On the validity of time-dependent AUC estimation in the presence of cure fraction

Kassu M. Beyene[1] | **Anouar El Ghouch**[1] | **Abderrahim Oulhaj**[2]

[1]Institute of Statistics, Biostatistics and Actuarial Sciences, Catholic University of Louvain, Louvain la Neuve, Belgium

[2]Institute of Public Health, College of Medicine and Health Sciences, UAE University, Al-Ain, United Arab Emirates

Abstract

During the last decades, several approaches have been proposed to estimate the time-dependent area under the receiver operating characteristic curve (AUC) of risk tools derived from survival data. The validity of these estimators relies on some regularity assumptions among which a survival function being proper. In practice, this assumption is not always satisfied because a fraction of the population may not be susceptible to experience the event of interest even for long follow-up. Studying the sensitivity of the proposed estimators to the violation of this assumption is of substantial interest. In this paper, we investigate the performance of a nonparametric simple estimator, developed for classical survival data, in the case when the population exhibits a cure fraction. Motivated from the current practice of deriving risk tools in oncology and cardiovascular disease prevention, we also assess the loss, in terms of predictive performance, when deriving risk tools from survival models that do not acknowledge the presence of cure. The simulation results show that the investigated method is valid even under the presence of cure. They also show that risk tools derived from survival models that ignore the presence of cure have smaller AUC compared to those derived from survival models that acknowledge the presence of cure. This was also attested with a real data analysis from a breast cancer study.

KEYWORDS

Cox model, mixture cure models, promotion time models, risk tools, time-dependent AUC

1 | INTRODUCTION

Prognostic tools derived from survival models are becoming more commonly used in many areas. In healthcare applications, predicting patient's risk of developing a disease of interest within a prespecified time horizon is usually of major interest. It assists in decision making such as optimizing clinical management of patients, defining better strategies for treatments, assessing cost effectiveness of potential therapies from health economic point of view, and assisting in recruitment of patients, in clinical trials, by identifying subjects with higher risk of developing the outcome of interest. In cardiovascular disease (CVD) research, for instance, many risk predictions tools based on survival models were developed and made publicly available during the last three decades; see, for example, Damen et al. (2016). The general aim of these tools is to predict the 10 years risk of developing any CVD event or some specific components of CVD such as stroke, myocardial infarction, or CV death. Some of these tools, derived from well-known prospective cohort studies such as the Framingham study, are currently playing a central role in CVD prevention guidelines such as the European guideline on cardiovascular disease prevention in clinical practice (ESC/EAS, 2016).

In oncology, similar publicly available tools were developed to predict risk of specific cancer type (e.g., breast cancer) or risk of developing more than one cancer type; see, for example, Rabin et al. (2013).

Due to the increasing role of risk predictions models in clinical practice, it is very important to assess their predictive performance before implementation. One common way of assessing their performance is by calculating some overall performance measures such as sensitivity, specificity, receiver operating characteristic (ROC) curve, and its corresponding area under the ROC curve (AUC). Prognostic tools derived from survival models usually aim at predicting the risk for the event of interest to occur within a fixed period of time usually chosen based on some clinical justifications. This makes sensitivity, specificity, ROC curve and its corresponding AUC become time dependent. Etzioni, Pepe, Longton, Hu, and Goodman (1999), Heagerty, Lumley, and Pepe (2000), Heagerty and Zheng (2005), and Slate and Turnbull (2000), among others, proposed several definitions for the time-dependent sensitivity and specificity in the case of survival models such as cumulative sensitivity and dynamic specificity (C/D), incident and dynamic and incident and static . The C/D is, however, the most widely used definition in clinical applications as it has more clinical relevance (Blanche, Dartigues, & Jacqmin-Gadda, 2013; Lambert & Chevret, 2016). Therefore, in this article we focus on the C/D definition for investigating predictive accuracy in survival models.

Let us assume that a prognostic tool Z is available for predicting the binary outcome $D(t) = I(T \leq t)$, where T is the time to event and t is a given time horizon. The prognostic tool Z could be a known single marker, a combination of known markers or a risk tool derived from a working statistical model. Without loss of generality, we use the conventional assumption that increasing values of Z are indicative of higher chances of experiencing the event of interest. Heagerty et al. (2000) used the C/D sensitivity and specificity and defined the time-dependent true-positive rate (TPR_t) and false-positive rate (FPR_t) as

$$\begin{aligned} TPR_t(z) &= \mathrm{P}(Z > z | T \leq t), \\ FPR_t(z) &= \mathrm{P}(Z > z | T > t), \end{aligned} \tag{1}$$

where z is a fixed threshold and t is a time horizon. The time-dependent ROC curve of Z, denoted by $ROC_{Z,t}$, is defined as

$$ROC_{Z,t}(u) = TPR_t\{FPR_t^{-1}(u)\}, \tag{2}$$

where $u \in (0, 1)$ and $z = FPR_t^{-1}(u)$ is the threshold values such that $u = FPR_t(z)$.

The corresponding area under the ROC curve that measures the discrimination ability of the prognostic tool Z can be derived as

$$AUC_Z(t) = \int_0^1 ROC_{Z,t}(s)ds.$$

The $AUC_Z(t)$ ranges between 0.5 and 1 with a value 1 indicating a perfect discrimination between those who experienced the event and those who did not and a value 0.5 for random discrimination. In general, increasing values of AUC are indicative of better prognostic ability of Z (Pepe, 2003).

In almost every follow-up study, some individuals may be lost during the follow-up period due to various reasons, or may not develop the event of interest before the end of the study. Such situations lead to censored observations since complete information is not available for these subjects. As a result, it is not straightforward to estimate the time-dependent AUC as the event status at the prediction time t remains unknown among censored subjects. During the last decades, several authors have proposed estimation methods for the time-dependent AUC that takes censoring into account. Heagerty et al. (2000) introduced two methods for time-dependent ROC curve. The first one is based on the Kapan–Meier estimator, whereas the second is based on the nearest neighbor estimation (NNE) method. Chambless and Diao (2006) proposed an alternative method to the Kaplan–Meier approach. This is based on a recursive calculation of the ordered times of the event. The advantage of this method is that the ROC estimator is monotone, bounded in [0, 1] and does not involve any smoothing parameter. Uno, Cai, Tian, and Wei (2007) and Hung and Chiang (2010) independently proposed inverse probability of censoring weighting (IPCW) method to correct the Kaplan–Meier estimate of the sensitivity. Blanche et al. (2013) modified the IPCW and proposed a method called conditional IPCW. The advantage of the latter over IPCW is that it does not assume independence between the prognostic tool and the censoring time. Martínez-Camblor, Bayón, and Pérez-Fernández (2016) and Li, Greene, and Hu (2018) independently proposed another method for ROC curve estimation. In their method, the missing status indicator for the censored subjects is replaced by weights obtained from the conditional survival function. However, their method was dependent on how the conditional survival function is estimated. Martínez-Camblor et al. (2016) used a semiparametric Cox model and Kaplan–Meier estimator for conditional survival function estimation, whereas Li et al. (2018) used a weighted Kaplan–Meier estimator. In this paper, we use the estimator proposed by Li et al. and hereafter we call it nonparametric simple (NPS) estimator. The advantage

of this method is that it uses the available information more efficiently than many (if not all) other methods available in the literature like the ICPW and the NNE methods. Moreover, this method is relatively insensitive to the choice of the smoothing parameter. More details about this method and the results of a numerical study demonstrating its superiority can be found in Li et al. (2018). Many other methods to estimate the time-dependent AUC exist in the literature; see, for example, Martínez-Camblor and Pardo-Fernández (2018) and the references given therein. For a comprehensive summary of the time-dependent ROC curve estimation methods, the reader is referred to Blanche et al. (2013) and Kamarudin, Cox, and Kolamunnage-Dona (2017).

The estimators of the time-dependent AUC presented above were developed under the assumptions of classical survival models such as noninformative censoring and correctly specified model. Schmid and Potapov (2012) conducted some extensive simulations to study the effects of model misspecification and nonrandom censoring on the behaviors of the estimators of the classification accuracy. They found that violations of these assumptions may generate nonignorable bias. Schmid, Kestler, and Potapov (2013) also conducted a similar study to assess the validity of time-dependent AUC estimators, and they indicated that violations of these regularity assumptions may result in overoptimistic estimates of prediction accuracy.

In the classical survival models, proportional hazards and random censoring are not the only assumptions. Another strong assumption is that all subjects of the study population are susceptible to the event of interest and will eventually experience the event if the follow-up period is sufficiently long. This assumption may not be valid in many situations. In oncology, for example, if the treatment is successful (which might be the case sometimes) the subject will be cured and hence will not have a recurrent event even if the follow-up period is large. This leads to an infinite time to event for cured subjects. In order to obtain valid estimates and inference, it is then necessary to take this into account when modeling survival data. During the last decades, different methods for modeling survival data with cure fraction that yield valid regression coefficient estimates and inferences were proposed. For a recent review on cure models, see, for example, Amico and Van Keilegom (2018) and Taweab and Ibrahim (2014).

Many studies have shown that fitting standard survival models when the data include cured subjects might result in invalid inference (Taweab & Ibrahim, 2014). To the best of our knowledge, no study has yet addressed the effect of the presence of a cure fraction on the properties of the estimates of the time-dependent AUC. Our main aim in this paper is to address this issue by investigating the following two questions: (1) Is it still valid to use the NPS estimator when there are cured subjects in the study population? The reasons to use this method is that it is a nonparametric method, it efficiently uses the available information, and it outperforms many of other existing methods including the ICPW method (Li et al., 2018). (2) What are the consequences of deriving a prognostic tool from a survival model that does not acknowledge the presence of cured subject? More precisely, we would like to see what is the cost, in terms of optimality/efficiency, of using a survival working model that do not acknowledge the presence of cured subjects. This question is motivated by the fact that in the current practice, even though a proportion of subjects is not susceptible to the event of interest, it is common to use classical survival models such as the Cox proportional hazard model to derive prognostic tools. So the main question here is about the robustness (to cure) of the classical survival model, which is the most commonly used in practice, as a working model for deriving prognostic tools. This paper is organized in five sections including the introduction. Section 2 introduces the methodology we used in this study and also defines and discusses time-dependent classification accuracy in the presence of cure fraction. In Section 3, the simulation design and results are presented, and in Section 4, we provide real example using data from breast cancer study. Finally, discussions and conclusions are presented in Section 5.

2 | METHOD

In this section, we first briefly describe the nonparametric simple time-dependent ROC curve and the AUC estimation in the presence of cure fraction. We afterwards give a short description regarding prognostic tools derivation and survival models with a cure rate.

Let us assume that a data set of size n is available in order to assess the predictive accuracy of a prognostic tool of interest. More specifically, the data will consist of $\{(Y_i, \Delta_i, \mathbf{X}_i); i = 1, 2, \dots, n\}$, where $Y_i = \min(T_i, C_i)$ is the observed time, T_i the time to event, C_i the censoring time, and $\Delta_i = I(T_i \leq C_i)$ the censoring indicator. $\mathbf{X}_i = (X_{1i}, \dots, X_{di})$, $d \geq 1$, is a d-dimensional vector of patient-specific covariates thought to have a potential influence on T. The event status at time t is denoted by $D_i(t) = I(T_i \leq t)$ with 1 indicating that the event happened by time t. The prognostic tool Z, for which the predictive ability is to be investigated, can be any univariate function of $\mathbf{X}$. This function can be either known (e.g., subjective summary rule of expert physicians) or derived (estimated) from a working statistical model fitted to the observed data. More details about prognostic tools derivation are given in Section 2.2.

2.1 | The estimation of ROC curve and AUC in survival models: The nonparametric simple method

The NPS method was proposed for estimating the ROC curve and AUC after reformulating the method of Heagerty et al. (2000) by introducing a weight variable representing the event status of the subject during follow-up (Li et al., 2018; Martínez-Camblor et al., 2016). The idea behind this method is similar to the idea of imputation in the context of missing data. They proposed to impute the event status in censored subjects using the survival probability conditional to the observed data. More specifically, the weights are given by

$$W_i(t) = \mathrm{P}(T_i \leq t | Y_i, \Delta_i, Z_i) = \left[1 - (1 - \Delta_i)\frac{S(t|Z_i)}{S(Y_i|Z_i)}\right] I(Y_i \leq t),$$

where $S(t|Z) = \mathrm{P}(T > t|Z)$ denotes the conditional survival distribution of T given the prognostic tool Z, which can be estimated using the weighted Kaplan–Meier estimator of Beran with either nearest neighbor or kernel weights; see Beran (1981).

For a given time horizon t and given a sample (Y_i, Δ_i, Z_i) of size n, the time-dependent TPR and FPR are estimated as follows:

$$\hat{TPR}_t(z) = \hat{\mathrm{P}}(Z > z | T \leq t) = \frac{\sum_i \hat{W}_i(t) I(Z_i > z)}{\sum_i \hat{W}_i(t)},$$

$$\hat{FPR}_t(z) = \hat{\mathrm{P}}(Z > z | T > t) = \frac{\sum_i (1 - \hat{W}_i(t)) I(Z_i > z)}{\sum_i (1 - \hat{W}_i(t))},$$

where $I(.)$ is the indicator function and $\hat{W}_i(t)$ is the Beran's estimate of the weight function defined above.

The time-dependent ROC curve can be obtained by plotting the sensitivity versus 1-specificity for all possible cutoff values z of the prognostic tool Z, and the corresponding time-dependent AUC can be calculated by using numerical integration such as the trapezoidal method. An alternative way for calculating the AUC without involving numerical integration is to use the Mann–Whitney U statistic as it has close connection with the AUC. In fact, Hanley and McNeil (1982) showed that the AUC is equivalent to the probability that a randomly selected subject with the event of interest has a higher prognostic tool than a randomly chosen subject without the event of interest. Using this, the AUC estimate can be obtained using the following formula:

$$\hat{AUC}_Z(t) = \frac{\sum_i \sum_j \hat{W}_i(t)(1 - \hat{W}_j(t))[I(Z_i > Z_j) + 0.5 \times I(Z_i = Z_j)]}{\sum_i \sum_j \hat{W}_i(t)(1 - \hat{W}_j(t))}, \tag{3}$$

where i and j are the indexes of two independent subjects (Li et al., 2018).

In the presence of cure fraction, the aforementioned time-dependent classification accuracy measures can be defined as follows: Let $B_i = I(T_i < \infty)$ be the latent variable, indicating the uncure status taking values 1 or 0 depending on whether the subject i is uncured or not. In this work, for a given time horizon t, the case and the control groups are taken to be $\{B_i = 1 \text{ and } T_i \leq t\} = \{T_i \leq t\}$ and $\{B_i = 0 \text{ or } T_i > t\} = \{T_i > t\}$, respectively. With these definitions, one gets the so-called cumulative sensitivity and dynamic specificity (see Equation 1), that are largely used in the classical survival analysis (without a cure fraction). The resulting ROC (see Equation 2) is known as the cumulative/dynamic (C/D) ROC; see Heagerty et al. (2000) and Heagerty and Zheng (2005) for more details. This suggests that the NPS estimators described above can be used to estimate the time-dependent classification accuracy measures in the presence of cure fraction.

Note that since the Beran estimator is the main ingredient of NPS estimator, the consistency of the latter holds if the former is consistent. Without a cure fraction, the consistency and asymptotic normality of Beran's estimator were established by several researchers, just to mention a few Beran (1981), Dabrowska (1989), Gonzalez-Manteiga and Cadarso-Suarez (1994), and Van Keilegom and Veraverbeke (1997). With a cure fraction, due to the identifiability issues related to censoring and cure, for consistency to hold, we need, among other technical assumptions, that the support of the censoring variable to be not contained in the support of T. More precisely, we need that $\tau_u < \tau_G(Z)$, with probability one, where $\tau_u = \sup_z\{\inf\{t : S_u(t|z) = 0\}\}$ and $S_u(t|z) = \mathrm{P}(T \geq t | Z = z, T < \infty)$ is the conditional survival function of the uncured subjects, and $\tau_G(z) = \inf\{t : G(t|z) = 0\}$ where $G(t|z)$ is conditional survival function of the censoring time. For this reason, one should restrict the analysis to a time horizon t not larger than τ_u; for more details, see Xu and Peng (2014) and López-Cheda, Cao, Jácome, and Van Keilegom (2017).

2.2 | Derivation of prognostic tools in survival analysis

As stated in the previous section, a prognostic tool can be either known or derived. An example of the first case is the Glasgow Coma Scale (GCS) routinely used in emergency medicine; see, for example, Baršić, Marton, Himbele, and Ravlić (1996). The GCS measures the severity of coma in patients admitted to intensive care unit and is usually used to predict the 24 or 48 hr in-hospital risk of death. It is a straightforward test that assesses the patients ability to open their eyes, move, and speak. The GCS test assesses three main vital components: The motor responses ranging from 1 to 6, the verbal responses ranging from 1 to 5, and the eye opening ranging from 1 to 4 resulting in a total score ranging from 3 to 15. In the second case, no known prognostic tool is available. The common practice is then to derive a prognostic tool by combining information from different potential predictors. Several researchers proposed optimal ways to combine information from different sources (see, e.g., Cai & Cheng, 2008; McIntosh & Pepe, 2002; Qin & Zhang, 2010). The most common way is by fitting a working survival model $\mathcal{W}$ to a set of data collected on the event of interest and potential predictors. An example of this is the United Kingdom prospective diabetes study (UKPDS) risk engine used to predict the 10 years risk of developing coronary heart disease in patients with type 2 diabetes; see Stevens, Kothari, Adler, Stratton, and Holman (2001). This prognostic tool was derived by fitting a parametric survival model ($\mathcal{W}$) to data prospectively collected on type 2 diabetes patients.

It has been shown that, given a set of predictors $\mathbf{X} = (X_1, \ldots, X_d)$, the conditional probability $\mathrm{P}(T \leq t|\mathbf{X})$, or any monotonic increasing transformation of it, is the optimal prognostic tool for predicting $I(T \leq t)$. Optimal here means that it maximizes the AUC (McIntosh & Pepe, 2002). For example, if the data are generated from the model

$$h(T) = \alpha + \boldsymbol{X}^T\boldsymbol{\beta} + \epsilon, \tag{4}$$

where h is an increasing function and ϵ is a random error independent of $\boldsymbol{X}$, then the prognostic tool $Z = \boldsymbol{X}^T\boldsymbol{\beta}$ is optimal. This is also the case for the classical Cox model

$$S^{\mathcal{COX}}(t|\boldsymbol{X}) = \exp(-\Lambda_0(t)\exp(\boldsymbol{X}^T\boldsymbol{\beta})), \tag{5}$$

where $\Lambda_0(t)$ is the cumulative baseline hazard function.

In practice, the parameters of the working model used to define the prognostic tool, for example, $\boldsymbol{\beta}$ in the Cox model, are unknown and need to be estimated from a training/learning data set, using, for example, the (partial)likelihood method. In the following, we denote by $\hat{Z}^{\mathcal{W}}$ the estimated prognostic tool based on the working model $\mathcal{W}$, for example, $\hat{Z}^{\mathcal{COX}} = \mathbf{X}^T\hat{\boldsymbol{\beta}}^{\mathcal{COX}}$, where $\hat{\boldsymbol{\beta}}^{\mathcal{COX}}$ is the partial likelihood estimator of $\boldsymbol{\beta}$ in the Cox model. The prognostic tool $\hat{Z}^{\mathcal{COX}}$ is optimal when the Cox model holds but may have poor accuracy when the proportional hazards assumption fails; for more details, see Cai and Cheng (2008) and Zheng, Cai, and Feng (2006).

2.3 | Survival models with cure

Standard survival models, such as the Cox model or the accelerated failure time model, implicitly assume proper survival functions. This means that any subject is susceptible to the event of interest and will eventually experience that event if the follow period is long enough. This assumption, however, might not be appropriate in many situations. This is, for instance, the case of cancer research when the main endpoint is time to relapse. In this scenario, a fraction of patients can be cured and hence will never relapse. Informally, long and stable plateau at the tail of the Kaplan–Meier estimate is an indication for the existence of cured fraction of subjects. Fitting standard survival models when the data include cured subjects might result in invalid inference (Taweab & Ibrahim, 2014). To overcome this, several models widely called *cure models* were proposed for analyzing survival data with cure fraction. The two well-known cure models are the mixture cure model (MCM) and the promotion time model.

The MCM was first introduced by Boag (1949) to study mouth cancer. In this model, it is assumed that the population is a mixture of a cured and noncured subjects. The MCM is a combination of two models, one for the probability of cure sometimes called "the incidence model" and one for the event time of uncured subjects sometimes called "the latency model." This model basically allows to investigate the effect of the covariates both on the cure probability and on the survival function of noncured subjects. The survival function in the MCM is given as

$$S^{\mathcal{MCM}}(t|\mathbf{X}) = \mathrm{P}(\mathbf{X})S_u(t|\mathbf{X}) + (1 - \mathrm{P}(\mathbf{X})), \tag{6}$$

where $\mathbf{X}$ is the vectors of covariates, $S^{\mathcal{MCM}}(t|\mathbf{X}) = \mathrm{P}(T > t|\mathbf{X})$ is the conditional population survival function, and $\mathrm{P}(\mathbf{X}) = \mathrm{P}(B = 1|\mathbf{X})$ is the conditional probability of being uncured. The function $S_u(t|\mathbf{X}) = \mathrm{P}(T > t|\mathbf{X}, B = 1)$ is the conditional sur-

vival function of the susceptible subjects. This function can be assumed to belong to a parametric or a semiparametric distribution family like, for example, the widely used family of Cox proportional hazard models.

Yakovlev et al. (1993) developed an alternative to the MCM called a promotion time model also known as a "proportional hazard cure model" or "bounded cumulative hazard model" (BHM). This model was strongly motivated from a biological mechanism assuming the number of cancer cells remaining after treatment to follow a Poisson distribution (Taweab & Ibrahim, 2014; Yakovlev et al., 1993). The survival function in the BHM is given as

$$S^{\mathcal{BHM}}(t|\mathbf{X}) = \exp\{-\theta(\mathbf{X})F(t)\}, \tag{7}$$

where $F(t)$ is a proper cumulative distribution function and $\theta(\mathbf{X})$ is a positive function which captures the effect of the covariates $\mathbf{X}$ on the population survival function and is usually expressed as $\exp(\mathbf{X}'\boldsymbol{\beta})$. Using the equation above, the cure fraction can be derived as $\mathrm{P}(B = 0|\mathbf{X} = \mathbf{x}) = \lim_{t\to\infty} S(t|\mathbf{X} = \mathbf{x}) = \exp\{-\theta(\mathbf{x})\}$. The advantage of this class of cure models, in comparison to the MCM, is that they still allow the proportional hazard structure since the covariates are involved only through the cure rate parameter. Furthermore, the results from this class of models have more meaningful interpretation for some studies and it has a simple structure for survival function which makes the computations attractive (Taweab & Ibrahim, 2014).

For the reasons explained above and for easiness of interpretation, the prognostic tools considered in all the following sections will be in the form $1 - S^{\mathcal{W}}(t|\boldsymbol{X})$, where $\mathcal{W}$ is one of the working models under study, that is COX, MCM, or BHM, and $S^{\mathcal{W}}$ is the corresponding estimated survival function.

3 | SIMULATIONS

The main objective of this simulation study is to assess the validity of the NPS estimator when there are cured subjects in the population under study. We also investigated the loss in terms of predictive ability when deriving prognostic tools from survival models that do not acknowledge the presence of cured subjects. Details of these simulations are provided in the upcoming subsections.

To generate the improper survival time T, we use the following equation:

$$T = T^u + \frac{1 - B}{B},$$

where T^u is a proper survival time (i.e., the time to event variable of the uncured subpopulation) and B is the cure indicator generated from a Bernoulli distribution with probability p. The distribution of T^u and the values of p will change according to our simulation settings and will be specified in the following subsections. In all our simulations, we used the NPS method implemented in `R package tdROC` with the default value (i.e., 0.1) for the smoothing parameter that is needed to calculate the weighted Kaplan–Meier estimate with 0/1 nearest neighbor weights. We also tried different smoothing values and observed minimal differences in the performance measures of the AUC estimates. This confirms that the NPS method is relatively insensitive to the choice of the smoothing parameter as shown in Li et al. (2018).

3.1 | The validity of the NPS estimator under cure: The case of a known prognostic tool

The objective of this section is to investigate the performance of the NPS estimator of the time-dependent AUC when the risk tool is known or fixed. The performance of the estimator will be assessed by calculating the percentage bias and the mean square error. Two scenarios were considered. In the first scenario, we assume that the probability of cure does not depend on the covariates whereas the second scenario assumes this probability to be dependent on the covariates. For both of these scenarios, the prognostic tool is defined as $Z = 1 - S^{\mathcal{MCM}}(t|X)$ and the data were generated from a MCM as follows:

Scenario 1: As in Heagerty and Zheng (2005), Li et al. (2018), and Martínez-Camblor et al. (2016), we generate $(X, \log(T^u), \log(C))$ from a trivariate normal distribution (MVN). More precisely,

$$\begin{bmatrix} X_i \\ \log(T_i^u) \\ \log(C_i) \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} 0 \\ 0 \\ \mu_c \end{bmatrix}, \begin{bmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right),$$

where $\rho = -0.9$, so that a higher value of the covariate X is an indication of a shorter duration T. To control the censoring and cure proportions, different values of μ_c and P (uncured probability) were considered. These values together with the chosen time

TABLE 1 Parameters used to generate the data for Scenario 1 and 2 and the resulting percentages of censoring (%cens), percentages of cure (%cure), true values of $P(T > t)$, and true values of $AUC_Z(t)$

			Scenario 1			Scenario 2			
%cens	%cure	t	μ_c	$P(T > t)$	$AUC_Z(t)$	λ	γ_0	$P(T > t)$	$AUC_Z(t)$
40	20	0.2	0.94	0.96	0.966	0.0550	1.65	0.80	0.902
60	20	0.8	0.00	0.67	0.886	0.4400	1.65	0.65	0.857
40	30	0.2	1.45	0.96	0.964	0.0100	1.00	0.87	0.893
60	30	0.8	0.25	0.71	0.863	0.1740	1.00	0.74	0.838
40	40	0.2	3.80	0.97	0.961	0.0001	0.50	0.89	0.896
60	40	0.8	0.60	0.75	0.844	0.0620	0.50	0.81	0.826

horizon t and the resulting percentages of censoring (%cens), P($T > t$) and $AUC_Z(t)$ can be found in Table 1 (Scenario 1). Note that in this scenario, the uncured probability P does not depend on the covariate X.

Scenario 2: The data are generated from a logistic-exponential MCM, where the probability of being uncured is given by

$$\mathrm{P}(X) = \frac{\exp(\gamma_0 + X)}{1 + \exp(\gamma_0 + X)},$$

and the conditional survival function of the uncured subjects is given by

$$S_u(t|X) = \exp(-\exp(-3X)t).$$

The covariate X is assumed to follow a standard normal distribution. The censoring time was generated from an exponential distribution with parameter λ. We consider different values for the parameters λ and γ_0 to get different censoring and cure fractions. These parameters are summarized in Table 1 (Scenario 2). Note that contrary to Scenario 1, the uncured probability P depend here on X. The true $AUC_Z(t)$ in Table 1 was calculated empirically, from a sample $(T_i, Z_i), i = 1, \ldots, 4 \times 10^6$, generated as shown in Scenarios 1 and 2, by using Equation (3) with $\hat{W}_i = I(T_i \leq t)$.

For each scenario, we generated $N = 2,000$ samples of size $n = 200$ and 400 and evaluated the performance of the AUC estimator using the empirical percent bias (% bias) and empirical mean squared error (MSE), given, respectively, by

$$\%\ \text{bias} = \frac{\hat{\mathrm{E}}(\hat{AUC}_Z(t)) - AUC_Z(t)}{AUC_Z(t)} \times 100, \quad \text{MSE} = N^{-1} \sum_{j=1}^{N} (\hat{AUC}_{Z,j}(t) - AUC_Z(t))^2.$$

where $\hat{\mathrm{E}}(\hat{AUC}_Z(t)) = N^{-1} \sum_{j=1}^{N} \hat{AUC}_{Z,j}(t)$ and $\hat{AUC}_{Z,j}(t)$ is the NPS estimate calculated from the jth ($j = 1, \ldots, N$) simulated data $(Y_{i,j}, \Delta_{i,j}, Z_{i,j})$, $i = 1, \ldots, n$.

Table 2 shows the results obtained for the sample sizes $n = 200$ and $n = 400$, censoring rates 40% and 60%, cure rates 20%, 30%, and 40%, and time horizons $t = 0.2$ and $t = 0.8$. From these results, it can be seen that, for both scenarios, the NPS AUC estimator performed very well, with small percent bias (less than 0.3%) and small MSE (less than 0.9×10^{-3}). Furthermore, as the sample size increased from 200 to 400, the MSE was reduced by around 50%, which indicates that the NPS estimator is consistent for data with cured subjects. The results also show that, as the percentage of censored and cure rate increases, both the percent bias and the MSE increase. A similar effect is also observed when we increase the time horizon t. From these results, we can conclude that, when the prognostic tool is known, whether the probability of cure depends on covariate X or not, the simple method gives very good results. This may be explained by the fact that, as we mention before, the NPS relies on Beran's estimator, which is consistent even in the presence of a cure fraction in the data (López-Cheda et al., 2017). However, care should be taken for the choice of the time horizon t as Beran's estimator is not consistent everywhere.

3.2 | The validity of the NPS estimator under cure: The case of an estimated prognostic tool

The objective of this section is to assess the performance of the NPS estimator in the case where the prognostic (risk) tool needs to be estimated from the data using a given working survival model. In order to avoid optimistic bias of future performance, we followed the established rule which consists in estimating (or deriving) the risk tool from a learning set and then evaluating this tool and comparing its prediction accuracy on an independent validation set. This method helps to avoid overoptimistic

TABLE 2 % bias and MSE of the NPS AUC estimate for data generated according to Scenarios 1 and 2

				Scenario 1		Scenario 2	
t	%cure	%cens	n	% bias	MSE ($\times 10^{-3}$)	% bias	MSE ($\times 10^{-3}$)
0.2	20	40	200	−0.042	0.358	0.076	0.419
0.2	20	40	400	0.014	0.167	0.098	0.200
0.2	20	60	200	−0.064	0.368	0.110	0.435
0.2	20	60	400	0.005	0.180	0.105	0.207
0.2	30	40	200	−0.108	0.444	−0.043	0.576
0.2	30	40	400	−0.054	0.207	0.018	0.287
0.2	30	60	200	−0.111	0.450	−0.046	0.578
0.2	30	60	400	−0.059	0.214	0.015	0.291
0.2	40	40	200	0.001	0.615	−0.108	0.609
0.2	40	40	400	0.032	0.246	−0.052	0.293
0.2	40	60	200	−0.009	0.617	−0.112	0.610
0.2	40	60	400	0.024	0.250	−0.053	0.294
0.8	20	40	200	0.042	0.534	0.112	0.565
0.8	20	40	400	0.092	0.266	0.088	0.285
0.8	20	60	200	0.206	0.703	0.186	0.669
0.8	20	60	400	0.194	0.345	0.139	0.340
0.8	30	40	200	0.062	0.652	−0.002	0.636
0.8	30	40	400	0.133	0.318	0.079	0.336
0.8	30	60	200	0.210	0.784	0.029	0.684
0.8	30	60	400	0.233	0.378	0.098	0.357
0.8	40	40	200	−0.022	0.757	−0.171	0.847
0.8	40	40	400	0.024	0.357	−0.139	0.432
0.8	40	60	200	0.004	0.853	−0.164	0.850
0.8	40	60	400	0.096	0.385	−0.140	0.437

prognoses, and it is more appropriate than using the same data for both learning and validation (Krzanowski & Hand, 2009; Schmid & Potapov, 2012).

Two approaches were used to answer this objective. In the first approach, we compared $\hat{AUC}_{\hat{Z}^{\mathcal{W}}}(t)$ to $AUC_{Z^{\mathcal{W}}}(t)$ where $\hat{AUC}_{\hat{Z}^{\mathcal{W}}}(t)$ is a random variable representing the NPS AUC estimate of the risk tool $\hat{Z}^{\mathcal{W}}$ derived from fitting a working survival model $\mathcal{W}$ to a finite survival data set and $AUC_{Z^{\mathcal{W}}}(t)$ is the true AUC of the risk tool $Z^{\mathcal{W}}$, where $Z^{\mathcal{W}}$is derived from fitting a working survival model $\mathcal{W}$ to a huge survival data set (population). The percent bias and MSE according to approach 1 are defined as follows:

$$\%\,\text{bias} = \frac{\text{E}(\hat{AUC}_{\hat{Z}^{\mathcal{W}}}(t)) - AUC_{Z^{\mathcal{W}}}(t)}{AUC_{Z^{\mathcal{W}}}(t)} \times 100, \quad \text{MSE} = \text{E}[\hat{AUC}_{\hat{Z}^{\mathcal{W}}}(t) - AUC_{Z^{\mathcal{W}}}(t)]^2. \tag{8}$$

In the second approach, we compared $\hat{AUC}_{\hat{Z}^{\mathcal{W}}}(t)$ to $AUC_{\hat{Z}^{\mathcal{W}}}(t)$, where $\hat{AUC}_{\hat{Z}^{\mathcal{W}}}(t)$ is defined as above and $AUC_{\hat{Z}^{\mathcal{W}}}(t)$ is the true population AUC of the risk tool $\hat{Z}^{\mathcal{W}}$, where $\hat{Z}^{\mathcal{W}}$ is derived by fitting a working survival model $\mathcal{W}$ to a finite survival data set. The percent bias and MSE according to Approach 2 are defined as follows:

$$\%\,\text{bias} = \frac{\text{E}(\hat{AUC}_{\hat{Z}^{\mathcal{W}}}(t)) - AUC_{\hat{Z}^{\mathcal{W}}}(t)}{AUC_{\hat{Z}^{\mathcal{W}}}(t)} \times 100, \quad \text{MSE} = \text{E}[\hat{AUC}_{\hat{Z}^{\mathcal{W}}}(t) - AUC_{\hat{Z}^{\mathcal{W}}}(t)]^2. \tag{9}$$

In Approach 2, we try to eliminate the error due to the estimation of the marker. By doing so, we focus only on the performance of the AUC estimator (given an estimated marker).

Approach 1: Comparing $\hat{AUC}_{\hat{Z}^{\mathcal{W}}}(t)$ to $AUC_{Z^{\mathcal{W}}}(t)$.

For the first approach, we proceeded with the following simulation plan:

TABLE 3 Parameters used to generate the data for Cases 1 and 2 and the resulting percentages of censoring (%cens), percentages of cure (%cure), true values of $P(T > t)$, and true values of $AUC_{Z^{\mathcal{MCM}}}(t)$

				Case 1			Case 2		
%cens	%cure	t	$P(T > t)$	λ	γ_0	$AUC_{Z^{\mathcal{MCM}}}(t)$	λ	γ_0	$AUC_{Z^{\mathcal{MCM}}}(t)$
40	20	1	0.648	0.200	2.17	0.826	0.130	1.170	0.701
60	20	2	0.511	0.730	2.17	0.836	0.500	1.170	0.691
40	30	1	0.710	0.100	1.53	0.830	0.050	0.520	0.662
60	30	2	0.590	0.560	1.53	0.832	0.310	0.520	0.654
40	40	1	0.764	0.001	1.00	0.833	0.001	0.001	0.633
60	40	2	0.661	0.380	1.00	0.826	0.180	0.001	0.629

(i) We generated a learning sample of size n_l that is, $(Y_i, \Delta_i, \mathbf{X}_i),\ i = 1, \ldots, l$ using a MCM.

(ii) From the generated data in (i), we fitted a working survival model $\mathcal{W}$, extracted the required estimated parameters, and derived an estimated risk tool $\hat{Z}^{\mathcal{W}}(\mathbf{X}) \equiv \hat{Z}^{\mathcal{W}} = 1 - \hat{S}^{\mathcal{W}}(t|\mathbf{X})$, where $\hat{S}^{\mathcal{W}}$ is the estimated conditional survival function of T based on $\mathcal{W}$.

(iii) Using the same MCM as in step (i), we generated a validation sample, $(Y_i, \Delta_i, \mathbf{X}_i)$, of size n_v and calculated the risk tool $\hat{Z}_i^{\mathcal{W}} = \hat{Z}^{\mathcal{W}}(\mathbf{X}_i),\ i = 1, \ldots, n_v$.

(iv) Based on $(Y_i, \Delta_i, \hat{Z}_i^{\mathcal{W}}),\ i = 1, \ldots, n_v$, we calculated $\hat{AUC}_{\hat{Z}^{\mathcal{W}}}(t)$, the NPS estimator of the estimated risk tool $\hat{Z}^{\mathcal{W}}$.

In step (ii) of the simulations plan above, two working survival models $\mathcal{W}$ were fitted: The Cox proportional hazard model (COX) and the MCM. The MCM model used to generate the data in steps (i) and (iii) is a mixture of logistic and exponential, where

$$\mathrm{P}(\mathbf{X}) = \frac{\exp(\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2)}{1 + \exp(\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2)}$$

is the probability of uncured and

$$S_u(t|\mathbf{X}) = \exp(-\exp(\beta_1 X_1 + \beta_2 X_2)t)$$

is the conditional survival function of the uncured subjects and the covariates $X_1 \sim N(0, 1)$ and $X_2 \sim Unif[0, 1]$. The censoring time was generated from an exponential distribution with parameter λ. Note that the covariates (X_1, X_2) affect both the probability of being uncured and the survival function of uncured subjects. We considered two cases in parameter settings:

Case 1: $(\beta_1, \beta_2) = (\gamma_1, \gamma_2) = (-1, -1)$,
Case 2: $(\beta_1, \beta_2) = -(\gamma_1, \gamma_2) = (-1, -1)$.

Note that in step (ii), we could have used the linear predictor function $\boldsymbol{\beta}^T\mathbf{X}$ as a potential prognostic tool instead of the distribution function $Z^{\mathcal{W}} = 1 - S^{\mathcal{W}}(t|\mathbf{X})$. This will indeed lead to the same AUC when $\mathcal{W} = COX$ or when $\mathcal{W} = MCM$ but only under case 1. This is, however, not true for case 2 since the survival function is not monotonic in $\boldsymbol{\beta}^T\mathbf{X}$ in that case.

To control the censoring and cure proportions, we considered different values for the parameters λ and γ_0. These parameters are summarized in Table 3.

For each case and each combination of the parameters in Table 3, we repeated steps (i)–(iv) described above $N = 2,000$ times with sample sizes $n = n_l = n_v = 200, 400, 600$, and evaluated the bias and MSE. To calculate the $AUC_{Z^{\mathcal{MCM}}}(t)$, the true optimal AUC, we generated a sample $(T_i, \mathbf{X}_i)$ of size 4×10^6 from the MCM and then calculated the AUC of $Z = 1 - S^{\mathcal{MCM}}(t|\mathbf{X})$ empirically, from $(T_i, Z_i^{\mathcal{MCM}})$, using Equation (3) with $\hat{W}_i = I(T_i \le t)$. For the $AUC_{Z^{\mathcal{COX}}}(t)$, the *pseudo* true AUC, we run $1 - 3$ once with $n_l = n_v = 4 \times 10^6$ and then calculated $AUC_{Z^{\mathcal{COX}}}(t)$ empirically from $(T_i, Z_i^{\mathcal{COX}})$. Note that the MSE was calculated according to the formula $\mathrm{MSE} = N^{-1}\sum_{j=1}^{N}(\hat{AUC}_{\hat{Z}_j^{\mathcal{W}},j}(t) - AUC_{Z^{\mathcal{W}}}(t))^2$, where $\hat{AUC}_{\hat{Z}_j^{\mathcal{W}},j}(t)$ is the NPS estimate calculated from the jth ($j = 1, \ldots, N$) validation sample using the estimated prognostic tool $\hat{Z}_j^{\mathcal{W}}$ calculated from the jth learning sample.

We also repeated all computations in steps (i)–(iv) using a BHM (bounded hazard model) instead of the MCM. More precisely, we generated the survival times from $S(t|\mathbf{X}) = \exp(-\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)F(t))$, where $(\beta_1, \beta_2) = (1, 0.6)$, $F(t) = 1 - \exp(-0.2t)$, $X_1 \sim N(0, 1)$, and $X_2 \sim Unif[0, 1]$. The censoring time was generated from an exponential distribution with parameter λ. We varied the values of λ and β_0 to get the same censoring and cure rates as the one obtained for the MCM. We

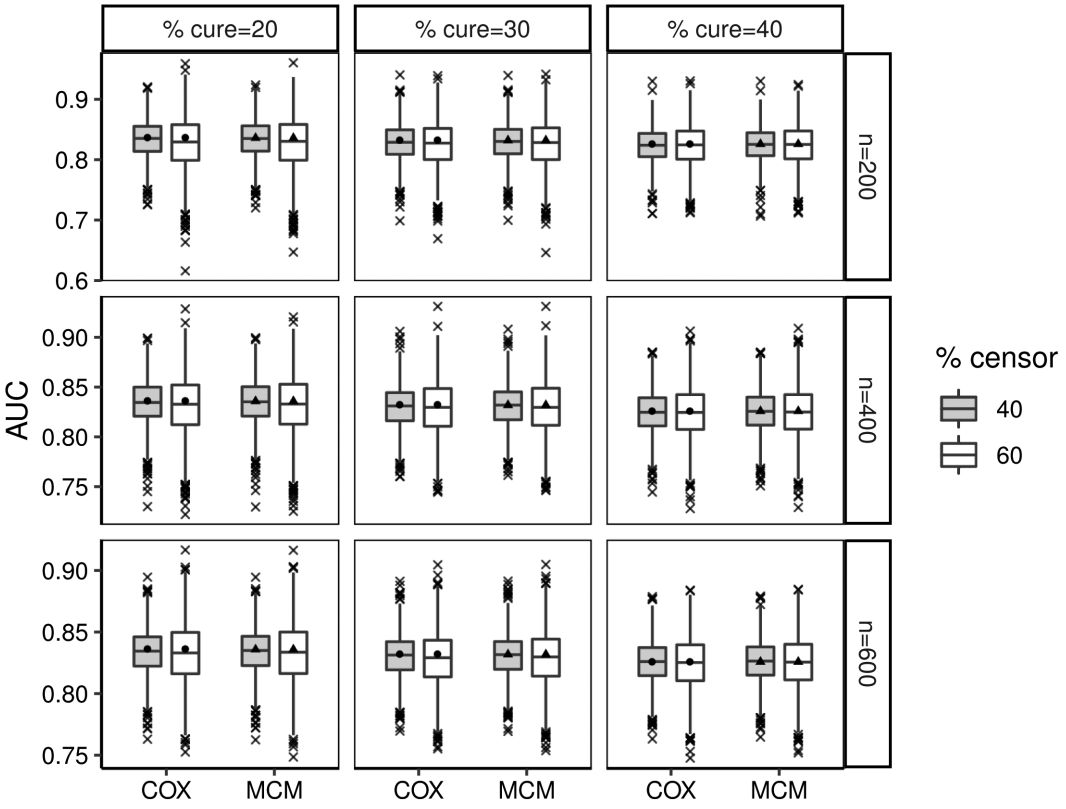

FIGURE 1 Boxplots of $A\hat{U}C_{\hat{Z}^W}(t)$ in Case 1 for the time horizon $t = 2$. The solid circles and triangles indicate the "true" AUC values $AUC_{Z^W}(t)$ of the Cox and mixture cure working models, respectively

also used the same time horizon ($t = 1, 2$) as for the MCM. For the sake of brevity, the results of the BHM are not reported here because they were very similar to the results obtained with the MCM in Case 1.

In Figures 1 and 2, we provide the boxplots of the resulting estimates for Cases 1 and 2, respectively, whereas in Table 4 we show the percent bias and the empirical MSE of these estimates for both Cases 1 and 2.

Let us first look at Case 1. The boxplots in Figure 1 and the results in Table 4 reveal that the prognostic tool estimate based on the Cox model and on the MCM result in almost exactly the same estimates and so the same performances. This is true for all cure rates, censoring rates, and sample sizes. We can also see that the bias and MSE increase when the censoring and/or the cure rates increase. In contrast, the bias and MSE decrease as the sample size increases. We obtained exactly the same conclusions when we analyzed the results of the BHM (not shown here). We can therefore conclude that under Case 1, the NPS method still provides valid and consistent estimates of the true AUC even in the situation where the prognostic tool is derived from a misspecified model such as the Cox proportional hazard model. This may be explained by the fact that, in this case, the misspecified Cox model estimates the vector of parameters $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ consistently. This is indeed the case for BHM as it has been shown in Portier, Van Keilegom, and El Ghouch (2016).

For Case 2, the most common situation (the coefficients in the cure probability are different from those in the survival component), the story is completely different. In fact, both the results presented in Figure 2 and Table 4 indicate that deriving (estimating) a prognostic tool from a survival data exhibiting cured subjects using a Cox proportional hazard model results in larger bias and larger MSE in the AUC estimates compared to a MCM for almost all studied cases. This is true for all cure rates, censoring rates, and sample sizes. This bias is noticeable even when the sample size is large ($n = 600$), especially when the cure rate is high. For example, for $t = 1$, when %cure = 40 and %cens = 60, the %bias using the COX model is almost 10 times larger than the one obtained using MCM. In contrast, the prognostic tool estimate using the MCM generally results in an AUC estimate with very small bias (less than 1.5%) and MSE (less than 0.001) that decrease with an increasing sample size. The possible reason for the large bias and MSE observed in the AUC estimates when using the misspecified Cox working model, in this case, can be explained by the bias in estimating the survival function for data with a cure fraction. This was well documented in Amico and Van Keilegom (2018). In their paper, when comparing the MCM and Cox model in a simulation study, the authors concluded that "not taking into account the presence of a cure fraction in survival data has important consequences that may lead

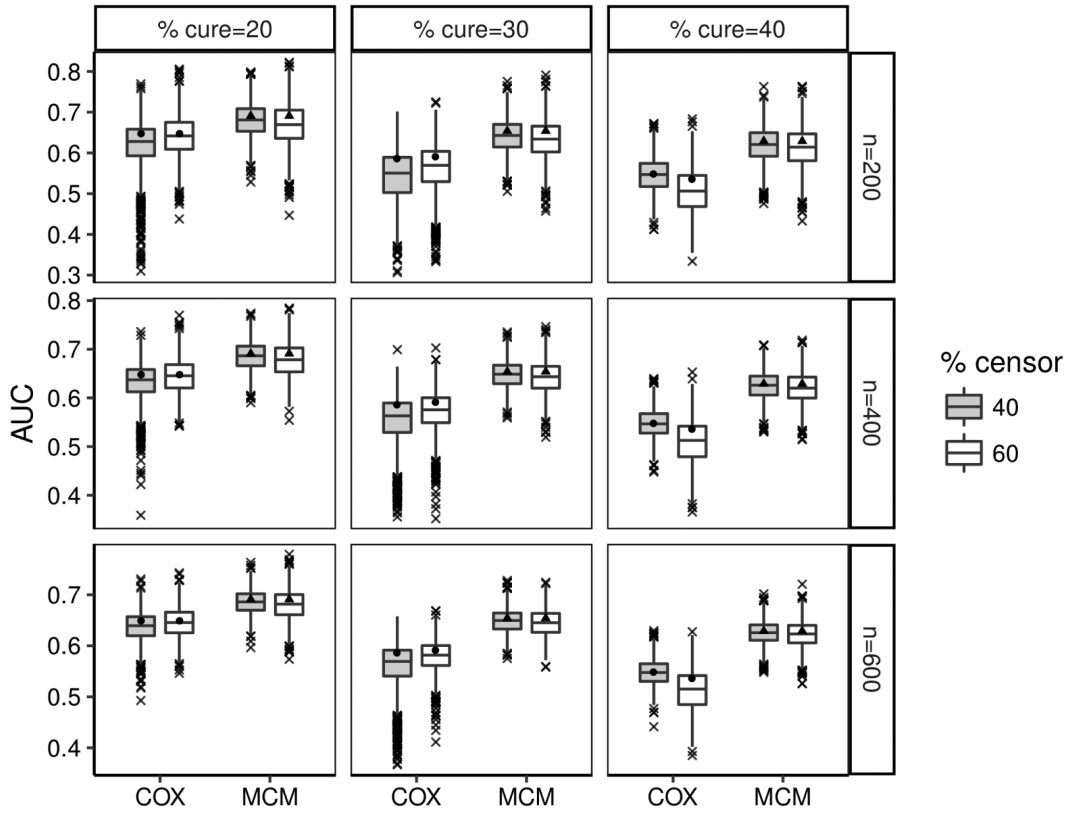

FIGURE 2 Boxplots of $A\hat{U}C_{\hat{Z}^{\mathcal{W}}}(t)$ in Case 2 for the time horizon $t = 2$. The solid circles and triangles indicate the "true" AUC values $AUC_{Z^{\mathcal{W}}}(t)$ of the Cox and mixture cure working models, respectively

to wrong conclusions." In our simulations study, we also investigated the case when the data are generated for the classical Cox model (without cure) and the AUC is estimated using either the Cox or the mixture cure working model. In all the investigated cases, both these models lead to almost exactly the same AUC estimates. For this reason and for sake of brevity, we do not show these results here.

Approach 2: Comparing $A\hat{U}C_{\hat{Z}^{\mathcal{W}}}(t)$ to $AUC_{\hat{Z}^{\mathcal{W}}}(t)$.

For the second approach and to reduce the burden of calculations, we investigated only Case 2. We proceeded with the following simulation plan:

(i) We generated a learning sample of size n_l from the MCM-Case 2 given above.

(ii) From the generated data, we estimated the working model $\mathcal{W}$ and calculated the prognostic tool $\hat{Z}^{\mathcal{W}}(\mathbf{X}) \equiv \hat{Z}^{\mathcal{W}} = 1 - \hat{S}^{\mathcal{W}}(t|\mathbf{X})$.

(iii) We generated $B = 1,000$ independent validation samples of size n_v from the same MCM-Case 2.

(iv) From each validation sample, we calculated the NPS estimator $A\hat{U}C_{j,\hat{Z}^{\mathcal{W}}}$, $j = 1, \ldots, B$.

Similar to Approach 1, two working survival models were fitted in step 2 namely, the Cox proportional hazard model and the MCM. The specification of the MCM is similar to the one described in Approach 1. For each combination of the parameters in Case 2 (see Table 3), we repeated steps (i)−(iv) described above $N = 200$ times ($k = 1, \ldots, N$) and evaluated, for each k, the percent bias and MSE according to Equation (9). For a given $\hat{Z}_k^{\mathcal{W}}$, $AUC_{\hat{Z}_k^{\mathcal{W}}}$, the "true" or population version of the AUC, was calculated from a validation sample of size 10^6. Note that, unlike the simulation design carried out in Approach 1 where the prognostic tool and its AUC were estimated for each pair of the simulated learning and validation data sets, here in Approach 2, once estimated from a given learning data, the prognostic tool Z was kept fixed and its AUC evaluated using B validation data sets. To account for the sampling variability, we repeated this process N times and report the averaged performance measures.

Table 5 reports the empirical averaged % bias and MSE, given, respectively, by $\%\ \text{bias} = \frac{1}{N}\sum_{k=1}^{N} \%\ \text{bias}_k$

TABLE 4 The percent bias and MSE for Cases 1 and 2 (Approach 1) MCM is MCM and COX is classical Cox model

				Case 1				Case 2			
				$n = 200$		$n = 600$		$n = 200$		$n = 600$	
t	% cure	% cens	Model	% bias	MSE ($\times 10^{-3}$)	% bias	MSE ($\times 10^{-3}$)	% bias	MSE ($\times 10^{-3}$)	% bias	MSE ($\times 10^{-3}$)
1	20	40	MCM	−0.126	1.000	−0.119	0.320	−0.855	1.561	−0.359	0.517
1	20	40	COX	−0.175	1.000	−0.154	0.322	−4.894	5.685	−1.816	1.065
1	20	60	MCM	−0.360	1.383	−0.217	0.415	−1.418	1.960	−0.709	0.664
1	20	60	COX	−0.336	1.364	−0.242	0.416	−1.209	2.122	−0.251	0.652
1	30	40	MCM	−0.096	0.965	−0.018	0.300	−1.340	1.802	−0.444	0.574
1	30	40	COX	−0.206	0.983	−0.059	0.300	−10.65	12.18	−6.000	5.312
1	30	60	MCM	−0.315	1.248	−0.115	0.384	−2.077	2.148	−0.799	0.706
1	30	60	COX	−0.339	1.258	−0.159	0.384	−5.995	6.326	−2.294	1.465
1	40	40	MCM	−0.016	0.861	0.021	0.288	−1.422	2.117	−0.530	0.661
1	40	40	COX	−0.188	0.885	−0.035	0.290	−0.782	1.993	−0.166	0.664
1	40	60	MCM	−0.158	1.076	−0.044	0.354	−2.071	2.388	−0.892	0.764
1	40	60	COX	−0.226	1.086	−0.096	0.357	−11.09	10.54	−8.569	6.946
2	20	40	MCM	−0.180	0.980	−0.156	0.322	−1.486	1.805	−0.715	0.561
2	20	40	COX	−0.237	0.979	−0.202	0.324	−4.122	4.205	−1.550	0.956
2	20	60	MCM	−1.008	2.015	−0.353	0.616	−3.154	3.226	−1.514	0.984
2	20	60	COX	−1.043	2.027	−0.390	0.614	−1.026	2.590	−0.280	0.862
2	30	40	MCM	−0.292	0.915	−0.125	0.290	−1.798	1.788	−0.764	0.545
2	30	40	COX	−0.400	0.931	−0.174	0.291	−7.455	6.075	−4.240	2.628
2	30	60	MCM	−0.678	1.486	−0.297	0.494	−3.162	2.675	−1.381	0.804
2	30	60	COX	−0.742	1.497	−0.361	0.495	−4.455	4.054	−1.832	1.080
2	40	40	MCM	−0.117	0.794	0.054	0.283	−1.471	1.848	−0.517	0.554
2	40	40	COX	−0.289	0.814	−0.005	0.286	−0.412	1.715	−0.086	0.597
2	40	60	MCM	−0.271	1.172	−0.062	0.424	−2.537	2.525	−0.982	0.702
2	40	60	COX	−0.353	1.182	−0.128	0.427	−5.450	3.881	−4.292	2.131

and $\text{MSE} = \frac{1}{N}\sum_{k=1}^{N}\text{MSE}_k$, where $\%\ \text{bias}_k = 100 \times (\frac{1}{B}\sum_{j=1}^{B} \hat{AUC}_{j,\hat{Z}_k^w}(t) - AUC_{\hat{Z}_k^w}(t))/AUC_{\hat{Z}_k^w}(t)$ and $\text{MSE}_k = \frac{1}{B}\sum_{j=1}^{B}(\hat{AUC}_{j,\hat{Z}_k^w}(t) - AUC_{\hat{Z}_k^w}(t))^2$, for sample sizes $n = n_l = n_v = 200, 400$. We also calculated the resulting averaged AUC estimate given by $\hat{AAUC}_{\hat{Z}_k^w} = B^{-1}\sum_{j=1}^{B}\hat{AUC}_{j,\hat{Z}_k^w}$, for $k = 1, \ldots, N$, and the averaged true AUC given by $AAUC_{\hat{Z}^w} = \frac{1}{N}\sum_{k=1}^{N} AUC_{\hat{Z}_k^w}$ and reported theme using boxplots in Figure 3. As done previously, all the results are stratified according to the percentage of censoring and percentage of cure.

The boxplots in Figure 3 and the values in Table 5 both show that, when the estimated prognostic tool is held fixed, the NPS AUC estimator provides small bias and MSE, especially for a large sample size ($n = 400$), for both the Cox and the MCM models. This is true for all cure rates, censoring rates, and sample sizes. Furthermore, the bias and MSE decrease when the sample size increases for both the MCM and the Cox working models. We can also see from both results that the estimated prognostic tool based on the Cox working survival model resulted in an AUC estimate that converge to its limiting value $AUC_{\hat{Z}^{COX}}(t)$ despite the working model is misspecified. Therefore, this simulation study confirms the consistency of the NPS estimator regardless the working survival model from which the prognostic tool was derived. The simulations results, however, gave the impression that the Cox proportional hazard model systematically leads to a smaller AUC compared to the MCM. This important difference is investigated in more detail in the next section.

3.3 | Deriving risk tools from survival models that ignore the presence of cure

The objective of this section is to investigate the loss, in terms of predictive performance, when deriving risk tools from a working survival model that does not acknowledge the presence of cured subjects. This objective is motivated from the current practice of deriving risk tools in cardiovascular disease prevention and oncology by fitting standard survival models, mainly

TABLE 5 The percent bias and MSE for MCM-case 2 (Approach 2)

				$n = 200$			$n = 400$		
t	% cure	% cens	Model	$AAUC_{\hat{Z}^w}$	% bias	MSE ($\times 10^{-3}$)	$AAUC_{\hat{Z}^w}$	% bias	MSE ($\times 10^{-3}$)
1	20	40	MCM	0.694	0.197	1.475	0.697	0.015	0.765
1	20	40	COX	0.647	0.141	1.596	0.665	0.060	0.800
1	20	60	MCM	0.689	0.252	1.804	0.695	−0.034	0.911
1	20	60	COX	0.674	0.355	1.828	0.680	0.178	0.920
1	30	40	MCM	0.654	0.164	1.675	0.657	0.040	0.884
1	30	40	COX	0.558	0.072	1.824	0.581	0.003	0.945
1	30	60	MCM	0.649	0.117	1.887	0.654	0.026	1.022
1	30	60	COX	0.602	0.183	1.969	0.613	0.111	1.050
1	40	40	MCM	0.625	0.069	1.969	0.628	0.156	0.981
1	40	40	COX	0.594	0.072	1.814	0.596	0.048	0.995
1	40	60	MCM	0.621	−0.018	2.107	0.626	0.022	1.064
1	40	60	COX	0.509	−0.044	2.278	0.522	−0.074	1.194
2	20	40	MCM	0.680	0.216	1.638	0.686	−0.009	0.818
2	20	40	COX	0.617	0.107	1.758	0.631	0.014	0.857
2	20	60	MCM	0.669	−0.042	2.497	0.680	−0.127	1.245
2	20	60	COX	0.638	0.171	2.494	0.643	0.121	1.214
2	30	40	MCM	0.644	0.070	1.577	0.648	0.053	0.745
2	30	40	COX	0.538	0.161	1.692	0.553	−0.025	0.851
2	30	60	MCM	0.635	−0.095	2.092	0.644	−0.123	1.010
2	30	60	COX	0.567	0.130	2.139	0.574	0.008	1.092
2	40	40	MCM	0.621	0.040	1.585	0.625	0.066	0.815
2	40	40	COX	0.545	0.177	1.577	0.547	−0.028	0.853
2	40	60	MCM	0.614	−0.056	1.978	0.622	−0.144	0.989
2	40	60	COX	0.504	0.005	1.975	0.509	−0.065	1.039

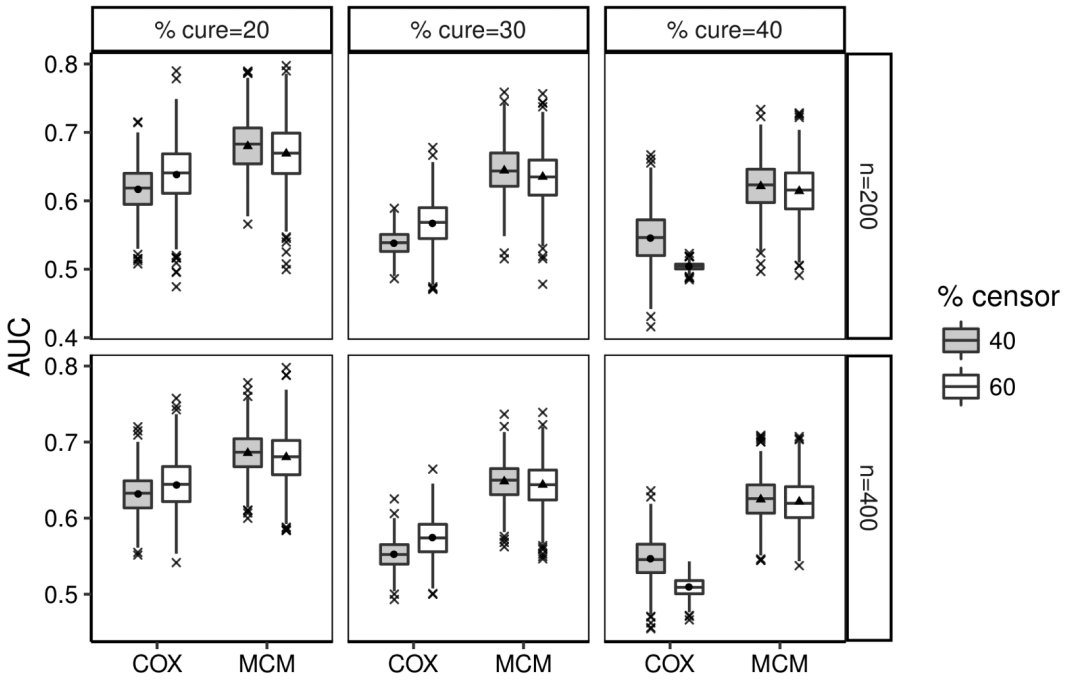

FIGURE 3 The boxplots of the averaged NPS AUC estimates $AA\hat{U}C_{\hat{Z}_k^w}$ obtained with MCM-Case 2 for a time horizon $t = 2$. The solid circles and triangles indicate the "true" or the population version of the AUC ($AUC_{\hat{Z}_k^w}$) of the Cox and mixture cure working models, respectively

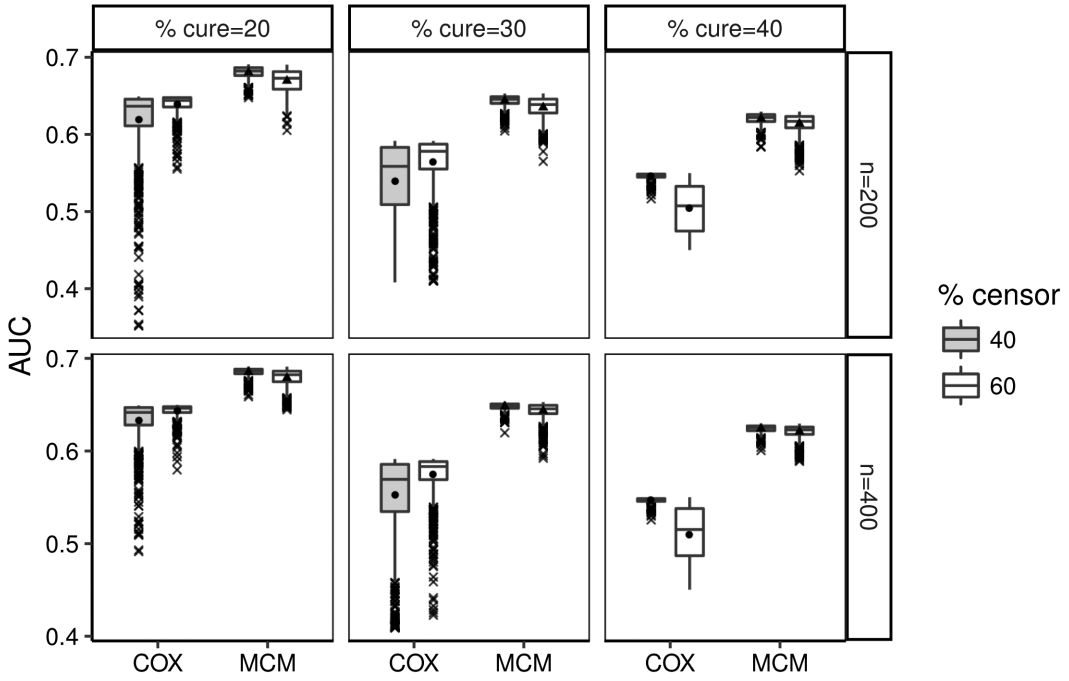

FIGURE 4 The boxplots of $AUC_{\hat{Z}_k^{\mathcal{W}}}$, the population version of AUC of the prognostic tool $\hat{Z}_k^{\mathcal{W}}$ obtained with MCM-Case 2 for a time horizon $t = 2$ with $\mathcal{W} = MCM, COX$

the Cox proportional hazard model, even though a large fraction of the population is nonsusceptible to experience that event during their lifetime. Two criteria were used to assess the loss in predictive ability. In the first criterion, we compared the distribution of the AUC estimates of a risk tool derived from a Cox proportional hazard model (the misspecified model) to the distribution of the AUC estimate of a risk tool derived from a MCM (the true model). In both models, the derived risk tools aim at estimating $P(T \leq t|\mathbf{X})$ for a fixed time horizon t, and both include exactly the same covariates. In the first criteria we aimed to show that, in average and for a given data set, a risk tool derived from a working survival model that does not ignore the presence of cure (i.e., the MCM model) produces better predictive performance in terms of the AUC estimate compared to a working survival model that ignores the presence of cure (,i.e., the Cox proportional hazard model). More specifically, we want to show that

$$\mathrm{E}(\hat{AUC}_{\hat{Z}^{MCM}}(t)) \geq \mathrm{E}(\hat{AUC}_{\hat{Z}^{COX}}(t)). \tag{10}$$

In the second criterion, we compared the distribution of the true AUC of a risk tool derived from a Cox proportional hazard model (the misspecified model) to the distribution of the true AUC of a risk tool derived from a MCM (the true model). Here we want to show that the true AUC of a risk tool derived from a working survival model that acknowledge the presence of cure is higher than the one derived from a working survival model that ignore the presence of cure. More specifically, we aimed to show that

$$\mathrm{E}(AUC_{\hat{Z}^{MCM}}(t)) \geq \mathrm{E}(AUC_{\hat{Z}^{COX}}(t)). \tag{11}$$

For both criteria, we used some of the results provided in the previous simulations and therefore no additional simulation plans were needed. The question in the first criteria as described in Equation (10) can be directly answered from the boxplots in Figures 2 and 3. In fact, one can clearly see from these two figures that, for a fixed censoring rate, the estimated AUC of the risk tool derived from the Cox proportional hazard model is far smaller than the one derived from an MCM model. This is true in all situations. For the second criterion as described in Equation (11), Figure 4 provides the boxplots of the true area under the ROC curve $AUC_{\hat{Z}_k^{\mathcal{W}}}$ obtained in the previous section with MCM-case 2 and time horizon $t = 2$ but with $N = 1000$ iterations compared to 200 iterations in the previous simulations. From this figure, one can clearly see that, compared to the Cox proportional hazard model, the average true AUC of the MCM is always higher and also exhibits smaller variability.

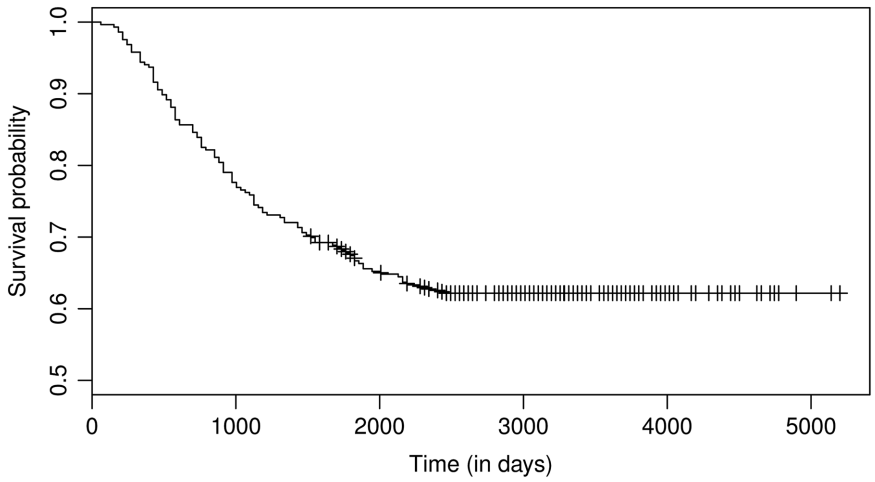

FIGURE 5 Kaplan–Meier survival estimate of the lymph-node-negative breast cancer patients (+ : censored observations)

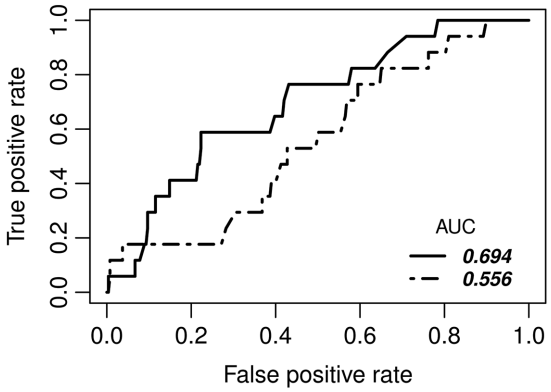

FIGURE 6 ROC curve and the corresponding AUC estimates for the MCM (solid line) and Cox model (dashed line) obtained for $t = 1$ year and using all data as training and validation set

4 | PREDICTING NODE-NEGATIVE BREAST CANCER: SHOULD CURE BE TAKEN INTO ACCOUNT?

The objective of this section is to compare, using real data example, the AUC of a survival model that does not acknowledge the presence of cured subjects with the AUC of a survival model that do acknowledge the cure. We use the data from a study on patients with lymph-node-negative breast cancer treated during 1980–1995. This data set was first reported by Wang et al. (2005), and, therein, it was used to derive a prognostic tool from a Cox model. During the follow-up period, 179 out of a total of 286 patients were relapse-free (censored), and 107 of them developed distant metastasis. This is the event of interest, and T, the time (measured in days) to this event, is the survival time of interest. Several factors are known to increase the risk of distant relapse. These factors include tumor size, involvement of lymph nodes and vascular invasion, tumor type and grade, hormone receptor status, etc. In this example, we consider three factors. These are age of the patients which range, in the analyzed data, from 26 to 83 with median of 52 years; the estrogen receptor (ER) status with levels ER− (less than 10 fmol per mg protein) and ER+ (at least 10 fmol per mg protein) and the tumor size that range from 1 to 4 with median of 1.

To explore the possible presence of cured patients, we plot the Kaplan–Meier survival curve estimate in Figure 5. This figure shows that the curve levels off above 0.6 with an obvious "plateau." This, together with medical evidence, clearly indicates the presence of a potential cure fraction. Modeling these data ignoring this information and considering that all patients will relapse may lead to wrong conclusion in explaining and predicting the event of interest. To investigate this, we first estimated the conditional survival distribution of T given age, ER, and tumor size using (i) the standard Cox model and (ii) the MCM and then we evaluated the predictive ability of each obtained prognostic tool (survival probability) using ROC curve and AUC. We followed two approaches, in a first analysis we used all the patients as learning and validation set, whereas, in a second analysis we randomly split the 286 patients into a learning set and a validation set with 65/35 ratio. The learning set was used to estimate the prognostic tools (Cox and MCM) and the validation set to estimate the ROC and the AUC. Figures 6 and 7 show the ROC curve and AUC derived according to the first approach for $t = 1$ year and $t = 3$ years, respectively. For the second approach, we generate 1,000 pairs of randomly selected learning and validation sets. From each pair, we estimate the tools and

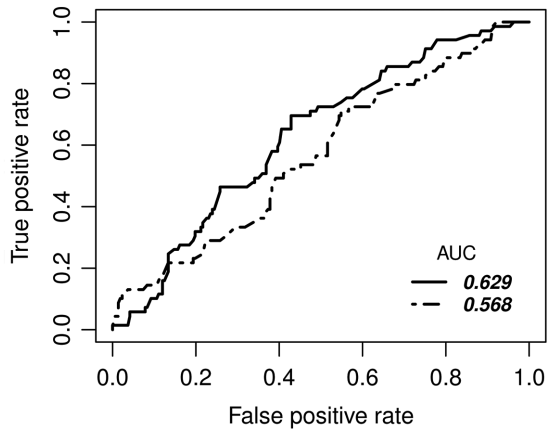

FIGURE 7 ROC curve and the corresponding AUC estimates for the MCM (solid line) and Cox model (dashed line) obtained for $t = 3$ years and using all data as training and validation set

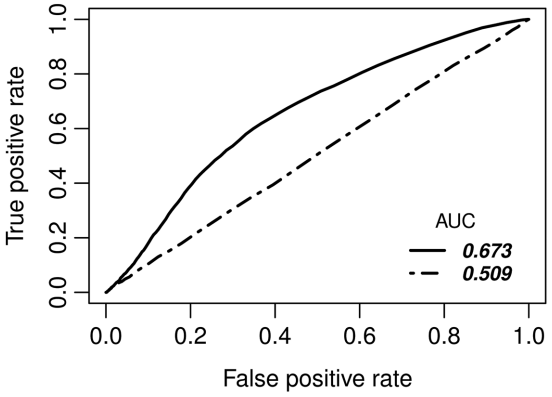

FIGURE 8 ROC curve and the corresponding AUC estimates for the MCM (solid line) and Cox model (dashed line) obtained for $t = 1$ year and by splitting the data to training and validation set

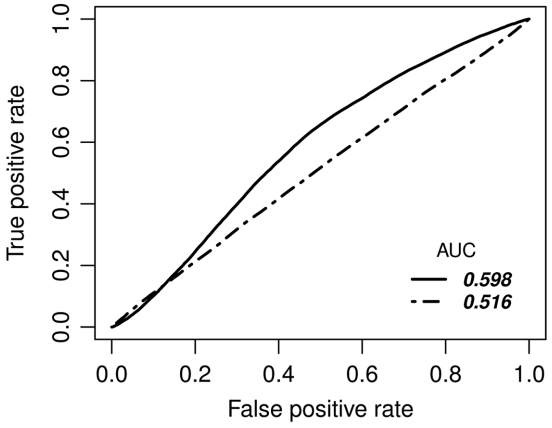

FIGURE 9 ROC curve and the corresponding AUC estimates for the MCM (solid line) and Cox model (dashed line) obtained for $t = 3$ years and by splitting the data to training and validation set

their corresponding AUC and ROC curve. The average of these estimates for $t = 1$ year and $t = 3$ years are presented in Figures 8 and 9, respectively. For the first approach, the estimated AUC values for the prognostic tool derived from MCM and Cox model are 0.694 and 0.556 for $t = 1$ year and 0.629 and 0.568 for $t = 3$ years, and for the second approach, the resulting estimates are 0.673 and 0.509 for $t = 1$ year and 0.598 and 0.516 for $t = 3$ years for MCM and Cox model, respectively. This analysis shows that the risk tool derived from the MCM has a better discrimination ability compared to the one derived from the standard Cox model as the ROC curve of the latter is clearly below the ROC of the former. This is also attested by the small AUC obtained from the COX model. This data analysis confirms our findings in the simulation study (see Section 3.2).

5 | DISCUSSION AND CONCLUSIONS

In this article, we studied the validity of a classification accuracy measure estimator based on right-censored survival data with a cured fraction. More precisely, we assessed the performance of the NPS AUC estimator by means of a comprehensive simulation with different censoring rates, cure rates, sample sizes, and time horizons. We considered both the case of a known and estimated prognostic tools. In both cases, we found that the NPS method provides a consistent estimator of the true AUC for the time-to-event data with a cure fraction. This is not surprising as the method is mainly based on Beran's estimator, and the latter is consistent even when there are cured subjects in the data (López-Cheda et al., 2017). Note that, in real life, it is often difficult to tell whether the data truly contain a cure fraction or not. Therefore, having a method for estimating the time-dependent AUC that works regardless of whether the cure fraction exists or not is very useful from a practical perspective.

We also investigated the loss, in terms of predictive performance, when deriving prognostic tools from a working survival model that does not acknowledge the presence of nonsusceptible subjects. This objective was mainly motivated from the current practice of deriving risk tools in CVD research and oncology. This current practice proceeds by fitting standard survival models, mainly the Cox proportional hazard model, even though a large fraction of the population are nonsusceptible to experience that event during their lifetime. We showed that a prognostic (risk) score derived from a working survival model that does not ignore the presence of cure (e.g., the MCM) produces better predictive performance in terms of the ROC and AUC estimates compared to a working survival model that ignores the presence of cure (e.g., the Cox proportional hazard model).

In conclusion, in order to avoid bias and suboptimal performances, we advocate the use of cure models instead of classical survival models in deriving prognostic tools when data exhibit the presence of a cure fraction.

ACKNOWLEDGMENTS

Abderrahim Oulhaj acknowledge support from the College of Medicine and Health Sciences grant 31M335-NP-17-31 (United Arab Emirates University). Kassu Mehari Beyene would like to thank the National Fund for Scientific Research (FRNS) of Belgium grant PDR(FRNS)-T0080.16 for supporting financially. The authors also thank the editor, associate editor, and two anonymous referees for many constructive and helpful comments that substantially improved the manuscript.

REFERENCES

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

Baršić, B., Marton, E., Himbele, J., & Ravlić, Ž. (1996). Evaluation of the Glasgow coma scale score in critically ill infectious disease patients. *Infection*, *24*, 297–300.

Beran, R. (1981). *Nonparametric regression with randomly censored survival data*. Technical Report, University of California, Berkeley, CA.

Blanche, P., Dartigues, J.-F., & 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 (Methodological)*, *11*, 15–53.

Cai, T., & Cheng, S. (2008). Robust combination of multiple diagnostic tests for classifying censored event times. *Biostatistics*, *9*, 216–233.

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

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. *The Annals of Statistics*, *17*, 1157–1167.

Damen, J. A. A. G., Hooft, L., Schuit, E., Debray, T. P. A., Collins, G. S., Tzoulaki, I., …, Moons, K. G. M. (2016). Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ*, *353*, 1–11. https://www.bmj.com/content/353/bmj.i2416.

Etzioni, R., Pepe, M., Longton, G., Hu, C., & Goodman, G. (1999). Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer. *Medical Decision Making*, *19*, 242–251.

Gonzalez-Manteiga, W., & Cadarso-Suarez, C. (1994). Asymptotic properties of a generalized Kaplan–Meier estimator with some applications. *Journal of Nonparametric Statistics*, *4*, 65–78.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36.

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

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

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

Kamarudin, A. N., Cox, T., & Kolamunnage-Dona, R. (2017). Time-dependent ROC curve analysis in medical research: Current methods and applications. *BMC Medical Research Methodology*, *17*, 53.

Krzanowski, W. J., & Hand, D. J. (2009). ROC curves for continuous data *(1st ed.)*. London, UK: Chapman and Hall.

Lambert, J., & Chevret, S. (2016). Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves. *Statistical Methods in Medical Research*, *25*, 2088–2102.

Li, L., Greene, T., & 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.

López-Cheda, A., Cao, R., Jácome, M. A., & Van Keilegom, I. (2017). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Computational Statistics & Data Analysis*, *105*, 144–165.

Martínez-Camblor, P., Bayón, G. F., & Pérez-Fernández, S. (2016). Cumulative/dynamic ROC curve estimation. *Journal of Statistical Computation and Simulation*, *86*, 3582–3594.

Martínez-Camblor, P., & Pardo-Fernández, J. C. (2018). Smooth time-dependent receiver operating characteristic curve estimators. *Statistical Methods in Medical Research*, *27*, 651–674.

McIntosh, M. W., & Pepe, M. S. (2002). Combining several screening tests: Optimality of the risk score. *Biometrics*, *58*, 657–664.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford, UK: Oxford University Press.

Portier, F., Van Keilegom, I., & El Ghouch, A. (2016). *On proportional hazards cure models*. ISBA.

Qin, J., & Zhang, B. (2010). Best combination of multiple diagnostic tests for screening purposes. *Statistics in Medicine*, *29*, 2905–2919.

Rabin, B. A., Gaglio, B., Sanders, T., Nekhlyudov, L., Dearing, J. W., Bull, S., …, Marcus, A. (2013). Predicting cancer prognosis using interactive online tools: a systematic review and implications for cancer care providers. *Cancer Epidemiology and Prevention Biomarkers*, *22*, 1645–1656.

Schmid, M., Kestler, H. A., & Potapov, S. (2013). On the validity of time-dependent AUC estimators. *Briefings in Bioinformatics*, *16*, 153–168.

Schmid, M., & Potapov, S. (2012). A comparison of estimators to evaluate the discriminatory power of time-to-event models. *Statistics in Medicine*, *31*, 2588–2609.

Slate, E. H., & Turnbull, B. W. (2000). Statistical models for longitudinal biomarkers of disease onset. *Statistics in Medicine*, *19*, 617–637.

Stevens, R. J., Kothari, V., Adler, A. I., Stratton, I. M., & Holman, R. R. (2001). The UKPDS risk engine: A model for the risk of coronary heart disease in Type II diabetes (UKPDS 56). *Clinical Science*, *101*, 671–679.

Taweab, F., & Ibrahim, N. A. (2014). Cure rate models: A review of recent progress with a study of change-point cure models when cured is partially known. *Journal of Applied Sciences*, *14*, 609–616.

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

Van Keilegom, I., & Veraverbeke, N. (1997). Estimation and bootstrap with censored data in fixed design nonparametric regression. *Annals of the Institute of Statistical Mathematics*, *49*, 467–491.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., …, Foekens, J. A. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, *365*, 671–679.

Xu, J., & Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Canadian Journal of Statistics*, *42*, 1–17.

Yakovlev, A. Y., Asselain, B., Bardou, V. J., Fourquet, A., Hoang, T., Rochefediere, A., & Tsodikov, A. D. (1993). A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. *Biometrie et analyse de donnees spatio-temporelles*, *12*, 66–82.

Zheng, Y., Cai, T., & Feng, Z. (2006). Application of the time-dependent ROC curves for prognostic accuracy with multiple biomarkers. *Biometrics*, *62*, 279–287.