Discrimination measures for survival outcomes: connection between the AUC and the predictiveness curve

Vivian Viallon, Aurélien Latouche

Discrimination measures for survival outcomes: connection between the AUC and the predictiveness curve

Vivian Viallon and Aurélien Latouche

Abstract

Finding out biomarkers and building risk scores to predict the occurrence of survival outcomes is a major concern of clinical epidemiology, and so is the evaluation of prognostic models. In this paper, we are concerned with the estimation of the time-dependent AUC – area under the receiver operating curve – which naturally extends standard AUC to the setting of survival outcomes and enables to evaluate the discriminative power of prognostic models. We establish a simple and useful relation between the predictiveness curve and the time-dependent AUC – AUC(t). This relation confirms that the predictiveness curve is the key concept for evaluating calibration and discrimination of prognostic models. It also highlights that accurate estimates of the conditional absolute risk function should yield accurate estimates for AUC(t). From this observation, we derive several estimators for AUC(t) relying on distinct estimators of the conditional absolute risk function. An empirical study was conducted to compare our estimators with existing ones and assess the effect of model misspecification – when estimating the conditional absolute risk function– on the AUC(t) estimation. We further illustrate the methodology on the Mayo PBC and the VA lung cancer data sets.

1 Introduction

Finding out biomarkers and developing risk scores to predict the occurrence of survival outcomes is a major concern of clinical epidemiology (Steyerberg et al., 2010). As a result, a very active domain of research in clinical epidemiology and biostatistics is the development of evaluation criteria for prognostic tools (Steyerberg et al., 2010; Schemper, 2003). Overall, most criteria proposed in the literature are either devoted to calibration or discrimination evaluation (Gail and Pfeiffer, 2005). Calibration evaluates the goodness-of-fit of, mostly, risk prediction tools by comparing average predicted risks with the observed proportion of events in groups of patients (it is usually evaluated on the whole sample and/or each decile of predicted risks). On the other hand, discrimination measures the ability of a risk prediction tool or a single biomarker to distinguish the individuals who developed the disease and those who did not. Most of the existing criteria were originally defined for diagnostic tests and rely on the observation of a binary outcome representing, for instance, disease status. Extending these criteria to survival outcomes is generally not straightforward, especially because of the presence of censored data (see, *e.g.*, Viallon et al. (2009) for the calibration of risk prediction tools). In the present work we will focus on methods that assess discrimination of prognostic tools. More precisely, we are concerned here with an extension of the area under the receiver operating curve (AUC) to survival outcomes.

The AUC is a standard tool for evaluating the discrimination of diagnostic models. Originally, the ROC curve was designed for a continuous (bio)marker X and a binary outcome D. In this simple case, it plots sensitivity, $\mathrm{P}(X > c|D = 1)$, against 1 minus specificity, $1 - \mathrm{P}(X \leq c|D = 0)$, for all possible values c; the AUC is then simply computed as the area under this curve. Several extensions of the AUC have been developed to account for survival outcomes. We shall notably evoke Harrel's concordance index (Harrell et al., 1982), which is the fraction of pairs of patients whose predicted survival times are correctly ordered among all pairs that can actually be ordered. More recently Gönen and Heller (2005) derived an analytical expression of the c-index under the Cox model (Cox, 1972) leading to an

estimator that is not affected by censoring. Another recently proposed approach consists in considering time-dependent AUC (see Heagerty et al. (2000); Heagerty and Zheng (2005); Chambless and Diao (2006); Pepe et al. (2008b)). Indeed, in prospective cohort studies, binary outcomes such as disease status can change over time, and it is therefore legitimate to consider time-dependent ROC curves. This concept requires to define time-dependent sensitivity and specificity accordingly. Various such definitions have been proposed in the literature, leading, on turn, to various definitions for the time-dependent ROC curve and time-dependent AUC, AUC(t) (Heagerty and Zheng, 2005). Heagerty and Zheng (2005) suggested extensions of the standard cross-sectional sensitivity and specificity based on extended definitions of *cases* and *controls* for survival outcomes. According to Heagerty and Zheng's terminology and 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 $T_i > t^\star$ for a fixed $t^\star \geq t$ or $T_i > t$ is used for defining controls at time t, they are said to be *static* or *dynamic* controls. In the sequel we focus on the setting of cumulative cases and dynamic controls, originally developed by Heagerty et al. (2000) and further studied by Chambless and Diao (2006). The static controls setting is detailed in Pepe et al. (2008b), while the incident/dynamic one is described in depth in Heagerty and Zheng (2005).

Recently, the predictiveness curve, which describes the distribution of the predicted disease risk, was advocated as a unifying approach for discrimination (Pepe et al., 2008a; Gu and Pepe, 2009). Indeed, many criteria used for evaluating discrimination, such as the proportion of explained variation, the standardized total gain and some recently proposed risk reclassification measures (Pencina et al., 2007), were shown to express as simple functions of the predictiveness curve (Gu and Pepe, 2009). On the contrary, no relation has been obtained between the AUC and the predictiveness curve. In this paper, we derive such a relation, show that is still holds in the setting of survival outcomes and, accordingly, propose an estimator of AUC(t) relying on the estimation of the conditional absolute risk. Our result highlights that proper estimation of the conditional absolute risk function would yield accurate estimates for AUC(t).

The paper is organized as follows. In Section 2, we first recall some basics about time-dependent AUC for survival outcomes following Heagerty and Zheng's terminology (Heagerty and Zheng, 2005). Then, we establish a useful relation between the cumulative/dynamic AUC(t), that we shall denote by $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ hereafter, and the predictiveness curve. This relation allows us to propose a bunch of new estimators for $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ based on several estimators of the conditional absolute risk function. Section 3 shows some results from an empirical comparative study. In Section 4, we illustrate the use of the proposed methods on the Mayo PBC and the VA lung cancer data sets. We close with a discussion.

2 Time–dependent ROC curves and AUC(t)

2.1 Notations

Let T_i and C_i denote survival and censoring times respectively for subject i, $i = 1, \ldots, n$, so that the only available information about T_i is (Z_i, δ_i) where $Z_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$ stand for the follow-up time and the censoring indicator respectively. Further denote by $D_i(t)$ the time-dependent outcome status for subject i at time t, $t \geq 0$; $D_i(t)$ will be defined precisely hereafter.

In this paper, our concern is the estimation of $\text{AUC}^{\mathbb{C},\mathbb{D}}(t_0)$ for a marker X and some fixed time t_0. Throughout, we will use the terminology "marker" when referring to X, but X may also denote a risk score computed from a given regression model or a published score, in which case, generally, $X = X(t_0) = \hat{P}(T \leq t_0 | \mathbf{Z})$, where $\mathbf{Z}$ stands for a vector of covariates. It is noteworthy that, even in this setting, X can be considered as fixed since t_0 is fixed. We then denote by G, G^{-1} and g the cumulative distribution function, the corresponding quantile function and the density function of marker X, respectively. Further denote by $F(t) = \text{P}(T \leq t)$ the absolute risk and let $F(t; X = x) = \text{P}(T \leq t | X = x)$ be the conditional absolute risk, $S(t; X = x) = 1 - F(t; X = x)$ the corresponding conditional survival function and set $f(t; X = x) = \partial F(t; X = x)/\partial t$.

For any threshold c, the true positive and false positive rates are time-dependent functions defined as $\mathrm{TPR}(c,t) = \mathrm{P}(X > c|D(t) = 1)$ and $\mathrm{FPR}(c,t) = \mathrm{P}(X > c|D(t) = 0)$. The time-dependent ROC curve $\mathrm{ROC}(t)$ plots $\mathrm{TPR}(c,t)$ vs $\mathrm{FPR}(c,t)$ for any threshold c, so that the time-dependent AUC evaluated at t_0 is given by

$$\mathrm{AUC}(t_0) = \int_{-\infty}^{\infty} \mathrm{TPR}(c,t_0) d\left[\mathrm{FPR}(c,t_0)\right], \tag{1}$$

where $d\left[\mathrm{FPR}(c,t_0)\right] = \partial c \times (\partial \mathrm{FPR}(c,t_0)/\partial c)$.

The setting of cumulative cases and dynamic controls corresponds to defining $D_i(t) = N_i^\star(t)$, where $N_i^\star(t) = I\{T_i \leq t\}$ is related to the counting process attached to T_i (Aalen et al., 2009). It follows that cumulative true positive rates and dynamic false positive rates are respectively defined as

$$\begin{aligned} \mathrm{TPR}^{\mathbb{C}}(c,t) &= \mathrm{P}(X > c|T \leq t) = \mathrm{P}(X > c|N^\star(t) = 1) & (2)\\ \mathrm{FPR}^{\mathbb{D}}(c,t) &= \mathrm{P}(X > c|T > t) = \mathrm{P}(X > c|N^\star(t) = 0). & (3) \end{aligned}$$

For survival outcomes, estimators of the cumulative true positive and dynamic false positive rates, and on turn of $\mathrm{AUC}(t)$, can not be directly derived from the above definitions because quantities $N_i^\star(t)$ involved in (2) and (3) are generally not fully observable due to drop-outs (*i.e.*, due to right censoring). To get round this issue, Bayes' theorem has to be used. When combined with (1), (2) and (3), $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t_0)$ can notably be shown to be

$$\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t_0) = \int_{-\infty}^{\infty}\int_{c}^{\infty} \frac{F(t_0; X = x)[1 - F(t_0; X = c)]}{[1 - F(t_0)]F(t_0)} g(x)g(c)dxdc. \tag{4}$$

From this equation, Chambless and Diao (2006) proposed, an estimator of $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$ partly based on a primary estimator of the conditional absolute risk $F(t_0; X = c)$ (see paragraph 2.3 below for more details). Equation (4) can further be refined to get a useful relation between $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$ and the time-dependent predictiveness curve introduced in the next paragraph.

Before that, and as a complement, we shall add that the setting of incident cases and dynamic controls corresponds to defining

$$D_i(t) = \begin{cases} 1 & \text{if } dN_i^\star(t) = 1;\\ 0 & \text{if } N_i^\star(t) = 0;\\ \texttt{NA} & \text{if } N_i^\star(t) = 0 \text{ and } dN_i^\star(t) = 0, \end{cases}$$

where $dN_i^\star(t) = N_i^\star(t) - N_i^\star(t-)$ refers to the increment of the counting process attached to T_i (Aalen et al., 2009). An analog of (4) for the incident/dynamic AUC evaluated at time t_0, $\mathrm{AUC}^{\mathbb{I},\mathbb{D}}(t_0)$, is then given by

$$\mathrm{AUC}^{\mathbb{I},\mathbb{D}}(t_0) = \int_{-\infty}^{\infty}\int_{c}^{\infty} \frac{f(t_0; X = x)[1 - F(t_0; X = c)]}{[1 - F(t_0)]f(t_0)} g(x)g(c)dxdc.$$

2.2 A new estimator for $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$

In this paragraph, we state a useful relation between the AUC and the predictiveness curve. This relation will enable us to derive estimators for $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$ relying on estimates of the conditional absolute risk function. For the sake of clarity, we first consider the binary outcomes setting for which the presentation is easier. The survival outcomes setting follows as a natural extension.

For a binary outcome D, let $R(q) = P\left[D = 1|X = G^{-1}(q)\right]$ denote the conditional absolute risk associated to the q-th quantile $G^{-1}(q)$ of marker X. The predictiveness curve plots $R(q)$ versus q and describes the distribution of $\mathrm{P}(D = 1|X)$ (Huang et al., 2007; Pepe et al., 2008a). It can be shown that

$$\mathrm{AUC} = \frac{\int_0^1 qR(q)dq - p^2/2}{p(1-p)}, \tag{5}$$

where we set $p = \mathrm{P}(D = 1) = \int_0^1 R(q)dq$. The proof of Equation (5) follows from arguments similar to those used to establish Equation (6) below (see the Appendix) and is then omitted. Equation (5) states that the AUC is a simple function of the predictiveness curve R and completes the results of Gu and Pepe (2009). It is easy to check that AUC=1/2 if $R(q) = p$ while AUC=1 if $R(q) = I(q \geq 1 - p)$. Figure 1 shows some examples of predictiveness curves and the corresponding AUC values in the case where $p = 1/2$: the closer the predictiveness curve from the step function $I(q \geq 1 - p)$, the higher the value of the corresponding AUC.

It is noteworthy that Equation (5) suggests that accurate estimates of $R(q)$ (especially for high values of q) should yield accurate estimates of AUC.

We now turn our attention to the survival outcomes setting. Set $R(t; q) = \mathrm{P}(D(t) = 1|X = G^{-1}(q)) = F(t|X = G^{-1}(q))$ (since we recall that $D_i(t) = I(T_i \leq t)$ in the cumulative/dynamic setting). For any t, the function $R(t; \cdot)$ is a natural extension of the predictiveness curve defined above. Accordingly, it will be referred to as the time-dependent predictiveness curve at time t. In this context, we have

$$\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t_0) = \frac{\int_0^1 qR(t_0; q)dq - F^2(t_0)/2}{F(t_0)[1 - F(t_0)]}. \tag{6}$$

The proof of (6) is deferred to the Appendix. Observing that $F(t_0) = \int_0^1 R(t_0; q)dq$, it follows from (6) that $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t_0)$ is a simple function of the time-dependent predictiveness curve.

The relation stated in Equation (6) will be the basis of our estimating procedure. We now give more details about the estimation of each component appearing in the right hand-side of (6). Assume we are given an estimator $\widehat{F}_n(t_0; x)$ of the conditional absolute risk $F(t_0; x)$ and recall that G and g denote the cumulative distribution function and the density function of X. Since $\int_0^1 qR(t_0; q)dq = \int_{-\infty}^{\infty} G(x)F(t_0; x)g(x)dx$, the empirical counterpart of the quantity $\int_0^1 qR(t_0; q)dq$ is given by

$$\sum_{i=1}^{n} \frac{i}{n} \widehat{F}_n(t_0; X_{(i)}),$$

where $X_{(i)}$ denotes the i-th order statistic attached to the sample $X_1, ..., X_n$. As for the marginal absolute risk function F, it can be directly estimated using Kaplan-Meier estimator $\widehat{F}_{n,(1)}(t_0)$. Observing that $F(t_0) = \int F(t_0; x)g(x)dx$, an alternative to $\widehat{F}_{n,(1)}(t)$ relying on the conditional risk estimate is the quantity

$$\widehat{F}_{n,(2)}(t_0) = \sum_{i=1}^{n} \widehat{F}_n(t_0; X_i).$$

This yields two estimators for $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t_0)$, namely, for $k = 1, 2$,

$$\mathrm{AUC}^{\mathbb{C},\mathbb{D}}_{n,(k)}(t_0) = \frac{\sum_{i=1}^{n} \frac{i}{n} \widehat{F}_n(t_0; X_{(i)}) - \widehat{F}^2_{n,(k)}(t_0)/2}{\widehat{F}_{n,(k)}(t_0)\big[1 - \widehat{F}_{n,(k)}(t_0)\big]}. \tag{7}$$

For illustration, we now will work under standard survival models and recall how to obtain estimates $\widehat{F}_n(t_0; x)$ of the conditional absolute risk which, when combined with Equation (7) above, will lead to estimates for $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t_0)$.

Denote by $\lambda(t; X = x)$ the conditional hazard rate and $\Lambda(t; X = x) = \int_0^t \lambda(u; X = x)du$ the conditional cumulative hazard rate of T given X. The conditional absolute risk function expresses as $F(t; X = x) = 1 - \exp\big\{-\Lambda(t; X = x)\big\}$. First consider the Cox model (Cox, 1972) under which the conditional hazard rate $\lambda(t; X = x)$ is of the form $\lambda_0(t) \exp(\alpha_0 + \alpha x)$, where λ_0 denotes the baseline hazard rate, α_0 is an intercept and α is the log hazard ratio pertaining to X. Denote by $\hat{\Lambda}_0(t)$, $\hat{\alpha}_0$ and $\hat{\alpha}$ the estimators of the quantities $\int_0^t \lambda_0(u)du$, α_0 and α respectively. Then a estimator of $F(t; X = x)$ is given by

$$\widehat{F}_{n,\mathrm{Cox}}(t_0; X = x) = 1 - \exp\Big\{ - \hat{\Lambda}_0(t_0) \exp(\hat{\alpha}_0 + \hat{\alpha} x)\Big\}. \tag{8}$$

Next, under Aalen's additive model (Aalen, 1989), the conditional hazard rate $\lambda(t; X = x)$ is of the form $\beta_0(t) + \beta_1(t)x$. Thus $F(t; x) = 1 - \exp\{-B_0(t) - B_1(t)x\}$ with $B_i(t) = \int_0^t \beta_i(u)du, i = 0, 1$. Given estimates $\widehat{B}_0(t)$ and $\widehat{B}_1(t)$ of $B_0(t)$ and $B_1(t)$ respectively, we can then define

$$\widehat{F}_{n,\text{Aalen}}(t_0; X = x) = 1 - \exp\Big\{ - \widehat{B}_0(t_0) - \widehat{B}_1(t_0)x\Big\}. \tag{9}$$

Checking that the assumptions of any statistical model are not violated is often tricky, especially when the sample size is small (and the statistical power is low). Consequently, it is often legitimate to use nonparametric estimators, at least for comparison matters. Several local or conditional versions of the Kaplan-Meier estimator have been proposed and studied in the literature (Beran, 1981; Akritas, 1994). Any of them can be used to estimate the conditional absolute risk for survival outcomes. For instance, denoting the empirical distribution function of X by $\widehat{G}$, a nearest-neighbor type-estimator of $F(t_0; x)$ is defined as

$$\widehat{F}_{n,\text{KMcond}}(t_0; X = x) = 1 - \prod_{Z_i \leq t_0, \delta_i = 1} \left\{ 1 - \frac{K_{\ell_n}(X_i, x)}{\sum_j I(Z_j \geq Z_i -) K_{\ell_n}(X_j, x)} \right\}, \tag{10}$$

where ℓ_n is the smoothing parameter of the 0/1 symmetric nearest-neighbor kernel K_{ℓ_n}, *e.g.*, $K_{\ell_n}(x, y) = I(|\widehat{G}(x) - \widehat{G}(y)| < \ell_n)$ (Akritas, 1994). The smoothing parameter ℓ_n needs to be carefully chosen to ensure good balance between bias and variance. Data-driven rules are generally employed for selecting ℓ_n in an optimal manner .

It is noteworthy that, in most applications, several evaluation times may be of interest: for instance, in the setting of cancer risk prediction, t_0 might be set to $1, 5$ and 10 years. Should $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ be estimated for several t's in the situation where $X = X(t)$, the estimation procedure of the conditional absolute risk have in general to be performed for every value of t (while, of course, a single estimation is needed in the context of a purely constant marker or score X).

2.3 Existing estimators for $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$

Following the proposal of Heagerty et al. (2000), several estimators of $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ have been presented in the literature. We briefly recall their principles in this paragraph.

Heagerty et al. (2000) developed a nonparametric estimator for $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ based on the nearest-neighbor bivariate distribution estimator of Akritas (1994). Rewriting sensitivity $\text{P}(X > c|D(t) = 1) = F(t|X > c)\text{P}(X > c)/F(t)$ and specificity $\text{P}(X \leq c|D(t) = 0) = S(t|X \leq c)\text{P}(X \leq c)/\{1 - F(t)\}$, the authors first observed that "naive" estimators of sensitivity and specifity obtained by pluging in the Kaplan-Meier estimator for S and $\hat{G}(c) = \sum I(X_i \leq c)/n$ for $\text{P}(X \leq c)$ may not be monotone in c. Proper estimates follow from first expressing sensitivity and specificity as functions of the bivariate survival function $S(c, t) = \text{P}(X > c, T > t)$, that is

$$\text{P}(X > c|D(t) = 1) = \frac{1 - G(c) - S(c, t)}{F(t)} \quad \text{and} \quad \text{P}(X \leq c|D(t) = 0) = 1 - \frac{S(c, t)}{1 - F(t)}.$$

Heagerty et al. (2000) then proposed to use the Nearest Neighbor Estimator of Akritas (1994) of $S(c, t)$, which relies on the conditional representation of the bivariate survival function $S(c, t) = \int_c^\infty S(t|X = s)dG_X(s)$. That is, the NNE estimator is given by $\widehat{S_{\ell_n}}(c, t) = \sum\{1 - \widehat{F}_{n,\text{KMcond}}(t_0; X = X_i)\} \times I(X_i > c)/n$ with $\widehat{F}_{n,\text{KMcond}}(t_0; X = x)$ as in (10). Given these proper estimators for sensitivity and specificity, an estimator of $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ follows from (1) by simple numerical integration (using the trapezoidal rule for instance). This method will be referred to as HLP.

Chambless and Diao (2006) suggested a recursive calculation over the ordered times of events for $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$, analogous in spirit to Kaplan Meier approach for the estimation of the survival function.

Given two random individuals i and j, it can be shown that $\text{AUC}^{\mathbb{C},\mathbb{D}}(t) = \text{P}(X_i > X_j | D_i(t) = 1, D_j(t) = 0)$, with $D_i(t) = N_i^\star(t)$. Then, applying Bayes' theorem leads to the expression

$$\text{AUC}^{\mathbb{C},\mathbb{D}}(t) = \frac{\text{P}(X_i > X_j, D_i(t) = 1, D_j(t) = 0)}{\text{P}(D_i(t) = 1)\text{P}(D_j(t) = 0)}.$$

Further let t_k be the unique ordered survival times $t_1 < t_2 < \ldots < t_n$. At a given time t_m with $1 \leq m < n$, the numerator of $\text{AUC}^{\mathbb{C},\mathbb{D}}(t_m)$ expresses as

$$\sum_{k \leq m} w_k^{(1)} \lambda(t_k)\{1 - \lambda(t_k)\}S(t_{k-1})^2 - \sum w_k^{(2)} \lambda(t_k) S(t_{k-1})\{1 - S(t_{k-1})\},$$

the denominator being $S(t_m)(1 - S(t_m)$. Chambless and Diao (2006) establish that the weights

$$\begin{aligned} w_k^{(1)} &= \text{P}(X_i > X_j | D_i(t_k) = 1, D_j(t_{k-1}) = 0, D_j(t_k) = 0) \\ w_k^{(2)} &= \text{P}(D_i(t_{k-1}) = 1, D_j(t_{k-1}) = 0, D_j(t_k) = 1) \end{aligned}$$

can then be estimated recursively. This method will be referred to hereafter as CD1. A nice property of this nonparametric estimator is that it does not involve any smoothing parameter, unlike the one proposed by Heagerty et al. (2000) or the one we propose using a conditional Kaplan-Meier estimator for the conditional risk function.

Another estimator, based on the estimation of the conditional survival function, was derived by Chambless and Diao (2006). From Equation (4) above, the authors observe that

$$\text{AUC}^{\mathbb{C},\mathbb{D}}(t_0) = \frac{\text{E}\big[\{1 - S(t;U)\}S(t;V)I(V < U)\big]}{\text{E}\{1 - S(t;X)\}\text{E}\{S(t;X)\}}, \tag{11}$$

where U and V are independent observations of X. They then suggest to estimate the conditional survival functions under a Cox model, while the bivariate expectation can be estimated as the mean over all (U, V) pairs of distinct observations. The corresponding method, which will be referred to as CD2, is very similar to our approach, when using a Cox model to estimate conditional risks. However, our approach does not involve the computation of the bivariate expectation: because the function L defined in the appendix is symmetric in its arguments, we were able to rewrite (11) in the more appealing form (6) which does not involve any bivariate expectation.

3 Simulation study

In this section, we present results we obtained from an empirical study, the main objectives of which being (i) to compare our estimators of $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ with those proposed in the literature and (ii) to assess the effect of a misspecified model – when estimating the conditional absolute risk– on the $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ estimation. Towards this end, we compared six estimators of $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ on synthetic data : Heagerty's estimator relying on Akritas' estimator for the bivariate survival function (method HLP which is implemented in the `survivalROC R` package; Heagerty et al. (2000)), the two estimators proposed by Chambless and Diao (2006) (CD1 and CD2, which are both implemented in a `SAS` macro available at `http://www.cscc.unc.edu/aric/addresms/`), and three estimates we derived by combining equation (7) with three distinct estimates of the conditional absolute risk. More precisely, the conditional absolute risk was estimated under a standard Cox proportional hazard model (see (8)), an Aalen additive model (see (9)) and using the conditional Kaplan-Meier estimator (see (10)). From our experimental results (not shown), we observed better performances for estimates obtained with $k = 2$ in (7) and we therefore only present results obtained with $k = 2$ in the sequel. The three corresponding methods will be referred to as VL Cox, VL Aalen and VL KM. Of note, we used the `basehaz` function of the `survival` package to obtain estimates for the baseline hazard rate under a Cox model. To estimate

$F(t;x)$ under Aalen's additive risk model, we used the function `aalen` of the `timereg` `R` package which returns estimates for both cumulative coefficients B_0 and B. Lastly, the conditional Kaplan-Meier estimator was computed using the `prodlim` package. Regarding the choice of the smoothing parameter for method VL KM, we used the default option of the `prodlim` package which employs a direct plug-in method. As for the smoothing parameter for method HLP, it was set to $0.25 \times n^{-1/5}$, following the guidelines of the `survivalROC` package.

3.1 Simulation design

For generating a random survival time variable with cumulative hazard rate Λ, it suffices to invert an exponential random variate. More precisely, given an exponential random variate E, the survival time $T = \Lambda^{-1}(E)$ is a random variable with cumulative hazard rate Λ (Leemis et al., 1990). In this simulation study, we chose to consider three distinct conditional distributions for the survival time T given marker X, that is three distinct cumulative hazard rates $\Lambda_1(\cdot;X), \Lambda_2(\cdot;X)$ and $\Lambda_3(\cdot;X)$. Denoting by $\lambda_1(\cdot;X)$, $\lambda_2(\cdot;X)$ and $\lambda_3(\cdot;X)$ the corresponding hazard rates, we considered

$$\lambda_1(t;X) = \frac{\exp(\beta X)}{1+t}, \quad \text{for some } \beta \in \mathbb{R}; \tag{12}$$

$$\lambda_2(t;X) = t\exp\left(\frac{\beta X t^2}{2}\right), \quad \text{for some } \beta \in \mathbb{R}; \tag{13}$$

$$\lambda_3(t;X) = \beta_0 t + \frac{\beta}{t+1}X, \quad \text{for some } \beta_0, \beta \in \mathbb{R}. \tag{14}$$

In (12), the model corresponds to a standard Cox model with a decreasing baseline hazard rate $\lambda_0(t) = 1/(t+1)$. Equation (13) describes a time-varying coefficient Cox model with time-varying coefficient $\beta(t) = (\beta X t^2)/2$ and increasing baseline hazard rate $\lambda_0(t) = t$. Finally, the last case is that of an Aalen's additive model with time varying coefficients $\beta_0 t$ and $\beta/(t+1)$. The conditional absolute risk functions for models (12) and (13), and (14) are respectively given by $F_1(t;X) = 1-(t+1)^{-\exp(\beta X)}$, $F_2(t;X) = 1-\exp[-\{\exp(\beta X t^2/2)-1\}/(\beta X_i)]$ and $F_3(t;X) = 1-\exp\{-\beta_0 t^2/2 + (\beta\log(t+1))X\}$. Moreover, the inverse functions of Λ_1 and Λ_2 can be derived analytically and are given by

$$\Lambda_1^{-1}(t;X) = \exp\{t\exp(-\beta X)\} - 1;$$

$$\Lambda_2^{-1}(t;X) = \sqrt{\frac{2\log(\beta X t)}{\beta X}}.$$

Observe that $\Lambda_2^{-1}(\cdot;X)$ is only defined for positive X. In this case, values of X were drawn from an exponential distribution, while X was generated according to a $\mathcal{N}(0,1)$ distribution under model (12). As for model (14), the `uniroot` `R` function was used to solve equations $\Lambda_3(T_i;X_i) = E_i$, for every observation $i = 1,...,n$ ($X_i, i = 1,...,n$, was drawn from a exponential distribution again).

Under each of the three aforementioned models, we applied an "administrative censoring" occurring at the time corresponding to the 80% percentile of the survival time distribution. Besides this administrative censoring, we considered three censoring schemes: (i) no additional censoring, (ii) $C_i \sim \mathcal{E}(\tau_1)$ and (iii) $C_i \sim \mathcal{E}(\tau_2)$, where rates τ_1 and τ_2 of the exponential distribution $\mathcal{E}(\cdot)$ were respectively chosen so that censoring rate attained 25% and 75% respectively.

3.2 Comparison of the $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ estimators

Sample size was set to 500. Estimates for $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ were computed at times t_{q1}, t_{q2} and t_{q3} corresponding to the first, second and third quartile of the survival time distribution respectively. Theoretical values of $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ were computed according to Equation (6). The `integrate` `R` function was used to compute terms of the form $\int_0^1 qR(t;q)dq = \int_0^1 qF(t;X = G^{-1}(q))dq$ (`qnorm` and `qexp` `R`

functions were used to compute quantities $G^{-1}(q)$) and $F(t; X = x)$ was computed according to (8), (9) or (10).

Average bias and mean squared-error (MSE) were computed for each method and under each design over 100 runs (see Table 1). Overall, higher censoring rates lead to poorer estimates of $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$, especially for late evaluation times (*i.e.*, when censoring is the most sensitive). Methods CD2 and VL Cox – both relying on a primary Cox estimate of the conditional risk function – achieved similar performances in most examples. Their performances highly depended on the true underlying model. They both performed the best under a Cox model and performed the worst under each of the other considered models. This highlights that misspecifying the model for conditional risk estimations has an important effect on the $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ estimation accuracy. In the same spirit, the three nonparametric methods (CD1, HLP and VL KM) achieved similar performances in most examples. As expected, they were less sensitive to the underlying model than methods CD2 and VL Cox. Method CD1 slightly outperformed its competitors for low censoring rates and for early evaluation times. Two main reasons may be put forward to explain this result. First, CD1 does not rely on any smoothing parameter, and, maybe more importantly, CD1 directly estimates $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$, while alternative methods use plug-in estimations of the conditional risk function. However, we observed chaotic performances for method CD1 when increasing censoring rates, especially for late evaluation times. This suggests that this method is inadvisable in these situations. On the contrary, VL KM appeared to perform the best (compared to CD1 and HLP) in the situations of high censoring rates and late evaluation times. Lastly, results attached to method VL Aalen were rather surprising: in terms of MSE, VL Aalen outperformed VL KM under a Cox model, but was worse under an Aalen's additive model (it was worse under a time-varying coefficients Cox model too but this was somehow expected). This suggests that estimates of $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ derived under VL Aalen may be less biased but, even if the underlying model is an Aalen's additive model, generally present larger variances than those derived under VL KM.

3.3 Assessing the accuracy of $\text{AUC}^{\mathbb{C},\mathbb{D}}$ estimates using predictiveness curves

As mentioned above, Equation (6) shows that accurate estimates of $R(t_0; q)$ (especially for high values of q) should yield accurate estimates for $\text{AUC}^{\mathbb{C},\mathbb{D}}(t_0)$. We then compared accuracies of the predictiveness curve and $\text{AUC}^{\mathbb{C},\mathbb{D}}(t_0)$ estimations on one sample generated under each of the three simulation designs described above. Two evaluation times were considered: the first quartile t_{q1} and the median t_{q2} of the survival time distribution. The three same models as above were considered to estimate the conditional absolute risk (and then $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$). Figure 2 shows the corresponding graphs, as well as the curve $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ and its estimate. Overall, these graphs confirm that if the predictiveness curve (or, equivalently, the conditional risk function) is accurately estimated, then $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ is accurately estimated too. They also confirm that errors made when estimating $R(t, q)$ for high values of q are predominant: for instance, under the time-varying coefficient Cox model, when a Cox model is used to estimate the conditional absolute risk function, the true predictiveness curve is underestimated on the quantiles interval $[0, 0.85]$ and slightly overestimated on the interval [0.85,1] while $\text{AUC}^{\mathbb{C},\mathbb{D}}(t_1)$ is largely overestimated.

A natural question then arises: how to check the accuracy of the time-dependent predictiveness curve estimation? In the binary outcomes setting Pepe et al. (2008a) showed that this problem was related to that of checking the goodness-of-fit of the underlying risk model. In a sense, the predictiveness curve can indeed be viewed as a graphical representation of each component of the Hosmer-Lemeshow statistic (Lemeshow and Hosmer, 1982): at the midpoint of each decile of predicted risk, the observed proportion of cases can be superimposed on and visually compared to the predictiveness curve. To our knowledge, the Hosmer-Lemeshow statistic has never been extended to survival outcomes. However, as mentioned in Viallon et al. (2009), this statistic can be computed with observed counts replaced by their estimates based on Kaplan-Meier estimator. Some more study would be needed to assess the asymptotical distribution of the resulting statistic under the null hypothesis in order to derive a proper statistical test. However, we can develop on this idea to propose a graphical tool, useful for

checking the goodness-of-fit of the risk model and then the accuracy of the $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$ estimation. The principle for a fixed evaluation time t is simple. Namely, (i) compute Kaplan-Meier estimator of the (unconditional) absolute risk for each decile of predicted risk and (ii) superimpose these values on the graph of the predictiveness curve (see Figure 2 for illustration).

Generally (and as can be seen on Figure 2), nonparametric estimators fit data well, and checking the goodness-of-fit is mostly relevant for parametric models. For such models, an alternative to the aforementioned extension of the Hosmer-Lemeshow statistic would be to compare the predictiveness curve obtained under the considered parametric (or semi-parametric) model to the one obtained with a nonparametric estimator of the conditional absolute risk (*e.g.*, using a conditional Kaplan-Meier estimator). The corresponding test would consist of an extension of the test proposed by Härdle and Mammen (1993). The comparison can also be visual (see Figure 2). For instance, under the time-varying coefficient Cox model, when a Cox model is used to estimate the conditional absolute risk function, the predictiveness curve is quite different from the one derived from a conditional Kaplan-Meier estimator of the conditional risk function for $t = t_{q1}$. The predictiveness curves estimates are much closer for $t = t_{q2}$ though. As a result, the estimator of $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t_{q2})$ obtained using a Cox estimator of the conditional risk function is much better than the estimator of $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t_{q1})$.

The complete study of either methods is beyond the scope of this paper and will be performed elsewhere. Of course, more standard statistical tests might also be advocated, such as those relying on Cox-Snell residuals (Cox and Snell, 1968).

4 Real examples

In this Section, we present results we obtained on two classical real data sets: the VA lung cancer data and the Mayo PBD data. These two data sets are freely available through the `MASS` and `survivalROC R` packages respectively. These data were especially used in Heagerty and Zheng (2005).

4.1 Mayo PBC data

As a first illustration, we evaluated the $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$ of a widely used score for predicting survival after a primary biliary cirrhosis (PBC). The data set consists of 312 subjects, enrolled at the Mayo Clinic between 1974 and 1984, and randomized in a placebo controlled trial of the drug D-penicillamine (Fleming and Harrington, 1991). About 40% of the patients died during the study. The score was constructed by including 5 covariates – log(bilirubin), albumin, log(prothrombin time), edema and age (see Heagerty and Zheng (2005))– into a Cox model. Our aim here was to evaluate the predictive performances of this score as well as checking whether its accuracy changes over time. We then computed estimates of $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$ according to to the method of Heagerty et al. (2000) (HLP) and our approach (see the right panel in Figure 3). The two left panels in Figure 3 present the predictiveness curves estimates obtained with t set to the first quartile and 35% percentile ($t_{q1} \approx 1481$ and $t_{p35} \approx 2365$) of the survival time respectively. For $t = t_{p35}$, the predictiveness curve obtained with a Cox model is very similar to that obtained with a conditional Kaplan-Meier estimator. This is reflected on the curves of the estimated $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$: estimates of $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t_{p35})$ obtained with either estimator of the conditional absolute risk are similar. For $t = t_{q1}$, the difference between the predictiveness curve obtained with a Cox model and that obtained with a conditional Kaplan-Meier estimator is bigger, and such is the difference between the estimates of $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t_{p35})$. In other respect, we observed that estimates of $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$ obtained using VL KM and HLP were close to each other. According to either methods, the effect of time on $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$ is moderate, but $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$ appeared to slightly decrease with time.

4.2 VA Lung Cancer Data

The VA lung cancer data set was presented and analyzed in Kalbfleisch and Prentice (2002) for instance. Overall, 137 males with inoperable cancer were randomized to a standard or a test chemotherapy. Death was considered as the endpoint, and more than 93% of the participants died during the study. Predictors of mortality include type of treatment, age, histological type of tumor and the Karnofsky score (which is a performance status measure). As in Heagerty and Zheng (2005), we considered a 500-day follow-up and a Cox model was used to build a risk score out of these baseline covariates. Our objective here was to estimate the $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ attached to this score. As above, we computed estimates of $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ according to the method of Heagerty et al. (2000) (HLP) and our approach (see the right panel in Figure 4). The two left panels in Figure 4 present the predictiveness curves obtained with t set to the first and third estimated quartiles (t_{q1} and t_{q3}) of the survival time respectively. For both $t = t_{q1}$ and $t = t_{q3}$, the predictiveness curve obtained with a Cox model is quite different from that obtained with the conditional Kaplan-Meier estimator and this is once again reflected on the curves of the estimated $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$, where the estimation relying on the Cox model is sensibly different from that obtained via the conditional Kaplan-Meier estimator. In other respect, we still observed good agreement between estimators obtained using either HLP or VL KM. As for the Mayo score, $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ for the VA lung score appeared to be a slightly decreasing function of the evaluation time t.

5 Discussion

In this paper, we derived a useful relation between the predictiveness curve and the AUC and showed that this relation still holds when considering extensions of these concepts to survival outcomes. This relation enabled us to propose new estimators for the cumulative/dynamic AUC, relying on primary estimators of the conditional absolute risk function. These estimators are similar, in spirit, to one of the two estimators formerly proposed by Chambless and Diao (2006). Through an empirical study, we further showed that our estimation procedure attained performances similar to that reached by existing estimates. This simulation study also highlighted that much attention had to be paid when selecting the form of the model used to estimate the conditional risk function. Working under an appropriate parametric model usually yields more accurate estimates (for both the conditional risk function and $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$) than those obtained from purely nonparametric approaches, but misspecifying the model generally leads to dramatically biased estimates. This observation leads us to recommend to always use nonparametric estimators of the conditional absolute risk function at least to visually check the goodness-of-fit of parametric models, for instance by comparing estimates of the predictiveness curve.

It is noteworthy that the proposed estimators of $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ are straightforward to implement: standard survival packages indeed return estimates of the conditional absolute risk function from which estimates of $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ are readily obtained in view of Equation (7). Moreover, because of their "plug-in" nature, their theoretical properties should follow from those established for estimators of the conditional absolute risk function. Closed form expressions might further be obtained for confidence intervals, but sub-sampling techniques (bootstrap for instance) can already be used to provide such intervals.

We shall also recall that the nonparametric estimator of Chambless and Diao (2006) was observed to slightly outperform its two nonparametric competitors (including our approach) in most of our empirical examples, except for high censoring rates and late evaluation times (where our approach appeared to perform the best). This might be due to the fact that the nonparametric proposal of Chambless and Diao (2006) directly estimates $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ instead of using plug-in estimates. Some theoretical study would be needed to confirm this observation. We may also recall here that most estimators of the conditional absolute risk function rely on some independence assumption between censoring time C and the pair (T, X), and so does our proposal for estimating $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$. Because it is based on the estimation of the bivariate survival function, the nearest-neighbor estimator of Heagerty et al. (2000) does not rely on any such assumption, and this method is then advisable in situations where these assumptions might be violated.

Another important remark is that if X actually stands for a risk score, the methodology presented here only applies for external validation. That is, we supposed that the risk score had been constructed on some sample and our goal was to evaluate its discriminative power on an external sample. Should the score be evaluated on the sample used to construct it, standard sub-sampling methods are further needed (Harrell et al., 1996).

In other respect, the results we obtained for (time-dependent) AUC are easy to extend to (time-dependent) partial AUC, p-AUC, which has recently gained popularity in epidemiology (Dodd and Pepe, 2003). The definition of p-AUC is similar to (1), with the interval of integration restricted to $(c_{\min}, c_{\max})$, for some $c_{\min} \geq -\infty$ and $c_{\max} \leq \infty$. In the binary outcome setting, an analog of Equation (5) is then obtained by replacing the integration interval $(0,1)$ by $(G(c_{\min}), G(c_{\max}))$. The survival outcome setting can be handled proceeding to the same replacement, and derivations of empirical counterparts are also straightforward. Therefore, it follows from our results that (time-dependent) partial AUC is also directly related to the predictiveness curve.

To conclude, this paper completes in some sense the work of Gu and Pepe (2009), confirming that the conditional risk function, through the predictiveness curve, is the key when assessing discrimination of prognostic tools.

Appendix

Proof of (6). Considering the numerator of (4), and using the changes of variables $x = G^{-1}(u)$ and $c = G^{-1}(v)$, we have

$$\begin{aligned}
& \int_{-\infty}^{\infty}\int_{c}^{\infty} F(t;X=x)[1-F(t;X=c)]g(x)g(c)dxdc \\
= & \int_0^1\int_v^1 F(t;X=G^{-1}(u))[1-F(t;X=G^{-1}(v))]dudv \\
= & \int_0^1\int_v^1 [1-S(t;X=G^{-1}(u))]S(t;X=G^{-1}(v))dudv \\
= & \int_0^1\int_v^1 [S(t;X=G^{-1}(v))-S(t;X=G^{-1}(u))S(t;X=G^{-1}(v))]dudv \\
= & \int_0^1 (1-v)S(t;X=G^{-1}(v))dv - \int_0^1\int_v^1 S(t;X=G^{-1}(u))S(t;X=G^{-1}(v))dudv \\
= & \int_0^1 (1-v)S(t;X=G^{-1}(v))dv - \int_0^1\int_0^1 S(t;X=G^{-1}(u))S(t;X=G^{-1}(v))I(u\geq v)dudv.
\end{aligned}$$

Setting

$$L(u,v) = S(t;X=G^{-1}(u))S(t;X=G^{-1}(v))I(u\geq v),$$

we have $L(u,v) = L(v,u)$ so that

$$\begin{aligned}
& \int_{-\infty}^{\infty}\int_{c}^{\infty} F(t;X=x)[1-F(t;X=c)]g(x)g(c)dxdc \\
= & \int_0^1 (1-v)S(t;X=G^{-1}(v))dv - \frac{1}{2}\int_0^1\int_0^1 S(t;X=G^{-1}(u))S(t;X=G^{-1}(v))dudv \\
= & \int_0^1 (1-v)S(t;X=G^{-1}(v))dv - \frac{1}{2}\left(\int_0^1 S(t;X=v)dv\right)^2.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t) &= \frac{\int_0^1 (1-v) S(t; X = G^{-1}(v)) dv - \frac{[1-F(t)]^2}{2}}{F(t)[1-F(t)]} \\
&= \frac{1 - F(t) - \int_0^1 v[1 - F(t; X = G^{-1}(v))] dv - \frac{[1-F(t)]^2}{2}}{F(t)[1-F(t)]} \\
&= \frac{1 - F(t) - \frac{[1-F(t)]^2}{2} - 1/2 + \int_0^1 v F(t; X = G^{-1}(v)) dv}{F(t)[1-F(t)]} \\
&= \frac{\int_0^1 cR(t;c) dc - \frac{F(t)^2}{2}}{F(t)[1-F(t)]},
\end{aligned}$$

which is (6).

References

Aalen, O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, 8:907–925.

Aalen, O., Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (2009). History of applications of martingales in survival analysis. *Electronic Journal for History of Probability and Statistics*, 5(1).

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

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, University of California, Berkley.

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

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 34:187–220.

Cox, D. R. and Snell, E. (1968). A general definition of residuals. *Journal of the Royal Statistical Society, Series B*, 30:248–275.

Dodd, L. E. and Pepe, M. S. (2003). Partial AUC estimation and regression. *Biometrics*, 59:614–623.

Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and survival analysis*. Wiley, New York.

Gail, G. H. and Pfeiffer, R. M. (2005). On criteria for evaluating models of absolute risk. *Biostatistics*, 6:227–239.

Gönen, M. and Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965–970.

Gu, W. and Pepe, M. S. (2009). Measures to summarize and compare the predictive capacity of markers. *International Journal of Biostatistics*, 5:Art. 27, 49.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, 21:1926–1947.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247:2543–2546.

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387.

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

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

Huang, Y., Pepe, M. S., and Feng, Z. (2007). Evaluating the predictiveness of a continuous marker. *Biometrics*, 63:1181–1188.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical analysis of failure time data*. Wiley, Hoboken.

Leemis, L. M., Shih, L.-H., and Reynertson, K. (1990). Variate generation for accelerated life and proportional hazards models with time dependent covariates. *Statistics & Probability Letters*, 10(4):335 – 339.

Lemeshow, S. and Hosmer, D. W. (1982). A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology*, 115:92–106.

Pencina, M. J., Larson, M. G., and D'Agostino, R. B. (2007). Evaluating the added predictive ability of a new marker:from area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, 27:157–172.

Pepe, M. S., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I. M., and Zheng, Y. (2008a). Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology*, 167:362–368.

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

Schemper, M. (2003). Predictive accuracy and explained variation. *Statistics in Medicine*, 22:2299–2308.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, 21:128–138.

Viallon, V., Ragusa, S., Clavel-Chapelon, F., and Bénichou, J. (2009). How to evaluate the calibration of a disease risk prediction tool. *Statistics in Medicine*, 28:901–916.

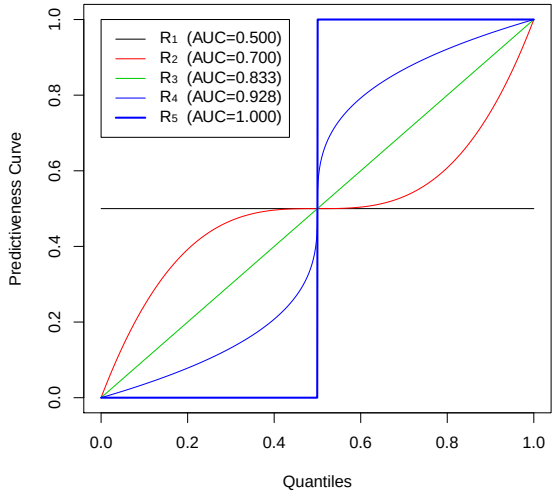

Figure 1: Generic predictiveness curves and their corresponding AUC values. A flat predictiveness curve, $R(q) = p$, where p is the proportion of events, is associated to an AUC of 0.5. The closer the predictiveness curve is from the step function $I(q \geq 1 - p)$, the closer the corresponding AUC is to 1.

Table 1: Results of the simulation study. Comparisons between several estimators of $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$. Averaged bias (multiplied by 100) and MSE (multiplied by 1000) obtained from 100 runs are reported.

Eval.	100× **Bias**						1000× **MSE**					
Time	CD2	VL Cox	VL Aalen	CD1	Heag	VL KM	CD2	VL Cox	VL Aalen	CD1	Heag	VL KM
	Standard Cox model											
	Censoring scheme 1											
t_{q1}	-0.388	-0.168	-0.361	0.131	-0.638	-1.052	0.203	0.190	0.356	0.570	0.623	0.683
t_{q2}	-0.358	-0.082	0.301	-0.239	-0.930	-1.262	0.189	0.176	0.317	0.465	0.565	0.640
t_{q3}	-0.359	0.103	1.485	-0.598	-1.319	-1.413	0.168	0.155	0.533	0.416	0.557	0.584
	Censoring scheme 2											
t_{q1}	-0.103	0.117	-0.288	0.031	-1.191	-1.111	0.263	0.262	0.395	0.631	0.798	0.723
t_{q2}	-0.104	0.170	0.423	-0.159	-1.304	-1.115	0.266	0.267	0.308	0.427	0.700	0.578
t_{q3}	-0.042	0.415	2.132	-0.280	-0.853	-0.774	0.252	0.269	0.910	0.730	0.735	0.645
	Censoring scheme 3											
t_{q1}	0.167	0.384	-0.459	-0.166	-1.609	-1.361	0.626	0.636	0.614	0.977	1.277	1.125
t_{q2}	0.128	0.367	-0.134	-1.023	-2.427	-1.571	0.636	0.645	1.403	2.401	2.833	2.573
t_{q3}	-2.434	-1.963	-1.533	-4.471	-7.270	-5.485	1.450	1.331	7.333	13.981	8.949	8.327
	Time-varying Cox model											
	Censoring scheme 1											
t_{q1}	6.686	6.906	2.882	0.163	-0.468	-0.731	4.651	4.949	2.291	1.159	1.184	1.174
t_{q2}	-2.388	-2.107	6.199	0.274	-0.197	-0.556	0.714	0.587	4.510	0.524	0.534	0.571
t_{q3}	-9.126	-8.629	7.377	-0.047	-0.417	-0.317	8.464	7.582	5.812	0.338	0.347	0.335
	Censoring scheme 2											
t_{q1}	5.716	5.927	2.528	-0.229	-2.457	-1.196	3.564	3.803	2.042	1.272	1.752	1.365
t_{q2}	-3.272	-3.004	5.867	0.071	-2.828	-0.881	1.310	1.135	4.268	0.832	1.683	0.909
t_{q3}	-10.005	-9.535	7.536	0.492	-1.343	-0.057	10.232	9.304	6.190	0.587	0.741	0.487
	Censoring scheme 3											
t_{q1}	3.044	3.268	2.152	0.042	-2.198	-0.845	1.307	1.448	2.616	2.156	2.775	2.198
t_{q2}	-5.532	-5.250	5.327	0.746	-2.799	-0.592	3.372	3.067	4.459	1.407	2.414	1.348
t_{q3}	-11.932	-11.460	6.576	1.724	-2.637	-0.200	14.562	13.462	6.399	4.178	2.912	1.896
	Aalen additive model											
	Censoring scheme 1											
t_{q1}	-7.898	-7.674	0.603	0.496	-0.190	-0.554	6.558	6.209	1.113	0.582	0.571	0.590
t_{q2}	-5.245	-4.955	0.248	-0.015	-0.551	-0.779	3.020	2.724	0.625	0.464	0.489	0.516
t_{q3}	-2.269	-1.778	0.621	0.294	-0.128	-0.099	0.762	0.564	0.777	0.671	0.657	0.640
	Censoring scheme 2											
t_{q1}	-7.847	-7.624	-0.204	-0.416	-2.247	-1.420	6.514	6.168	1.172	0.715	1.293	0.879
t_{q2}	-5.186	-4.898	-0.070	-0.199	-1.638	-0.718	2.993	2.703	0.842	0.720	1.059	0.755
t_{q3}	-2.188	-1.703	-0.022	-1.342	-1.791	-0.593	0.750	0.563	0.877	1.036	1.126	0.729
	Censoring scheme 3											
t_{q1}	-6.831	-6.615	-0.271	-0.187	-2.124	-1.110	5.349	5.051	1.994	1.231	1.796	1.316
t_{q2}	-4.273	-3.998	0.308	-0.666	-1.506	-0.285	2.396	2.164	2.728	2.490	2.880	2.315
t_{q3}	-1.291	-0.944	3.252	-7.243	-2.914	-0.877	0.764	0.679	13.687	31.585	6.311	7.095

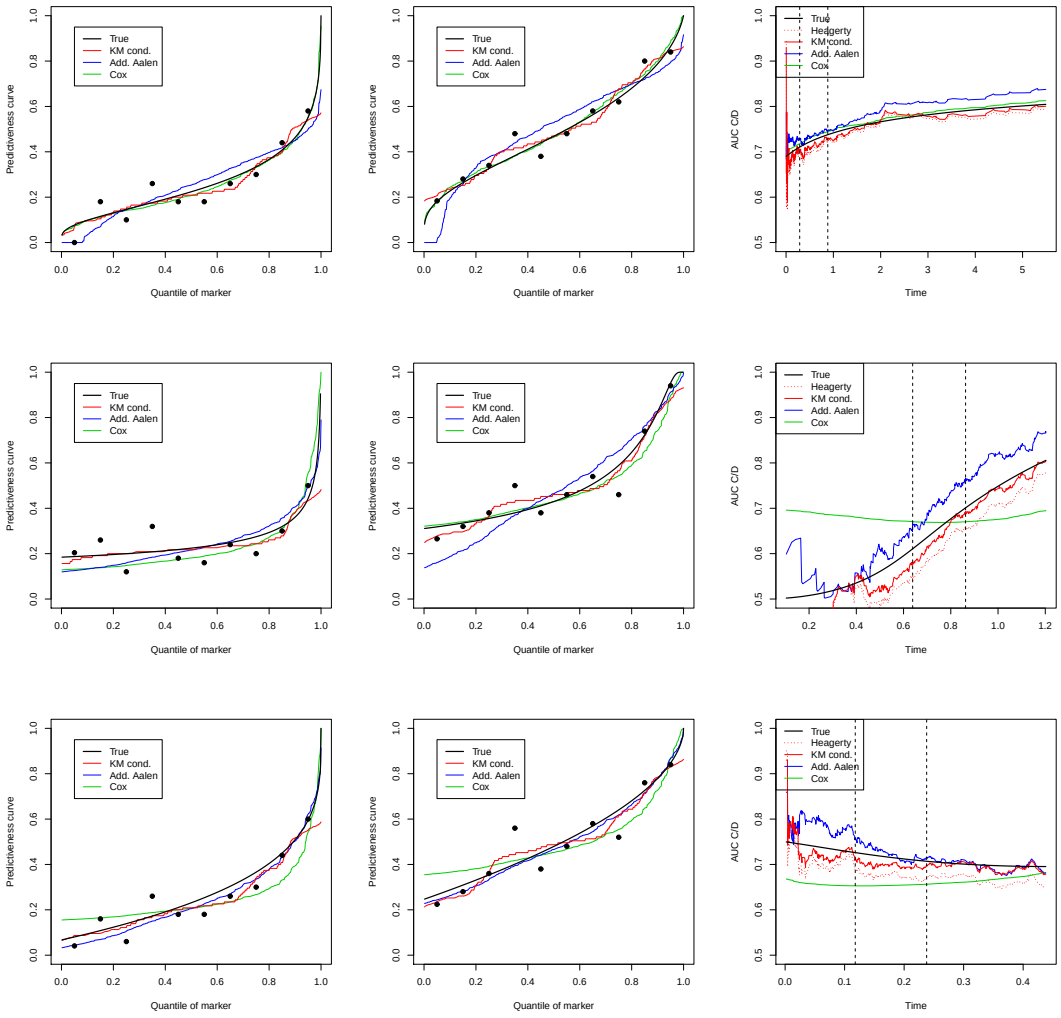

Figure 2: Time-dependent predictiveness curves (left panel) and estimates of $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ (right panel) under the three simulation designs considered in this paper: Cox model (top), time-varying Cox model (center) and Aalen's additive model (bottom). Results were obtained on one sample of size $n = 500$. In each case, time-dependent predictiveness curves were computed at the times corresponding to the first quartile and median of the survival time distribution (represented by the dotted vertical lines on the right panel). In addition, black bullets represent Kaplan-Meier estimators of the (unconditional) absolute risk for each decile of predicted risk.

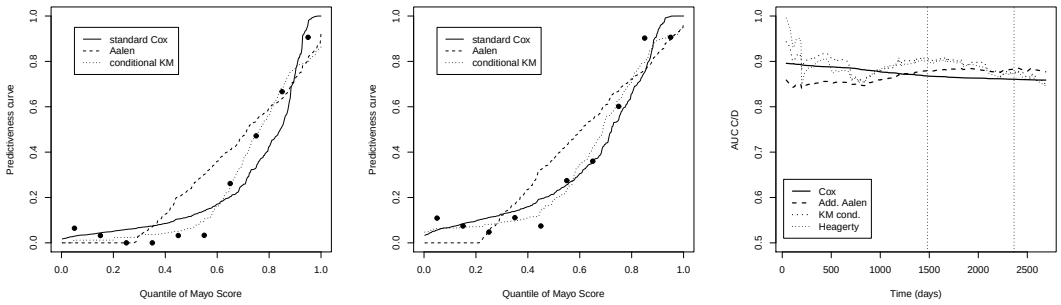

Figure 3: Time-dependent predictiveness curves (left panel) and estimates of $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$ (right panel) on the Mayo data. Time-dependent predictiveness curves were computed at the times corresponding to the 25% and 35% percentiles of the survival time distribution (represented by the dotted vertical lines on the right panel). Black bullets represent Kaplan-Meier estimators of the (unconditional) absolute risk for each decile of predicted risk.

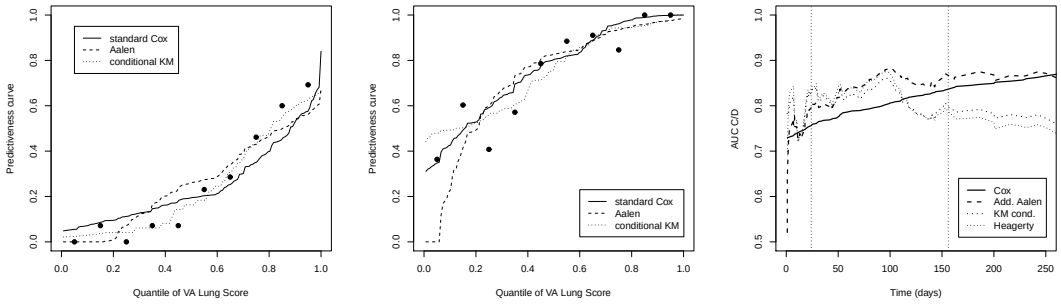

Figure 4: Time-dependent predictiveness curves (left panel) and estimates of $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$ (right panel) on the VA Lung data. Time-dependent predictiveness curves were computed at the times corresponding to the 25% and 75% percentiles of the survival time distribution (represented by the dotted vertical lines on the right panel). Black bullets represent Kaplan-Meier estimators of the (unconditional) absolute risk for each decile of predicted risk.