Special Issue Paper

Statistics in Medicine

Received 6 November 2012, Accepted 2 August 2013 Published online 12 September 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5958

Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks

Paul Blanche,[a,b*†] Jean-François Dartigues[a,b] and Hélène Jacqmin-Gadda[a,b]

The area under the time-dependent ROC curve (AUC) may be used to quantify the ability of a marker to predict the onset of a clinical outcome in the future. For survival analysis with competing risks, two alternative definitions of the specificity may be proposed depending of the way to deal with subjects who undergo the competing events. In this work, we propose nonparametric inverse probability of censoring weighting estimators of the AUC corresponding to these two definitions, and we study their asymptotic properties. We derive confidence intervals and test statistics for the equality of the AUCs obtained with two markers measured on the same subjects. A simulation study is performed to investigate the finite sample behaviour of the test and the confidence intervals. The method is applied to the French cohort PAQUID to compare the abilities of two psychometric tests to predict dementia onset in the elderly accounting for death without dementia competing risk. The 'timeROC' R package is provided to make the methodology easily usable. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: AUC; competing risks; discrimination; inverse probability of censoring weighting; prognosis; survival analysis

1. Introduction

It is often useful to identify markers that enable the discrimination between subjects at high and low risk of a disease in the future. Markers with high predictive accuracy help clinicians for early medical decisions and thus may reduce morbidity and mortality in the high-risk population. Identification of such markers are therefore a major concern in medical research. For instance, the level of prostate-specific antigen is often used to identify subjects at high risk of prostate cancer recurrence and to guide treatment management. In Alzheimer's disease, the most frequent dementia in the elderly, some studies suggest that the decline in cognitive functions begins long before all the criteria for the clinical diagnosis of dementia are reached [1]. These results suggest that cognitive tests could be useful for identification of subjects with high risk of Alzheimer's disease in the coming years. As treatment against Alzheimer's disease have only demonstrated a modest efficiency, current research is focusing on preventive treatments that could be administered in the pre-diagnosis phase to subjects at high risk of dementia [2]. In this context, relevant questions are as follows. Are psychometric tests good markers for discrimination of subjects with respect to their risk of Alzheimer's disease in the few years following a test result? Are some psychometric tests better than others for this task? Data from large prospective cohort studies can be used to answer these questions. To evaluate prognostic abilities on such data, statistical methodology

[a] *University Bordeaux, ISPED, Centre INSERM U897-Epidémiologie-Biostatistique, F-33000 Bordeaux, France*
[b] *INSERM, ISPED, Centre INSERM U897-Epidémiologie-Biostatistique, F-33000 Bordeaux, France*
[*] *Correspondence to: Paul Blanche, INSERM U897 - Equipe de biostatistique, ISPED, Université Bordeaux Segalen, 146 rue Leo Saignat, 33076 Bordeaux cedex, France.*
[†] *E-mail: Paul.Blanche@isped.u-bordeaux2.fr*

have (i) to account for censoring due to lost of follow-up and (ii) to deal with the competing risk of death without dementia, which is a major issue in the elderly population.

Motivated by the prediction of Alzheimer's disease, the goal of this paper is to propose a method to estimate and compare the predictive accuracy of different rival markers using the ROC methodology, accounting for censoring and competing events.

The ROC methodology was originally introduced in medicine for the evaluation of diagnostic accuracy of quantitative markers. The ROC curve displays the sensitivity (true positive rate) versus 1-specificity (false positive rate) for all possible cutpoints that define a binary test, by dichotomizing a quantitative marker. The area under the ROC curve (AUC) is often used to summarize and compare diagnostic accuracy of several markers. The AUC may be interpreted as a concordance index. Indeed, the AUC is equal to the probability that the marker value of a randomly chosen diseased subject is above the marker value of a randomly chosen healthy subject.

For the evaluation of the predictive accuracy, Heagerty *et al.* [3, 4] introduced the time-dependent ROC curves and proposed several definitions of cases and controls. In particular, they introduced the so-called *cumulative/dynamic* definition, where a case is a subject diagnosed before a time point t and a healthy subject is free of the disease at time t. More recently, Saha and Heagerty [5] and Zheng *et al.* [6] extended this definition for the competing risks setting. As noticed by Zheng *et al.* [6], two definitions of the specificity can be considered depending of the way to deal with subjects who undergo the competing events. Such subjects can be considered as controls or not, depending on the clinical setting.

Without censoring, sensitivity and specificity can be estimated by empirical true positive and true negative fractions [7] and AUCs are often compared with the test of DeLong *et al.* [8]. These approaches assume that all subjects can be classified as cases or controls. However, when data are censored, the status of subjects lost of follow-up before time t is unknown. In the standard situation without competing risks, Uno *et al.* [9] and Hung and Chiang [10] proposed a nonparametric estimator of the time-dependent ROC curve using the inverse probability of censoring weighting (IPCW) approach, and Chiang and Hung [11] proposed a test for comparing AUCs. Besides, simulation studies have shown the good behaviour of the IPCW approach for various censoring scenarios [11, 12].

In the competing risks setting, only estimators that rely on parametric [13], semiparametric [6] or nonparametric smooth estimators that require a bandwidth selection [5] were proposed. In addition, no test was proposed to compare AUCs of different markers. In this paper, we first extend the IPCW estimators for ROC curves and AUCs to the competing risks setting for both definitions of specificity. By contrast to previous nonparametric proposed methods, this nonparametric approach does not require any bandwidth selection and has practical computational advantages, avoiding bootstrapping for making inference and being implemented in the `timeROC` R package. Moreover, whereas previous works were mainly focused on ROC curve estimation, this paper mainly focusses on providing and studying a test for comparing AUCs of two rival markers on censored data with competing events. We also propose some extensions such as estimators that properly account for marker dependent censoring when comparing AUCs or estimation of confidence bands.

The paper is organized as follows. Section 2 presents the notations and the definitions of the time-dependent ROC curves in presence of competing risks. Section 3 describes the proposed estimators, the asymptotic results and the inference procedures. The finite sample performances of the inference procedures are evaluated by simulations in Section 4. Section 5 presents the application of the proposed method to the comparison of two cognitive tests to predict dementia onset in the elderly. Finally, Section 6 concludes the paper contrasting our work with the previous ones and discussing some perspectives.

2. Receiver operating characteristic curves with competing risks

2.1. Notations

Let M denote a marker that is measured at baseline. Let T denote the event-time, C the censoring time and $\Delta = \mathbb{1}_{(T \leqslant C)}$ the censoring indicator, with $\mathbb{1}_{(\cdot)}$ denoting the indicator function. Let K denote the number of competing event and η the type of event. Let $\widetilde{T} = \min(T, C)$, the observed time and $\widetilde{\eta} = \Delta\eta$, which indicates either the type of event (when $\widetilde{\eta} \in \{1, \ldots, K\}$) or a censored observation (when $\widetilde{\eta} = 0$). Hereafter, for all time point t, we denote $S_T(t) = \mathbb{P}(T > t)$ and $G(t) = \mathbb{P}(C > t)$ and $\widehat{S}_T(t)$ and $\widehat{G}(t)$ the Kaplan–Meier estimators of $S_T(t)$ and $G(t)$, $S_{\widetilde{T}}(t) = \mathbb{P}(\widetilde{T} > t)$ and $\widehat{S}_{\widetilde{T}}(t)$ its empirical estimator. To make formulae easier to read, we also introduce the weight $W(t) = 1/G(t)$ and its estimator

$\widehat{W}(t) = 1/\widehat{G}(t)$.
We observe the independent and identically distributed (i.i.d.) sample of n subjects $\left\{\left(\widetilde{T}_i, \Delta_i, \widetilde{\eta}_i, M_i\right),\ i = 1, \ldots, n\right\}$. To simplify the presentation of the estimators, we assume that the marker M is measured on a continuous scale without ties. Adaptation for ties is discussed in Section 3.5.

2.2. *Definitions of receiver operating characteristic curves and area under the receiver operating characteristic curves*

With competing events, definition of cases is clear but for controls, Zheng *et al.* [6] considered two definitions leading to two different definitions of the time-dependent specificity.

For clarity, we suppose that we are interested in the assessment of the predictive accuracy for the first type of event (called main event) corresponding to $\eta = 1$. Cases at time t are defined as subjects who undergo the main event $\eta = 1$ before time t, that is, subjects i with $T_i \leqslant t, \eta_i = 1$. Without loss of generality, we assume that larger values of the marker M are associated with higher risks of events. Then, for a threshold $c \in \mathbb{R}$ the sensitivity at time t is defined by

$$Se(c,t) = \mathbb{P}\left(M > c | T \leqslant t, \eta = 1\right).$$

Controls at time t were originally defined as event-free subjects at time t, that is, subjects j with $T_j > t$ [5]. In the following, we denote '*control**' (with an asterix) the controls for this definition that leads to a specificity at time t defined by

$$Sp^*(c,t) = \mathbb{P}\left(M \leqslant c | T > t\right). \tag{1}$$

According to this definition, subjects who experience another type of event before time t, that is, subjects j with $T_j \leqslant t, \eta_j \neq 1$, are neither cases nor controls.
Alternatively, a control may be defined as a subject who is not a case, that is, a subject j with $T_j > t$ or $T_j \leqslant t, \eta_j \neq 1$ [6]. In the following, we denote '*control*' (without an asterix) the controls for this definition, leading to the specificity at time t:

$$Sp(c,t) = \mathbb{P}\left(M \leqslant c | \{T > t\} \cup \{T \leqslant t, \eta \neq 1\}\right). \tag{2}$$

Two different time-dependent ROC curves can be obtained plotting $Se(c,t)$ versus either $1 - Sp^*(c,t)$ or $1 - Sp(c,t)$. As with usual ROC curve, the AUC can be shown to be the probability that the marker of a case is greater than the marker of a control [7]. As a consequence, the AUC at time t for both definitions are

$$AUC^*(t) = \mathbb{P}\left(M_i > M_j | T_i \leqslant t, \eta_i = 1, T_j > t\right) \tag{3}$$

$$\text{and} \quad AUC(t) = \mathbb{P}\left(M_i > M_j \middle| T_i \leqslant t, \eta_i = 1, \left\{T_j > t\right\} \cup \left\{T_j \leqslant t, \eta_j \neq 1\right\}\right) \tag{4}$$

with i and j the indexes of two independent subjects. In the following, we will use AUC for the generic term and $AUC^*(t)$ or $AUC(t)$ for definitions (3) or (4).

Thus, subjects who meet one of the competing events before time t do not contribute to $AUC^*(t)$, although they are considered as control in $AUC(t)$. Applied to the example of dementia (as main event) and death without dementia (as competing event), this means $AUC^*(t)$ evaluates discrimination between demented subjects and subjects alive and non-demented at the time point t of interest, whereas $AUC(t)$ evaluates discrimination between demented subjects and subjects non-demented at the time point t or dead without dementia in the window of prediction. Given that many predictors of dementia are also predictors of death, the two measures may give very different results.

3. Inverse probability of censoring weighting estimators and inference procedures

3.1. *Inverse probability of censoring weighting estimators*

Without competing risks, Uno *et al.* [9] and Hung and Chiang [10] proposed IPCW estimators for the ROC curve and the AUC with censored data. A more general theory about IPCW can be found in [14] or [15]. The rational of the IPCW approach is to mainly use the observed cases and controls and to weight them by their probability of being observed. In the competing risk setting, *observed cases* are subjects i with $\widetilde{T}_i \leqslant t$ and $\widetilde{\eta}_i = 1$; *observed controls** are the uncensored event-free subjects at t, that is, subjects

j with $\widetilde{T}_j > t$, and *observed controls* are either uncensored and event-free or subjects who met a competing event before t, that is, subjects j with $\widetilde{T}_j > t$ or $\widetilde{T}_j \leqslant t, \widetilde{\eta}_j \notin \{0,1\}$. Subjects censored before t, that is, subjects i with $\widetilde{T}_i \leqslant t, \widetilde{\eta}_i = 0$, are only used to estimate the weights.

Assuming that the censoring C is independent of (T, η, M), we propose to estimate sensitivity $Se(c,t)$ by

$$\widehat{Se}(c,t) = \frac{\sum_{i=1}^n \mathbb{1}_{(M_i>c)} \mathbb{1}_{(\widetilde{T}_i \leqslant t, \widetilde{\eta}_i=1)} \widehat{W}\left(\widetilde{T}_i\right)}{\sum_{i=1}^n \mathbb{1}_{(\widetilde{T}_i \leqslant t, \widetilde{\eta}_i=1)} \widehat{W}\left(\widetilde{T}_i\right)}. \tag{5}$$

Heuristically, the weighting $\widehat{W}\left(\widetilde{T}_i\right) = 1/\widehat{G}\left(\widetilde{T}_i\right)$ is justified by the fact that as $n \to \infty$, the value of $n^{-1}\times$ numerator of $\widehat{Se}(c,t)$ converges to

$$\begin{aligned}\mathbb{E}\left\{\frac{\mathbb{1}_{(M_i>c)}\mathbb{1}_{(\widetilde{T}_i \leqslant t, \widetilde{\eta}_i=1)}}{G\left(\widetilde{T}_i\right)}\right\} &= \mathbb{E}\left\{\mathbb{E}\left(\frac{\mathbb{1}_{(M_i>c)}\mathbb{1}_{(T_i\leqslant t)}\mathbb{1}_{(\eta_i=1)}\mathbb{1}_{(T_i\leqslant C_i)}}{G(T_i)}\middle| T_i, \eta_i, M_i\right)\right\} \\ &= \mathbb{E}\left\{\frac{\mathbb{1}_{(M_i>c)}\mathbb{1}_{(T_i\leqslant t)}\mathbb{1}_{(\eta_i=1)}}{G(T_i)}\mathbb{E}\left(\mathbb{1}_{(T_i\leqslant C_i)}\middle| T_i, \eta_i, M_i\right)\right\} \\ &= \mathbb{E}\left\{\frac{\mathbb{1}_{(M_i>c)}\mathbb{1}_{(T_i\leqslant t)}\mathbb{1}_{(\eta_i=1)}}{G(T_i)}G(T_i)\right\} = \mathbb{P}\left(M > c, T \leqslant t, \eta = 1\right)\end{aligned}$$

and similarly the value of $n^{-1}\times$ denominator of $\widehat{Se}(c,t)$ converges to $\mathbb{P}\left(T \leqslant t, \eta = 1\right)$, both together leading to a consistent estimator of sensitivity.

By analogy, specificity $Sp^*(c,t)$ is estimated by

$$\widehat{Sp}^*(c,t) = \frac{\sum_{i=1}^n \mathbb{1}_{(M_i\leqslant c)}\mathbb{1}_{(\widetilde{T}_i>t)}\widehat{W}(t)}{\sum_{i=1}^n \mathbb{1}_{(\widetilde{T}_i>t)}\widehat{W}(t)} = \frac{\sum_{i=1}^n \mathbb{1}_{(M_i\leqslant c)}\mathbb{1}_{(\widetilde{T}_i>t)}}{\sum_{i=1}^n \mathbb{1}_{(\widetilde{T}_i>t)}} \tag{6}$$

and specificity $Sp(c,t)$ by

$$\widehat{Sp}(c,t) = \frac{\sum_{i=1}^n \mathbb{1}_{(M_i\leqslant c)}\left(\mathbb{1}_{(\widetilde{T}_i>t)}\widehat{W}(t) + \mathbb{1}_{(\widetilde{T}_i\leqslant t, \widetilde{\eta}_i\notin\{0,1\})}\widehat{W}(\widetilde{T}_i)\right)}{\sum_{i=1}^n \left\{\mathbb{1}_{(\widetilde{T}_i>t)}\widehat{W}(t) + \mathbb{1}_{(\widetilde{T}_i\leqslant t, \widetilde{\eta}_i\notin\{0,1\})}\widehat{W}(\widetilde{T}_i)\right\}}. \tag{7}$$

The resulting estimated ROC curves are increasing step functions. As with usual empirical ROC curve for uncensored data [7, p. 103], summing rectangular areas of widths and heights corresponding to increases in specificity and sensitivity, it can easily be shown that the area under the estimated ROC curves corresponding to the two specificity estimators (6) and (7) are given by

$$\widehat{AUC}^*(t) = \frac{\sum_{i=1}^n\sum_{j=1}^n \mathbb{1}_{(\widetilde{T}_i\leqslant t,\widetilde{\eta}_i=1)}\widehat{W}(\widetilde{T}_i)\mathbb{1}_{(\widetilde{T}_j>t)}\widehat{W}(t)\mathbb{1}_{(M_i>M_j)}}{\left(\sum_{i=1}^n \mathbb{1}_{(\widetilde{T}_i\leqslant t,\widetilde{\eta}_i=1)}\widehat{W}(\widetilde{T}_i)\right)\left(\sum_{j=1}^n \mathbb{1}_{(\widetilde{T}_j>t)}\widehat{W}(t)\right)} \tag{8}$$

and

$$\widehat{AUC}(t) = \frac{\sum_{i=1}^n\sum_{j=1}^n \mathbb{1}_{(\widetilde{T}_i\leqslant t,\widetilde{\eta}_i=1)}\widehat{W}(\widetilde{T}_i)\left(\mathbb{1}_{(\widetilde{T}_j>t)}\widehat{W}(t) + \mathbb{1}_{(\widetilde{T}_j\leqslant t,\widetilde{\eta}_j\notin\{0,1\})}\widehat{W}(\widetilde{T}_j)\right)\mathbb{1}_{(M_i>M_j)}}{\left(\sum_{i=1}^n \mathbb{1}_{(\widetilde{T}_i\leqslant t,\widetilde{\eta}_i=1)}\widehat{W}(\widetilde{T}_i)\right)\left(\sum_{j=1}^n \left\{\mathbb{1}_{(\widetilde{T}_j>t)}\widehat{W}(t) + \mathbb{1}_{(\widetilde{T}_j\leqslant t,\widetilde{\eta}_j\notin\{0,1\})}\widehat{W}\left(\widetilde{T}_j\right)\right\}\right)} \tag{9}$$

Let us recall that the IPCW estimator of the cumulative incidence function $F_1(t) = \mathbb{P}\left(T \leqslant t, \eta = 1\right)$ defined by

$$\widehat{F}_1(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(\widetilde{T}_i\leqslant t,\widetilde{\eta}_i=1)}\widehat{W}\left(\widetilde{T}_i\right) \tag{10}$$

is equal to the usual nonparametric maximum likelihood estimator, defined by $\widehat{F}_1(t) = \sum_{i=1}^n \widehat{\lambda}_1\left(\widetilde{T}_i\right)\widehat{S}_T\left(\widetilde{T}_i\right)$ where $\widehat{\lambda}_1(t) = \frac{\sum_{j=1}^n \mathbb{1}_{(\widetilde{\eta}_j=1,\widetilde{T}_j=t)}}{\sum_{j=1}^n \mathbb{1}_{(\widetilde{T}_j\geqslant t)}}$ [16, 17]. Besides, by the equality $\widehat{S}_T(t) =$

$1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(\widetilde{T}_i \leqslant t)} \Delta_i \widehat{W}\left(\widetilde{T}_i\right)$ that represents the Kaplan–Meier estimator as an IPCW estimator [18], then $1 - \widehat{F}_1(t) = \frac{1}{n}\sum_{j=1}^{n}\left\{\mathbb{1}_{(\widetilde{T}_j > t)}\widehat{W}(t) + \mathbb{1}_{(\widetilde{T}_j \leqslant t, \widetilde{\eta}_j \notin \{0,1\})}\widehat{W}\left(\widetilde{T}_j\right)\right\}$. As a consequence, the denominator of formulae (8) and (9) are respectively equal to the more compact form $\widehat{S}_T(t)\widehat{F}_1(t)$ and $\left(1 - \widehat{F}_1(t)\right)\widehat{F}_1(t)$.

In both formulae (8) and (9), the AUC estimator is the ratio of the estimated probability of observing a pair of a case and a control with ordered markers over the estimated probability of observing a pair with a case and a control.

3.2. Large sample properties

Only results about $\widehat{AUC}(t)$ are presented in the following. Results about $\widehat{AUC}^*(t)$, based on lemma 2 provided in the web Supporting Information[‡] B, are similar.

Assume that $\tau_1 > \inf\{u : F_1(u) > 0\}$ and $\tau_2 < \sup\left\{u : S_{\widetilde{T}}(u) > 0\right\}$. Thus, $[\tau_1, \tau_2]$ represents a period of times t in which there is both a non-null probability of observing a main event before time t and a non-null probability of observing someone at risk at time t.

Lemma 1

Let us assume that the censoring time C is independent of (T, η, M), then for all time t in $[\tau_1, \tau_2]$:

$$\sqrt{n}\left(\widehat{AUC}(t) - AUC(t)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathrm{IF}\left(\widetilde{T}_i, \widetilde{\eta}_i, M_i, t\right) + o_p(1)$$

where $\mathbb{E}\left[\mathrm{IF}\left(\widetilde{T}, \widetilde{\eta}, M, t\right)\right] = 0$. The influence function of the estimator $\mathrm{IF}(\cdot)$ is detailed in the appendix.

Proof

The proof is the adaptation to the competing risks setting of the proof of Theorem 1 of [10] and is given in the web Supporting Information A. □

From the decomposition of the estimator $\widehat{AUC}(t)$ as a sum of i.i.d. terms in lemma 1, it follows that

$$\sqrt{n}\left(\widehat{AUC}(t) - AUC(t)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_t^2\right).$$

The variance of the influence function σ_t^2 can be consistently estimated by the empirical estimator

$$\widehat{\sigma}_t^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathrm{IF}}\left(\widetilde{T}_i, \widetilde{\eta}_i, M_i, t\right)^2 \tag{11}$$

where $\widehat{\mathrm{IF}}\left(\widetilde{T}, \widetilde{\eta}, M, t\right)$ is a simple plug-in estimator that is detailed in the Appendix. Therefore, we obtain the asymptotic $(1-\alpha)$-level confidence interval (CI)

$$\left[\widehat{AUC}(t) - z_{1-\alpha/2}\frac{\widehat{\sigma}_t}{\sqrt{n}}\,,\ \widehat{AUC}(t) + z_{1-\alpha/2}\frac{\widehat{\sigma}_t}{\sqrt{n}}\right]$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the univariate standard normal distribution.

3.3. Test for comparing the area under the receiver operating characterisitc curves

Let M_1 and M_2 denote two rival markers measured on the same subject, $AUC_{M_1}(t)$ and $AUC_{M_2}(t)$ the areas under their time-dependent ROC curve and $\widehat{AUC}_{M_1}(t)$ and $\widehat{AUC}_{M_2}(t)$ their estimators. Assuming that the censoring C is independent of (T, η, M_1, M_2), under the null hypothesis $\mathcal{H}_0 : AUC_{M_1}(t) = AUC_{M_2}(t)$, it follows from lemma 1 that

$$\frac{\sqrt{n}}{\widehat{\sigma}_{t12}}\left(\widehat{AUC}_{M_1}(t) - \widehat{AUC}_{M_2}(t)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

[‡] *Supporting information may be found in the online version of this article.*

where

$$\widehat{\sigma}^2_{t12} = \frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{\mathrm{IF}}\left(\widetilde{T}_i, \widetilde{\eta}_i, M_{1i}, t\right) - \widehat{\mathrm{IF}}\left(\widetilde{T}_i, \widetilde{\eta}_i, M_{2i}, t\right)\right\}^2. \tag{12}$$

Alternatively, σ^2_{t12} may also be estimated by bootstrapping.

3.4. *Extension for marker-dependent censoring*

The previous methodology assumes that the censoring time C is independent of (T, η, M). It is sometimes more realistic to assume that C is only independent of (T, η) given the marker M. Indeed, the marker may be associated with the risk of censoring. The IPCW estimators can be adapted to deal with this lighter assumption by replacing in the weights, the marginal Kaplan–Meier estimator $\widehat{G}(t)$ by any consistent estimator of the conditional survival function $G(t|M) = \mathbb{P}(C > t|M)$. Nonparametric estimators proposed by Beran [19] and Akritas [20] can be used for this task.

However, to build a test for comparing two or more rival markers measured on the same subject, it is necessary to account for possible dependency of censoring on all the markers being compared. Thereby, in presence of marker-dependent censoring, assuming that the censoring time C is independent of (T, η) given $\mathbf{M}$, where $\mathbf{M}$ is the vector of rival markers ($\mathbf{M} = (M_1, M_2)^t$ for two rival markers), the weights should be based on an estimator $\widehat{G}(t|\mathbf{M})$ of $G(t|\mathbf{M}) = \mathbb{P}(C > t|\mathbf{M})$. As nonparametric estimators are often not efficient with moderate sample size and several explanatory variables, semiparametric estimators are therefore favoured.

Estimators of sensitivity, specificity and AUCs can be adapted replacing the marginal Kaplan–Meier $\widehat{G}(\cdot)$ by any semiparametric estimator of $G(\cdot|\mathbf{M})$ in the weights of formulae (5) to (9). Note that the weights no longer cancel in formula (6) in this case, because for $i \neq j$ then $G(\cdot|\mathbf{M}_i) \neq G\left(\cdot|\mathbf{M}_j\right)$ in general. Under the assumption that the censoring mechanism is correctly specified, these estimators are unbiased. Decompositions of these estimators as a sum of i.i.d terms as presented in lemma 1 can also be obtained for most of usual semiparametric estimators of $G(\cdot|\mathbf{M})$, including those derived from proportional or additive hazards models. However, the influence function of these estimators of $G(\cdot|\mathbf{M})$ is complex [21] and so are the influence functions of the AUC estimators. Therefore, a bootstrap resampling method is required to estimate the variances.

3.5. *Adaptation for ties and markers with ordered discrete results*

In practice, samples often include ties in the marker values, especially when the marker is measured on a discrete scale. Definitions and estimators of sensitivity and specificities are still valid in this case. However, in all formulae of estimators of $AUC(t)$ and $AUC^*(t)$, the term $\mathbb{1}_{(M_i > M_j)}$ need to be replaced by $\mathbb{1}_{(M_i > M_j)} + \frac{1}{2}\mathbb{1}_{(M_i = M_j)}$. This leads to a consistent estimator of the area under the ROC curve defined by the linear interpolation between two points. In this case, the AUC has a slightly different interpretation: this is the probability that the marker of a case is greater than the marker of a control, plus half the probability that the marker of a case is equal to the marker of a control. We refer to Section 4.5 of [7] for a more thorough discussion about the ROC curve for ordinal marker.

3.6. *Adjusted tests for multiple comparisons*

In prognostic studies, several time points t_l, $l = 1, \ldots, L$ are often of interest for clinicians who aim at evaluating the predictive accuracy of biomarkers in more or less long term. Thus, ROC curves may be estimated for different windows of prediction t_l, and the corresponding AUC of two rival markers may be compared at all these different times t_l. When several tests are performed, the p-values must be adjusted for multiple testing. Using the same approach as in the proof of lemma 1, under the null hypotheses $\mathcal{H}_0^l : AUC_{M_1}(t_l) = AUC_{M_2}(t_l), \quad l = 1, \ldots, L$, it can be shown that the vector of the L test statistics, which is the vector of L standardized differences of AUC estimated at time points t_l, as in Section 3.3, converges to a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Sigma_L)$. The covariance terms of the covariance-matrix Σ_L may be estimated similarly to the variance terms in Sections 3.2 and 3.3, either by empirical estimator and estimated influence functions $\widehat{\mathrm{IF}}(\cdot)$ or by Bootstrap. These results may be used either for performing a multivariate test or for computing asymptotically exact adjusted p-value for multiple univariate tests. The asymptotically exact adjusted p-value for testing $\mathcal{H}_0^l : AUC_{M_1}(t_l) = AUC_{M_2}(t_l)$ is computed as

$$p\text{-value}(t_l) = \mathbb{P}\left(\max|\mathbf{Z}| > |z(t_l)|\right)$$

where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma_L)$ and $z(t_l)$ is the value of the realization of the test statistic at time point t_l [22].

3.7. *Simultaneous confidence bands*

When an interval of time points is of interest, say $[t_{\min}, t_{\max}]$, it may be desirable to compute a confidence band. A so-called $(1-\alpha)$-simultaneous confidence band for the curve $\{(t, AUC(t)), t \in [t_{\min}, t_{\max}]\}$ is a region containing this curve with probability level $1-\alpha$. Note that the simultaneous confidence band is by definition larger that the band of pointwise CIs. We propose to compute an asymptotic $(1-\alpha)$-simultaneous confidence band by

$$\left[\widehat{AUC}(t) - \widehat{q}_{1-\alpha}\frac{\widehat{\sigma}_t}{\sqrt{n}}\,,\ \widehat{AUC}(t) + \widehat{q}_{1-\alpha}\frac{\widehat{\sigma}_t}{\sqrt{n}}\right], \quad t \in [t_{\min}, t_{\max}]$$

where $\widehat{q}_{1-\alpha}$ is estimated by the following simulation technique, paralleling the approaches of [10,21,23,24] among others, and making use of lemma 1:

Step 1: Generate a random sample $\left(\omega_1^b, \ldots, \omega_n^b\right)$ from n independent standard normal distributions.

Step 2: Using step 1 and the estimator $\widehat{\mathrm{IF}}(\cdot)$ detailed in the appendix, compute

$$\Theta^b = \sup_{t_{\min} \leqslant t \leqslant t_{\max}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega_i^b \frac{\widehat{\mathrm{IF}}\left(\widetilde{T}_i, \widetilde{\eta}_i, M_i, t\right)}{\widehat{\sigma}_t} \right|$$

Step 3: Repeat steps 1 and 2 a large number of times, say $B = 2000$ times for instance, and compute $\widehat{q}_{1-\alpha}$ as the $100(1-\alpha)$th percentile of $\left\{\Theta^1, \ldots, \Theta^B\right\}$.

Similarly, we are able to compute confidence band for the curve $\left\{\left(t, AUC_{M_1}(t) - AUC_{M_2}(t)\right), t \in [t_{\min}, t_{\max}]\right\}$ replacing $\widehat{AUC}(t)$ by $\widehat{AUC}_{M_1}(t) - \widehat{AUC}_{M_2}(t)$, and $\frac{\widehat{\mathrm{IF}}\left(\widetilde{T}_i, \widetilde{\eta}_i, M_i, t\right)}{\widehat{\sigma}_t}$ by $\frac{\widehat{\mathrm{IF}}\left(\widetilde{T}_i, \widetilde{\eta}_i, M_{1i}, t\right) - \widehat{\mathrm{IF}}\left(\widetilde{T}_i, \widetilde{\eta}_i, M_{2i}, t\right)}{\widehat{\sigma}_{t12}}$ in the previous expressions. This latter confidence band is of particular interest for testing by observing whether or not the zero function is contained within the band.

3.8. *Inference for points on receiver operating characteristic curve*

Although we mainly focus on AUC in this paper, inference procedures for points on ROC curve are also desirable. In particular, CIs for sensitivity and specificity at a clinically relevant threshold c are useful. We therefore provide lemma 3 in the web Supporting Material C that gives large sample results for sensitivity and specificity estimators, similarly to lemma 1 for the $AUC(t)$ estimator. Then, CI or other confidence regions may be derived using the approach described in Section 3.2.

4. Simulation studies

We conducted numerical investigations to assess the performances of the proposed estimators and of the test of comparison.

4.1. *Data generation*

We generated two rival markers M_1 and M_2 through standard normal distribution $\mathcal{N}(0,1)$ with correlation 0.5. We generated two competing events with proportional cause-specific hazards models

$$\begin{aligned}\alpha_j(t) &= \lim_{dt \to 0} P\left(T \in [t, t+dt), \eta = j \,|\, T \geqslant t\right)/dt \\ &= \alpha_{0j} \exp\left(\beta_{j1} M_1 + \beta_{j2} M_2\right),\end{aligned}$$

for $j = 1, 2$, following the algorithm described in [25]. We generated the censoring time with a Cox proportional hazards model $\lambda_C(t) = \lambda_{0C} \exp(\gamma_1 M_1 + \gamma_2 M_2)$. We considered several scenarios: two sample sizes $n = 200$ or $n = 400$, a censoring that does not depend on any markers ($HR_{C_1} = \exp(\gamma_1) = 1$

and $HR_{C_2} = \exp(\gamma_2) = 1$) or that depends on the marker M_1 only with hazard ratio $HR_{C_1} = 1.35$ (and $HR_{C_2} = 1$), and four values 0, 0.22, 0.45 or 0.68 for parameter Δ_{β_1}, where $\beta_{11} = 1 + \Delta_{\beta_1}$ and $\beta_{12} = 1 - \Delta_{\beta_1}$, to make the differences of AUCs of M_1 and M_2 at time $t = 1$ equal to 0, 0.05, 0.10 or 0.15. Both markers were also associated with the competing event ($\beta_{21} = \beta_{22} = 0.2$). To more thoroughly investigate the behaviour of the test when censoring depends on the markers, we also performed simulations under the null hypothesis where $\left(HR_{C_1}, HR_{C_2}\right) = (2, 1)$ and $(1.35, 0.6)$. Other parameters were chosen to have approximately 33%, 17% and 38% of *observed cases, controls* and controls* (as defined in Section 3.1) and 28% of censored subjects at time $t = 1$. These frequencies can be 1% to 3% greater or smaller depending on the scenario generated.
The choice of the hazard ratio $HR_{C_1} = 1.35$, the correlation between M_1 and M_2 equal to 0.5 and the proportion of censored subjects equal to 27.5% was guided by the PAQUID cohort data.

For each data set, we computed the IPCW estimators of $AUC(t)$ and $AUC^*(t)$ for the first event-type, the 95%-level CIs and the tests of comparison with α-level equal to 5%. We used the two weighting procedures: assuming C is independent of (T, η, M_1, M_2), estimating $G(t)$ by Kaplan–Meier (KM-weights), or relaxing this assumption, assuming that C is independent of (T, η) given (M_1, M_2), estimating $G\left(t|M_1, M_2\right)$ by a Cox model (Cox-weights). We also computed the nonparametric nearest neighbour estimators (NNE) of Saha and Heagerty [5] for comparison with our approach. As no optimal rule exists, the choice of a bandwidth $\lambda_n = O\left(n^{-1/3}\right)$ required for NNE was set equal to $\lambda_n = 0.33 \times n^{-1/3}$, that leads to approximately $2\lambda_n = 11.3\%$ and 8.9% of subjects included in each neighbourhood, when $n = 200$ and $n = 400$ respectively.

The standard error estimators based on the influence function defined at equations (11) and (12) were computed for the KM-weights estimator, and for all estimators, standard errors were also computed by bootstrapping (400 replications).

4.2. Simulation results

Tables I–III summarize the main simulation results on the basis of 1000 replications for each scenario including $n = 400$ subjects. Similar tables of results with $n = 200$ are provided in the web Supporting Information D. First, with the KM-weights, the bootstrap and influence function estimators of the standard errors lead to very similar results.

When the censoring does not depend on the markers (Table I), estimators based on both weightings perform as well. As expected, they are unbiased, and the coverage probabilities of the CIs are close to the nominal value 95%. Under the null hypothesis that AUCs are equal for both markers, estimated α-level is close to the nominal value 5%, and the power of the test tends to one when the difference between AUCs increases. The two weighting procedures also lead to similar results in terms of efficiency.

When the censoring depends moderately on the marker M_1 (Table II), the AUC estimates for M_1 are slightly biased, and the coverage probabilities of their CIs decrease when C is assumed independent of M_1 (KM-weights). Because of the positive correlation between the two markers, the bias for the KM-estimates of AUC for marker M_2 is in the same direction but much smaller (with no impact on the coverage rates). As a consequence, the test of comparison remains valid in these simulation scenarios even if the weights are not well adapted to the censoring.

However, Table III shows that when the dependency between C and M_1 is stronger ($HR_{C_1} = 2$, $HR_{C_2} = 1$), or when there is a positive association between C and M_1 and a negative association between C and M_2 ($HR_{C_1} = 1.35$, $HR_{C_2} = 0.6$), the test is biased. AUC estimators and all inference procedures perform still very well in any cases when weights are computed from a Cox model.

The NNE performs also well for independent censoring with only slight underestimations, probably because of smoothing (Table II). With moderate marker-dependent censoring (Table II), the bias for NNE increases little, and this has no consequences on the validity of inference procedures even when censoring depends on marker M_1, whereas the ROC curve is estimated for M_2, setting which is not handled by NNE. The main reason is that slight underestimation due to smoothing compensates slight overestimation due to marker-dependent censoring. Even when the association between censoring and marker becomes stronger, or when censoring is associated with both markers with opposite directions, NNE appears quite robust (Table III).

Finally, these simulations illustrate the difference between the two definitions of the specificity. Defining all subjects who met the competing event as controls decreases the AUC and increases the power of the test of comparison. The decrease of AUC is due to the positive dependence of both types of event on the markers. Indeed, discrimination between subjects who will experience the main event and subjects

Table I. Simulation results with sample size $n = 400$ and independent censoring, that is, HR_{C_1}=HR_{C_2}=1 (1000 replications).

			AUC(t)				Coverage probability					
			True		Bias		M_1		M_2		Type I error or power	
	Method	ΔAUC(t)	M_1	M_2	M_1	M_2	IF	B	IF	B	IF	B
$AUC^*(t)$	IPCW KM	0.0	85.6	85.6	0.0	0.1	93.2	93.3	94.5	94.7	5.0	4.8
	IPCW Cox				0.0	0.1	—	94.0	—	95.5	—	5.0
	NNE				−0.3	−0.2	—	95.0	—	95.2	—	4.7
$AUC(t)$	IPCW KM	0.0	79.8	79.8	0.0	0.0	95.7	95.7	95.1	94.9	4.7	4.6
	IPCW Cox				0.0	0.0	—	95.1	—	94.9	—	4.4
	NNE				−0.3	−0.3	—	94.4	—	95.3	—	4.1
$AUC^*(t)$	IPCW KM	5.0	87.9	82.9	0.1	0.1	93.7	93.8	95.1	95.6	26.7	25.2
	IPCW Cox				0.1	0.0	—	93.3	—	95.9	—	26.7
	NNE				−0.2	−0.2	—	94.6	—	95.6	—	30.2
$AUC(t)$	IPCW KM	5.1	82.2	77.1	0.1	0.0	94.4	94.4	96.8	96.5	34.8	34.3
	IPCW Cox				0.1	0.0	—	94.8	—	96.2	—	35.7
	NNE				−0.2	−0.2	—	94.9	—	95.4	—	36.0
$AUC^*(t)$	IPCW KM	10.0	89.9	79.9	0.1	0.1	93.5	93.7	94.8	94.8	77.7	78.1
	IPCW Cox				0.1	0.1	—	94.4	—	95.2	—	79.3
	NNE				−0.2	−0.2	—	94.5	—	95.3	—	82.4
$AUC(t)$	IPCW KM	10.3	84.4	74.1	0.1	0.1	94.9	94.9	95.9	95.8	89.8	90.0
	IPCW Cox				0.1	0.1	—	94.8	—	95.9	—	90.8
	NNE				−0.2	−0.2	—	95.5	—	95.2	—	90.3
$AUC^*(t)$	IPCW KM	14.7	91.4	76.8	0.1	0.1	94.3	94.3	94.8	95.1	98.4	98.4
	IPCW Cox				0.1	0.1	—	94.7	—	96.0	—	98.6
	NNE				−0.2	−0.2	—	94.7	—	95.3	—	99.3
$AUC(t)$	IPCW KM	15.1	86.0	71.0	0.1	0.0	94.0	94.6	95.3	95.8	99.7	99.8
	IPCW Cox				0.1	0.0	—	94.4	—	96.3	—	99.7
	NNE				−0.2	−0.2	—	94.5	—	95.8	—	99.7

Bias of estimates for two markers M_1 and M_2 (multiplied by 100), empirical coverage probabilities of 95% confidence intervals and type I errors and powers of the test of $\mathcal{H}_0 : \mathrm{AUC}_{M_1}(t) = \mathrm{AUC}_{M_2}(t)$, depending on $\Delta\mathrm{AUC}(t) = \mathrm{AUC}_{M_1}(t) - \mathrm{AUC}_{M_2}(t)$. Inverse probability of censoring weighting (IPCW) estimators using the Kaplan–Meier estimator (IPCW KM) or a Cox proportional hazards model with covariates M_1 and M_2 (IPCW Cox) for weighting, and the nearest neighbour estimator (NNE). Variances are computed from the estimated influence function (IF) for IPCW KM, or by bootstrapping 400 times (B).

who will undergo the competing event is more difficult than discrimination between subjects who will undergo the main event and subjects who will experience none of the event. As it will be illustrated in Section 5, it is frequent that both events are associated with the markers. The increase of the power is because subjects who met the competing event are more informative for estimating $AUC(t)$, for which they are controls, than for estimating $AUC^*(t)$ for which they are only used for estimating the weights.

The good behaviour of the inference procedures for points on the ROC curve discussed in Subsection 3.8 was also assessed, and some simulation results are provided in the web Supporting Information Table 4.

5. Application to dementia prediction

5.1. Objective

The objective of this analysis is the estimation and the comparison of the abilities of two cognitive tests to predict the risk of dementia within the 3, 5 and 10 years following the test result, in the elderly population accounting for death without dementia competing risk. A window of prediction of 3 or 5 years could correspond to the duration of a preventive clinical trial, and thus after validation, these cognitive tests could be used to select the population at risk to include in a trial. A window of 10 years is probably too long for a preventive trial but may be of interest for general practitioner for reassuring worried patients. The two tests compared are the mini-mental state examination (MMSE) [26] and the

Table II. Simulation results with sample size $n = 400$ and moderate M_1-dependent censoring, that is, HR_{C_1}=1.35, HR_{C_2}=1 (1000 replications).

			AUC(t)				Coverage probability					
			True		Bias		M_1		M_2		Type I error or power	
	Method	ΔAUC(t)	M_1	M_2	M_1	M_2	IF	B	IF	B	IF	B
$AUC^*(t)$	IPCW KM	0.0	85.6	85.6	1.5	0.6	87.6	87.8	93.7	93.7	5.2	5.6
	IPCW Cox				0.0	0.2	—	94.1	—	95.0	—	4.8
	NNE				−0.5	−0.2	—	94.2	—	95.9	—	5.5
$AUC(t)$	IPCW KM	0.0	79.8	79.8	1.1	0.6	92.0	92.0	92.9	93.0	5.2	5.4
	IPCW Cox				0.0	0.1	—	95.6	—	94.3	—	5.4
	NNE				−0.4	0.0	—	94.9	—	94.3	—	5.4
$AUC^*(t)$	IPCW KM	5.0	87.9	82.9	1.3	0.6	86.2	86.0	94.1	94.4	36.7	36.2
	IPCW Cox				0.1	0.1	—	93.4	—	95.8	—	26.5
	NNE				−0.4	−0.2	—	94.3	—	95.5	—	29.5
$AUC(t)$	IPCW KM	5.1	82.2	77.1	1.0	0.7	91.5	91.2	94.6	94.3	41.2	41.2
	IPCW Cox				0.1	0.0	—	94.2	—	95.6	—	33.3
	NNE				−0.3	0.0	—	94.9	—	95.5	—	33.8
$AUC^*(t)$	IPCW KM	10.0	89.9	79.9	1.1	0.8	88.0	87.9	94.1	93.7	86.7	86.2
	IPCW Cox				0.1	0.1	—	94.2	—	95.3	—	78.8
	NNE				−0.4	−0.2	—	95.5	—	95.3	—	81.8
$AUC(t)$	IPCW KM	10.3	84.4	74.1	0.9	0.8	92.1	92.1	94.5	94.6	92.7	92.5
	IPCW Cox				0.1	0.1	—	95.4	—	95.6	—	89.5
	NNE				−0.3	0.0	—	94.9	—	95.1	—	90.3
$AUC^*(t)$	IPCW KM	14.7	91.4	76.8	0.9	0.9	87.8	88.5	93.4	93.0	99.5	99.4
	IPCW Cox				0.1	0.1	—	93.7	—	94.8	—	99.0
	NNE				−0.4	−0.2	—	95.5	—	94.9	—	99.0
$AUC(t)$	IPCW KM	15.1	86.0	71.0	0.8	0.8	91.2	90.9	93.8	93.6	99.9	99.8
	IPCW Cox				0.1	0.0	—	94.6	—	95.0	—	99.8
	NNE				−0.2	0.0	—	94.9	—	95.0	—	99.8

Bias of estimates for two markers M_1 and M_2 (multiplied by 100), empirical coverage probabilities of 95% confidence intervals and type I errors and powers of the test of $\mathcal{H}_0 : \mathrm{AUC}_{M_1}(t) = \mathrm{AUC}_{M_2}(t)$, depending on $\Delta\mathrm{AUC}(t) = \mathrm{AUC}_{M_1}(t) - \mathrm{AUC}_{M_2}(t)$. Inverse probability of censoring weighting (IPCW) estimators using the Kaplan–Meier estimator (IPCW KM) or a Cox proportional hazards model with covariates M_1 and M_2 (IPCW Cox) for weighting, and the nearest neighbour estimator (NNE). Variances are computed from the estimated influence function (IF) for IPCW KM, or by bootstrapping 400 times (B).

digit symbol substitution test (DSST) [27]. The MMSE is a sum-score evaluating various dimensions of cognition (memory, calculation, orientation in space and time, language and word recognition), which is often used as an index of global cognitive performance and for screening of dementia. MMSE score ranges from 0 to 30. The DSST explores attention and psychomotor speed. Given a code table displaying the correspondence between pairs of digits and symbols, the subjects have to fill in blank squares with the symbol, which is paired to the digit displayed above the square. The subjects have to fill in as many squares as possible in 90 s. The maximum value is 90.

As larger values of DSST and MMSE are associated with lower risks of dementia, let us note that, following the previous notations, ROC analyses were performed for minus DSST and minus MMSE to reverse the associations.

5.2. *The PAQUID data*

Paquid is a population-based study including 3777 subjects aged 65 years and older, living at home in the south-west of France at enrollment in 1988. Individuals were seen at home by psychologists trained in home interviews at the initial visit and at 1, 3, 5, 8, 10, 13, 15, 17 and 20 years later. Each visit included a neuropsychological evaluation through a battery of psychometric tests and a standardized diagnosis of dementia [28].

We used MMSE and DSST collected at baseline to predict dementia diagnosis over the first 3, 5 or 10 years of follow-up. We excluded from the sample subjects demented, blind, deaf or confined to bed

Table III. Simulation results with sample size $n = 400$ and strong dependent censoring (1000 replications).

	Method	ΔAUC(t)	AUC(t)				Coverage probability					
			True		Bias		M_1		M_2		Type I error or power	
			M_1	M_2	M_1	M_2	IF	B	IF	B	IF	B
HR$_{C_1}$=2, HR$_{C_2}$=1												
$AUC^*(t)$	IPCW KM	0.0	85.6	85.6	2.6	0.8	74.1	74.8	91.8	91.6	9.2	9.3
	IPCW Cox				0.1	0.2	—	94.5	—	94.2	—	5.0
	NNE				−1.0	−0.3	—	94.2	—	95.3	—	5.8
$AUC(t)$	IPCW KM	0.0	79.8	79.8	2.0	1.0	84.4	84.3	90.3	90.5	6.7	6.6
	IPCW Cox				0.0	0.1	—	95.5	—	93.8	—	5.6
	NNE				−0.6	0.0	—	95.3	—	93.9	—	6.1
HR$_{C_1}$=1.35, HR$_{C_2}$=0.6												
$AUC^*(t)$	IPCW KM	0.0	85.6	85.6	1.2	−2.5	90.3	90.7	91.7	91.5	15.8	14.9
	IPCW Cox				0.1	0.1	—	92.5	—	94.2	—	4.8
	NNE				−0.2	0.2	—	94.8	—	94.6	—	5.6
$AUC(t)$	IPCW KM	0.0	79.8	79.8	0.4	−1.7	94.0	94.1	92.5	92.3	9.7	9.9
	IPCW Cox				0.0	0.0	—	94.9	—	94.0	—	5.1
	NNE				−0.5	0.3	—	94.7	—	93.6	—	6.2

Bias of estimates for two markers M_1 and M_2 (multiplied by 100), empirical coverage probabilities of 95% confidence intervals and type I errors and powers of the test of $\mathcal{H}_0$: AUC$_{M_1}(t)$ = AUC$_{M_2}(t)$, depending on ΔAUC(t) = AUC$_{M_1}(t)$−AUC$_{M_2}(t)$. Inverse probability of censoring weighting (IPCW) estimators using the Kaplan–Meier estimator (IPCW KM) or a Cox proportional hazards model with covariates M_1 and M_2 (IPCW Cox) for weighting, and the nearest neighbour estimator (NNE). Variances are computed from the estimated influence function (IF) for IPCW KM, or by bootstrapping 400 times (B).

Table IV. Comparison of $AUC(t)$ of digit symbol substitution test and mini-mental state examination at times $t = 3, 5$ and 10 years.

		$t = 3$	$t = 5$	$t = 10$
Demented		70(2.7%)	122(4.8%)	318(12.4%)
Died without dementia		194(7.6%)	313(12.2%)	545(21.3%)
Censored		180(7.0%)	292(11.4%)	591(23.1%)
KM weights				
$AUC(t)$	DSST	79.9 [74.9,84.9]	77.8 [74.0,81.6]	72.2 [69.2,75.2]
	MMSE	74.7 [68.7,80.8]	72.0 [67.2,76.8]	66.9 [63.6,70.2]
	p-value	0.03	0.01	< 0.01
	Adjusted p-value	0.09	0.02	< 0.01
$AUC^*(t)$	DSST	80.9 [76.0,85.8]	79.7 [76.0,83.5]	76.7 [73.8,79.7]
	MMSE	75.4 [69.4,81.4]	73.2 [68.4,78.0]	69.9 [66.6,73.3]
	p-value	0.02	< 0.01	< 0.01
	Adjusted p-value	0.06	0.01	< 0.01
Cox weights				
$AUC(t)$	DSST	79.8 [74.8,84.8]	77.5 [73.7,81.4]	71.7 [68.8,74.7]
	MMSE	74.7 [68.6,80.7]	71.9 [67.0,76.7]	66.6 [63.3,69.9]
	p-value	0.04	0.01	< 0.01
	Adjusted p-value	0.10	0.03	< 0.01
$AUC^*(t)$	DSST	80.8 [75.9,85.7]	79.4 [75.6,83.2]	76.1 [73.2,79.1]
	MMSE	75.2 [69.3,81.4]	73.0 [68.2,77.8]	69.6 [66.2,73.0]
	p-value	0.03	< 0.01	< 0.01
	Adjusted p-value	0.07	0.01	< 0.01

Estimates (multiplied by 100), confidence intervals and p-values of the test of $\mathcal{H}_0$: AUC$_{DSST}(t)$ = AUC$_{MMSE}(t)$. PAQUID cohort, $n = 2561$.

DSST, digit symbol substitution test; MMSE, mini-mental state examination; AUC, area under the receiver operating characterisitic; KM weights, Kaplan–Meier weights.

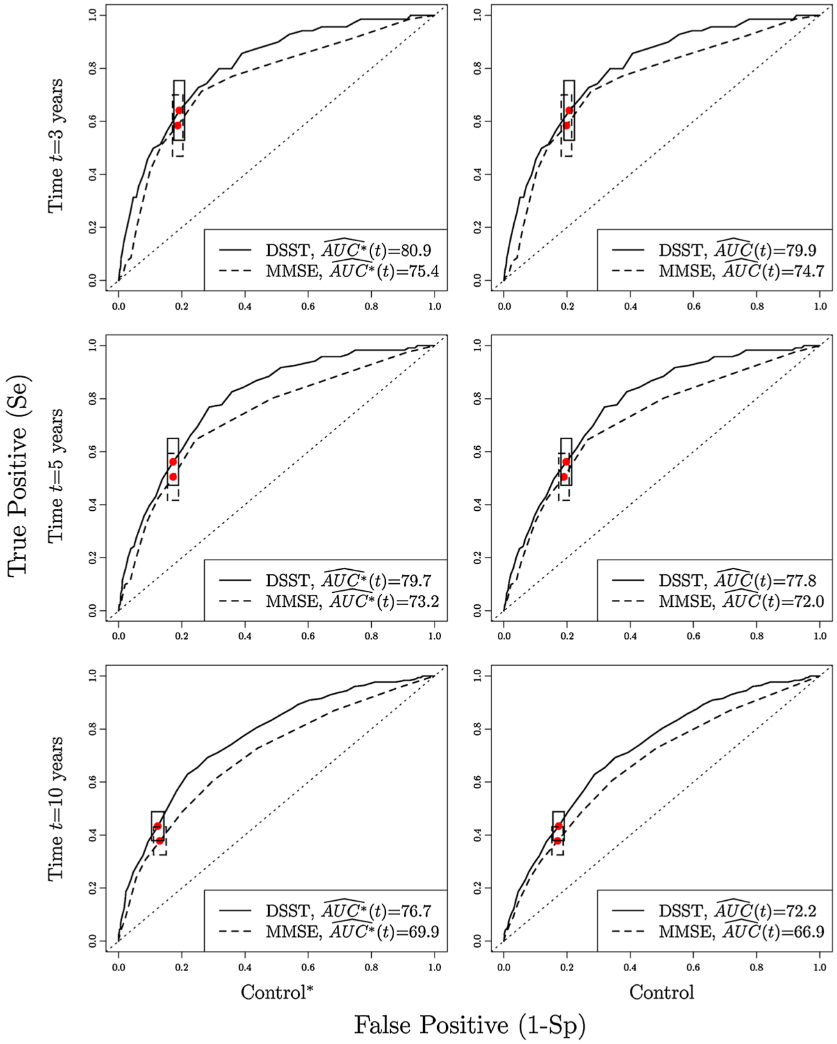

Figure 1. Time-dependent receiver operating characteristic curves estimated by inverse probability of censoring weighting estimators (KM-weights) at $t = 3, 5$ and 10 years for digit symbol substitution test and mini-mental state examination with the two definitions of the specificity. PAQUID, $n = 2561$ subjects. Points and rectangles display estimates and 95% confidence intervals for sensitivity and one minus specificity estimates for a threshold value equal to the third quartile of the marker distribution (19 for digit symbol substitution test, 25 for mini-mental state examination).

at the initial visit, subjects with missing values for the MMSE and the DSST and subjects who dropped out just after the baseline visit. The final sample included $n = 2561$ subjects. The time-to-dementia was computed as the mid-point between the time of the visit of diagnosis and the time of the last visit without dementia. Subjects who died without a dementia diagnosis were considered as free of dementia at their death if the last visit was less than 2 years before the death and were considered as censored at the last visit if duration between the last visit and death was longer. Using a threshold of 3 years instead of 2 years led to the same conclusions (results not shown).

5.3. Results

The mean DSST score was 27.4 (standard deviation: 11.5, inter quartile range: 19–35), and the mean MMSE was 26.4 (standard deviation: 2.9, inter quartile range: 25–28). The correlation between the two markers was estimated at 0.55.

We fitted a multivariate Cox model in order to explore the dependence of the censoring time on both psychometric tests. The estimated hazard ratios were 1.019 for DDST (95% CI: (1.010, 1.028)) and 1.023 for MMSE (95% CI :(0.989, 1.057)). As in our simulation study, only one marker is significantly

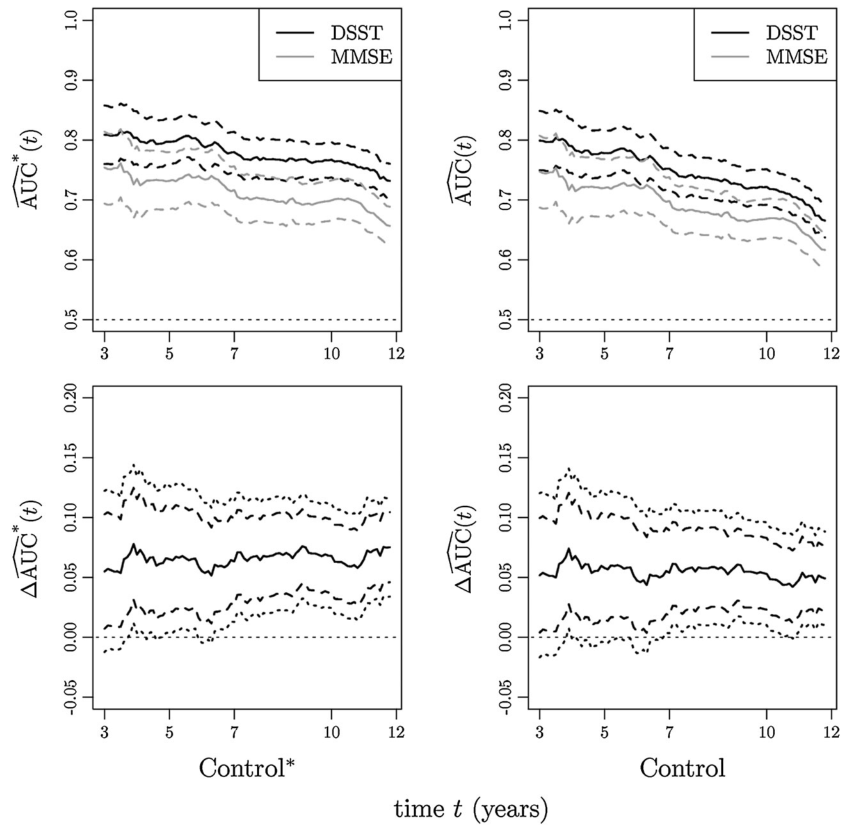

Figure 2. Comparisons of the evolutions of $\widehat{AUC}(t)$ and $\widehat{AUC}^*(t)$ over time for digit symbol substitution test and mini-mental state examination. PAQUID cohort, $n = 2561$. $\Delta\widehat{AUC}(t)$ and $\Delta\widehat{AUC}^*(t)$ mean the differences of estimates between digit symbol substitution test and mini-mental state examination. Dashed and dotted lines display respectively 95% pointwise confidence intervals and 95% simultaneous confidence bands.

associated with the censoring time. The hazard ratio for the increase of one standard error of the DSST was 1.25, smaller than the one in the simulation study.

The frequencies of *observed cases, controls* and controls* (as defined in Section 3.1) and censored subjects before each time point t, for $t = 3, 5, 10$ years, are given in Table IV. Although the proportions of *observed cases* at time $t = 3, 5, 10$ are small, respectively 2.7%, 4.8% and 12.4%, the frequencies are big enough, respectively 70, 122 and 328 because of the large sample size ($n = 2561$). Respectively, 7.0%, 11.4% and 23.1% of subjects are lost of follow-up at time 3, 5 and 10 years.

Figure 1 displays the ROC curves estimated at time $t = 3, 5, 10$ years for the two definitions of specificity, assuming independence between censoring and marker (KM-weights). For both definitions, the ROC curves go down and become closer to the diagonal as time increases, confirming that the discrimination decreases as the window of prediction enlarges. This is illustrated on Figure 2 that displays the estimates of $AUC(t)$ and $AUC^*(t)$ for time t from 3 to 12 years and their 95% pointwise CIs. With both definition of controls, Figure 1 shows that the discrimination ability of DSST appears always better than the one of MMSE, and more specifically, the three ROC curves of DSST are always above the ROC curves of MMSE. Consequently, whatever the time $t = 3, 5, 10$ years and whatever the cutpoint chosen for the marker (or equivalently the value of the specificity), the sensitivity is always estimated greater for DSST. As an illustration, Figure 1 also depicts CIs for sensitivity and specificity estimates, for a threshold value equal to the third quartile of the marker distribution. Table IV displays AUC estimates with 95% CI and tests for comparison between MMSE and DSST for the two definitions of controls and the two weightings (with KM-weights, estimated influence functions were used to compute the variances; with Cox-weights, variances were estimated by bootstrapping 5000 times). According to the unadjusted p-values, the differences between the AUC were significant for the three windows of prediction; adjusting for multiple testing (one test for each time point), only the differences at 5 and 10 years were still significant. The 95% pointwise CIs of the differences between AUCs of MMSE and DSST displayed in

Figure 2 show that the differences are significant for all times t from 3 to 12 years. However, the 95% simultaneous confidence bands of these differences intersect the zero line. As the association between censoring and marker is weak, results obtained with the Cox-weights and with the NNE of Saha and Heagerty [5] were nearly identical, as expected (results not shown).

Figure 2 and Table IV also show that the discrimination is better when subjects died without dementia are not defined as controls ($AUC^*(t)$ versus $AUC(t)$). However, the differences between DSST and MMSE are similar for both definitions, and the test results are concordant. The great difference between $AUC^*(t)$ and $AUC(t)$ is due to the association between the two markers and the competing event: the hazard ratio for the association between DSST or MMSE and death without dementia are 1.037 for DSST (95% CI : (1.030, 1.044)) and 1.080 for MMSE (95% CI : (1.055, 1.107)). Thus, it is easier to discriminate future demented subjects from subjects alive and non-demented rather than from subjects alive and non-demented or died without dementia.

To conclude, although MMSE is largely used as a screening test for dementia, all these analyses found that the DSST has better performances to detect future demented subjects in a period of time of 3, 5 or 10 years compared with the MMSE. Moreover, whatever the specificity, the sensitivity of DSST is always estimated better than the one of MMSE.

6. Discussion

In this manuscript, we used the IPCW approach to estimate time-dependent ROC curve and AUC for censored events with competing risks. Large sample theory of the estimators was established, and CIs, simultaneous confidence band and tests of comparison were derived. Simulation results show that the proposed procedures work well for moderate sample sizes. The practical interest of the procedure for estimating and comparing psychometric tests for dementia onset was illustrated with data from the PAQUID cohort. Significant results suggest that the DSST is better than the MMSE to discriminate subjects at high risk of Alzheimer's disease in the next few years.

We proposed estimators for the two definitions of specificity because both definitions may be useful depending the clinical setting. For instance, if the prognostic marker is used as inclusion criteria in a clinical trial for Alzheimer's disease, it is essential to be able to discriminate subjects at high risk of dementia from all subjects and especially from subjects who have a high risk of death without dementia during the trial duration. Similarly, in prostate cancer treatment of very old patients, the main criteria for treatment is the risk of progression of the disease before death from other causes. Thus, for both setting the marker must be validated with $AUC(t)$ defined at equation (4). On the other hand, from the point of view of a patient worried about his cognitive decline, the main question of interest may be : If I survive 10 years, am I at high risk of becoming demented during this period? Then, $AUC^*(t)$ defined at equation (3) is more relevant. In any case, as the estimated value may be very different, it is essential in the clinical application to clearly state which definition is used for the specificity.

As pointed out by one referee, our analyses of the predictiveness of the cognitive tests on the PAQUID data do not account for the age of the subjects. Consequently, the predictiveness of the cognitive tests has to be carefully interpreted as a marginal predictiveness among the population of subjects aged 65 years and older enrolled in the PAQUID study. Thus, the estimated predictive accuracy is of interest for clinicians who are interested in quantifying how informative can be the result of a cognitive test to evaluate the risk of dementia of a patient in the t-years following a test result.

To account for age in the estimation of the ROC curves and AUCs, we could have fitted semiparametric models for ROC curves or AUC given age, using semiparametric methods proposed by Zheng *et al.* [6] or adapting a method of Hung and Chiang [10]. Alternatively, our nonparametric methods could also have been computed on several age groups, performing stratified analyses. Another possible approach would have been to fit a multivariate prediction model accounting for age, using a binomial regression model [29] for example, to then estimate the ROC curve and AUC of the resulting prediction tool. For this latter approach, a careful correction for overfitting and optimism bias is required [30], which is outside the scope of this manuscript. Kerr and Pepe [31] have recently discussed and contrasted these two strategies in detail.

To deal with interval censoring of dementia in the PAQUID data, we used a simple imputation rule of the status of subjects deceased without dementia diagnosis. Changing this rule (3 years instead of 2 years between the last visit and the death) did not change the results. To our knowledge, only parametric estimators based on an illness-death model can account for this particular kind of interval-censoring [32]. In an ongoing work, we developed such model-based estimators. However, simulation results suggest that the bias for the IPCW estimators are small and may be smaller than the bias of the model-based estimators when the model is misspecified.

Compared with earlier papers that focused on estimation, this work mainly focused on tests for comparing AUCs. The test of comparison that we proposed is the extension to the competing risks setting with censored data of the popular test proposed by DeLong *et al.* [8]. Indeed, without censoring, weights are all equal to one.

Assuming that censoring does not depend on the markers, the estimation and the test do not require any parametric assumptions by contrast to the approach of Foucher *et al.* [13] or Zheng *et al.* [6], nor bandwidth selection due to kernel smoothing as in Saha and Heagerty [5] or as in the smooth method of Zheng *et al.* [6].

Moreover, nonparametric approaches previously proposed cannot properly deal with censoring depending on several markers. Indeed, with the method of Saha and Heagerty [5] or the smooth method of Zheng *et al.* [6], the cause specific hazards are estimated only conditionally on the marker under study, by non-parametric kernel-based methods. Therefore, dependence between censoring and other markers are not taken into account, whereas when modelling hazard in survival analysis, the assumption is that the censoring mechanism can only depend on covariates included in the model [33, Section 2.2.8]. Although they could be extended to deal with censoring depending on several markers, using multidimensional kernels, their practical interest would be limited by the so-called curse of dimensionality phenomenon [34]. To our mind, they are therefore not suitable for comparing several markers and making tests with marker dependent censoring, even though they appeared robust in our realistic simulation scenarii with two correlated markers and moderate associations between censoring and markers. By contrast, when censoring depend on the markers, the IPCW approach enables such tests of comparison under the additional assumption that the censoring mechanism is well specified. Nevertheless, when the dependence between censoring and markers is light, our simulations suggest that the estimator using KM-weights can be robust.

The proposed methodology is implemented in the `timeROC` package, written in R [35], that is publicly available on the Comprehensive R Archive Network (CRAN) site. The PAQUID data presented in Section 5 is attached to the package, and most of the analyses presented in Section 5 are presented as examples, making them easily reproducible.

Appendix A

Definition of IF$(\cdot)$

Let $M_{C_i}(t) = \mathbb{1}_{(\widetilde{\eta}_i=0,\widetilde{T}_i\leqslant t)} - \int_0^t \mathbb{1}_{(\widetilde{T}_i\geqslant t)} d\Lambda_C(u)$, where $\Lambda_C(\cdot)$ is the cumulative hazard function of the censoring variable C, $h_{tij,1} = \frac{\mathbb{1}_{(\widetilde{T}_i\leqslant t,\widetilde{\eta}_i=1)}\mathbb{1}_{(\widetilde{T}_j>t)}}{G(\widetilde{T}_i)G(t)}\mathbb{1}_{(M_i>M_j)}$, $h_{tij,2} = \frac{\mathbb{1}_{(\widetilde{T}_i\leqslant t,\widetilde{\eta}_i=1)}\mathbb{1}_{(\widetilde{T}_j\leqslant t,\widetilde{\eta}_j\notin\{0,1\})}}{G(\widetilde{T}_i)G(\widetilde{T}_j)}\mathbb{1}_{(M_i>M_j)}$, $f_{i1t} = \frac{\mathbb{1}_{(\widetilde{T}_i\leqslant t,\widetilde{\eta}_i=1)}}{G(\widetilde{T}_i)}$ and $h_t = \mathbb{E}\left[h_{tij,1}+h_{tij,2}\right]$. Then,

$$\mathrm{IF}\left(\widetilde{T}_i,\widetilde{\eta}_i,M_i,t\right) = \frac{\mathbb{E}\left(\Psi_{ijkl}(t)+\Psi_{jikl}(t)+\Psi_{jkil}(t)+\Psi_{jkli}(t)\middle|\left(\widetilde{T}_i,\widetilde{\eta}_i,M_i\right)\right)}{F_1(t)\left(1-F_1(t)\right)}$$

where

$$\begin{aligned}\Psi_{ijkl}(t) =& h_{tij,1}\left(1+\int_0^{\widetilde{T}_i}\frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right)\left(1+\int_0^t\frac{dM_{C_l}(u)}{S_{\widetilde{T}}(u)}\right)\\ &+h_{tij,2}\left(1+\int_0^{\widetilde{T}_i}\frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right)\left(1+\int_0^{\widetilde{T}_j}\frac{dM_{C_l}(u)}{S_{\widetilde{T}}(u)}\right)\\ &-h_t-\frac{h_t\left(1-2F_1(t)\right)}{F_1(t)\left(1-F_1(t)\right)}\left[f_{i1t}\left(1+\int_0^{\widetilde{T}_i}\frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right)-F_1(t)\right]\end{aligned}$$

Estimator $\widehat{IF}(\cdot)$

Let $\widehat{M}_C(\cdot)$ be the estimator defined by plugging in the usual Nelson–Aalen estimator of the cumulative incidence function of the censoring C and $\widehat{h}_{tij,1}$, $\widehat{h}_{tij,2}$ and $\widehat{f}_{i1t}$ be defined by plugging in the

Kaplan–Meier estimator $\widehat{G}(\cdot)$. Let $\widehat{h}_t = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \widehat{h}_{tij,1} + \widehat{h}_{tij,2}$. Then,

$$\widehat{\mathrm{IF}}\left(\widetilde{T}_i, \widetilde{\eta}_i, M_i, t\right) = \frac{\frac{1}{n(n-1)(n-2)}\sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n \widehat{\Psi}_{ijkl}(t) + \widehat{\Psi}_{jikl}(t) + \widehat{\Psi}_{jkil}(t) + \widehat{\Psi}_{jkli}(t)}{\widehat{F}_1(t)\left(1 - \widehat{F}_1(t)\right)}$$

where $\widehat{\Psi}_{ijkl}(t)$ is defined plugging in the estimators $\widehat{h}_t$, $\widehat{h}_{tij,1}$, $\widehat{h}_{tij,2}$, $\widehat{f}_{i1t}$, $\widehat{M}_C(\cdot)$, $\widehat{F}_1(t)$ and $\widehat{S}_{\widetilde{T}}(\cdot)$.

Supporting Information

Web Appendix referenced in Sections 3.2, 3.8 and 4.2 is available with this paper at the Statistics in Medicine website on Wiley Online Library. It contains details about large sample results and additional simulation results.

Acknowledgements

The MCIA (Mésocentre de Calcul Intensif Aquitain) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour provided computer time for this study with its computing facilities. France Alzheimer partly funded this work by a grant awarded to Hélène Jacqmin-Gadda in 2009. IPSEN and Novartis laboratories funded the PAQUID study, managed by Jean-François Dartigues. We are also grateful to the associate editor and two anonymous referees for their helpful comments and suggestions.

References

1. Amieva H, Jacqmin-Gadda H, Orgogozo J, Le Carret N, Helmer C, Letenneur L, Barberger-Gateau P, Fabrigoule C, Dartigues J. The 9-year cognitive decline before dementia of the Alzheimer type: a prospective population-based study. *Brain* 2005; **128**(5):1093. DOI: 10.1093/brain/awh451.
2. Aisen P, Andrieu S, Sampaio C, Carrillo M, Khachaturian Z, Dubois B, Feldman H, Petersen R, Siemers E, Doody R, *et al.* Report of the task force on designing clinical trials in early (predementia) AD. *Neurology* 2011; **76**(3):280–286. DOI: 10.1212/WNL.0b013e318207b1b9.
3. Heagerty P, Lumley T, Pepe M. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000; **56**(2):337–344. DOI: 10.1111/j.0006-341X.2000.00337.x.
4. Heagerty P, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005; **61**(1):92–105. DOI: 10.1111/j.0006-341X.2005.030814.x.
5. Saha P, Heagerty P. Time-dependent predictive accuracy in the presence of competing risks. *Biometrics* 2010; **66**(4):999–1011. DOI: 10.1111/j.1541-0420.2009.01375.x.
6. Zheng Y, Cai T, Jin Y, Feng Z. Evaluating prognostic accuracy of biomarkers under competing risk. *Biometrics* 2012; **68**(2):388–396. DOI: 10.1111/j.1541-0420.2011.01671.x.
7. Pepe M. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, 2003.
8. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves : a nonparametric approach. *Biometrics* 1988; **44**(3):837–845.
9. Uno H, Cai T, Tian L, Wei LJ. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association* 2007; **102**(478):527–537. DOI: 10.1198/016214507000000149.
10. Hung H, Chiang CT. Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics* 2010; **38**(1):8–26. DOI: 10.1002/cjs.10046.
11. Chiang C, Hung H. Non-parametric estimation for time-dependent AUC. *Journal of Statistical Planning and Inference* 2010; **140**(5):1162–1174. DOI: 10.1016/j.jspi.2009.10.012.
12. Blanche P, Dartigues JF, Jacqmin-Gadda H. Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal* 2013. DOI: 10.1002/bimj.201200045. in press.
13. Foucher Y, Giral M, Soulillou JP, Daures JP. Time-dependent ROC analysis for a three-class prognostic with application to kidney transplantation. *Statistics in Medicine* 2010; **29**(30):3079–3087. DOI: 10.1002/sim.4052.
14. Van der Laan MJ, Robins JM. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Verlag: New York, 2003.
15. Tsiatis AA. *Semiparametric Theory and Missing Data*. Springer Verlag: New York, 2006.
16. Jewell NP, Lei X, Ghani AC, Donnelly CA, Leung GM, Ho LM, Cowling BJ, Hedley AJ. Non-parametric estimation of the case fatality ratio with competing risks data: an application to severe acute respiratory syndrome (SARS). *Statistics in Medicine* 2007; **26**(9):1982–1998. DOI: 10.1002/sim.2691.
17. Antolini L, Biganzoli EM, Boracchi P. Crude cumulative incidence in the form of a Horvitz-Thompson like and Kaplan-Meier like estimator, (October 2006). COBRA Preprint Series. Working Paper 10. http://biostats.bepress.com/cobra/art10.
18. Satten GA, Datta S. The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician* 2001; **55**(3):207–210. DOI: 10.1198/000313001317098185.

19. Beran R. Nonparametric regression with randomly censored survival data. *Unpublished technical report*, University of California, Berkeley, 1981.
20. Akritas MG. Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics* 1994; **22**:1299–1327.
21. Martinussen T, Scheike TH. *Dynamic Regression Models for Survival Data*. Springer: New York, 2006.
22. Bretz F, Hothorn T, Westfall P. *Multiple Comparisons Using R*. CRC press, Boca Raton, 2010.
23. Cai T, Pepe MS. Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *Journal of the American Statistical Association* 2002; **97**(460):1099–1107. DOI: 10.1198/016214502388618915.
24. Lin DY, Fleming TR, Wei LJ. Confidence bands for survival curves under the proportional hazards model. *Biometrika* 1994; **81**(1):73–81.
25. Beyersmann J, Latouche A, Buchholz A, Schumacher M. Simulating competing risks data in survival analysis. *Statistics in Medicine* 2009; **28**(6):956–971. DOI: 10.1002/sim.3516.
26. Folstein MF, Folstein SE, McHugh PR, *et al.* "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* 1975; **12**(3):189–198.
27. Wechsler D. *Wechsler Adult Intelligence Scale (rev. ed.)* Psychological Corporation: New York, 1981.
28. Dartigues JF, Gagnon M, Barberger-Gateau P, Letenneur L, Commenges D, Sauvel C, Michel P, Salamon R. The PAQUID epidemiological program on brain ageing. *Neuroepidemiology* 1992; **11**(1):14–18.
29. Scheike T, Zhang M, Gerds T. Predicting cumulative incidence probability by direct binomial regression. *Biometrika* 2008; **95**:205–220. DOI: 10.1093/biomet/asm096.
30. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer: New York, 2009.
31. Kerr KF, Pepe MS. Joint modeling, covariate adjustment, and interaction: contrasting notions in risk prediction models and risk prediction performance. *Epidemiology* 2011; **22**(6):805–812. DOI: 10.1097/EDE.0b013e31823035fb.
32. Joly P, Commenges D, Helmer C, Letenneur L. A penalized likelihood approach for an illness–death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics* 2002; **3**(3):433. DOI: 10.1093/biostatistics/3.3.433.
33. Aalen O, Borgan Ø, Gjessing HK, Gjessing S. *Survival and Event History Analysis: A Process Point of View*. Springer, 2008.
34. Robins JM, Ritov Y. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine* 1997; **16**:285–319.
35. R Core Team. R: A language and environment for statistical computing, R foundation for statistical computing, Vienna, Austria, 2013. Available from: http://www.R-project.org.