Time-dependent AUC with right-censored data: a survey study

Paul Blanche, Aurélien Latouche, Vivian Viallon

Abstract The ROC curve and the corresponding AUC are popular tools for the evaluation of diagnostic tests. They have been recently extended to assess prognostic markers and predictive models. However, due to the many particularities of time-to-event outcomes, various definitions and estimators have been proposed in the literature. This review article aims at presenting the ones that accommodate to right-censoring, which is common when evaluating such prognostic markers.

1 Introduction

In the medical literature, a variety of general criteria have been used to assess diagnostic tests [24, 14]. Among them, the ROC curve and the area under it – the AUC – are popular tools, originally aimed at evaluating the discriminant power of continuous diagnostic tests. In this simple situation, the outcome status D is a binary variable (typically, $D = 1$ for cases and $D = 0$ for controls) and the ROC curve for a continuous diagnostic test X plots the true positive rate, or sensitivity, $\mathrm{TPR}(c) = \mathbb{P}(X > c | D = 1)$ against the false positive rate, or one minus the specificity, $\mathrm{FPR}(c) = \mathbb{P}(X > c | D = 0)$, when making threshold c vary. The AUC, which is the area under this curve, is a commonly used summary measure of the information contained in the sequences $(\mathrm{TPR})_{c \in \mathbf{R}}$ and $(\mathrm{FPR})_{c \in \mathbf{R}}$. As such, it inherits some of the properties of true and false positive rates. In particular, it is not affected by the disease prevalence (unlike positive and negative predictive values) and it can be evaluated from random samples of cases and controls. It also has a nice interpretation since it corresponds to the probability that the marker value of a randomly selected case exceeds that of a randomly selected control.

Paul Blanche
Univ. Bordeaux, ISPED and INSERM U897, Bordeaux, FRANCE e-mail: Paul.Blanche@isped.u-bordeaux2.fr

Aurélien Latouche
Conservatoire national des arts et métiers, Paris, FRANCE e-mail: aurelien.latouche@cnam.fr

Vivian Viallon
UMRESTTE (Univ. Lyon 1 and IFSTTAR), Bron, FRANCE e-mail: viallon@math.univ-lyon1.fr

The extension of the AUC (and other evaluation criteria) to the setting of prognostic markers has raised several issues. In particular, when evaluating such markers, the outcome status typically changes over time: in a cohort study for instance, patients are diseased-free when entering the study and may develop the disease during the study. This leads to three major differences with the evaluation of diagnostic tests. First, this time-dependent outcome status (which may be defined in several ways, as will be seen in Section 3) naturally implies that sensitivity, specificity, ROC curves, their AUC values and, more generally, any extension of the criteria used in the diagnostic setting, are functions of time as well. Second, the time-to-event, i.e. the time between the entry in the study and the disease onset, is usually *censored* and not fully observed, requiring dedicated inference.Third, the time lag between the entry in the study and the disease onset also leads to two further refinements: (i) the marker can be repeatedly measured over time and (ii) competing events (in addition to censoring) may be observed between the entry and the putative disease onset.

These particularities has led to the development of numerous methods aimed at estimating the time-dependent AUC for prognostic markers. In this paper, we review those that accommodate to right-censoring. Some notations are introduced in the following Section 2. Then, in Section 3 we will present several definitions and estimators of the time-dependent AUC in the "standard" setting of a baseline marker and univariate survival data. Section 4 will cover the case of longitudinal markers which corresponds to the marker being repeatedly measured over time, while we will discuss the competing events setting in Section 5. Finally, concluding remarks will be given in Section 6.

2 Notations

Let T_i and C_i denote survival and censoring times for subject i, $i = 1, \ldots, n$. We further let $Z_i = \min(T_i, C_i)$ and $\delta_i = \mathbb{I}(T_i \leq C_i)$ denote the observed time and the status indicator respectively. We will denote by $D_i(t)$ the time-dependent outcome status for subject i at time t, $t \geq 0$. Several definitions for $D_i(t)$ will be given hereafter, but we will always have $D_i(t) = 1$ if subject i is considered as a case at time t and $D_i(t) = 0$ if subject i is considered as a control at time t.

We will denote by X the marker under study, which can be a single biological marker or several biological markers combined into a predictive model (in this case, it is assumed throughout this article that the predictive model has been constructed on an independent data set; otherwise sub-sampling techniques are needed [21]). Without loss of generality, we will suppose that larger values of X are associated with greater risks (otherwise, X can be recoded to achieve this). We will denote by g and G^{-1} the probability density function and the quantile function of marker X. In Section 3, we assume that marker X is measured once at $t = 0$, and we will denote by X_i the marker value for subject i. In Section 4, which treats the longitudinal setting, the marker is measured repeatedly over time, and we will denote by $X_i(s)$ the marker value at time s for subject i.

3 Time-dependent AUCs in the standard setting

Definitions of time-dependent ROC curves, $\text{ROC}(t)$, follow from definitions of usual ROC curves and thus rely on first defining time-dependent true and false positive rates. For any threshold c, these two functions of time are defined as $\text{TPR}(c,t) = \mathbf{P}(X > c|D(t) = 1)$ and $\text{FPR}(c,t) = \mathbf{P}(X > c|D(t) = 0)$. $\text{ROC}(t)$ then simply plots $\text{TPR}(c,t)$ against $\text{FPR}(c,t)$ making threshold c vary. The time-dependent AUC at time t is then defined as the area under this curve,

$$\text{AUC}(t) = \int_{-\infty}^{\infty} \text{TPR}(c,t) \left| \frac{\partial \text{FPR}(c,t)}{\partial c} \right| dc. \tag{1}$$

As a matter of fact, these definitions deeply rely on that of the outcome status at time t, $D(t)$. Heagerty and Zheng [18] described several definitions of *cases* and *controls* in this survival outcome setting. According to Heagerty and Zheng's terminology and still denoting by T_i survival time for subject i, cases are said to be *incident* if $T_i = t$ is used to define cases at time t, and *cumulative* if $T_i \leq t$ is used instead. Similarly, depending on whether $T_i > \tau$ for a large time $\tau > t$ or $T_i > t$ is used for defining controls at time t, they are said to be *static* or *dynamic* controls. Depending on the definition retained for cases and controls at time t, four definitions of the time-dependent AUC value may be put forward. In the following paragraphs, we will present formulas and estimators for the most commonly used ones and will discuss their respective interests.

3.1 The cumulative dynamic AUC: $AUC^{\mathbb{C},\mathbb{D}}(t)$

The setting of cumulative cases and dynamic controls may be regarded as the most natural choice for planning enrollment criteria in clinical trials or when specific evaluation times are of particular interest. It simply corresponds to defining cases at time t as subjects who experienced the event prior to time t, and controls at time t as patients who were still event-free at time t. In other words, it corresponds to setting $D_i(t) = \mathbb{I}(T_i \leq t)$. Cumulative true positive rates and dynamic false positive rates are then respectively defined as

$$\text{TPR}^{\mathbb{C}}(c,t) = \mathbf{P}(X > c|T \leq t) \quad \text{and} \quad \text{FPR}^{\mathbb{D}}(c,t) = \mathbf{P}(X > c|T > t). \tag{2}$$

The *cumulative/dynamic* AUC at time t is then obtained by using these definitions of true and false positive rates in (1). Usually, however, $\mathbb{I}(T \leq t)$ is not observed for all subjects due to the presence of censoring before time t and simple contingency tables can therefore not be used to return estimates of $\text{TPR}^{\mathbb{C}}(c,t)$, $\text{FPR}^{\mathbb{D}}(c,t)$ and $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$. To handle censoring, Bayes' theorem can be used to rewrite $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ as a function of the conditional survival function $\mathbf{P}(T > t|X = x)$ (see paragraph 3.1.1 below). Other approaches rely on so-called Inverse Probability of Censoring

Weighted (IPCW) estimates (see paragraph 3.1.2 below). Before describing these two approaches in more details below, we shall add that Chambless and Diao [6] developed an alternative method – which will not be described here – based on an idea similar to the one used to derive the Kaplan-Meier estimator of the cumulative distribution function in the presence of censoring. Among all these methods, only those relying on primary estimates of $\mathbf{P}(T > t|X = x)$ (and a recent extension of IPCW estimates proposed in [2]) may account for the dependence between censoring and the marker (since they basically only assume that T and C are independent given X and not that T and C are independent). We refer the reader to [2, 19, 37] for empirical comparisons and illustrations of these various methods.

3.1.1 Methods based on primary estimates of $\mathbf{P}(T > t|X = x)$

Bayes' theorem yields the following expressions for $\mathrm{TPR}^{\mathbb{C}}(c,t)$ and $\mathrm{FPR}^{\mathbb{D}}(c,t)$

$$\mathrm{TPR}^{\mathbb{C}}(c,t) = \frac{\int_c^{\infty} \mathbf{P}(T \leq t|X=x)g(x)dx}{\mathbf{P}(T \leq t)}, \quad \mathrm{FPR}^{\mathbb{D}}(c,t) = \frac{\int_c^{\infty} \mathbf{P}(T > t|X=x)g(x)dx}{\mathbf{P}(T > t)}.$$

From (1), it readily follows that

$$\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t) = \int_{-\infty}^{\infty} \int_c^{\infty} \frac{\mathbf{P}(T \leq t|X=x)\mathbf{P}(T > t|X=c)}{\mathbf{P}(T \leq t)\mathbf{P}(T > t)} g(x)g(c)dxdc. \quad (3)$$

Since $\mathbf{P}(T > t) = \int_{-\infty}^{\infty} \mathbf{P}(T > t|X = x)g(x)dx$, any estimator $\widehat{S}_n(t|x)$ of the conditional survival function $\mathbf{P}(T > t|X = x)$ yields an estimator of $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$:

$$\widehat{\mathrm{AUC}}^{\mathbb{C},\mathbb{D}}(t) = \frac{\sum_{i=1}^n \sum_{j=1}^n \widehat{S}_n(t|X_j)[1 - \widehat{S}_n(t|X_i)]\, \mathbb{I}(X_i > X_j)}{\sum_{i=1}^n \sum_{j=1}^n \widehat{S}_n(t|X_j)[1 - \widehat{S}_n(t|X_i)]}.$$

In [6], the authors suggested to use a Cox model to derive estimates $\widehat{S}_n(t|x)$, while one of the methods described in Heagerty *et al.* [17] reduces to using the conditional Kaplan-Meier estimator as in [1]. Some theoretical results for these two methods can be found in [33] and [3, 7, 20] respectively.

We shall add that Viallon and Latouche [37] related $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$ to the quantity $\mathbf{P}(T \leq t|X = G^{-1}(q))$ – a time-dependent version of the predictiveness curve:

$$\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t) = \frac{\int_0^1 q\mathbf{P}(T \leq t|X = G^{-1}(q))dq - [\mathbf{P}(T \leq t)]^2/2}{\mathbf{P}(T > t)\mathbf{P}(T \leq t)}.$$

This confirms that most standard statistical summaries of predictability and discrimination can be derived from the predictiveness curve, as pointed out in [14, 15].

3.1.2 Methods based on IPCW estimators

In [19] and [36], the authors independently suggested to use IPCW-type estimates:

$$\widehat{\text{TPR}}^{\mathbb{C}}(c,t) = \frac{\sum_{i=1}^{n} \mathbb{1}(X_i > c, Z_i \le t)\frac{\delta_i}{n\hat{S}_C(Z_i)}}{\sum_{i=1}^{n} \mathbb{1}(Z_i \le t)\frac{\delta_i}{n\hat{S}_C(Z_i)}}, \quad \widehat{\text{FPR}}^{\mathbb{D}}(c,t) = \frac{\sum_{i=1}^{n} \mathbb{1}(X_i > c, Z_i > t)}{\sum_{i=1}^{n} \mathbb{1}(Z_i > t)},$$

where $\hat{S}_C(\cdot)$ is the Kaplan-Meier estimator of the survival function of the censoring time C. The expression of the false positive rate estimator is more compact because weights all equal $1/(n\hat{S}_C(t))$ under the assumption of independence between C and X, and then vanish. $\widehat{\text{FPR}}^{\mathbb{D}}(c,t)$ corresponds to 1 minus the empirical distribution function of X among individuals for whom $Z_i > t$. In the absence of censoring before time t, $\widehat{\text{TPR}}^{\mathbb{C}}(c,t)$ also reduces to the usual empirical version of $\text{TPR}^{\mathbb{C}}(c,t)$, i.e., 1 minus the empirical distribution function of X among individuals for whom $T_i \le t$. It can be shown (see [19, 30]) that an estimator of $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ is then given by

$$\widehat{\text{AUC}}^{\mathbb{C},\mathbb{D}}(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}(Z_i \le t)\,\mathbb{1}(Z_j > t)\,\mathbb{1}(X_i > X_j)\frac{\delta_i}{\hat{S}_C(Z_i)\hat{S}_C(t)}}{n^2\hat{S}(t)[1 - \hat{S}(t)]},$$

where $\hat{S}(t)$ is the Kaplan-Meier estimator of $\mathbb{P}(T > t)$.

Theoretical guarantees for these estimators can be found in [19] and [36]. These estimators are in a sense more flexible than those presented in paragraph 3.1.1 above: they are model-free and they do not rely on any bandwidth selection (unlike the estimator of Heagerty *et al.* [17] for instance, which is based on a local version of the Kaplan-Meier estimator). However, when censoring may depend on marker X, quantities like the conditional survival function of C given the marker X, $S_C(\cdot|X)$, have to be estimated [2], which implies either to work under some (semi-)parametric model or the selection of some parameter if nonparametric estimation is prefered.

3.2 The incident static AUC: $AUC^{\mathbb{I},\mathbb{S}}(t)$

Using the dynamic definition of controls, the control group varies with time, and so does the x-axis of the corresponding ROC curves: in situations where trends over time are of particular interest, this renders their interpretation more difficult (since such trends may be partly due to changing control groups). Moreover, the group of static controls is interesting in that it tries to mimic the group of individuals who never develop the disease, which can be seen as an ideal control group in some situations. In particular, patients with preclinical diseases are eliminated from the control group as far as possible, if τ is large enough.

Regarding cases, the incident definition has several advantages over the dynamic definition [26]. The cumulative TPR does not distinguish between events that occur

early versus late, and it shows redundant information over time (since early events are also included in the cumulative TPR for late evaluation times). Moreover, as pointed out by [4], the cumulative TPR can be computed from the incident TPR when the distribution of the event time is known.

Putting all this together, several authors have proposed estimators of the time-dependent ROC curve relying on the incident definition of cases and static definition of controls. Standard numerical integration techniques are then used to compute an estimate for $\text{AUC}^{\mathbb{I},\mathbb{S}}(t)$ from the estimators of the ROC curve.

Incident true positive rates and static false positive rates are defined as

$$\text{TPR}^{\mathbb{I}}(c,t) = \mathbb{P}(X > c | T = t) \quad \text{and} \quad \text{FPR}_\tau^{\mathbb{S}}(c) = \mathbb{P}(X > c | T > \tau). \qquad (4)$$

Applying Bayes' theorem, they can further be rewritten (see, *e.g.*, [33])

$$\text{TPR}^{\mathbb{I}}(c,t) = \frac{\int_c^\infty f(t|x)g(x)dx}{\int_{-\infty}^\infty f(t|x)g(x)dx} \quad \text{and} \quad \text{FPR}_\tau^{\mathbb{S}}(c) = \frac{\int_c^\infty \mathbb{P}(T > \tau | X = x)g(x)dx}{\int_{-\infty}^\infty \mathbb{P}(T > \tau | X = x)g(x)dx},$$

where $f(t|x) = \partial \mathbb{P}(T \leq t | X = x)/\partial t$ is the conditional density function of T given $X = x$. Under a standard Cox model of the form $\lambda(t;X) = \lambda_0(t)\exp(\beta X)$ – here $\lambda(t;X)$ stands for the conditional hazard rate of T given X while λ_0 is the unspecified baseline hazard rate – Song and Zhou [33] deduced that

$$\text{TPR}^{\mathbb{I}}(c,t) = \frac{\int_c^\infty \exp(\beta x)\exp\{-\Lambda_0(t)\exp(\beta x)\}g(x)dx}{\int_{-\infty}^\infty \exp(\beta x)\exp\{-\Lambda_0(t)\exp(\beta x)\}g(x)dx}$$

$$\text{FPR}_\tau^{\mathbb{S}}(c) = \frac{\int_c^\infty \exp\{-\Lambda_0(\tau)\exp(\beta x)\}g(x)dx}{\int_{-\infty}^\infty \exp\{-\Lambda_0(\tau)\exp(\beta x)\}g(x)dx},$$

where $\Lambda_0(t) = \int_{-\infty}^t \lambda_0(u)du$ is the cumulative baseline hazard function. Estimation of $\text{TPR}^{\mathbb{I}}(c,t)$ and $\text{FPR}_\tau^{\mathbb{S}}(c)$ can then be achieved by plug-in methods. We shall add that Song and Zhou actually considered a slightly more general set-up where additional covariates can be accounted for.

In [18], Heagerty and Zheng adopted a slightly different approach. To estimate $\text{TPR}^{\mathbb{I}}(c,t)$, they used a (possibly time-varying-coefficients) Cox model of the form $\lambda(t;X) = \lambda_0(t)\exp(\beta(t)X)$ in combination with the fact that the distribution of $X \cdot \exp(\beta X)$ for subjects in the risk set at time t is equal to the conditional distribution of X given $T = t$ (see, *e.g.,* [38]). Setting $R(t) = \{i : Z_i \geq t\}$, this leads to

$$\widehat{\text{TPR}}^{\mathbb{I}}(c,t) = \frac{\sum_{i \in R(t)} \mathbb{I}(X_i > c)\exp\{\beta(t)X_i\}}{\sum_{i \in R(t)} \exp\{\beta(t)X_i\}}.$$

As for the estimation of $\text{FPR}_\tau^{\mathbb{S}}(c)$, they proposed a model-free approach using the empirical distribution function for marker values among the control set $S_\tau := \{i : Z_i > \tau\}$. Namely, denoting by n_τ the cardinality of S_τ, they proposed

$$\widehat{\text{FPR}}_\tau^\mathbb{S}(c) = \frac{1}{n_\tau} \sum_{i \in S_\tau} \mathbb{I}(X_i > c),$$

which is $\widehat{\text{FPR}}^\mathbb{D}(c,t)$ of Section 3.1.2, except τ is used instead of t.

Cai *et al.* [4] proposed another approach in the context of longitudinal markers; it will be described in more details in Section 4. In addition, two non parametric approaches were recently proposed (see [32] and [29]).

Note also that estimators for the time-dependant incident/dynamic AUC, $\text{AUC}^{\mathbb{I},\mathbb{D}}(t)$, can be obtained by simply replacing τ by t in the definitions of $\widehat{\text{FPR}}_\tau^\mathbb{S}(c)$ above [18]. A global accuracy measure has further been derived from the definition of $\text{AUC}^{\mathbb{I},\mathbb{D}}(t)$, which is particularly appealing when no a priori time t is identified and/or when trends over time are not of interest [18].

4 Time-dependent ROC curve and AUCs with longitudinal marker

In this section, we review extensions of the above estimators for longitudinally collected subject measurements. For instance some authors would assess the discrimination performance of CD4 counts repeatedly measured every week on time from seroconvertion to progression to AIDS [42]. Therefore, time-dependent sensitivities and specificities have been extended to deal with the fact that (*i*) the time at which marker X is measured can vary and (*ii*) marker can be repeatedly measured on the same subject. Let s denote the timing of marker measurement and $X(s)$ the marker value at time s. For $t \geq s$, Zheng and Heagerty [42] extended *cumulative/dynamic* definitions

$$\text{TPR}^\mathbb{C}(c,s,t) = \mathbb{P}(X(s) > c | T \in [s,t]), \qquad \text{FPR}^\mathbb{D}(c,s,t) = \mathbb{P}(X(s) > c | T > t).$$

For a fixed time $\tau \geq s$, Zheng and Heagerty [40] extended *incident/static* definitions:

$$\text{TPR}^\mathbb{I}(c,s,t) = \mathbb{P}(X(s) > c | T = t) \quad \text{and} \quad \text{FPR}^\mathbb{S}(c,s,\tau) = \mathbb{P}(X(s) > c | T > \tau).$$

Although others approaches have been proposed to estimate these quantities, we only review estimators that deal with censored data here. We should also mention that in this longitudinal context, estimators of the AUC are obtained by numerically integrating estimators of the ROC curve.

4.1 Cumulative dynamic estimators with longitudinal marker

Rizopoulos [27] recently proposed a joint modeling approach. The marker trajectory is modeled by usual linear mixed model for longitudinal data, and a parametric pro-

portional hazard is used to model the time-to-event given the marker trajectory. The two submodels are linked with shared random effects to capture the intra-subject correlation. Standard maximum likelihood estimation is used to fit the joint model. Then, $\text{TPR}^{\mathbb{C}}$ and $\text{FPR}^{\mathbb{D}}$ are computed from the estimated parameters and Monte Carlo simulations are used to make inference. As this approach is fully parametric, its main advantage is its efficiency. This approach also allows censoring to depend on the marker [35]. The counterpart is that the parametric model must be carefully chosen, and checking model fit is not straightforward.

A more flexible methodology was proposed in [42], with fewer parametric assumptions. Setting $T^* = T - s$, the "residual failure time", and $t^* = t - s$, they rewrote

$$\text{TPR}^{\mathbb{C}}(c,s,t^*) = \frac{1 - F_{X|s}(c) - S(c,t^*|s)}{1 - S(t^*|s)}, \qquad \text{FPR}^{\mathbb{D}}(c,s,t^*) = 1 - \frac{S(c,t^*|s)}{S(t^*|s)},$$

with $F_{X|s}(c) = \mathbb{P}(X(s) < c|s, T^* > 0)$ the conditional distribution of marker given measurement time, $S(t^*|s) = \mathbb{P}(T^* > t^*|s, T^* > 0)$ the survival probability for individuals who survived beyond s and $S(c,t^*|s) = \mathbb{P}(X(s) > c, T^* > t^*|s, T^* > 0)$. They proposed to estimate $F_{X|s}(c)$ with the semiparametric estimator proposed by Heagerty and Pepe [16]. Therefore, only the location and scale of the conditional distribution of marker given measurement time are parametrized. To estimate the survival terms $S(c,t^*|s)$ and $S(t^*|s)$, they proposed the use of a "partly conditional" hazard function to model the residual failure time $T^* = T - s$. For subject i at measurement time s_{ik}, this function is modeled by

$$\lambda_{ik}(t^*|X_i(s_{ik}), 0 \le s_{ik} \le T_i) = \lambda_0(t^*)\exp\left[\beta(t^*)X_i(s_{ik}) + \alpha^T f(s_{ik})\right]$$

where $f(s)$ are vectors of spline basis functions evaluated at measurement time s, and $\lambda_0(t^*)$ is left unspecified. Estimators of $\beta(\cdot)$ and α have been previously proposed [41]. As this approach is semiparametric, its main advantage is its flexibility. However, by contrast to the approach of Rizopoulos [27], this one is less efficient and does not allow marker-dependent censoring.

4.2 Incident static estimators with longitudinal marker

Several authors consider the incident/static definition of AUC [31, 12, 4, 34]. However, censored data are only accounted for by Cai *et al.* [4] who proposed to model

$$\text{TPR}^{\mathbb{I}}(c,s,t) = g_D\left(\eta_\alpha(t,s) + h(c)\right), \quad t \le \tau$$
$$\text{FPR}^{\mathbb{S}}(c,s,\tau) = g_\tau\left(\xi_a(s) + d(c)\right)$$

where g_D and g_τ are specified inverse link functions and $h(\cdot)$ and $d(\cdot)$ are unspecified baseline functions of threshold c. The dependence in time is parametrically modeled by $\eta_\alpha(t,s) = \alpha^T \eta(t,s)$ and $\xi_a(s) = a^T \xi(s)$ where $\eta(t,s)$ and $\xi(s)$ are vectors of polynomial or spline basis functions. These models are semiparametric with

respect to the marker distribution in cases and nonparametric in regards to controls. As pointed out in [26], this model is very flexible as it does not specify any distributional form for the distribution of the marker given the event-time, but only model the effect of time-to-event on the marker distribution with a parametric form. Model estimation is performed by solving some estimating equations and large sample theory was established allowing a resampling method to construct confidence bands and make inference [4]. Interestingly, the authors of [4] also showed that covariates can easily be included in $\text{TPR}^{\mathbb{I}}$ and $\text{FPR}^{\mathbb{S}}$, enabling to directly quantify how performances of the marker vary with these covariates.

5 Time-dependent AUC and Competing risks

We now consider the setting where a subject might experience multiple type of failures: in this section, we review extensions of time-dependent AUCs to competing risks. For example, we may want to assess the discrimination of a given score on death from prostate cancer with death from other causes acting as a competing event. For the sake of simplicity, we will assume there are only two competing events, and we let $\delta_i = j$ denote that subject i experienced the competing event of type j ($j = 1, 2$, with $j = 1$ for the event of interest). The observed data consists of a failure time and a failure type (Z_i, δ_i) with $\delta_i = 0$ denoting a censored observation.

In their review paper [26] sketched most potential extensions and introduced event-specific sensitivity and specificity. They also highlighted that the crucial point was to determine whether patients experimenting a competing event should be treated as a control when evaluating the discrimination of the marker under study with respect to the event of interest. More precisely, two settings can be considered.

First, if marker X is potentially discriminatory for both the event of interest and the competing event, then both event specific AUCs should be considered simultaneously [28, 13]. For illustration, in the cumulative/dynamic setting, cases at time t can be stratified according to the event type, $\text{Case}_1 = \{i : T_i \leq t, \delta_i = 1\}$ and $\text{Case}_2 = \{i : T_i \leq t, \delta_i = 2\}$, and controls at time t are the event-free group at time t, $\text{Control} = \{i : T_i > t\}$. Following these lines Saha and Heargerty [28] proposed event specific versions of (2)

$$\text{TPR}_j^{\mathbb{C}}(c,t) = \mathbb{P}(X > c | T \leq t, \delta = j), \ \text{FPR}^{\mathbb{D}}(c,t) = \mathbb{P}(X > c | T > t, \delta \in \{1,2\}). \ (5)$$

Estimation follows from [17], using the conditional cumulative incidence associated to competing event j, $\mathbb{P}(T \leq t, \delta = j | X)$, instead of the conditional survival function of $\mathbb{P}(T \leq t | X)$. In the context of renal transplantation, Foucher *et al.* [13] considered a slight modification of definitions (5), where controls can also be "event specific". In addition, an extension of the incident/dynamic AUC to the competing events setting was proposed by Saha and Heargerty [28].

The other option is to consider both event-free patients and patients with the competing event [39, 21] as controls. For instance, dynamic controls at time t can be defined as the group $\{i : T_i > t\} \cup \{i : T_i \le i, \delta_i = 2\}$. This leads to the estimation of only one ROC curve, for the event of interest. In [39], Zheng *et al.* based their approach on initial estimates of the conditional cumulative incidence function for the event of interest. Their initial method provides consistent estimators if the proportional hazard assumption holds for each cause specific hazard. To relax this assumption a smooth estimator was also proposed. Another approach was described in [21], which follows the lines of DeLong *et al.* [9]. However, the suitability of this method to deal with censored data is not established.

We shall add that, as pointed out in [28, 39], employing a direct regression model for the conditional cumulative incidence would lead to a simpler estimation of the cumulative/dynamic AUC and a less convoluted interpretation of the marker effect. However, the extension to the setting of a longitudinal marker [8] as well as the evaluation of a risk score (which is usually built with a cause-specific hazard approach) would not be straightforward.

6 Discussion

While the AUC is uniquely defined in the context of the evaluation of diagnostic tests, its extension to prognostic markers has led to the development of a variety of definitions: these definitions vary according to the underlying definitions of cases (incident or cumulative) and controls (static or dynamic), and also depend on the study characteristics (the marker can be measured only once or repeatedly and competing events may be considered, or not). Regarding the choice of the retained definition for cases and controls, no clear guidance has really emerged in the literature. It seems however that the cumulative/dynamic definition may be more appropriate for clinical decisions making (enrollment in clinical trials for instance) while the incident/static definition may be more appropriate for "pure" evaluation of the marker (if interpretation of trends of AUC values over time is of particular interest for instance). Once this definition has been chosen, appropriate estimators are available, depending on various assumptions (independence of the marker and censoring, proportional hazards, ...), and we presented most of them in this review article.

For the sake of brevity, we were not able to cover some interesting extensions of time-dependent AUCs. In particular, covariate specific time-dependent ROC curves and AUCs have been studied in order to adjust the discrimination of a marker for external covariates (age, gender, ...). We refer the reader to [19, 33] for the standard setting, [4] for the longitudinal setting and [39] for the competing events setting. In addition, some authors advocate that not the entire ROC curve is of interest and the area under only a portion of it should be computed, leading to the so-called *partial AUC* [11]. In the context of prognostic markers, Hung and Chiang [20] proposed a nonparametric estimator of the cumulative/dynamic time-dependent version of the partial AUC. Other interesting extensions include diverse censoring patterns [22]

(only right-censoring was considered in this review) and the combination of results from multiple studies [3] which is particularly useful in genomic studies.

Another closely related topic is the evaluation of the added predictive ability of a new marker: for instance, we may wonder how better a risk score would be if we added some biological markers (SNPs, genes, ...). We refer the reader to the works in [5, 10, 23, 25] for some insights, noticing though that most of these works do not cover the right-censored setting considered in our review.

References

1. M. Akritas. Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics*, 22:1299–1327, 1994.
2. P. Blanche, J. F. Dartigues, and H. Jacqmin-Gadda. Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Submitted*, 2012.
3. T. Cai, T.A. Gerds, Y. Zheng, and J. Chen. Robust prediction of t-year survival with data from multiple studies. *Biometrics*, 67:436–444, Jun 2011.
4. T. Cai, M.S. Pepe, Y. Zheng, T. Lumley, and N.S. Jenny. The sensitivity and specificity of markers for event times. *Biostatistics*, 7:182–197, Apr 2006.
5. L.E. Chambless, C.P. Cummiskey, and G. Cui. Several methods to assess improvement in risk prediction models: extension to survival analysis. *Statistics in medicine*, 30(1):22–38, 2011.
6. L.E. Chambless and G. Diao. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in Medicine*, 25:3474–3486, Oct 2006.
7. C.T. Chiang and H. Hung. Nonparametric estimation for time-dependent AUC. *Journal of Statistical Planning and Inference*, 140(5):1162 – 1174, 2010.
8. G. Cortese and P. K. Andersen. Competing risks and time-dependent covariates. *Biometrical Journal*, 52(1):138–158, 2010.
9. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, 1988.
10. O.V. Demler, M.J. Pencina, and R.B. D'Agostino Sr. Equivalence of improvement in area under ROC curve and linear discriminant analysis coefficient under assumption of normality. *Statistics in Medicine*, 30(12):1410–1418, 2011.
11. L.E. Dodd and M.S. Pepe. Partial AUC estimation and regression. *Biometrics*, 59(3):614–623, 2003.
12. R. Etzioni, M. Pepe, G. Longton, C. Hu, and G. Goodman. Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Medical Decision Making*, 19(3):242–251, 1999.
13. Y. Foucher, M. Giral, J.P. Soulillou, and J.P. Daures. Time-dependent ROC analysis for a three-class prognostic with application to kidney transplantation. *Statistics in Medicine*, 29:3079–3087, Dec 2010.
14. M.H. Gail and R.M. Pfeiffer. On criteria for evaluating models of absolute risk. *Biostatistics*, 6(2):227–239, 2005.
15. W. Gu and M.S. Pepe. Measures to summarize and compare the predictive capacity of markers. *International Journal of Biostatistics*, 5:27–49, 2009.
16. P. J. Heagerty and M. S. Pepe. Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):533–551, 1999.
17. P.J. Heagerty, T. Lumley, and M.S. Pepe. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56:337–344, Apr 2000.
18. P.J. Heagerty and Y. Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61:92–105, 2005.

19. H. Hung and C.T. Chiang. Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics*, 38(1):8–26, 2010.

20. H. Hung and C.T. Chiang. Nonparametric methodology for the time-dependent partial area under the ROC curve. *Journal of Statistical Planning and Inference*, 141(12):3829 – 3838, 2011.

21. M. Lee, K. A. Cronin, M. H. Gail, and E. J. Feuer. Predicting the absolute risk of dying from colorectal cancer and from other causes using population-based cancer registry data. *Statistics in Medicine*, 31(5):489–500, 2012.

22. J. Li and S. Ma. Time-dependent ROC analysis under diverse censoring patterns. *Statistics in Medicine*, 30:1266–1277, 2011.

23. M.J. Pencina, R.B. D'Agostino Sr, R.B. D'Agostino Jr, and R.S. Vasan. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine*, 27(2):157–172, 2008.

24. M.S. Pepe. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA, 2004.

25. M.S. Pepe, K.F. Kerr, G. Longton, and Z. Wang. Testing for improvement in prediction model performance. *Preprint available at http://www.bepress.com/uwbiostat/paper379/*, 2011.

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

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

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

29. P. Saha-Chaudhuri and PJ Heagerty. Non-parametric estimation of a time-dependent predictive accuracy curve. *Biostatistics*, 2012.

30. G.A. Satten and S. Datta. The kaplan-meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, 55(3):207–210, 2001.

31. E.H. Slate and B.W. Turnbull. Statistical models for longitudinal biomarkers of disease onset. *Statistics in medicine*, 19(4):617–637, 2000.

32. X. Song, X. H. Zhou, and S. Ma. Nonparametric receiver operating characteristic-based evaluation for survival outcomes. *Statistics in Medicine*, 31(23):2660–2675, 2012.

33. X. Song and X.H. Zhou. A semi-parametric approach for the covariate specific roc curve with survival outcome. *Statistica Sinica*, 18:947–965, 2008.

34. F. Subtil, C. Pouteil-Noble, S. Toussaint, E. Villar, and M. Rabilloud. A Simple Modeling-free Method Provides Accurate Estimates of Sensitivity and Specificity of Longitudinal Disease Biomarkers. *Methods of Information in Medicine*, 48:299–305, 2009.

35. A.A. Tsiatis and M. Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834, 2004.

36. H. Uno, T. Cai, L. Tian, and L.J Wei. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537, 2007.

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

38. R. Xu and J. O'Quigley. Proportional hazards estimate of the conditional survival function. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):667–680, 2000.

39. Y. Zheng, T. Cai, Y. Jin, and Z. Feng. Evaluating Prognostic Accuracy of Biomarkers under Competing Risk. *Biometrics*, Dec 2011.

40. Y. Zheng and P.J. Heagerty. Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics*, 5(4):615–632, 2004.

41. Y. Zheng and P.J. Heagerty. Partly conditional survival models for longitudinal data. *Biometrics*, 61(2):379–391, 2005.

42. Y. Zheng and P.J. Heagerty. Prospective accuracy for longitudinal markers. *Biometrics*, 63:332–341, 2007.