KU LEUVEN

FACULTEIT ECONOMIE EN BEDRIJFSWETENSCHAPPEN

Assessing cure status prediction from survival data using ROC curves

Amico M, Van Keilegom I.

KBI_1827

Assessing cure status prediction from survival data using ROC curves

Maïlis Amico and Ingrid Van Keilegom*

Research Center for Operations Research and Business Statistics, KU Leuven, Leuven, Belgium
(mailis.amico@kuleuven.be, ingrid.vankeilegom@kuleuven.be)

July 31, 2018

Abstract

Survival analysis relies on the hypothesis that, if the follow-up will be long enough, the event of interest will eventually be observed for all observations. This assumption, however, is often not realistic. In fact, when interest lies in the time until a relapse from breast cancer or the time until the occurrence of a certain disease for example, a fraction of the patients may never experience the event of interest. The survival data then contain a 'cure' fraction or long-term survivors, usually associated with infinite survival times. A common approach to model and analyse this type of data consists in using cure models. Two types of information can therefore be obtained from survival data: the survival at a given time and the cure status, both possibly modelled as a function of the covariates. The cure status is often of interest for medical practitioners, and one is usually interested in predicting it based on markers. ROC curves are one way to evaluate these predicting performances. However, the 'classical' ROC curve method is not appropriate because the cure status is partially unobserved due to the presence of censoring in survival data. In this research, we propose a ROC curve estimator aiming to evaluate the cured / non-cured status classification performance from cure survival data. This estimator, which handles the presence of censoring, decomposes sensitivity and specificity by means of Bayes' theorem, and estimates these two quantities by means of weighted empirical distribution functions. The mixture cure model is used to calculate the weights. Based on simulations, we demonstrate the good performance of the proposed method and compare it with the 'classical' ROC curve nonparametric estimator that would be obtained if the cure status was fully observed. Finally, we illustrate the methodology on a cancer data set.

Key Words: Area under the ROC curve, cure models, survival analysis, ROC curves, sensitivity, specificity.

*Acknowledgements: The authors acknowledge support from the European Research Council (2016-2021, Horizon 2020 / ERC grant agreement No. 694409).

1 Introduction

A fundamental assumption of survival analysis is that all subjects under study will eventually experience the event of interest. In some situations, however, it is possible that some subjects never experience this particular event. Indeed, when interest lies in the time until a woman gets pregnant, for example, some women will never have a child. Likewise, when one is interested in the time until a patient relapses from a cancer, some of them may never experience a relapse, as it is the case, for example, for melanoma or breast cancer in early stages, among others. In both of these examples, the assumption stated above does not hold, and it seems reasonable to consider that the survival data do not only contain 'susceptible' observations, but that they rather are a combination of two types of subjects: those who experience the event, and those who do not, these latter subjects being considered as long-term survivors or as 'cured' subjects.

A common difficulty when working with survival data is the presence of right censoring, meaning that only a lower bound of the survival time is observed. Therefore, the exact event time is only observed for some observations, the remaining individuals being censored. In the presence of a cure fraction, cured subjects are always censored since they never experience the event of interest. As a result, censored observations can be cured or uncured, and the cure status is therefore not observed. In order to take such feature into account, classical survival analysis has been extended to cure models. Initially introduced by the works of Boag (1949) and Berkson and Gage (1952), the literature on cure models is mainly composed of two classes of models, namely, the mixture cure model, introduced by Farewell (1977, 1982), and the promotion time cure model proposed by Yakovlev et al (1996).

Two quantities can be obtained from cure survival data: the survival at a given time t as in classical survival analysis, and the cure status. The literature on cure models mainly focused on modelling the effect of covariates on these two quantities (see Amico & Van Keilegom (2018) for a detailed literature review on that topic), while very little has been done on evaluating the performance of predicting these two outcomes based on cure survival data, even though good predictions are essential for practitioners. Indeed, when there exists a possible cure fraction, we can think of situations where one would be interested in predicting who is cured and who is not based on marker(s) in order to determine if a treatment is necessary to prevent a cancer relapse. Likewise, being able to correctly predict the survival probability of an uncured patient after a certain time by taking into account the presence of

cured subjects in the data is also important. A first contribution to that topic is due to Yu et al. (2008) who propose to validate individual prediction for patients with prostate cancer performed based on a joint longitudinal survival-cure model. Recently, Mehari Beyene et al. (2018) investigate the accuracy of time-dependent event prediction, extending to cure survival data the results that have been previously obtained for classical survival analysis (see for example Heagerty et al. (2000), Heagerty & Zheng (2005), Chambless & Diao (2006), Blanche et al. (2013), Li et al. (2018) among others). Zhang & Shao (2018) propose a concordance measure, in the spirit of the c-index proposed by Harrell et al. (1982, 1984), to assess the prediction accuracy of the overall survival for uncured patients by taking into account the presence of a cure fraction. They extend the work of Göner & Heller (2005) for the Cox (1972) proportional hazards (PH) model. For the cure status, on the contrary, nothing has been done to the best of our knowledge, while it is an important issue.

To evaluate if a classifier, corresponding to a single variable or a combination of variables and hereafter denoted by M, classifies correctly a set of subjects into two classes, called cases and controls, one usually considers jointly two quantities: the sensitivity which corresponds to the proportion of subjects classified as case when they are effectively a case, and the specificity, that is, the proportion of subjects classified as a control when they effectively belong to the control class. When the classifier M is measured on a continuous scale, it is necessary to dichotomise it in order to perform a binary classification. Suppose that the classes are represented by the binary variable D, such that $D = 1$ for a case, and $D = 0$ for a control. A common convention is to consider that a subject i is classified as a case when its classifier M_i is such that $M_i > k$, for k a threshold. As k can take on several values, there exist several possible sensitivities and specificities. To summarise all the information, one usually considers a Receiver Operating Characteristic (ROC) curve (see for example Pepe, 2003, and Krzanowski and Hand, 2009), which represents graphically all possible combinations of the sensitivity given by

$$Se(k) = P(M > k|D = 1), \tag{1.1}$$

and one minus the specificity,

$$1 - Sp(k) = P(M > k|D = 0), \tag{1.2}$$

that can be obtained from all possible dichotomised versions of M, based on the value of the threshold k. It plots the sensitivity against one minus the specificity for all possible values

of $k \in \mathbb{R}$ and its equation is given by

$$ROC(u) = Se\left[(1 - Sp)^{-1}(u)\right], \quad 0 < u < 1, \tag{1.3}$$

where u is an index. It is a monotone increasing function in the quadrant $(0,1) \times (0,1)$. The position of the curve indicates the ability of the classifier M to discriminate between the two classes. A perfect classifier is such as $P(M > k|D = 1) = 1$ and $P(M > k|D = 0) = 0$ for some k. In that case all observations are perfectly classified. It corresponds to a point of coordinate $(0,1)$. Conversely, an uninformative classifier is such that $P(M > k|D = 1) = P(M > k|D = 0)$, for all k. In this situation, the distribution of M is the same in the two classes. The ROC curve is therefore equal to the bisector.

Alongside the ROC curve, one usually computes the area under the curve (AUC) given by

$$AUC = \int_0^1 ROC(u)\ du, \tag{1.4}$$

which summarises into one single value the performance of M. An AUC equal to 1 corresponds to a perfect classifier, while an AUC equal to 0.5 is obtained for an uninformative classifier.

In this paper, we propose to develop a ROC curve approach in order to evaluate the accuracy of a (combination of) covariate(s) to predict the cure status based on cure survival data. Since the cure status is missing for censored observations, 'classical' ROC curve approaches, which rely on the knowledge of the classes of the observations, can not be directly implemented in this context. Therefore, an important issue to address is how to handle the latency of the cure status. Our proposal is presented in Section 2 alongside some important points related to the estimation of the sensitivity and the specificity. In Section 3, some asymptotic properties are presented, followed in Section 4 by the investigation of the finite sample performance of the proposed method through an extensive simulation study. Section 5 illustrates the practical use of our proposal on a melanoma dataset, while Section 6 concludes with some final remarks and discussion. Finally, Appendix 1 provides additional results on the finite sample performance of our estimators, while Appendix 2 gives the proofs of the asymptotic properties derived in Section 3.

2 Methodology

Let us consider a non-negative random variable, denoted by T, which represents the survival time, with survival function $S(t) = P(T > t)$, and let us assume that there exists a cure fraction. To further define this situation, let us consider that a cured subject is such that $T = \infty$, in order to represent the fact that the event never happens.

A popular model for cure survival data is the mixture cure model (Farewell, 1982), which assumes that the population of interest is a mixture of a cured and an uncured subpopulation, and which models the survival function for the entire population as a mixture model:

$$S_{pop}(t|\boldsymbol{x}, \boldsymbol{z}) = \{1 - p(\boldsymbol{x})\} + p(\boldsymbol{x})S_u(t|\boldsymbol{z}), \quad t \geq 0, \tag{2.1}$$

where $p(\boldsymbol{x}) = P(T < \infty|\mathbf{X} = \boldsymbol{x})$ is the probability of being uncured, referred to as the 'incidence', with $\mathbf{X}$ a vector of covariates, and $S_u(t|\boldsymbol{z}) = P(T > t|T < \infty, \mathbf{Z} = \boldsymbol{z})$ is the conditional survival function for uncured observations, referred to as the 'latency', with $\mathbf{Z}$ another vector of covariates that may share some or all components or be completely different from $\mathbf{X}$.

Let us further assume that T is subject to random right censoring, and that instead of observing T, we rather observe the follow-up time $Y = \min(T, C)$ and the censoring indicator $\Delta = I(T \leq C)$, where C denotes the censoring time that is supposed to be independent of T given $\mathbf{X}$ and $\mathbf{Z}$, and where $I(\cdot)$ is the indicator function. Let us consider that we have a random sample of n independently and identically distributed (i.i.d.) observations $(Y_i, \Delta_i, \mathbf{X}_i, \mathbf{Z}_i),\ i = 1, \ldots, n$, having the same distribution as $(Y, \Delta, \mathbf{X}, \mathbf{Z})$.

The objective is to derive a ROC curve estimator in order to evaluate the prediction accuracy of M for the cure status $D = I(T = \infty)$. Hereafter, we assume that $M = \gamma_0 + \boldsymbol{\gamma}^t\mathbf{X}$, where $\boldsymbol{\gamma}$ is a vector of parameters associated with $\mathbf{X}$ and γ_0 is an intercept term. We further assume that $\mathbf{X}$ can be unidimensional or multidimensional, and for this latter case, that the vector of parameters $(\gamma_0, \boldsymbol{\gamma})^t$ can be known, in such a case M is a known score such as a genetic score, for example, or unknown that needs therefore to be estimated.

2.1 Infeasible estimators

A simple and common nonparametric method to estimate a ROC curve consists in estimating the sensitivity and the specificity by their empirical distribution functions given by

$$\check{Se}(k) = 1 - \frac{1}{\check{N}_1} \sum_{i=1}^{n} \check{W}_{i1} I(M_i \leq k), \tag{2.2}$$

$$\check{Sp}(k) = \frac{1}{\check{N}_0} \sum_{i=1}^{n} \check{W}_{i0} I(M_i \leq k), \tag{2.3}$$

where $\check{W}_{i1} = I(D_i = 1)$, $\check{W}_{i0} = I(D_i = 0)$, $\check{N}_1 = \sum_{i=1}^{n} \check{W}_{i1}$ and $\check{N}_0 = n - \check{N}_1$. The ROC curve estimator takes therefore the form of a step function with jumps at each M_i. When working with cure survival data, however, these estimators cannot be used as the cure status is unobserved.

When dealing with cure survival data, Taylor (1995) proposes to consider as cured an observation with a follow-up time greater than the last uncensored follow-up time, denoted by τ. Referred hereafter as the *cure threshold*, this rule makes reasonably sense when there is a clear evidence for the presence of a cure fraction. In such a context, we consider the existence of two sub-populations, and it is reasonable to consider that, when the follow-up period is sufficiently long and when it goes well after the last uncensored event time τ, observations with a censored follow-up time greater than most event times can be categorised as cured. Based on this rule, it is therefore possible to distinguish three types of observations from cure survival data. In fact, since an uncensored subject experiences the event, it belongs to the non-cured population with certainty, that is, $D = 0$. Based on the cure threshold, censored observations can be separated into two groups, those with a follow-up time $Y > \tau$, for whom $D = 1$, and those with a follow-up time $Y \leq \tau$. For this latter case, a probability, given by $P(D = 1|\mathbf{X}, \mathbf{Z}, Y, T > Y)$, replaces the unobserved cure status. It follows that estimators for the sensitivity and the specificity are given by the following weighted empirical distribution functions:

$$\tilde{Se}(k) = 1 - \frac{1}{\tilde{N}_1} \sum_{i=1}^{n} \tilde{W}_{i1} \; I(M_i \leq k), \tag{2.4}$$

$$\tilde{Sp}(k) = \frac{1}{\tilde{N}_0} \sum_{i=1}^{n} \tilde{W}_{i0} \; I(M_i \leq k), \tag{2.5}$$

where $\tilde{W}_{i1} = (1 - \Delta_i) P(D = 1|\mathbf{X} = \mathbf{X}_i, \mathbf{Z} = \mathbf{Z}_i, Y = Y_I, T > Y_i)$, $\tilde{W}_{i0} = 1 - \tilde{W}_{i1}$, $\tilde{N}_1 = \sum_{i=1}^{n} \tilde{W}_{i1}$, and $\tilde{N}_0 = n - \tilde{N}_1$. Furthermore, when the cure threshold is assumed, $\tilde{W}_{i1}$ can

further be written as $\tilde{W}_{i1} = I(Y_i > \tau) + \ (1 - \Delta_i)\ I(Y_i \leq \tau)\ P(D = 1|\mathbf{X} = \mathbf{X}_i, \mathbf{Z} = \mathbf{Z}_i, Y = Y_i, T > Y_i)$. An infeasible estimator for the ROC curve is then given by

$$\widetilde{ROC}(u) = \tilde{Se}\{(1 - \tilde{Sp})^{-1}(u)\},\ 0 < u < 1. \tag{2.6}$$

This estimator is a monotone increasing function of u and is invariant to strictly increasing transformations of M, which are both required properties of ROC curves as described by Pepe (2003). Note that these estimators consider a random design. However, they can also be applied when the design is fixed. In such a case, notations will be different.

The development of this method relies on the following theoretical elements. By applying Bayes' theorem, the sensitivity (1.1) can be written as $Se(k) = P(M > k, T = \infty)/P(T = \infty)$. Let us consider the numerator:

$$\begin{aligned} P(M > k, T = \infty) &= E\left[I(M > k)\ I(T = \infty)\ I(\Delta = 0)\right] \\ &= E\left[I(M > k)\ P(T = \infty|\mathbf{X}, \mathbf{Z}, Y, T > Y)\ I(\Delta = 0)\right]. \end{aligned}$$

Hence,

$$Se(k) = \frac{E\left[I(M > k)\ P(T = \infty|\mathbf{X}, \mathbf{Z}, Y, T > Y)\ I(\Delta = 0)\right]}{E\left[P(T = \infty|\mathbf{X}, \mathbf{Z}, Y, T > Y)\ I(\Delta = 0)\right]}. \tag{2.7}$$

By assuming the cure threshold, and by replacing the expectation by a sum, it follows that a natural estimator for $Se(k)$ is given by (2.4).

Based on these derivations, an estimator for the AUC can also be obtained. The AUC defined in the introduction can be written as $AUC = \int_0^1 Se\big[(1-Sp)^{-1}(u)\big]\ du$. Define $f_0(k) = (d/dk)\ Sp(k)$. By proceeding to a change of variable (assuming that $(1 - Sp)^{-1}(u) = k$), we have for arbitrary $1 \leq i \neq j \leq n$ that

$$\begin{aligned} AUC &= \int_{-\infty}^{+\infty} Se(k)\ f_0(k)\ dk \\ &= E\Big[Se(M_i)|T_i < \infty\Big] \\ &= E\Big[E\{I(M_j > M_i)|T_j = \infty, \mathbf{X}_i, \mathbf{Z}_i\}|T_i < \infty\Big] \\ &= P(M_j > M_i|T_j = \infty, T_i < \infty). \end{aligned}$$

By applying Bayes' theorem, the AUC can be rewritten as

$$AUC = \frac{P(M_j > M_i, T_j = \infty, T_i < \infty)}{P(T = \infty)\ P(T < \infty)}.$$

Let us consider the numerator:

$$\begin{aligned}P(M_j > M_i, T_j = \infty, T_i < \infty) &= E\big[I(M_j > M_i)\ I(T_j = \infty)\ I(T_i < \infty)\big] \\ &= E\Big[I(M_j > M_i)\ P(T_j = \infty|\mathbf{X}_j, \mathbf{Z}_j, \Delta_j = 0, Y_j)I(\Delta_j = 0) \\ &\quad \times\ \Big\{P(T_i < \infty|\mathbf{X}_i, \mathbf{Z}_i, \Delta_i = 0, Y_i)I(\Delta_i = 0) + I(\Delta_i = 1)\Big\}\Big].\end{aligned}$$

By replacing the expectation by a sum and by assuming the cure threshold, it follows that an infeasible estimator of the AUC is given by

$$\widetilde{AUC} = \frac{1}{\tilde{N}_0\tilde{N}_1}\sum_{i=1}^{n}\sum_{j=1}^{n} I(M_j > M_i)\tilde{W}_{j1}\tilde{W}_{i0}. \tag{2.8}$$

2.2 Feasible estimators

The probability $P(D = 1|\mathbf{X}, \mathbf{Z}, Y, T > Y)$ is involved in the infeasible estimators (2.4) and (2.5) of the sensitivity and the specificity, as well as in the infeasible AUC estimator (2.8). It is therefore necessary to estimate this quantity in order to obtain estimators that can be used in practice. Based on Bayes' theorem, this probability can be written as

$$P(D = 1|\mathbf{X}, \mathbf{Z}, Y, T > Y) = \frac{P(T = \infty|\mathbf{X}, \mathbf{Z}, Y)}{P(T > Y|\mathbf{X}, \mathbf{Z}, Y)} = \frac{P(T = \infty|\mathbf{X}, \mathbf{Z})}{P(T > Y|\mathbf{X}, \mathbf{Z})}$$

since T and C are independent given $\mathbf{X}$ and $\mathbf{Z}$. Since we suppose that the data come from the mixture cure model (2.1), it can be further written as

$$\frac{P(T = \infty|\mathbf{X})}{P(T > Y|\mathbf{X}, \mathbf{Z})} = \frac{1 - p(\mathbf{X})}{\{1 - p(\mathbf{X})\} + p(\mathbf{X})S_u(Y|\mathbf{Z})}. \tag{2.9}$$

The literature on cure models offers various modelling approaches for the mixture cure model (2.1). The most common one is the logistic / Cox (LC) mixture cure model proposed by Kuk and Chen (1992), and further studied by Sy & Taylor (2000) and Peng & Dear (2000). This proposal assumes a logistic model for p, that is $p(\boldsymbol{x}) = \exp(\gamma_0+\boldsymbol{\gamma}^t\boldsymbol{x})/\{1+\exp(\gamma_0+\boldsymbol{\gamma}^t\boldsymbol{x})\}$ and considers a Cox PH model for S_u, where $S_u(t|\boldsymbol{z}) = S_0(t)^{\exp(\boldsymbol{\beta}^t\boldsymbol{z})}$, with $S_0(t) = P(T > t|T < \infty, \mathbf{Z} = 0)$, a baseline conditional survival function which remains totally unspecified, and $\boldsymbol{\beta}$ a vector of parameters associated with $\mathbf{Z}$. A drawback of this model, however, is that the estimator for $P(D = 1|\mathbf{X}, \mathbf{Z}, Y, T > Y)$ relies on a parametric assumption for p which may not be fulfilled by the data. An alternative model is the single-index / Cox (SIC)

mixture cure model proposed by Amico et al. (2018), which assumes a single-index structure for p, that is $p(\boldsymbol{x}) = g(\boldsymbol{\gamma}^t\boldsymbol{x})$, where g is a smooth unknown function, and a Cox PH model for S_u. This SIC cure model assumes a less restrictive model for p and it may therefore be more appropriate. Both approaches are considered and their respective finite sample performances are compared in Section 4. The estimators for $\tilde{W}_{i0}$ and $\tilde{W}_{i1}$ are given by

$$\hat{W}_{i1} = I(Y_i > \tau) + \ (1-\Delta_i)\ I(Y_i \leq \tau)\ \frac{1-\hat{p}(\mathbf{X}_i)}{\{1-\hat{p}(\mathbf{X}_i)\} + \hat{p}(\mathbf{X}_i)\hat{S}_u(Y_i|\mathbf{Z}_i)}$$
$$\hat{W}_{i0} = 1 - \hat{W}_{i1},$$

and are obtained by either a LC cure model or a SIC cure model. The feasible estimators of Se, Sp, ROC and AUC are now given by

$$\hat{Se}(k) = 1 - \frac{1}{\hat{N}_1}\sum_{i=1}^{n}\hat{W}_{i1}\ I(M_i \leq k), \tag{2.10}$$

$$\hat{Sp}(k) = \frac{1}{\hat{N}_0}\sum_{i=1}^{n}\hat{W}_{i0}\ I(M_i \leq k), \tag{2.11}$$

$$\widehat{ROC}(u) = \hat{Se}\{(1-\hat{Sp})^{-1}(u)\}, \quad 0 < u < 1, \tag{2.12}$$

$$\widehat{AUC} = \frac{1}{\hat{N}_0\hat{N}_1}\sum_{i=1}^{n}\sum_{j=1}^{n} I(M_j > M_i)\hat{W}_{j1}\hat{W}_{i0}, \tag{2.13}$$

where $\hat{N}_1 = \sum_{i=1}^{n}\hat{W}_{i1}$ and $\hat{N}_0 = n - \hat{N}_1$.

Both $\mathbf{X}$ and $\mathbf{Z}$ enter in the computation of $\hat{W}_0$ and $\hat{W}_1$, while M only relies on $\mathbf{X}$. For the choice of the covariates to include in $\mathbf{X}$, we consider those included in M. A more delicate question concerns the choice of the covariates to consider for $\mathbf{Z}$. When M only contains one covariate, and when there is only one covariate available in the data, it is easy to assume that $X = Z$. If there are several covariates in the data, or when M is a combination of covariates, on the contrary, the choice of $\mathbf{Z}$ will depend on the knowledge of the topic of the analysis, and on which covariates are thought to influence the survival of uncured subjects. In such contexts, $\mathbf{Z}$ can be partially or fully identical to $\mathbf{X}$, or completely different from $\mathbf{X}$. However, we are not free of misspecification. The influence of a misspecification of this vector on the estimation of the ROC curve is therefore investigated through simulations in Section 4.

3 Asymptotic theory

In this section we will develop the limiting distribution of the proposed estimators of the sensitivity, the specificity, the ROC curve and the AUC given in equations (2.10), (2.11), (2.12) and (2.13). In the previous section these estimators were constructed either based on a logistic/Cox mixture cure model or on a single-index/Cox mixture cure model. However, asymptotic theory for the estimation of these models has only been developed so far under the logistic/Cox model (see Lu, 2008), and so we restrict attention in this section to the latter model.

The proofs of the results of this section can be found in Appendix 2.

Theorem 3.1. *Assume that conditions 1–4 in Lu (2008) are satisfied and that the logistic/Cox mixture cure model is valid. Then,*

$$\begin{aligned}\hat{Se}(k) - Se(k) &= n^{-1}\sum_{i=1}^{n} \eta_{Se}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, k) + R_{n,Se}(k) \\ \hat{Sp}(k) - Sp(k) &= n^{-1}\sum_{i=1}^{n} \eta_{Sp}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, k) + R_{n,Sp}(k)\end{aligned}$$

where $\sup_k |R_{n,Se}(k)| = o_P(n^{-1/2})$, $\sup_k |R_{n,Sp}(k)| = o_P(n^{-1/2})$, *and* $\eta_{Se}(\mathbf{x}, \mathbf{z}, y, \delta, k)$ *and* $\eta_{Sp}(\mathbf{x}, \mathbf{z}, y, \delta, k)$ *are defined in (7.4) and (7.5) in Appendix 2.*

Moreover, the process $n^{1/2}(\hat{Se}(k) - Se(k))$ $(k \in \mathbb{R})$ *converges weakly to a Gaussian process* $Z_{Se}(k)$ *with zero mean and covariance function given by*

$$Cov(Z_{Se}(k_1), Z_{Se}(k_2)) = E\big[\eta_{Se}(\mathbf{X}, \mathbf{Z}, Y, \Delta, k_1)\ \eta_{Se}(\mathbf{X}, \mathbf{Z}, Y, \Delta, k_2)\big],$$

and the process $n^{1/2}(\hat{Sp}(k) - Sp(k))$ $(k \in \mathbb{R})$ *converges weakly to a Gaussian process* $Z_{Sp}(k)$ *with zero mean and covariance function given by*

$$Cov(Z_{Sp}(k_1), Z_{Sp}(k_2)) = E\big[\eta_{Sp}(\mathbf{X}, \mathbf{Z}, Y, \Delta, k_1)\ \eta_{Sp}(\mathbf{X}, \mathbf{Z}, Y, \Delta, k_2)\big].$$

As a corollary to the above result we now state the limiting distribution of the estimator $\widehat{ROC}(u)$ defined in (2.12) and of the estimator $\widehat{AUC}_\delta$, given by

$$\widehat{AUC}_\delta = \int_\delta^{1-\delta} \widehat{ROC}(u)\, du.$$

For technical reasons we need to restrict the integration to the internal $[\delta, 1-\delta]$ (for some small $\delta > 0$), which can however be made arbitrarily close to the interval $[0, 1]$. The corresponding theoretical AUC is denoted by $AUC_\delta = \int_\delta^{1-\delta} ROC(u)\, du$.

Corollary 3.1. *Assume that conditions 1–4 in Lu (2008) are satisfied and that the logistic/Cox mixture cure model is valid. Assume in addition that* $\inf_{k_1 \leq k \leq k_2} Sp'(k) > 0$*, where* $k_1 = (1-Sp)^{-1}(\delta)$ *and* $k_2 = (1-Sp)^{-1}(1-\delta)$ *for some* $\delta > 0$*, and that the functions Se and Sp are twice continuously differentiable on* $[k_1, k_2]$*. Then,*

$$\widehat{ROC}(u) - ROC(u) = n^{-1} \sum_{i=1}^{n} \eta_{ROC}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, u) + R_{n,ROC}(u),$$

where $\sup_{\delta<u<1-\delta} |R_{n,ROC}(u)| = o_P(n^{-1/2})$*, and*

$$\begin{aligned} \eta_{ROC}(\mathbf{x}, \mathbf{z}, y, \delta, u) &= \eta_{Se}\big(\mathbf{x}, \mathbf{z}, y, \delta, (1-Sp)^{-1}(u)\big) \\ &\quad + \frac{Se'\{(1-Sp)^{-1}(u)\}}{(1-Sp)'\{(1-Sp)^{-1}(u)\}} \eta_{Sp}\big(\mathbf{x}, \mathbf{z}, y, \delta, (1-Sp)^{-1}(u)\big). \end{aligned}$$

Moreover, the process $n^{1/2}(\widehat{ROC}(u) - ROC(u))$ $(u \in [\delta, 1-\delta])$ *converges weakly to a Gaussian process* $Z_{ROC}(u)$ *with zero mean and covariance function given by*

$$Cov(Z_{ROC}(u_1), Z_{ROC}(u_2)) = E\big[\eta_{ROC}(\mathbf{X}, \mathbf{Z}, Y, \Delta, u_1)\ \eta_{ROC}(\mathbf{X}, \mathbf{Z}, Y, \Delta, u_2)\big],$$

and

$$n^{1/2}(\widehat{AUC}_\delta - AUC_\delta) \xrightarrow{d} N(0, \sigma^2_{AUC}),$$

where

$$\sigma^2_{AUC} = \int_\delta^{1-\delta} \int_\delta^{1-\delta} E(Z_{ROC}(u_1) Z_{ROC}(u_2))\, du_1\, du_2.$$

4 Finite sample performances

4.1 Some preliminaries

In this section, an extensive simulation study is performed in order to evaluate the finite sample performances of the ROC curve estimator (2.12). Two versions of this estimator are considered:

LC : assuming a LC cure model for W_0 and W_1. The LC cure model is estimated assuming the method proposed by Sy & Taylor (2000) based on the EM algorithm,

SIC : assuming a SIC cure model for W_0 and W_1, where the model is estimated according to the maximum likelihood approach described in Amico et al. (2018).

Both the case of known and unknown M are investigated, and for both of them, the following points are analysed. First, we are interested in the general performance of the proposed estimators of the sensitivity and the specificity. Particular interest lies in the effect of censoring and of an incorrect specification of the vector $\mathbf{Z}$. Then, other points include a misspecification of the model for S_u and a non-logistic model for the cure proportion p.

To assess the performance of our proposed method, we consider two infeasible competitors:

CSK (Cure Status Known): corresponding to the ROC curve estimator that would be obtained if the cure status would be fully observed. Equations (2.2) and (2.3) give the estimators for the sensitivity and the specificity in that case. The objective here is to evaluate the effect of the imputation of the cure status described in Section 2.1,

TW (True Weights): corresponding to the estimator (2.6) combined with (2.9), based on the true values of p and S_u. This benchmark estimator allows to investigate the effect of estimating p and S_u by means of the LC or SIC cure model.

Two criteria are considered to compare the four estimators, namely, the L1 distance between the true and the estimated ROC curves and the AUC. The L1 distance is given by

$$L1 = V^{-1} \sum_{i=1}^{V} |\widehat{ROC}(u_i) - ROC(u_i)|,$$

where $\widehat{ROC}$ is one of the ROC curve estimates and ROC is the true ROC curve. It is computed over a grid of points $u_i = \frac{i}{100}$ for $i = 1, \ldots, V = 99$. For LC and SIC estimators, the AUC is given by

$$\widehat{AUC} = \frac{1}{\hat{N}_0 \hat{N}_1} \sum_{i=1}^{n} \sum_{j=1}^{n} \Big[\big\{ I(M_j > M_i) + 0.5 \times I(M_j = M_i) \big\} \hat{W}_{j1} \hat{W}_{i0} \Big].$$

For CSK and TW etimators, the formula is almost the same, but with different W_{i0}, W_{i1}, N_0 and N_1. For CSK, they are replaced by $\check{W}_{i0}$, $\check{W}_{i1}$, $\check{N}_0$ and $\check{N}_1$, while for TW, they are given by $\tilde{W}_{i0}$, $\tilde{W}_{i1}$, $\tilde{N}_0$ and $\tilde{N}_1$. Note that these formulas take into account possible ties in M with the added term $0.5 \times I(M_j = M_i)$.

4.2 Data generating process

Within this section, we assume that the data are generated from the mixture cure model (2.1). The data generating process is as follows.

1. First, the incidence is considered :

   (i) The uncure probability p is generated according to the model $p(\boldsymbol{x}) = g(\boldsymbol{\gamma}^t\boldsymbol{x})$, where $g(\cdot)$ is a link function. Primary interest lies in the logistic link function, that is, $g(a) = \exp(\gamma_0 + a)/\{1 + \exp(\gamma_0 + a)\}$, which gives the logistic regression model, but other link functions can also be assumed;

   (ii) The second step consists in generating, for given $\boldsymbol{x}$, the uncure status $(1 - D)$ from a Bernoulli distribution with parameter equal to $p(\boldsymbol{x})$.

2. Next, the latency is generated :

   (i) We consider two models for the survival function of the uncured observations. The first model is a Gompertz model with survival function $S(t|\boldsymbol{z}) = S_0(t)^{\exp(\boldsymbol{\beta}^t\boldsymbol{z})}$, $S_0(t) = \exp[-\theta\alpha^{-1}\{\exp(\alpha t) - 1\}]$, $\theta = 0.5$ and $\alpha = 0.03$. The second model is an Accelerated Failure Time (AFT) model assuming a log-logistic distribution for T and with survival function $S_u(t|\boldsymbol{z}) = [1 + \lambda\{t/\exp(\beta^t\boldsymbol{z})\}^{\kappa}]^{-1}$ where $\lambda = 0.05$ and $\kappa = 2.5$. Note that the AFT model does not respect the proportional hazards property contrarily to the Gompertz model. Since $\hat{W}_0$ and $\hat{W}_1$ are obtained from a mixture cure model assuming a Cox PH model for the latency, this allows us to verify whether a model misspecification of S_u affects the ROC curve estimate;

   (ii) Next, we generate the censoring time from an uniform distribution on $[U_{min}, U_{max}]$ that is independent of T, $\mathbf{X}$ and $\mathbf{Z}$. We further truncate the survival times of the susceptible observations at $U_{max} - 1$ so that the support of C is larger than the support of T;

   (iii) We finally generate the follow-up time $Y = \min(T, C)$ and the censoring indicator $\Delta = I(T \leq C)$.

4.3 Known classifier

First, we consider the case where the classifier takes the form of a single variable or of a known one-dimensional score denoted by X. Note that when $\dim(\mathbf{X}) = 1$, the single-index model reduces to a non-parametric model. We assume three different scenarios for the incidence. The first two scenarios assume a logistic regression model for $p(x)$ corresponding to different discriminations between the cured and the uncured sub-populations:

> *Scenario 1* : $X \sim N(2, 2.5)$, $\gamma_0 = 0$, $\gamma_1 = 1$, and $AUC = 0.9016$. This scenario corresponds to a a good discrimination with a cure proportion equal to 25.6%.
>
> *Scenario 2*: $X \sim N(1.2, 1)$, $\gamma_0 = 0$, $\gamma_1 = 1$, and $AUC = 0.7374$. This scenario is associated with a moderate separation between the two sub-populations, and the proportion of cured subjects is equal to 26.9%.

The third scenario assumes a non-logistic model for $p(x)$ with a non-monotone shape in order to evaluate the performance of the LC and SIC estimators in such a case. The link function is given by $g(a) = [\sin\{(3/2)\ \pi\ a\} + 1]/2$. Its characteristics are as follows:

> *Scenario 3* : $X \sim unif(0, 1)$, $\gamma_1 = 1$, and $AUC = 0.8124$, corresponding to a good separation between cured and uncured sub-populations. The cure proportion equals 39.4%.

The graphical representation of the respective ROC curves is given in Figure 1.

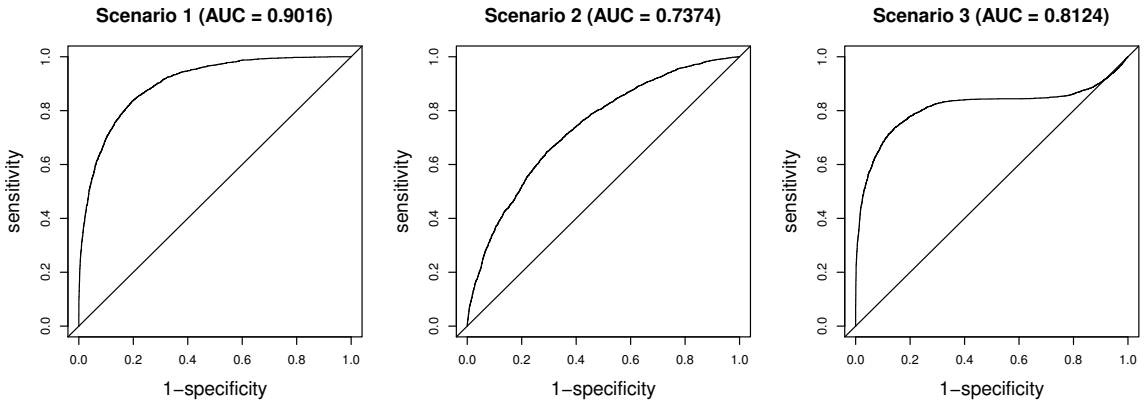

Figure 1: *True ROC curves for scenarios 1, 2 and 3 when the classifier is known.*

For the survival times, the Gompertz model and the AFT model have the following characteristics.

Gompertz model: We consider two covariates, Z_1 and Z_2, that are independent, following a Bernoulli distribution with parameter equal to 0.6 and 0.2, respectively. The associated vector of parameters is $\boldsymbol{\beta} = (1.5, -0.5)^t$. For the uniform distribution considered for the censoring time C, we assume that $U_{min} = 0$ and three different values are considered for U_{max} : 65, 25 and 10, corresponding to three different levels of censoring denoted by level 1, level 2 and level 3.

AFT model: Two independent covariates, Z_1 and Z_2, are considered, following a Bernoulli distribution with parameter equal to 0.6 and 0.3, respectively. The associated vector of parameters is $\boldsymbol{\beta} = (0.7, -0.3)^t$. As for the Gompertz model, the censoring time is generated from a uniform distribution with $U_{min} = 0$ and with three different values for U_{max}. These values are chosen such that the proportion of censored observations with a follow-up time lower than or equal to τ is the same as for the Gompertz model, in order to allow comparison between the two models.

We now consider the following five settings, corresponding to the three scenarios for the incidence and the two models for the latency considered above (note that the non-logistic scenario for the incidence is only considered in combination with the Gompertz model, since it serves to assess the performance of the LC estimator when the logistic model is not satisfied). Each setting has a particular objective:

- Scenario 1 / Gompertz model – to evaluate our proposal and the effect of censoring when the discrimination is good;
- Scenario 1 / AFT model – to verify whether a model misspecification of S_u affects the ROC curve estimate when the discrimination is good;
- Scenario 2 / Gompertz model – to evaluate our proposal and the effect of censoring when the discrimination is moderate;
- Scenario 2 / AFT model – to verify whether a model misspecification of S_u affects the ROC curve estimate when the discrimination is moderate;

- Scenario 3 / Gompertz model – to evaluate our proposal when the link function is not logistic. Note that for this latter setting, a fourth censoring rate is considered in order to further evaluate its impact in such a case.

To assess the effect of a misspecification of $\mathbf{Z}$, the estimators are further estimated assuming that $Z = X$, on scenario 1 / Gompertz and scenario 2 / Gompertz. Table 1 summarises the setting characteristics, comprising parameter values for the censoring distributions, cure rates, censoring rates, and the percentage of observations for which $Y \leq \tau$.

Table 1: *Setting characteristics when the classifier is known: parameters of the censoring distribution, cure rate, censoring rate and percentage of censored observations for which* $Y \leq \tau$.

	latency type							
	Gompertz model				*AFT model*			
incidence type	U_{max}	cure rate	censoring rate	% obs. $\Delta = 0$, $Y \leq \tau$	U_{max}	cure rate	censoring rate	% obs. $\Delta = 0$, $Y \leq \tau$
Scenario 1	65	25.6%	27.0%	5.3%	345	25.6%	27.0%	5.3%
	25	25.6%	29.1%	12.6%	121	25.6%	29.5%	12.6%
	10	25.6%	34.0%	25.4%	44	25.6%	35.9%	25.5%
Scenario 2	65	26.9%	28.2%	5.3%	360	26.9%	28.2%	5.3%
	25	26.9%	30.3%	13.0%	120	26.9%	30.8%	12.9%
	10	26.9%	35.1%	25.9%	45	26.9%	37.1%	26.1%
Scenario 3	65	39.4%	40.5%	6.8%				
	25	39.4%	42.2%	16.2%				
	10	39.4%	46.3%	31.4%				
	5	39.4%	51.8%	43.1%				

For each setting, we consider 500 datasets, and for each dataset, we assume two sample sizes, $n = 250$ and $n = 500$.

Figure 2 shows the boxplots of the L1 distance when $n = 250$ for the settings with

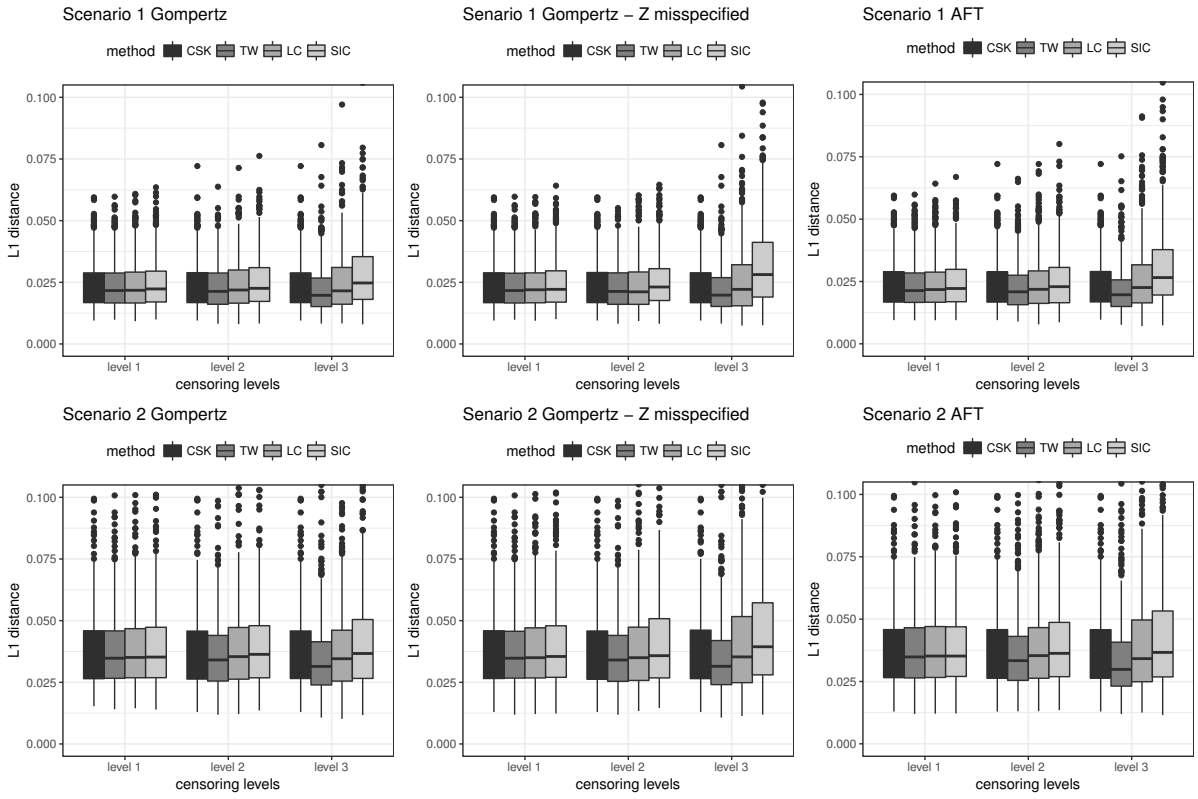

Figure 2: *Boxplots of the L1 distances for scenarios 1 and 2 for* $n = 250$.

scenarios 1 and 2 for the incidence. As it can be seen, when the censoring rate is close to the cure rate, and when everything is specified correctly, our proposals perform almost as well as the two infeasible competitors whatever the model assumed for W_1. In such a case, very few censored observations are below τ, which are those with weight equal to $P(D = 1|\mathbf{X}, \mathbf{Z}, Y, T > Y)$. A larger censoring rate is conversely associated with higher L1 distance and larger variance, particularly for SIC under the third censoring level. When the censoring rate gets larger, fewer censored observations are located in the plateau, meaning that less censored observations are considered as cured, that is, with $W_{i1} = 1$. Furthermore, as shown by Amico et al. (2018), the SIC cure model performs worse than the LC cure model when the true model for the incidence is a logistic model and when the censoring rate increases as it is the case for the third censoring level. Interestingly, LC is close to CSK even for the third censoring level. Another interesting point is that the L1 distance for TW decreases slightly when the censoring rate increases. It seems that having more censored observations below τ produces better results when considering the true W_0 and W_1. In such

a case, the size of the jumps are smaller and it seems that the ROC curve becomes smoother and closer to the true curve. Nevertheless, this feature is not observed for LC and SIC. By comparing scenario 1 and scenario 2, we observe that the L1 distance is larger for scenario 2. Indeed, it is more difficult to correctly separate cured from uncured sub-populations based on this scenario since the discrimination is moderate. The discrimination between cured and uncured sub-populations seems therefore to have an influence on the performance of the ROC curve estimators. However, the general conclusions are the same for both scenarios.

For the settings where $\mathbf{Z}$ is misspecified, we observe that when few censored observations are below τ, the L1 distance for our proposals is only slightly higher in comparison with the two infeasible competitors, while when the number of censored observations below τ is larger as for the third censoring level, both the LC and SIC estimators have higher L1 distance than when $\mathbf{Z}$ is correctly specified. Interestingly, for the second censoring level, the L1 distance for LC seems to be comparable to the L1 distance of CSK for both scenarios while SIC seems to already present some difficulties. Note that we consider the case where $\mathbf{Z}$ is completely misspecified, whereas it seems more likely to have only a partial misspecification of this vector of covariates in practical applications. We are therefore in an extreme case.

When the survival times are generated according to an AFT model, our proposals show a higher increase in the L1 distance in comparison with when there is no misspecification, especially when the censoring rate increases. SIC is still the least favourable estimator. However, a misspecification in $\mathbf{Z}$ affects the performance of our proposals more than a misspecification in the latency.

Figure 3 provides the boxplots of the AUC for scenarios 1 and 2 when $n = 250$. The same conclusions as for the L1 distances can be drawn. Note that SIC performs less good than LC, especially for the third censoring rate. LC on the contrary is close to CSK for all censoring rates considered, but we observe more variability. For all these settings, the true incidence is a logistic regression model.

When the true incidence is not a logistic regression model (scenario 3), LC has always higher L1 distances than SIC as it can be seen in Figure 4. When the censoring rate gets larger, the difference between LC and SIC increases, and we observe that SIC outperforms LC, especially for the third and the fourth censoring levels. It seems therefore that, when $\hat{p}(\boldsymbol{x})$ is inconsistent and when the proportion of censored observations below τ is large, LC performs badly. Based on the analysis of the AUC, the same conclusions can be drawn.

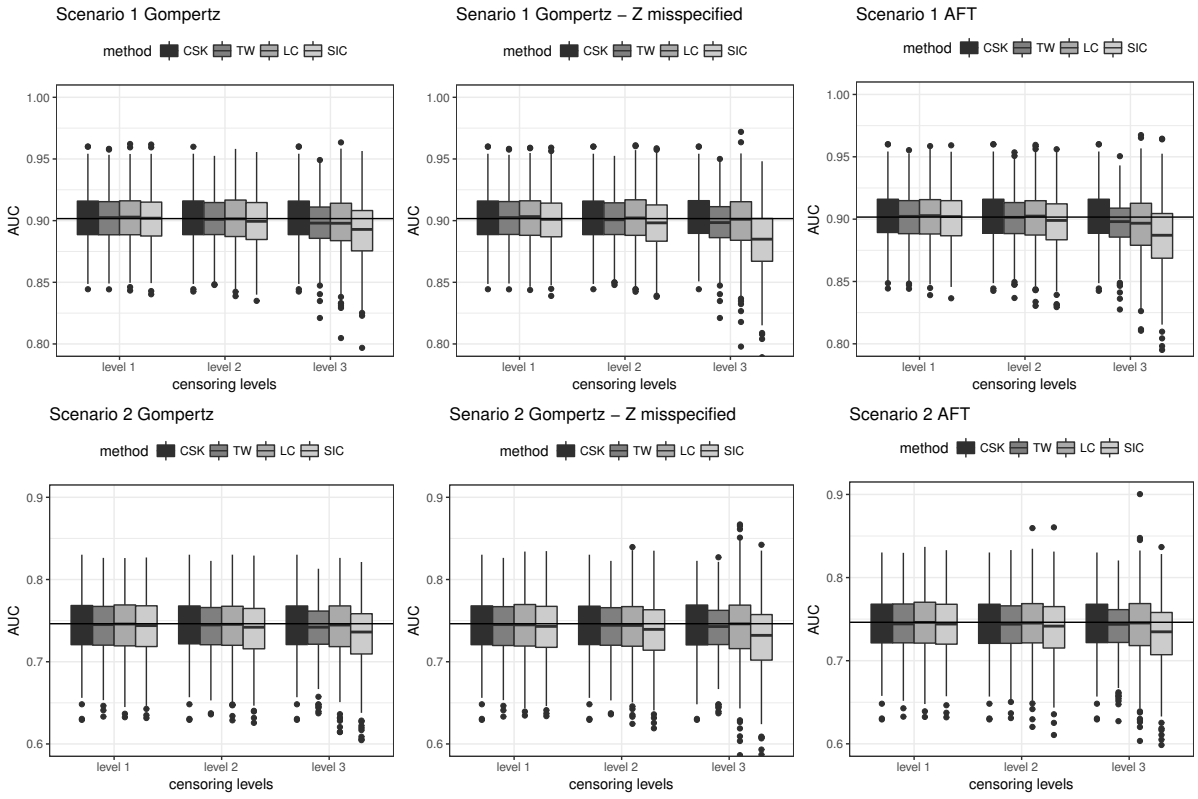

Figure 3: *Boxplots of the AUC for scenarios 1 and 2 for* $n = 250$.

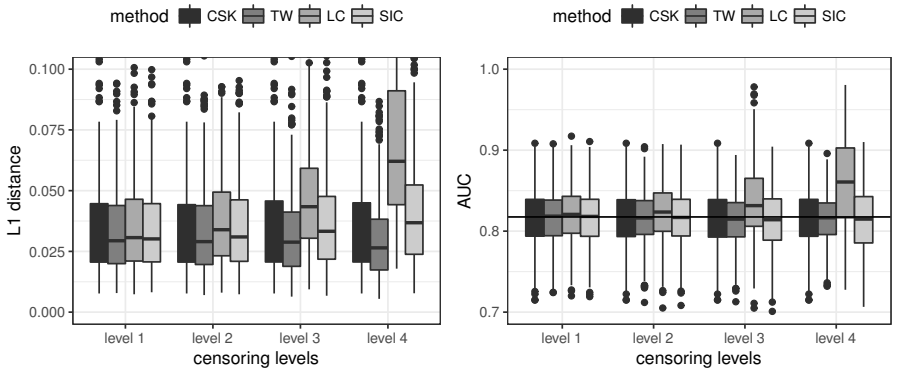

Figure 4: *Boxplots of the L1 distance and the AUC for scenario 3 Gompertz for* $n = 250$.

An interesting feature that was already observable from the previous settings and which is confirmed with the fourth censoring level here is that the more the censoring rate increases, the more the L1 distance of LC and SIC increases since many more observations have $\hat{W}_0$

and $\hat{W}_1$ relying on $P(D = 1|\mathbf{X}, \mathbf{Z}, Y, T > Y)$ in that case. Therefore, the censoring level is crucial in the performance of our proposals.

For both the L1 distance and the AUC, the same conclusions also apply when $n = 500$ (see Figures 10, 11 and 12 in Appendix 1). Furthermore, as expected, the L1 distances are smaller and less variable than for $n = 250$. For the AUC, the variability is also lower.

4.4 Unknown classifier

We next consider the case where the classifier is an unknown combination of variables. Two scenarios are considered for the incidence, both of them assuming a logistic regression model and corresponding to two different levels of discrimination between the two sub-populations. For both of them, we assume three independent variables. Each setting has the following characteristics:

> *scenario 4*: $X_1 \sim N(0, 1)$, $X_2 \sim Bernoulli(0.6)$, and $X_3 \sim Bernoulli(0.3)$. We take $\gamma_0 = 1$ and $\boldsymbol{\gamma} = (2, 3, -2)^t$. The cure rate is equal to 24.5% and $AUC = 0.9080$, corresponding to a good discrimination between cured ans uncured subjects.
>
> *scenario 5*: $X_1 \sim N(0.6, 1)$, $X_2 \sim Bernoulli(0.5)$, and $X_3 \sim Bernoulli(0.4)$. We take $\gamma_0 = -1.7$ and $\boldsymbol{\gamma} = (0.7, 1.1, 0.3)^t$. The cure rate is equal to 62.2% and $AUC = 0.7219$. With this scenario the discrimination between the two sub-populations is moderate. These characteristics have been chosen such that the scenario mimics the characteristics of the melanoma data on which our methodology is illustrated (see Section 5).

The graphical representation of the true ROC curves for these two scenarios is given in Figure 5.

For the latency, we consider the same model as when the classifier is known, that is, a Gompertz model or an AFT model, with the same characteristics. We consider the four following settings corresponding to all combinations of the two scenarios for the incidence and the two models for the latency. As for the known classifier case, each setting has the following objective:

- Scenario 4 / Gompertz model – to evaluate our proposal and the effect of censoring when the discrimination is good;

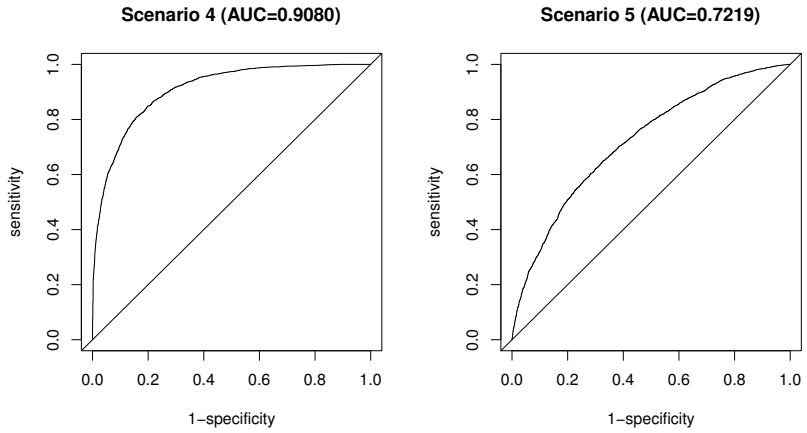

Figure 5: *True ROC curves for scenarios 4 and 5 when the classifier is unknown.*

- Scenario 4 / AFT model – to verify whether a model misspecification in S_u affects the ROC curve estimate when the discrimination is good;
- Scenario 5 / Gompertz model – to evaluate our proposal and the effect of censoring when the discrimination is moderate;
- Scenario 5 / AFT model – to verify whether a model misspecification in S_u affects the ROC curve estimate when the discrimination is moderate;

To further evaluate the performance of our proposal, the estimators are also estimated assuming that $\mathbf{Z} = \mathbf{X}$ in scenario 4 / Gompertz and scenario 5 / Gompertz. Note also that a fourth censoring rate is considered when scenario 5 is assumed for the incidence in order to evaluate the situation when many observations are censored before τ as it is the case for the melanoma data set. Table 2 summarises the setting characteristics, comprising parameter values for the censoring distributions, cure rates, censoring rates, and the percentage of observations below τ.

Since the classifier is unknown, it first needs to be estimated before the ROC curve can be computed. We consider a LC cure model, M being estimated by the score $\hat{\gamma}_0 + \hat{\boldsymbol{\gamma}}^t\mathbf{X}$ from the logistic model considered in the incidence. However, a difficulty remains. Indeed, if M and the ROC curve are estimated on the same dataset, the ROC curve can overestimate the classification performance of M as explained by Copas & Corbett (2002), and hence this would lead to misleading conclusion(s). It is therefore necessary to obtain generalizable

Table 2: *Setting characteristics when the classifier is unknown: parameters of the censoring distribution, cure rate, censoring rate and percentage of censored observations for which* $Y \leq \tau$.

	latency type							
	Gompertz model				*AFT model*			
incidence type	U_{max}	cure rate	censoring rate	% obs. $\Delta = 0$, $Y \leq \tau$	U_{max}	cure rate	censoring rate	% obs. $\Delta = 0$, $Y \leq \tau$
Scenario 4	65	24.5%	25.9%	4.9%	370	24.5%	25.9%	4.9%
	25	24.5%	28.0%	12.1%	126	24.5%	28.5%	12.1%
	10	24.5%	33.2%	24.5%	47	24.5%	34.9%	24.4%
Scenario 5	65	62.2%	63.3%	8.7%	337	62.2%	63.4%	8.7%
	25	62.2%	64.4%	21.3%	111	62.2%	64.8%	21.3%
	10	62.2%	66.9%	40.4%	40	62.2%	68.6%	40.5%
	5	62.2%	70.3%	54.4%	23	62.2%	72.6%	54.4%

conclusions about the performance of M. To do so, an ideal approach would consist in splitting the dataset into two groups, a training set on which the model is fitted and a test set on which predictions are made and then used to build the ROC curve. However, to split a dataset, a large sample size is required. When this is not the case, the use of other approaches exist as explained by Hastie et al. (2009) among others, among which *cross-validation*. To mimic real data settings, the finite sample performance of our proposal is therefore evaluated as if the sample size was not large enough to be split into two groups. We instead perform it using cross-validation. We proceed as follows. First, the initial dataset is split into K folds. Each fold is then considered as the test set successively, the other folds being considered as the training set. For example, if $K = 5$, and denote by $F1$, $F2$, $F3$, $F4$, and $F5$ the five folds, the cross-validation procedure will produce five different runs with the following composition for the test and training sets :

run	test set	training set
1	$F1$	$F2 + F3 + F4 + F5$
2	$F2$	$F1 + F3 + F4 + F5$
3	$F3$	$F1 + F2 + F4 + F5$
4	$F4$	$F1 + F2 + F3 + F5$
5	$F5$	$F1 + F2 + F3 + F4$

For each run, a LC cure model is fitted on the training set in order to estimate γ_0 and $\boldsymbol{\gamma}$. Then, predictions for M are performed on the test set and the ROC curves are estimated. Furthermore, the L1 distance and the AUC are computed for each ROC curve at each run. At the end of the K runs, the L1 distances and AUCs are averaged over the K folds. In what follows, we take $K = 5$. As for the known classifier case, 500 datasets are generated for each setting, and two sample sizes are considered, $n = 250$ and $n = 500$.

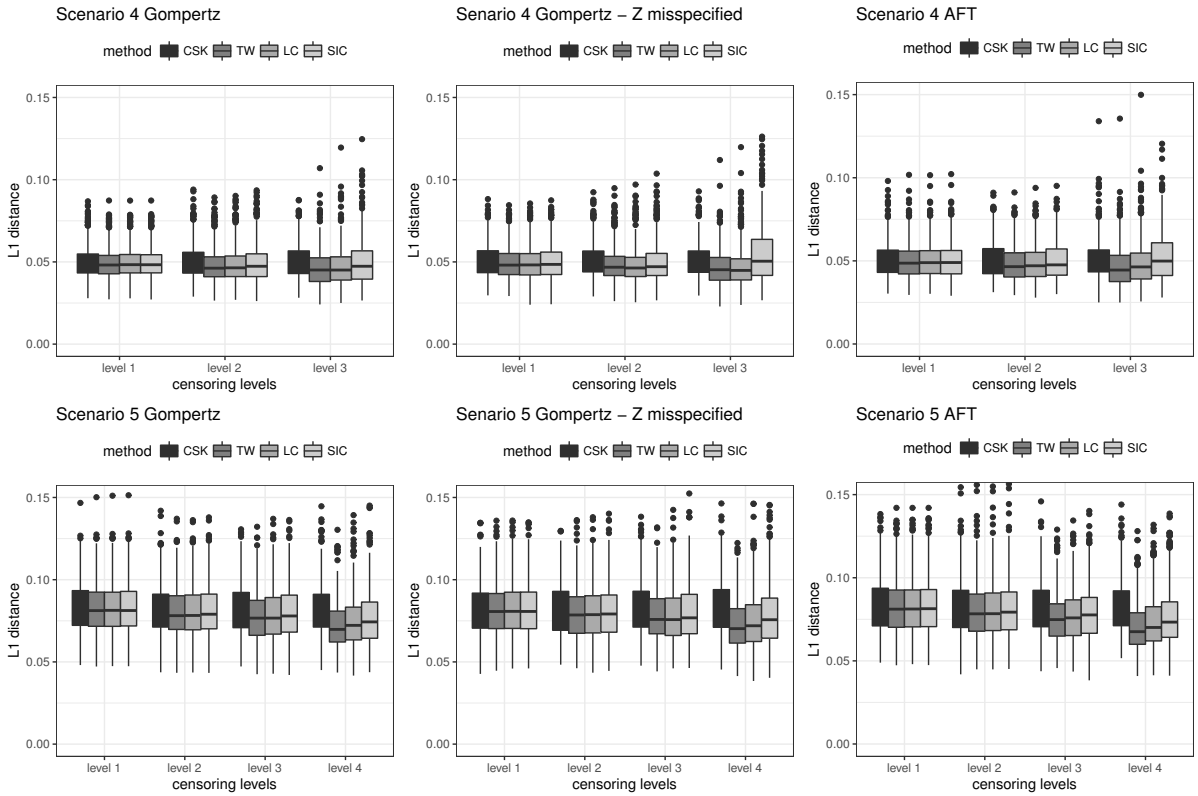

Figure 6: *Boxplots of the L1 distances for scenarios 4 and 5 for* $n = 250$.

The boxplots of the L1 distance for all settings and for $n = 250$ are given in Figure 6. As

it can be seen, when there is no misspecification on $\mathbf{Z}$ or on the latency, the conclusions are almost the same as when the classifier is known. We observe only slight differences between the four estimators when few observations are censored before τ, while when the censoring rate gets larger, LC and SIC present slightly larger L1 distance than TW, particularly for the third and the fourth (only for scenario 5) censoring level. We also notice that, as for the known classifier case, SIC presents higher L1 distance than LC in such a case. However, we also observe that compared to CSK, our proposals have ROC curves that are closer to the true one when the number of censored observations for which $Y \leq \tau$ increases. Since the ROC curves are computed based on cross-validation, only 50 observations are considered to build the ROC curves on each run of the cross-validation. Furthermore, as already mentioned for the known classifier case, when the censoring rate increases, more observations have a weight equal to $P(D = 1|\mathbf{X}, \mathbf{Z}, Y, T > Y)$ and it follows that our proposed method gives a smoother curve which is closer to the true one. Since the sample size is small the ROC curves have fewer jumps than previously and it seems that, for the unknown classifier case, not only TW presents better results in such a case, but also LC and SIC. As before, the L1 distances are large for scenario 5 compared to scenario 4. This is due to the lower discrimination in scenario 5.

When $\mathbf{Z}$ is misspecified, LC performs almost as well as TW and the results are comparable to the case where there is no misspecification and so for all censoring levels. Conversely, when the number of censored observations before τ increases, SIC presents slightly higher L1 distance compared to the situation where there is no misspecification. When the latency is misspecified, the same conclusions can be drawn. We only observe slight differences with the case where there is no misspecification. When $n = 500$, the conclusions stay the same as it can be seen in Figure 13 given in Appendix 1 even if we notice lower L1 distances, as expected.

Figure 7 (and Figure 14 in Appendix 1) contains the boxplots of the AUC for $n = 250$ (for $n = 500$). The same conclusions as for the L1 distance can be drawn. As for the known classifier, our proposals perform well in comparison with the two infeasible competitors, even when the censoring rate increases or when the weights are not correctly specified. As for the L1 distance, both LC and SIC present some difficulties when many censored observations are such that $Y \leq \tau$. SIC is still the one performing the least good in such a case.

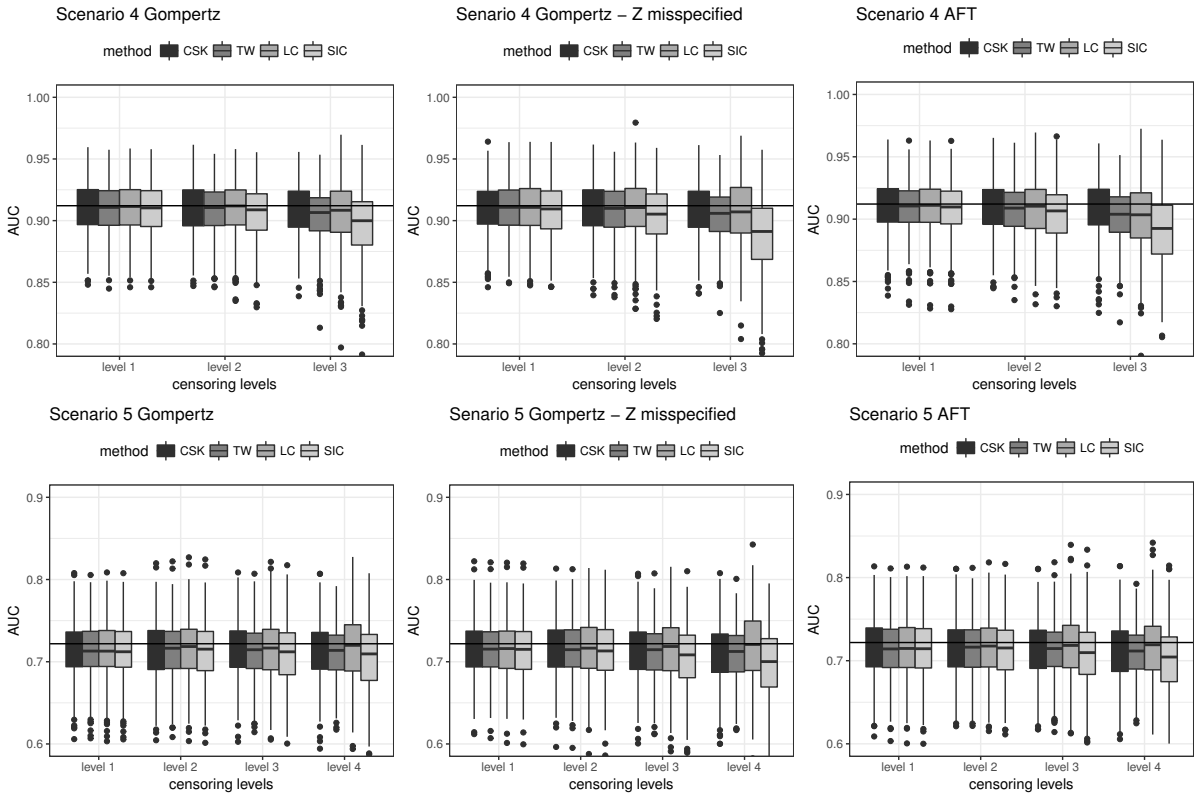

Figure 7: *Boxplots of the AUC for scenarios 4 and 5 for* $n = 250$.

5 Application

In order to illustrate our methodology on real data, we use a melanoma data set coming from the textbook by Andersen et al. (1993). This data set consists of 205 patients diagnosed with malignant melanoma (skin cancer) during the period 1962–1977, and who experienced a radical operation (complete removal of the tumour together with the skin within a distance of about 2.5cm around it) performed at the Plastic Surgery department of the University Hospital of Odense in Denmark. All patients have been followed since surgery and until the end of the year 1977. The endpoint of interest is the death from melanoma. Among the 205 patients, 57 died from the melanoma. Figure 8 represents the Kaplan-Meier estimator of the survival function. As it can be seen, it levels off at around 65% and there is a long plateau of 2227 days (approximately 6 years), which contains 23% of the censored observations. These two elements are indicative of the presence of a cure fraction alongside the contextual evidence. Indeed, melanoma belongs to the group of cancers for which it is known that some

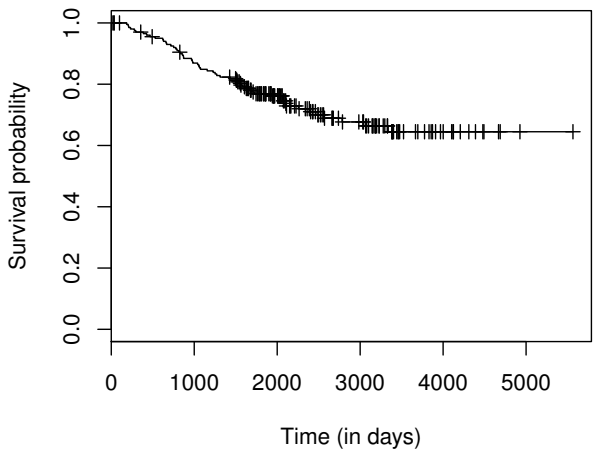

Figure 8: *Kaplan-Meier estimator of the survival function for the melanoma dataset.*

patients get cured. It seems therefore that there exists a cure fraction for patients suffering from melanoma and that this data set contains such subjects.

Alongside the survival time, three covariates are available, the thickness of the tumour (in millimetres (mm), ranges from 0.10 to 17.47 mm with a median value of 1.94mm), a binary variable indicating whether the tumour was ulcerated or not (0 = absence - 115 patients, 1 = presence - 90 patients) and the gender of the patients (0 = female - 126 patients, 1 = male - 79 patients). Our objective is to assess if these three variables are good predictors

Table 3: *Parameter estimates for the melanoma data set based on the LC cure model.*

	incidence			*latency*		
covariates	estimate	std error	p-value	estimate	std error	p-value
intercept	-1.7759	0.3079	< 0.0001	-	-	-
ulceration[pres. vs. abs.]	1.0493	0.4261	0.0138	0.2658	0.4132	0.5201
log(tum. thick.)	0.5329	0.2341	0.0282	0.6703	0.2751	0.0148
gender[male vs. female]	0.2971	0.3845	0.4397	0.6215	0.4035	0.1235

of the cure status. Before computing the ROC curve, we first estimate a LC mixture cure model with the three covariates included in both parts of the model. The results are given

in Table 3. Note that we fit the model considering the logarithm of the tumour thickness. As it can be seen, by considering a 5% level, both the ulceration and the thickness of the tumour significantly affect the cure probability. An absence of ulceration and a thinner tumour increase the cure probability.

We then compute the ROC curve to assess the predictive performance of a linear combination of the three covariates according to the cross-validation procedure described in Section 4. We consider a LC cure model to estimate the classifier and we compute the ROC based on the estimator (2.12), assuming both a LC cure model and a SIC cure model for W_0 and W_1. Figure 9 (a) provides the graphical representation of the two curves corresponding to

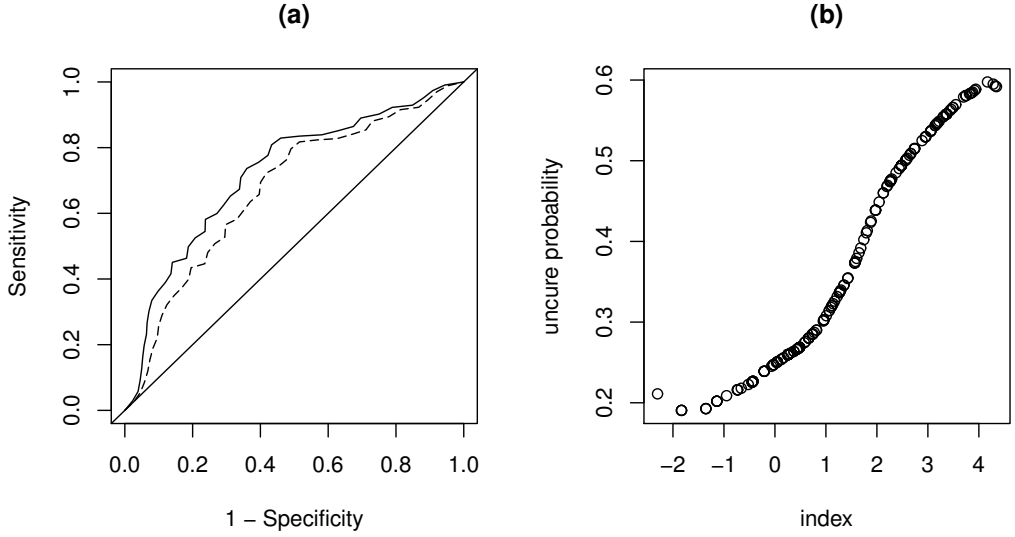

Figure 9: *(a) ROC curve estimates - solid curve: LC cure model, dashed curve: SIC cure model - (b) Estimated link function for the SIC cure model.*

the mean ROC curves over the five runs of the cross-validation. Their respective AUC are 0.7257 and 0.6753, showing that the ROC curve based on the SIC cure model is lower than the ROC curve obtained from the LC cure model. In fact, the proportion of observations censored before τ represents 55.6% of the total number of observations. Furthermore, as it can be seen from Figure 9 (b), the estimated link function for the SIC cure model shows that a logistic model for the incidence is appropriate. As it has been demonstrated in Section 4, when there is a substantial number of observations censored before τ and when the true link

function is a logistic one, the perfomance of the ROC curve based on a SIC cure model is less good than that of the ROC curve based on a LC cure model. Therefore, it seems that the former one has some more difficulties to correctly evaluate the discrimination ability of the classifier considered here. Nevertheless, the two ROC curves are quite close and they indicate that the tumour thickness, tumour ulceration and the gender of the patient only moderately discriminate cured from uncured patients.

6 Concluding remarks

In this paper we proposed a method to assess cure status prediction from survival data using ROC curves. Based on Bayes' theorem, we derived estimators for the sensitivity and the specificity taking the form of weighted empirical distribution functions. We proposed to estimate the weights based on the so-called mixture cure model, assuming both a LC and a SIC cure model. We further developed an estimator of the area under the curve, and we derived the asymptotic properties of the proposed estimators. Through an extensive simulation study we showed that our proposal performs well when the censoring rate is reasonably high and when not too many censored observations are below τ, both when the classifier is known and unknown. When many censored observations have a follow-up time lower than τ, however, our proposal shows some difficulties when a SIC cure model is considered to compute the weights and when the true model for the incidence is a logistic regression model. In such a case, assuming a LC cure model provides more accurate results compared to the infeasible competitors. We further investigated the effect of a misspecification of the weights, both at the covariate and at the modelling levels. We have seen that the performance of our proposal is only slightly affected and that when the proportion of censored observations below τ is not too large, the LC particularly still performs very well. In summary, censoring is the element affecting the most the performance of our proposal and we therefore recommend to be cautious when the censoring rate is high and when many observations are censored before τ. We further recommend to check the model in the incidence since, as we have seen, when the true link function is not logistic, the LC cure model can provide bad results when there are many censored observations below τ.

Throughout this article, we supposed that M is a linear combination of variables. However, it is possible to extend our proposal to the case where the classifier would be obtained

from a different model. Further investigation would be necessary to assess the impact of such a situation on the computation of the weights, but our proposal is not restricted to the linear case. Furthermore, we have considered mixture cure models to compute $\hat{W}_0$ and $\hat{W}_1$. However, a promotion time cure model, such as the model proposed by Tsodikov (1998), could also be considered to estimate a ROC curve for the cure status prediction from survival data.

7 References

Amico, M. and Van Keilegom, I. (2018). Cure models in survival analysis. *Annual Reviews of Statistics and Its Applications*, **5**, 311–342.

Amico, M., Van Keilegom, I., and Legrand, C. (2018). The single-index / Cox mixture cure model. *Under revision for Biometrics*.

Andersen, P. K., Borgan, R., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

Berkson, J. and Gage, R.P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**, 501–515.

Blanche, P., Dartigues, J.-F., and Jacqmin-Gadda, H. (2013). Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal*, **55**, 687–704.

Boag, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society - Series B*, **11**, 15–53.

Copas, J.B. and Corbett, P. (2002). Overestimation of the Receiver operating characteristic curve for logistic regression. *Biometrika*, **89**, 315–331.

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.

Farewell, V.T. (1977). A model for binary variable with time-censored observations. *Biometrika*, **64**, 43–46.

Farewell, V.T. (1982). The use of a mixture model for the analysis of survival data with long-term survivors. *Biometrics*, **38**, 1041–1046.

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

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., Califf, R. M., Pryor, D. B., and Rosati, R. A. (1984). Regression modeling strategies for improved prognostic prediction. *Statistics in Medicine*, **3**, 143–152.

Hastie, T., Tibshirani, R. and Friedman, R. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition.* Springer Series in Statistics, Springer.

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.

Krzanowski, W.J. and Hand, D.J. (2009). *ROC Curves for Continuous Cata. Monographs on Statistics and Applied Probability*, Chapman & Hall/CRC, Boca Raton.

Kuk, Y.C. and Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, **79**, 531-541.

Li, L., Greene, T., and Hu, B. (2018) A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data. *Statistical Methods in Medical Research*, **27**, 2264–2278.

Lu, W. (2008). Maximum likelihood estimation in the proportional hazards cure model. *Annals of the Institute of Statistical Mathematics*, **60**, 545–574.

Mehari Beyene, K., El Ghouch, A. and Oulhaj, A.(2018). On the validity of time-dependent AUC estimation in the presence of a cure fraction. *Submitted*.

Peng, Y. and Dear, K.B.G. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, **56**, 237–243.

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

Sherman, R.P. (1994). Maximal inequalities for degenerate U-processes with applications to optimization estimators. *The Annals of Statistics*, **22**, 439–459.

Sy, J.P. and Taylor, J.M.G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, **56**, 227–236.

Taylor, J.M.G. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics*, **51**, 899–907.

Tsodikov, A. (1998). A proportional hazards model taking account off long-term survivors. *Biometrics*, **54**, 1508–1516.

Van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer, New-York.

Yakovlev, A.Y., Tsodikov, A.D. , and Asselain, B. (1996). Stochastic models of tumor latency and their biostatistical applications. Vol. 1 of Mathematical Biology and Medicine, World Scientific, Singapore.

Zhang, Y. and Shao, Y. (2018). Concordance measure and discriminatory accuracy in transformation cure models. *Biostatistics*, **19**, 14–26.

Appendix 1

This appendix contains the boxplots of the L1 distance and the AUC for all settings when $n = 500$.

Known classifier

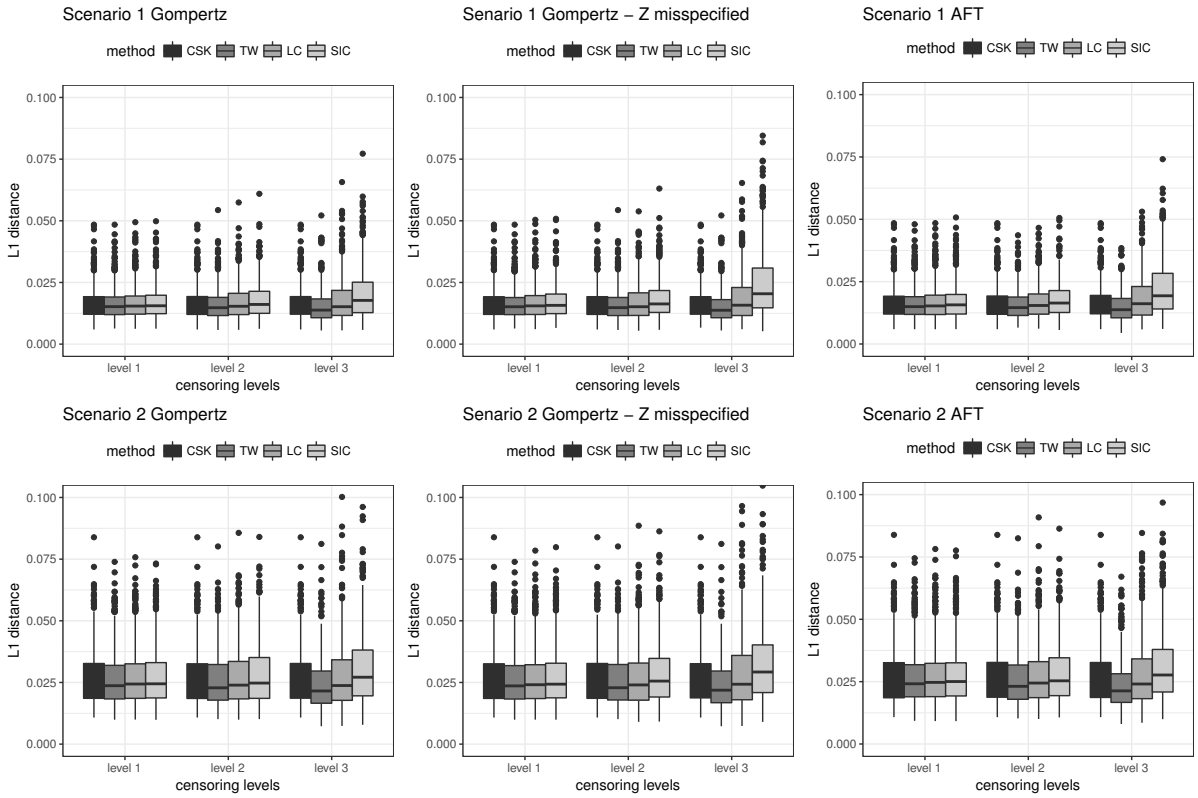

Figure 10: *Boxplots of the L1 distances for scenarios 1 and 2 for* $n = 500$.

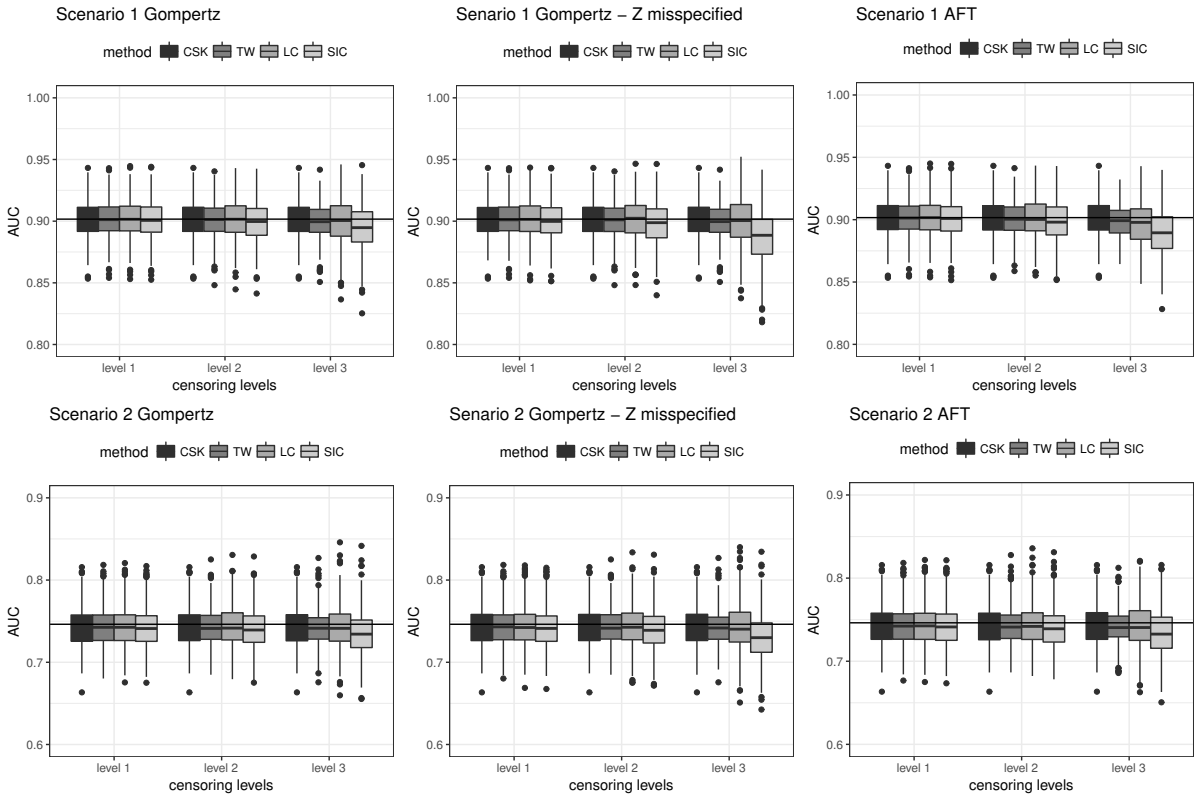

Figure 11: *Boxplots of the AUC for scenarios 1 and 2 for* $n = 500$.

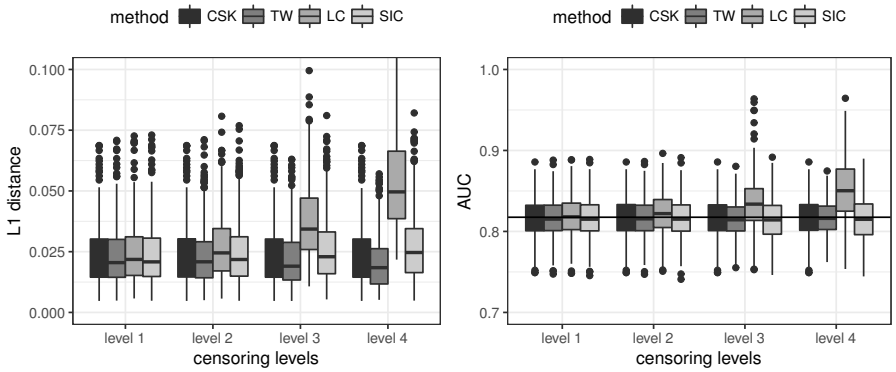

Figure 12: *Boxplots of the L1 distance and the AUC for scenario 3 Gompertz for* $n = 500$.

Unknown classifier

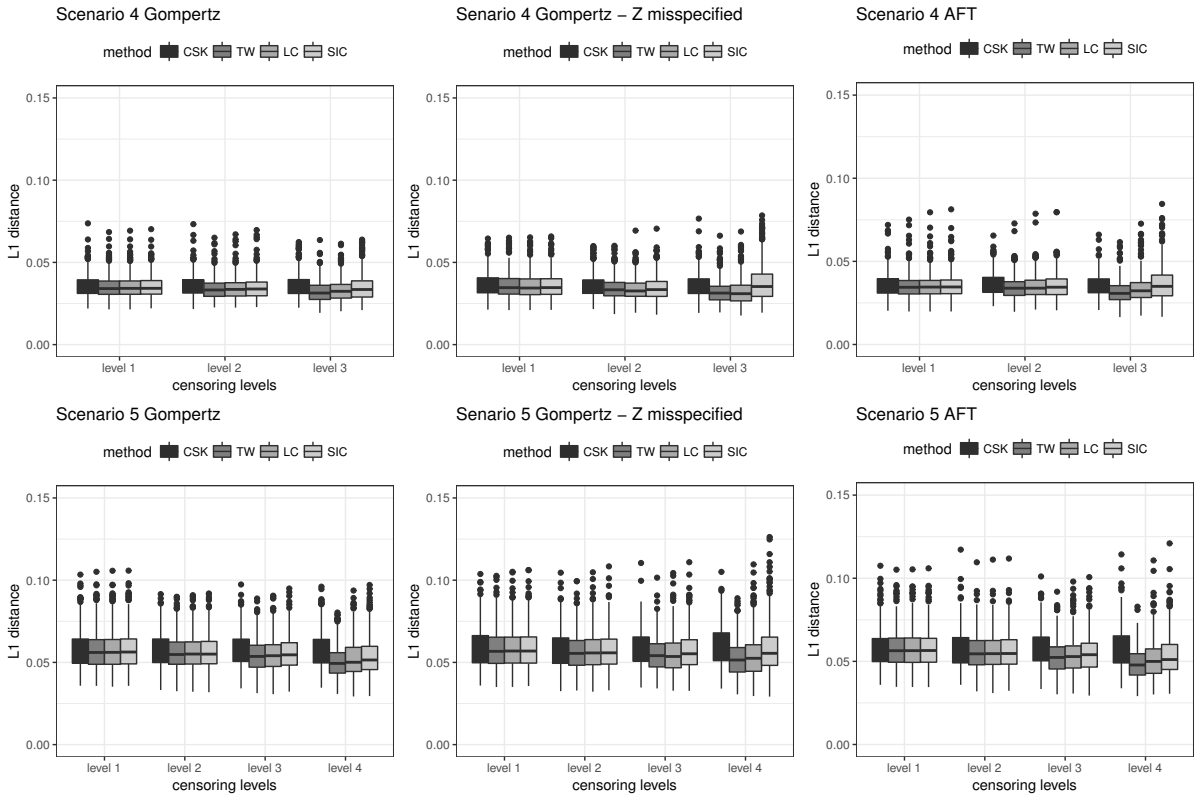

Figure 13: *Boxplots of the L1 distances for scenarios 4 and 5 for* $n = 500$.

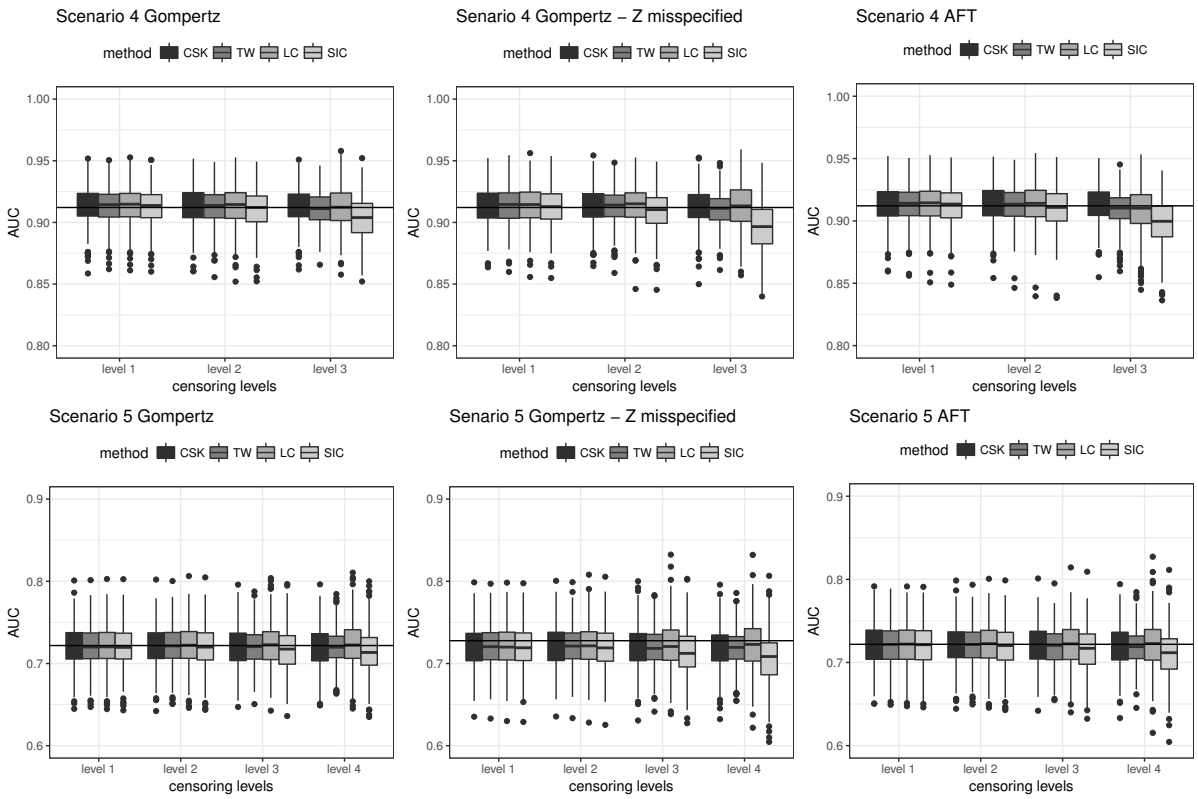

Figure 14: *Boxplots of the AUC for scenarios 4 and 5 for* $n = 500$.

Appendix 2 : Proofs

This appendix contains the proofs of Theorem 3.1 and Corollary 3.1.

Proof of Theorem 3.1. Write

$$\begin{aligned}\hat{S}e(k) - Se(k) &= [\hat{S}e(k) - \tilde{S}e(k)] + [\tilde{S}e(k) - Se(k)] \\ &= S_1(k) + S_2(k) \quad \text{(say)}.\end{aligned}$$

Note that it follows from (2.7) that under the LC mixture cure model,

$$Se(k) = \frac{E[W_1 I(M > k)]}{E(W_1)} := \frac{N(k)}{D}, \quad \text{with} \quad W_1 = \frac{\{1 - p(\mathbf{X})\}(1-\Delta)}{1 - p(\mathbf{X}) + p(\mathbf{X})S_u(Y|\mathbf{Z})}.$$

Similarly, write

$$\hat{S}e(k) = \frac{\hat{N}(k)}{\hat{D}} = \frac{n^{-1}\sum_{i=1}^n \hat{N}_i(k)}{n^{-1}\sum_{i=1}^n \hat{D}_i} \quad \text{and} \quad \tilde{S}e(k) = \frac{\tilde{N}(k)}{\tilde{D}} = \frac{n^{-1}\sum_{i=1}^n \tilde{N}_i(k)}{n^{-1}\sum_{i=1}^n \tilde{D}_i}.$$

Then,

$$\begin{aligned}S_2(k) &= \frac{\tilde{N}(k) - N(k)}{\tilde{D}} + \Big(\frac{1}{\tilde{D}} - \frac{1}{D}\Big)N(k) \\ &= \Big\{\frac{\tilde{N}(k) - N(k)}{D} - \frac{N(k)}{D^2}(\tilde{D} - D)\Big\}\big\{1 + o_P(1)\big\} \\ &= \Big\{\frac{1}{D}n^{-1}\sum_{i=1}^n(\tilde{N}_i(k) - E\tilde{N}(k)) - \frac{N(k)}{D^2}n^{-1}\sum_{i=1}^n(\tilde{D}_i - E\tilde{D})\Big\}\big\{1 + o_P(1)\big\}, \quad (7.1)\end{aligned}$$

since $D = E\tilde{D}$ and $N(k) = E\tilde{N}(k)$, which is a sum of zero-mean i.i.d. terms indexed by k.

Next, consider $S_1(k)$. Using a similar derivation as for $S_2(k)$, we have :

$$\begin{aligned}S_1(k) &= \Big\{\frac{1}{D}n^{-1}\sum_{i=1}^n(\hat{N}_i(k) - \tilde{N}_i(k)) - \frac{N(k)}{D^2}n^{-1}\sum_{i=1}^n(\hat{D}_i - \tilde{D}_i)\Big\}\big\{1 + o_P(1)\big\} \\ &= \Big\{S_{11}(k) + S_{12}(k)\Big\}\big\{1 + o_P(1)\big\}.\end{aligned}$$

Let us consider first $S_{11}(k)$. The term $S_{11}(k)$ depends on $\hat{W}_{i1} - \tilde{W}_{i1}$, which equals

$$
\begin{aligned}
&\hat{W}_{i1} - \tilde{W}_{i1} \\
&= (1-\Delta_i)\Big\{\frac{1-\hat{p}(\mathbf{X}_i)}{1-\hat{p}(\mathbf{X}_i)+\hat{p}(\mathbf{X}_i)\hat{S}_u(Y_i|\mathbf{Z}_i)} - \frac{1-p(\mathbf{X}_i)}{1-p(\mathbf{X}_i)+p(\mathbf{X}_i)S_u(Y_i|\mathbf{Z}_i)}\Big\} \\
&:= (1-\Delta_i)\Big\{\frac{\hat{A}_i}{\hat{B}_i} - \frac{A_i}{B_i}\Big\} \\
&= (1-\Delta_i)\Big\{\frac{\hat{A}_i - A_i}{B_i} - \frac{\hat{B}_i - B_i}{B_i^2}A_i\Big\}\big\{1+o_P(1)\big\} \\
&= \frac{1-\Delta_i}{B_i^2}\Big\{ -(\hat{p}(\mathbf{X}_i)-p(\mathbf{X}_i))(1-p(\mathbf{X}_i)+p(\mathbf{X}_i)S_u(Y_i|\mathbf{Z}_i)) \\
&\qquad\qquad +(1-S_u(Y_i|\mathbf{Z}_i))(\hat{p}(\mathbf{X}_i)-p(\mathbf{X}_i))(1-p(\mathbf{X}_i)) \\
&\qquad\qquad -p(\mathbf{X}_i)(1-p(\mathbf{X}_i))(\hat{S}_u(Y_i|\mathbf{Z}_i)-S_u(Y_i|\mathbf{Z}_i))\Big\}\big\{1+o_P(1)\big\} \\
&= \frac{1-\Delta_i}{B_i^2}\Big\{ -(\hat{p}(\mathbf{X}_i)-p(\mathbf{X}_i))S_u(Y_i|\mathbf{Z}_i) - (\hat{S}_u(Y_i|\mathbf{Z}_i)-S_u(Y_i|\mathbf{Z}_i))p(\mathbf{X}_i)(1-p(\mathbf{X}_i))\Big\} \\
&\qquad\qquad \times\big\{1+o_P(1)\big\}.
\end{aligned}
$$

It follows from the proof of Theorem 3 in Lu (2008) that

$$\hat{p}(\mathbf{x}) - p(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n}\xi(\mathbf{X}_j, \mathbf{Z}_j, Y_j, \Delta_j, \mathbf{x}) + o_p(n^{-1/2})$$

uniformly in $\mathbf{x}$, for a certain function ξ satisfying $E(\xi(\mathbf{X}, \mathbf{Z}, Y, \Delta, \mathbf{x})) = 0$ for all $\mathbf{x}$, and

$$\hat{S}_u(t|\mathbf{z}) - S_u(t|\mathbf{z}) = \frac{1}{n}\sum_{j=1}^{n}\zeta(\mathbf{X}_j, \mathbf{Z}_j, Y_j, \Delta_j, t|\mathbf{z}) + o_p(n^{-1/2})$$

uniformly in t and $\mathbf{z}$, for a certain function ζ satisfying $E(\zeta(\mathbf{X}, \mathbf{Z}, Y, \Delta, t|\mathbf{z})) = 0$ for all t and

z. Hence,

$$\begin{aligned}
S_{11}(k) &= \frac{1}{D}\frac{1}{n}\sum_{i=1}^{n}(\hat{W}_{i1}-\tilde{W}_{i1})I(M_i>k) \\
&= \frac{1}{D}\frac{1}{n}\sum_{i=1}^{n}\frac{(1-\Delta_i)I(M_i>k)}{B_i^2}\Big\{-S_u(Y_i|\mathbf{Z}_i)(\hat{p}(\mathbf{X}_i)-p(\mathbf{X}_i)) \\
&\qquad -p(\mathbf{X}_i)(1-p(\mathbf{X}_i))(\hat{S}_u(Y_i|\mathbf{Z}_i)-S_u(Y_i|\mathbf{Z}_i))\Big\}\big\{1+o_P(1)\big\} \\
&= \frac{1}{D}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{(1-\Delta_i)I(M_i>k)}{B_i^2}\Big\{-S_u(Y_i|\mathbf{Z}_i)\xi(\mathbf{X}_j,\mathbf{Z}_j,Y_j,\Delta_j,\mathbf{X}_i) \\
&\qquad -p(\mathbf{X}_i)(1-p(\mathbf{X}_i))\zeta(\mathbf{X}_j,\mathbf{Z}_j,Y_j,\Delta_j,Y_i|\mathbf{Z}_i)\Big\}+o_P(n^{-1/2}) \\
&:= \frac{1}{D}\frac{2}{n(n-1)}\sum_{i<j}\Big\{\frac{1}{2}\Big(h(V_i,V_j,k)+h(V_j,V_i,k)\Big)\Big\}+o_P(n^{-1/2}) \\
&:= \frac{1}{D}\frac{2}{n(n-1)}\sum_{i<j}\tilde{h}(V_i,V_j,k)+o_P(n^{-1/2}).
\end{aligned}$$

We have a U-process of order 2 with symmetric kernel $\tilde{h}$, where $V_i=(\mathbf{X}_i,\mathbf{Z}_i,Y_i,\Delta_i)$. It follows from Corollary 4 in Sherman (1994) that this U-process can be decomposed in its Hajek projection and a remainder term that is uniformly of smaller order :

$$\begin{aligned}
S_{11}(k) &= \frac{2}{D}\frac{1}{n}\sum_{j=1}^{n}E\big[\tilde{h}(V,V_j,k)|V_j\big]+o_P(n^{-1/2}) \\
&= \frac{1}{D}\frac{1}{n}\sum_{j=1}^{n}E\Big[\frac{(1-\Delta)I(M>k)}{B^2}\Big\{-S_u(Y|\mathbf{Z})\xi(\mathbf{X}_j,\mathbf{Z}_j,Y_j,\Delta_j,\mathbf{X}) \\
&\qquad -p(\mathbf{X})(1-p(\mathbf{X}))\zeta(\mathbf{X}_j,\mathbf{Z}_j,Y_j,\Delta_j,Y|\mathbf{Z})\Big\}\Big]+o_P(n^{-1/2}). \qquad (7.2)
\end{aligned}$$

In a similar way, we can show that

$$\begin{aligned}
S_{12}(k) &= -\frac{N(k)}{D^2}\frac{1}{n}\sum_{j=1}^{n}E\Big[\frac{1-\Delta}{B^2}\Big\{-S_u(Y|\mathbf{Z})\xi(\mathbf{X}_j,\mathbf{Z}_j,Y_j,\Delta_j,\mathbf{X}) \\
&\qquad -p(\mathbf{X})(1-p(\mathbf{X}))\zeta(\mathbf{X}_j,\mathbf{Z}_j,Y_j,\Delta_j,Y|\mathbf{Z})\Big\}\Big]+o_P(n^{-1/2}). \qquad (7.3)
\end{aligned}$$

We can now combine the expressions for $S_{11}(k)$, $S_{12}(k)$ and $S_2(k)$, which leads to

$$\begin{aligned}
\hat{Se}(k)-Se(k) &= S_{11}(k)+S_{12}(k)+S_2(k)+o_P(n^{-1/2}) \\
&= \frac{1}{n}\sum_{j=1}^{n}\eta_{Se}(\mathbf{X}_j,\mathbf{Z}_j,Y_j,\Delta_j,k)+o_P(n^{-1/2}), \qquad (7.4)
\end{aligned}$$

uniformly in k, where $\eta_{Se}(\mathbf{X}_j, \mathbf{Z}_j, Y_j, \Delta_j, k)$ is obtained by combining the expressions given in (7.2), (7.3) and (7.1).

Next, it can be shown that the class

$$\{(\mathbf{x}, \mathbf{z}, y, \delta) \to \eta_{Se}(\mathbf{x}, \mathbf{z}, y, \delta, k) : k \in \mathbb{R}\}$$

is Donsker, by decomposing the function η_{Se} in products and sums of subfunctions that are bounded and Donsker (see Examples 2.10.7 and 2.10.8 in Van der Vaart and Wellner, 1996, VW hereafter). For this, it suffices to show that the bracketing number of the classes corresponding to each of these subfunctions is small enough, in the sense of Theorem 2.5.6 in VW. For calculating these bracketing numbers, one can use the well known results about the bracketing number of classes of bounded and monotone functions (see Theorem 2.7.5 in VW), classes of sufficiently smooth functions (see Corollary 2.7.2 in VW), or related results.

Finally, in a similar way as for the specificity, it can be shown that the estimator $\hat{Sp}(k)$ of the specificity can be decomposed in a sum of iid terms, plus a remainder term that is uniformly of smaller order :

$$\hat{Sp}(k) - Sp(k) \;=\; \frac{1}{n}\sum_{j=1}^{n} \eta_{Sp}(\mathbf{X}_j, \mathbf{Z}_j, Y_j, \Delta_j, k) + o_P(n^{-1/2}), \tag{7.5}$$

where η_{Sp} is obtained by replacing in the expression of η_{Se} all indicators $I(M > k)$ by $I(M \leq k)$ and by noting that $W_0 = 1 - W_1$. □

Proof of Corollary 3.1. Write

$$\begin{aligned} \hat{ROC}(u) - ROC(u) &= \big[\hat{Se}\{(1-\hat{Sp})^{-1}(u)\} - Se\{(1-\hat{Sp})^{-1}(u)\}\big] \\ &\quad + \big[Se\{(1-\hat{Sp})^{-1}(u)\} - Se\{(1-Sp)^{-1}(u)\}\big] \\ &:= T_1(u) + T_2(u). \end{aligned}$$

We start with $T_2(u)$:

$$\begin{aligned} T_2(u) &= Se'\big\{(1-Sp)^{-1}(u)\big\}\Big\{(1-\hat{Sp})^{-1}(u) - (1-Sp)^{-1}(u)\Big\} + o_P(n^{-1/2}) \\ &= Se'\big\{(1-Sp)^{-1}(u)\big\}\Big[-\frac{(1-\hat{Sp})\{(1-Sp)^{-1}(u)\} - u}{(1-Sp)'\{(1-Sp)^{-1}(u)\}}\Big] + o_P(n^{-1/2}). \end{aligned}$$

We know that

$$\hat{Sp}(k) - Sp(k) \;=\; n^{-1}\sum_{i=1}^{n} \eta_{Sp}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, k) + o_P(n^{-1/2}),$$

uniformly in k. It follows that

$$T_2(u) \;=\; \frac{Se'\{(1-Sp)^{-1}(u)\}}{(1-Sp)'\{(1-Sp)^{-1}(u)\}} n^{-1} \sum_{i=1}^{n} \eta_{Sp}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, (1-Sp)^{-1}(u)) + o_P(n^{-1/2}),$$

uniformly in $\delta \le u \le 1-\delta$. Next,

$$T_1(u) \;=\; \hat{Se}\{(1-Sp)^{-1}(u)\} - Se\{(1-Sp)^{-1}(u)\} + o_P(n^{-1/2}),$$

since it follows from the weak convergence of $\hat{Se} - Se$ that

$$\sup_{|k_2-k_1| \le K n^{-1/2}} |\hat{Se}(k_2) - Se(k_2) - \hat{Se}(k_1) + Se(k_1)| = o_p(n^{-1/2})$$

for $0 < K < \infty$, and

$$\sup_{\delta \le u \le 1-\delta} |(1-\hat{Sp})^{-1}(u) - (1-Sp)^{-1}(u)| \;=\; O_P\Big(\sup_k |\hat{Sp}(k) - Sp(k)| \Big) = O_P(n^{-1/2}).$$

It follows that

$$T_1(u) \;=\; n^{-1} \sum_{i=1}^{n} \eta_{Se}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, (1-Sp)^{-1}(u)) + o_P(n^{-1/2}).$$

Hence,

$$\begin{aligned}
&\hat{ROC}(u) - ROC(u) \\
&= n^{-1} \sum_{i=1}^{n} \Big[\eta_{Se}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, (1-Sp)^{-1}(u)) \\
&\quad + \frac{Se'\{(1-Sp)^{-1}(u)\}}{(1-Sp)'\{(1-Sp)^{-1}(u)\}} \eta_{Sp}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, (1-Sp)^{-1}(u)) \Big] + o_P(n^{-1/2}),
\end{aligned}$$

uniformly in $\delta \le u \le 1-\delta$. We know that the classes

$$\big\{ (\mathbf{x}, \mathbf{z}, y, \delta) \to \eta_{Se}(\mathbf{x}, \mathbf{z}, y, \delta, k) : k \in \mathbb{R} \big\}$$

and

$$\big\{ (\mathbf{x}, \mathbf{z}, y, \delta) \to \eta_{Sp}(\mathbf{x}, \mathbf{z}, y, \delta, k) : k \in \mathbb{R} \big\}$$

are Donsker, and that the functions $u \to (1-Sp)^{-1}(u)$ and $k \to Se'(k)/(1-Sp)^{-1}(k)$ are continuously differentiable and bounded. Hence, the weak convergence of the process $n^{1/2}\{\widehat{ROC}(u) - ROC(u)\}$ $(\delta \le u \le 1-\delta)$ follows.

It remains to show the limiting distribution of $n^{1/2}(\hat{AUC}_\delta - AUC_\delta)$. Note that

$$\begin{aligned}
\widehat{AUC}_\delta - AUC_\delta &= \int_\delta^{1-\delta} \left\{\widehat{ROC}(u) - ROC(u)\right\} du \\
&= n^{-1} \sum_{i=1}^{n} \int_\delta^{1-\delta} \eta_{ROC}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i, u)\, du + o_P(n^{-1/2}) \\
&:= n^{-1} \sum_{i=1}^{n} \eta_{AUC}(\mathbf{X}_i, \mathbf{Z}_i, Y_i, \Delta_i) + o_P(n^{-1/2})
\end{aligned}$$

and that

$$\mathrm{Var}(\eta_{AUC}(\mathbf{X}, \mathbf{Z}, Y, \Delta)) = \int_\delta^{1-\delta} \int_\delta^{1-\delta} \mathrm{Cov}\{\eta_{ROC}(\mathbf{X}, \mathbf{Z}, Y, \Delta, u_1), \eta_{ROC}(\mathbf{X}, \mathbf{Z}, Y, \Delta, u_2)\}\, du_1\, du_2.$$

This finishes the proof. □

FACULTY OF ECONOMICS AND BUSINESS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be