estimators for CIPCW ("Naive" or IPCW), whereas when λ_n tends to 1, NNE tends to a degenerated estimate such as $Se(c,t) = 1 - Sp(c,t)$. Besides, simulation results suggest that CIPCW could be less affected by the choice of λ_n. As most of the ROC curve estimators, CIPCW estimator is also a step function that better highlights the information available in the data compared to a smooth function.

CIPCW can be adapted to other informative censoring than marker-dependent censoring. For instance, if it is more realistic to assume that censoring does not depend on the marker X but on another measured marker Z, we just need to replace the estimator $\widehat{S}_C(\cdot|X_i)$ by $\widehat{S}_C(\cdot|Z_i)$ in the weights of CIPCW to ensure consistency.

CIPCW may also be extended for more complicated diagnostic rules involving several markers or repeated measures of one marker. For instance, Rizopoulos (2011) suggested to consider as positive, subjects with both the baseline measure of the marker X_0 greater than a threshold c and the measure one year later X_1 greater than $0.5c$. Although Rizopoulos (2011) used parametric sensitivity and specificity estimators, CIPCW could be easily extended to this case to propose an alternative nonparametric estimator. Assuming censoring depends only on the last value of the marker computation of CIPCW would only require the estimate of $\widehat{S}_C(t|X_1)$.

Saha and Heagerty (2010) extended NNE for competing risks using cumulative incidence functions, instead of conditional survival functions, to compute sensitivity and specificity. The estimators IPCW and CIPCW may also be generalized for competing risks since cumulative incidence functions can be estimated by IPCW (Scheike et al., 2008). However, this cannot be directly applied to PAQUID data to account for competing risk due to death. Indeed, the censoring times are different for the two events. Dementia may be assessed only if the subject completes the follow-up visit while the vital status is known for every subject. The exact date of death may be collected for all the subjects who died before the final visit (20 years) while the dementia time is interval censored. To our knowledge, only parametric multistate models (Joly et al., 2002) can account for this kind of interval censoring and thus defining a nonparametric ROC curve estimator for this kind of design seems not feasible.

To conclude, we showed that marker-dependent censoring may bias dramatically some ROC curve estimators for independently censored data. Thus, we recommend to use estimators for marker-dependent censoring (CIPCW or NNE) or IPCW that appears quite robust.

Softwares We used the `survivalROC` R package to compute KM_{HLP} and NNE. R code for KM_{CD}, IPCW, CIPCW is available on request from the corresponding author.

Acknowledgments This work was partly funded by a grant from France Alzheimer awarded to Hélène Jacqmin-Gadda in 2009. The PAQUID study, managed by Jean-François Dartigues, is funded by IPSEN and Novartis laboratories. We thank the two anonymous referees and the associate editor for their helpful comments.

Conflict of interest

The authors have declared no conflict of interest.

Appendix: Bias of "Naive" estimators

For all subsets $\mathcal{A}, \mathcal{B}, \mathcal{C}$ with non null probability, we have the equality:

$$\begin{aligned}\mathrm{P}(\mathcal{A}|\mathcal{B}\cap\mathcal{C}) &= \frac{\mathrm{P}(\mathcal{A}\cap\mathcal{B}\cap\mathcal{C})}{\mathrm{P}(\mathcal{B}\cap\mathcal{C})} = \frac{\mathrm{P}(\mathcal{A}\cap\mathcal{B}\cap\mathcal{C})}{\mathrm{P}(\mathcal{A}\cap\mathcal{C})} \times \frac{\mathrm{P}(\mathcal{A}\cap\mathcal{C})}{\mathrm{P}(\mathcal{C})} \times \frac{\mathrm{P}(\mathcal{C})}{\mathrm{P}(\mathcal{B}\cap\mathcal{C})} \\ &= \mathrm{P}(\mathcal{A}|\mathcal{C}) \times \frac{\mathrm{P}(\mathcal{B}|\mathcal{A}\cap\mathcal{C})}{\mathrm{P}(\mathcal{B}|\mathcal{C})}. \qquad \star\end{aligned}$$

Due to the law of large numbers, as $n \to \infty$ we have for the "Naive" sensitivity estimator:

$$\begin{aligned}\widehat{Se}(c,t) = \frac{\frac{1}{n}\sum_{i=1}^{n}\delta_i I(X_i > c) I(T_i^* \le t)}{\frac{1}{n}\sum_{i=1}^{n}\delta_i I(T_i^* \le t)} &\xrightarrow{a.s} \frac{\mathrm{P}\left(X > c, T \le t, T \le C\right)}{\mathrm{P}\left(T \le t, T \le C\right)} \\ &= \mathrm{P}\left(X > c | T \le t, T \le C\right) \\ &= \mathrm{P}\left(X > c | T \le t\right) \times \frac{\mathrm{P}\left(T \le C | X > c, T \le t\right)}{\mathrm{P}\left(T \le C | T \le t\right)} \\ &= Se(c,t) \times \frac{\mathrm{P}\left(T \le C | X > c, T \le t\right)}{\mathrm{P}\left(T \le C | T \le t\right)}\end{aligned}$$

by the formula (*) with $\mathcal{A} = \{X > c\}$, $\mathcal{B} = \{T \le C\}$, and $\mathcal{C} = \{T \le t\}$. Similarly, the formula $(*)$ with $\mathcal{A} = \{X \le c\}$, $\mathcal{B} = \{C > t\}$, and $\mathcal{C} = \{T > t\}$ yields to the "Naive" specificity bias:

$$\begin{aligned}\widehat{Sp}(c,t) = \frac{\sum_{i=1}^{n} I(X_i \le c) I(T_i^* > t)}{\sum_{i=1}^{n} I(T_i^* > t)} &\xrightarrow{a.s} \frac{\mathrm{P}(X \le c, T > t, C > t)}{\mathrm{P}(T > t, C > t)} \\ &= Sp(c,t) \times \frac{\mathrm{P}(C > t | X \le c, T > t)}{\mathrm{P}(C > t | T > t)}.\end{aligned}$$

For the "Naive" AUC estimator, due to theory of U-statistics, when $n \to \infty$ we have

$$\begin{aligned}\widehat{AUC}(t) &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i I(T_i^* \le t) I(T_j^* > t) I(X_i > X_j)}{\sum_{i=1}^{n}\delta_i I(T_i^* \le t)\sum_{j=1}^{n} I(T_j^* > t)} \\ &\xrightarrow{a.s} \frac{\mathrm{P}(X_i > X_j, T_i \le C_i, T_i \le t, T_j > t, C_j > t)}{\mathrm{P}(T_i \le C_i, T_i \le t, T_j > t, C_j > t)} \\ &= \mathrm{P}(X_i > X_j | T_i \le C_i, T_i \le t, T_j > t, C_j > t)\end{aligned}$$

$$= \mathrm{P}(X_i > X_j | T_i \le t, T_j > t) \times \frac{\mathrm{P}(T_i \le C_i, C_j > t | T_i \le t, T_j > t, X_i > X_j)}{\mathrm{P}(T_i \le C_i, C_j > t | T_i \le t, T_j > t)}$$

by the formula (*) with $\mathcal{A} = \{X_i > X_j\}$, $\mathcal{B} = \{T_i \le C_i, C_j > t\}$, and $\mathcal{C} = \{T_i \le t, T_j > t\}$.

References

Akritas, M. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics* **22**, 1299–1327.

Amieva, H., Jacqmin-Gadda, H., Orgogozo, J., Le Carret, N., Helmer, C., Letenneur, L., Barberger-Gateau, P., Fabrigoule, C. and Dartigues, J. (2005). The 9 year cognitive decline before dementia of the Alzheimer type: a prospective population-based study. *Brain* **128**, 1093–1101.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Unpublished technical report, University of California, Berkeley.

Cai, T., Gerds, T., Zheng, Y. and Chen, J. (2011). Robust prediction of t-year survival with data from multiple studies. *Biometrics* **67**, 436–444.

Cai, T., Pepe, M., Zheng, Y., Lumley, T. and Jenny, N. (2006). The sensitivity and specificity of markers for event times. *Biostatistics* **7**, 182–197.

Chambless, L. and Diao, G. (2006). Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in Medicine* **25**, 3474–3486.

Chiang, C. and Hung, H. (2010). Non-parametric estimation for time-dependent AUC. *Journal of Statistical Planning and Inference* **140**, 1162–1174.

Foucher, Y., Giral, M., Soulillou, J. and Daures, J. (2010). Time-dependent ROC analysis for a three-class prognostic with application to kidney transplantation. *Statistics in Medicine* **29**, 3079–3087.

Gerds, T. and Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal* **48**, 1029–1040.

Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* **18**, 2529–2545.

Heagerty, P., Lumley, T. and Pepe, M. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344.

Heagerty, P. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61**, 92–105.

Hung, H. and Chiang, C. (2010a). Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics* **38**, 8–26.

Hung, H. and Chiang, C. (2010b). Optimal composite markers for time-dependent receiver operating characteristic curves with censored survival data. *Scandinavian Journal of Statistics* **37**, 664–679.

Hung, H. and Chiang, C.-T. (2011). Nonparametric methodology for the time-dependent partial area under the ROC curve. *Journal of Statistical Planning and Inference* **141**, 3829–3838.

Jacqmin-Gadda, H., Fabrigoule, C., Commenges, D., and Dartigues, J. (1997). A 5-year longitudinal study of the Mini-Mental State Examination in normal aging. *American Journal of Epidemiology* **145**, 498–506.

Joly, P., Commenges, D., Helmer, C. and Letenneur, L. (2002). A penalized likelihood approach for an illness–death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics* **3**, 433–443.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.

Pepe, M. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, Oxford.

Pepe, M., Zheng, Y., Jin, Y., Huang, Y., Parikh, C. and Levy, W. (2008). Evaluating the ROC performance of markers for future events. *Lifetime Data Analysis* **14**, 86–113.

Proust-Lima, C. and Taylor, J. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. *Biostatistics* **10**, 535–549.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67**, 819–829.

Saha, P. and Heagerty, P. (2010). Time-dependent predictive accuracy in the presence of competing risks. *Biometrics* **66**, 999–1011.

Satten, G. and Datta, S. (2001). The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician* **55**, 207–210.

Scheike, T., Zhang, M. and Gerds, T. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika* **95**, 205–220.

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

Uno, H., Cai, T., Tian, L. and Wei, L. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association* **102**, 527–537.

Vellas, B., Andrieu, S., Ousset, P., Ouzid, M. and Mathiex-Fortunet, H. (2006). The guidage study: methodological issues. A 5-year double-blind randomized trial of the efficacy of egb 761 (r) for prevention of alzheimer disease in patients over 70 with a memory complaint. *Neurology* **67**(Suppl 3), S6.

Viallon, V. and Latouche, A. (2011). Discrimination measures for survival outcomes: connection between the AUC and the predictiveness curve. *Biometrical Journal* **53**, 217–236.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale* (rev. ed.). Psychological Corporation, New York.

