# Nonparametric kernel estimation of the probability of cure in a mixture cure model when the cure status is partially observed

**Wende Clarence Safari[1], Ignacio López-de-Ullibarri[2], and María Amalia Jácome[3]**

**Abstract**

Cure models are a class of time-to-event models where a proportion of individuals will never experience the event of interest. The lifetimes of these so-called cured individuals are always censored. It is usually assumed that one never knows which censored observation is cured and which is uncured, so the cure status is unknown for censored times. In this paper, we develop a method to estimate the probability of cure in the mixture cure model when some censored individuals are known to be cured. A cure probability estimator that incorporates the cure status information is introduced. This estimator is shown to be strongly consistent and asymptotically normally distributed. Two alternative estimators are also presented. The first one considers a competing risks approach with two types of competing events, the event of interest and the cure. The second alternative estimator is based on the fact that the probability of cure can be written as the conditional mean of the cure status. Hence, nonparametric regression methods can be applied to esti-mate this conditional mean. However, the cure status remains unknown for some censored individuals. Consequently, the application of regression methods in this context requires handling missing data in the response variable (cure status). Simulations are performed to evaluate the finite sample performance of the estimators, and we apply them to the analysis of two datasets related to survival of breast cancer patients and length of hospital stay of COVID-19 patients requiring intensive care.

**Keywords**

Bootstrap bandwidth, censoring, cure model, cure status, kernel estimators

**How to cite**

Safari WC, López-de-Ullibarri I, Jácome MA. Nonparametric kernel estimation of the probability of cure in a mixture cure model when the cure status is partially observed. *Statistical Methods in Medical Research*. 2022;31(11):2164-2188. doi:10.1177/09622802221115880

## 1 Introduction

Standard survival models assume that, without censoring, all the subjects would experience the event of interest at some time. Such assumption may not be true in practice, for example, some patients could never experience a recur-rence of their primary cancer after treatment.[1] These individuals are called long-term survivors, or, simply, cured from the event. In such cases, it is important to estimate the probability of being cured from the event. Cure models[2] address this issue.

The mixture cure model (MCM)[2] has received much attention in recent years. It assumes that the population is a mixture of cured and susceptible individuals. Note that here a "cured" individual is defined as being free of experiencing the event of interest even after being followed for an arbitrarily long time, and not necessarily cured in medical terms. Extensive research has been conducted for the standard MCM with both (semi)parametric[3–5] and nonparametric[6–8] approaches.

The absence of an individual's cure status (i.e. cured, uncured) is an important challenge for cure models. A subject whose event is observed is known to be uncured. However, censoring prevents from observing whether a censored

[1]Department of Mathematics, Faculty of Computer Science, CITIC, University of A Coruña, A Coruña, Spain
[2]Department of Mathematics, Escuela Politécnica de Ingeniería de Ferrol, University of A Coruña, A Coruña, , Spain
[3]Department of Mathematics, Faculty of Science, CITIC, University of A Coruña, A Coruña, Spain

*wende.safari@udc.es

subject would eventually experience the event or not. This hinders the classification of censored observations as cured or uncured. It is customary to assume no additional information on the cure status of censored individuals, thus, to model the cure status as a latent variable. Nonetheless, there are situations where some of the censored individuals can be identified to be immune to the event of interest, that is, to be cured. For example, diagnostic procedures in medical studies can provide further information on whether a subject suffering a curable illness can be considered as cured and therefore will not die from that disease. Also, for some types of cancer, it is extremely unlikely to have any recurrence later than a given time after treatment, known as cure threshold. Consequently, those patients with observed time surpassing the cure threshold can be considered cured from recurrence.[9] In settings with a subgroup identified as cured from the event, there are three groups of observations: the event times (of individuals experiencing the event during the follow-up time); the "regular" censored times (of those who neither experienced the event nor were classified as cured); and a new third group, the "cured" censored times (of those acknowledged as cured from the event). Just modeling the data under the usual cure model framework, that considers the "cured" censored times as simple "regular" censored times, will not take advantage of this additional cure status information given by the third group.

Few authors have studied cure models when the cure status is known for some censored observations. Nonparametric cure probability estimation with cure status partially available was discussed without covariates by Laska and Meisner[10] using the relationship of the cure rate with the survival function, and Betensky and Schoenfeld[11] by a competing risk approach. In a conditional setting with covariates, Safari et al.[12] developed a generalized product-limit estimator of the conditional survival function in the MCM when the cure status is partially known. In the parametric and semiparametric modeling framework, the logistic regression is typically used to model the probability of cure, while the survival function of the uncured individuals is modeled using a Bayesian semiparametric approach in the presence of covariates,[13] a Cox proportional hazards cure model,[14] or an accelerated failure time model with errors distributed according to the seminonparametric distribution.[15]

In this paper, we propose a kernel estimator of the conditional cure probability with covariates in the MCM when the cure status is partially known. It extends the unconditional nonparametric estimator by Laska and Meisner[10] to handle covariates, and the conditional nonparametric estimator by Xu and Peng,[6] tailored for the usual cure model framework, by incorporating the cure status information.

A different approach for modeling survival data with a cure fraction when cure is randomly observed proceeds through a competing risks model where the event of interest and being identified as cured are competing risks, and only the minimum of the times of the two risks is observed.[11,16] The probability of cure is the limit as time tends to infinity of the cumulative incidence function (CIF) of the competing risk given by cure.

Alternatively, note that the cure status is a binary variable and the cure probability can be regarded as the conditional expectation of the cure status. As a consequence, standard regression techniques can be used to model the probability of cure. In the present MCM, the cure status is observed for the individuals with uncensored times (uncured) and for some censored individuals (cured). But censored individuals with unknown cure status will either experience the event (censored and susceptible) or never experience the event (censored and cured) in the future. Therefore, the cure status is only partially observed. This turns the estimation of the cure probability into a regression problem with missing response values.

The most commonly used regression-based techniques to deal with missing data in the response variable are inverse probability weighting (IPW)[17–19] and multiple imputation (MI)[20–23] methods. In the MI approach,[24] multiple complete data sets are produced from the incomplete data by imputing the missing data $m$ times by some reasonable method. Then, one can perform standard regression analysis with the imputed data and the resulting estimates are combined to provide the final estimation. Commonly used imputation methods for missing response values include semiparametric imputation,[25] nearest neighbor imputation,[26] and kernel-based techniques.[27,20,28] Aerts et al.[20] applied the nonparametric kernel regression imputation scheme to estimate the unconditional mean of the response variable. In this paper, their procedure is extended to handle covariates, and applied when the response variable is the cure status. The method results in an estimator of the conditional cure probability based on a regression fit with multiple imputation for the unknown cure status. Note that this estimator strongly depends on the proportion of missing cure status, giving unreliable estimates under substantial missingness.

All the aforementioned estimators are valid in the general context where the identification of the cured subjects is random and does not necessarily depend on a specified cure threshold. This is the case for our motivating examples.

The first example relates to a dataset of breast cancer patients from The Cancer Genome Atlas.[29]A total of 898 breast cancer female patients were diagnosed and followed over time between 1988 and 2013. Demographic and clinical characteristics were recorded at baseline. Our goal was to estimate the probability of surviving from breast cancer given specific patient characteristics. That is, the event of interest is death from breast cancer. The observed time-to-event was considered as censored if the patient was alive at the end of the study, cancer-free either alive or dead, or lost to follow-up. Patients who have been free of cancer for at least 10 years can be considered to be long-term survivors or "known cured observations."[30,31] Note that the observed times-to-event of these "cured" individuals are large, at least 10 years.

The second example considers the $n = 2484$ COVID-19 patients hospitalized in Galicia (NW of Spain) during the first few weeks of the outbreak. The data was collected by the Galician Healthcare Service[32] with the aim of improving hospital management. During the initial phase of the pandemic, it was extremely necessary to estimate the occupancy of both ward and ICU dependencies, in order to avoid overloads of the Galician healthcare system. To do this, it was required to model the length of stay of patients in hospital capacities, and, in particular, to estimate the probability that a patient admitted to the hospital would finally need ICU admission. That is, the event of interest was admission to ICU of hospital inpatients. The observed time-to-event was censored if the patient did not enter ICU during the follow-up. Some censored patients were discharged or died before admission to ICU so they would never require ICU admission. In our context, this means that they can be considered as cured from the event. Unlike the previous example, the observed times-to-event of these "cured" individuals are not necessarily large, as patients may die or be discharged after a short stay.

The rest of the paper is organized as follows. In Section 2, a kernel estimator of the cure probability based on the MCM is proposed and its asymptotic properties are studied. Also, a bootstrap bandwidth selector is proposed. In Section 3, alternative estimators of the cure probability based, respectively, on the competing risks and regression approaches are presented. The performance of these estimators is illustrated with a simulation study in Section 4. The estimators are applied to the breast cancer and COVID-19 datasets in Section 5. Section 6 concludes with a discussion and thoughts for future work. Additional simulation results and proofs are relegated to the Supplemental Material.

## 2 Mixture cure model when the cure status is partially known

### 2.1 Model notation

This section starts by introducing the notation in the MCM model when some censored individuals are randomly observed to be cured from the event of interest. Let $Y$ be the time until the event of interest and $\boldsymbol{X}$ a covariate vector. The conditional cumulative distribution and survival functions of $Y$ are $F(t \mid \boldsymbol{x}) = P(Y \leq t \mid \boldsymbol{X} = \boldsymbol{x})$, and $S(t \mid \boldsymbol{x}) = 1 - F(t \mid \boldsymbol{x}) = P(Y > t \mid \boldsymbol{X} = \boldsymbol{x})$, respectively. Suppose the survival time $Y$ is censored by a random variable $C$, and $Y$ and $C$ are assumed to be conditionally independent given $\boldsymbol{X} = \boldsymbol{x}$. Under right censoring, only the pair $(T, \delta)$ is observed, where $T = \min(Y, C)$ and $\delta = \mathbf{1}(Y < C)$. The conditional distribution of the observed time $T$ is $H(t \mid \boldsymbol{x}) = P(T \leq t \mid \boldsymbol{X} = \boldsymbol{x})$. Without loss of generality, hereafter, we consider a continuous covariate $X$ with density function $m(x)$.

When considering a cure fraction, it is assumed that the time-to-event for a cured subject is $Y = \infty$, in order to represent the fact that the event will never happen. Let $\nu = \mathbf{1}(Y = \infty)$ denote the indicator of being cured from the event. Note that $\nu$ is partially observed because $\delta = 1$ implies $\nu = 0$, but $\nu$ is usually unknown for the censored observations. When $t$ becomes large, a fraction of the observations is still event-free and the survival function is improper:

$$\lim_{t \to \infty} S(t \mid x) > 0.$$

The MCM considers that the population is a mixture of the cured and uncured subgroups. The probability of being cured is $1 - p(x) = P(Y = \infty \mid X = x)$, and the conditional survival function of the uncured individuals, also known as latency, is $S_0(t \mid x) = P(Y > t \mid Y < \infty, X = x)$. The MCM writes $S(t \mid x)$ as

$$S\big(t \mid x\big) = 1 - p(x) + p(x) S_0\big(t \mid x\big). \tag{1}$$

In standard cure models, the cure status is known only for a subject who experienced the event during the follow-up period ($\delta = 1$) and thus known to be uncured ($\nu = 0$). For a censored subject ($\delta = 0$) the cure status $\nu$ is unknown, thus $\nu = 1$ is never observed. To accommodate the possible availability of the cure status information, let the censoring distribution be an improper distribution function $G(t \mid x) = (1 - \pi(x)) G_0(t \mid x)$, so with probability $\pi(x)$ the censoring variable is $C = \infty$, and with probability $1 - \pi(x)$ the value of the censoring variable $C$ corresponds to a random censoring time $C_0$ with proper continuous distribution function $G_0(t \mid x) = P(C_0 \leq t \mid X = x)$.

Thereof, a cured subject ($Y_i = \infty$), whose observed lifetime $T_i$ is always censored, is known to be cured if $C_i = \infty$. So, the cure status $\nu_i = 1$ is observed for some censored individuals. Note that a cured individual is identified with probability $\pi(x)$. Let $\xi$ be a binary random variable which indicates whether the cure status $\nu$ is observed ($\xi = 1$) or not ($\xi = 0$). If $Y$

and $C$ are independent conditional on $X = x$, then the conditional probability of observing the cure status is

$$P(\xi = 1 \mid X = x) = P(C < Y \mid X = x)p(x) + \pi(x)(1 - p(x)).$$

The observed data $\{(X_i, T_i, \delta_i, \xi_i\nu_i), i = 1, \ldots, n\}$ is classified into three groups:

(a) the event is observed and the individual is known to be uncured $(X_i, T_i, \delta_i = 1, \xi_i\nu_i = 0)$,
(b) the lifetime is censored and the cure status is unknown $(X_i, T_i, \delta_i = 0, \xi_i\nu_i = 0)$,
(c) the lifetime is censored and the individual is known to be cured $(X_i, T_i, \delta_i = 0, \xi_i\nu_i = 1)$,

where

$$T_i = \min(Y_i, C_i)\big[1 - \mathbf{1}(Y_i = \infty, C_i = \infty)\big] + C_{0i}\mathbf{1}(Y_i = \infty, C_i = \infty). \tag{2}$$

In this context, when the observed times of the individuals known to be cured are random, the random variable $C_0$ models these observed *cured times*. In the unconditional setting in Betensky and Schoenfeld,[11] besides $Y$ and $C$, the so-called variable $U$ plays a similar role. In standard cure models, where the cure status is unknown for some censored observations ($\pi(x) = 0$ and $C_i < \infty$, $\xi_i\nu_i = 0$, for $i = 1, \ldots, n$), then $T_i = \min(Y_i, C_i)$, and only groups (a) and (b) are considered.

To further understand the relationship between the notation introduced above and the usual notation in survival analysis under right censoring, let $T_i$ in (2) be the *actual* observed times, and let $\widetilde{T}_i = \min(Y_i, C_i)$ denote the *usual* observed time, as it is usually defined in standard survival analysis. Note that $\widetilde{T}_i = T_i$ in groups (a) and (b) above. But if an observation is known to be cured ($\xi_i\nu_i = 1$), then $Y_i = \infty$ and $C_i = \infty$, and the *usual* observed time is $\widetilde{T}_i = \infty$. Nonetheless, the *actual* observed time for the individuals known to be cured is always finite and is recorded as $T_i = C_{0i}$. Therefore, when an individual is known to be cured it is guaranteed to observe a cure time $C_{0i}$, similarly as in Betensky and Schoenfeld.[11] In summary, if the observed times $T_i$ of the censored individuals known to be cured are replaced by an extremely large observed time, say infinity, we recover the observations using the usual definition as $\tilde{T}_i = \min(Y_i, C_i)$.

One key issue in cure models is identifiability. This arises because of the lack of cure status information at the end of the follow-up period, resulting in difficulties in distinguishing models with high incidence of susceptibles and long tails of the latency distribution from low incidence of susceptibles and short tails of the latency distribution.[33] Following Hanin and Huang,[34] who discussed the identifiability of MCM, model (1) is identifiable if the latency function is proper. Thus, it is assumed that $\lim_{t\to\infty} S_0(t \mid x) = 0$ for all $x$. This condition is similar to the zero-tail constraint in Taylor,[9] López-Cheda et al.[7] and other works.

## 2.2 Proposed estimator of the cure probability

Assuming model (1), the cure probability can be written in terms of the survival function:

$$1 - p(x) = P(Y = \infty \mid X = x) = \lim_{t\to\infty} P(Y > t \mid X = x) = \lim_{t\to\infty} S(t \mid x). \tag{3}$$

An estimator of $1 - p(x)$ could be derived from an estimator of $S(t \mid x)$ by computing its limit when $t$ tends to infinity. Consider the generalized product-limit estimator of the conditional survival function $S(t \mid x)$ when the cure is partially observed,[12]

$$\widehat{S}_h^c\big(t \mid x\big) = \prod_{i=1}^{n}\left(1 - \frac{\delta_{[i]}B_{h[i]}(x)\mathbf{1}\big(T_{(i)} \le t\big)}{\sum_{j=i}^{n} B_{h[j]}(x) + \sum_{j=1}^{i-1} B_{h[j]}(x)\mathbf{1}\big(\xi_{[j]}\nu_{[j]} = 1\big)}\right), \tag{4}$$

where $X_{[i]}$, $\delta_{[i]}$, $\xi_{[i]}$ and $\nu_{[i]}$ are the concomitants of the ordered observed times $T_{(1)} \le \cdots \le T_{(n)}$, $B_{h[i]}(x)$ are the Nadaraya-Watson[35,36] (NW) weights,

$$B_{h[i]}(x) = \frac{K_h\big(x - X_{[i]}\big)}{\sum_{j=1}^{n} K_h\big(x - X_j\big)}, \tag{5}$$

and $K_h(\cdot) = K(\,\cdot\,/h)/h$ is a kernel function $K(\cdot)$ rescaled with bandwidth $h \to 0$ as $n \to \infty$. The proposed estimator of the cure probability $1 - p(x)$ is

$$1 - \widehat{p}_h^c(x) = \widehat{S}_h^c\Big(T_{(n)}^1 \mid x\Big) = \prod_{i=1}^{n}\left(1 - \frac{\delta_{[i]}B_{h[i]}(x)}{\sum_{j=i}^{n} B_{h[j]}(x) + \sum_{j=1}^{i-1} B_{h[j]}(x)\mathbf{1}\big(\xi_{[j]}\nu_{[j]} = 1\big)}\right), \tag{6}$$

where $T^1_{(n)} = \max_{i:\delta_i=1} T_i$ is the largest uncensored observed lifetime. An important feature of the estimator is that one does not need to know the observed censored time for the individuals who are known to be cured, as this estimator is precisely the XP[6] estimator computed with the *usual* observed times, that is, if the observed times of the individuals known to be cured are replaced with an extremely large value (e.g. infinity).

**Proposition 1.** The estimator $1 - \widehat{p}^c_h(x)$ has the following basic properties.

1. When there are no censored observations known to be cured, that is, $\xi_i \nu_i = 0$ for $i = 1, \ldots, n$, $1 - \widehat{p}^c_h(x)$ reduces to the XP estimator by Xu and Peng[6]:

$$1 - \widehat{p}_h(x) = \prod_{i=1}^{n} \left( 1 - \frac{\delta_{[i]} B_{h[i]}(x)}{\sum_{j=i}^{n} B_{h[j]}(x)} \right). \tag{7}$$

2. When individuals are classified as cured if their observed survival time exceeds a known fixed cure threshold, $1 - \widehat{p}^c_h(x)$ reduces to the XP estimator.
3. When there is no censoring, all the cure status indicators $\nu_i$ are observed ($\xi_i = 1$, $i = 1, \ldots, n$). Then, $1 - \widehat{p}^c_h(x)$ reduces to the NW estimator:

$$1 - \widehat{p}^{\mathrm{NW}}_h(x) = \sum_{i=1}^{n} B_{hi}(x) \mathbf{1}(\nu_i = 1) = \frac{\sum_{i=1}^{n} K_h(x - X_i)\nu_i}{\sum_{j=1}^{n} K_h\left(x - X_j\right)}. \tag{8}$$

   In this case, the XP estimator will be zero.
4. In an unconditional setting, the proposed estimator is

$$1 - \widehat{p}^c_n = \prod_{i=1}^{n} \left( 1 - \frac{\delta_{[i]}}{n - i + 1 + \sum_{j=1}^{i-1} \mathbf{1}\left(\xi_{[j]}\nu_{[j]} = 1\right)} \right). \tag{9}$$

   In the case where an individual is known to be cured only if the observed time is greater than a known fixed time, $1 - \widehat{p}^c_n$ reduces to the generalized maximum likelihood estimator[10]:

$$1 - \widehat{p}_n = \prod_{i=1}^{n} \left( 1 - \frac{\delta_{[i]}}{n - i + 1} \right). \tag{10}$$

   Moreover, if there are no individuals known to be cured, then $1 - \widehat{p}^c_n$ becomes the unconditional version of the XP estimator.

The proof of these properties is outlined in Appendix B of the Supplemental Material.

**Proposition 2.** The estimator $1 - \widehat{p}^c_h(x)$ in (6) is the nonparametric local maximum likelihood estimator of $1 - p(x)$.
The proof of Proposition 2 follows the proof of Proposition 2 in Safari et al.[12], and it is omitted.

## 2.3 Asymptotic results

Here, the asymptotic properties of $1 - \widehat{p}^c_h(x)$ are studied. For brevity the required assumptions are in Appendix A of the Supplemental Material. Let us define the following (sub)distribution functions:

$$H\left(t \mid x\right) = P\left(T \le t \mid X = x\right), \tag{11}$$

$$H^1(t \mid x) = P\left(T \le t, \delta = 1 \mid X = x\right), \tag{12}$$

$$H^{11}(t \mid x) = P\left(T \le t, \xi\nu = 1 \mid X = x\right), \tag{13}$$

$$J(t \mid x) = 1 - H\left(t \mid x\right) + H^{11}\left(t \mid x\right). \tag{14}$$

The functions $H(t \mid x)$ and $H^1(t \mid x)$ are the conditional distribution function of the observed times $T_i$, and the conditional subdistribution function of the observed events, respectively. $H^{11}(t \mid x)$ provides insight on the distribution of the *cure times*, that is, the recorded times $T_i$ of the individuals known to be cured. $J(t \mid x)$ is the survival function of the observed times defined with the *usual* definition, that is, the conditional survival function of $\widetilde{T} = \min(Y, C)$.

The NW kernel estimators of (11)–(14) are, respectively

$$\widehat{H}_h(t \mid x) = \sum_{i=1}^{n} B_{hi}(x)\mathbf{1}(T_i \le t),$$

$$\widehat{H}_h^1(t \mid x) = \sum_{i=1}^{n} B_{hi}(x)\mathbf{1}(T_i \le t, \delta_i = 1),$$

$$\widehat{H}_h^{11}(t \mid x) = \sum_{i=1}^{n} B_{hi}(x)\mathbf{1}(T_i \le t, \xi_i\nu_i = 1),$$

$$\widehat{J}_h(t^- \mid x) = \sum_{i=1}^{n} B_{hi}(x)\mathbf{1}(T_i \ge t) + \sum_{i=1}^{n} B_{hi}(x)\mathbf{1}(T_i < t, \xi_i\nu_i = 1).$$

Further, define $\tau_H(x) = \inf\{t : H(t \mid x) = 1\}$, $\tau_{S_0}(x) = \inf\{t : S_0(t \mid x) = 0\}$ and $\tau_{G_0}(x) = \inf\{t : G_0(t \mid x) = 1\}$. Note that $\tau_H(x) = \max\{\tau_{S_0}(x), \tau_{G_0}(x)\}$. Let $\tau_0 = \sup_{x \in I} \tau_{S_0}(x)$, then it is required that

$$\tau_0 < \tau_{G_0}(x) \text{ for any } x \text{ with probability 1.} \tag{15}$$

Condition (15) relies on the assumption that the follow-up is long enough for the support of the latency function $S_0(t \mid x)$ to be contained within the support of $G_0(t \mid x)$. This implies that all observations censored after the largest uncensored observed lifetime correspond to cured subjects, as the susceptible subjects will experience the event within the follow-up period. This condition guarantees that the proposed estimator does not overestimate the true probability of cure. A similar condition has been used in the literature.[10,6–8] Xu and Peng[6] pointed out that if $G_0(t \mid x)$ has a heavier tail than $S_0(t \mid x)$, then condition (15) can be relaxed. Maller and Zhou[37] proposed a test to assess whether a condition analogous to (15) is fulfilled in an unconditional setting. It is based on the difference between the largest observed time $T_{(n)}$ and the largest uncensored time $T^1_{(n)}$. If this interval is large, then there is sufficient follow-up time and (15) can be assumed. With covariates, one may divide the data into subgroups according to the covariate values and apply this test in each subgroup. The next theorem establishes an asymptotic representation for $1 - \widehat{p}_h^c(x)$. The proof is in Appendix B of the Supplemental Material. Based on that result, we prove the strong consistency and asymptotic normality. In Section 2.4, we will provide evidence that $1 - \widehat{p}_h^c(x)$ has smaller asymptotic variance than the XP estimator $1 - \widehat{p}_h(x)$.

**Theorem 1** (**Asymptotic representation**). Suppose that Assumptions 1–8 and condition (15) hold, and the bandwidth $h = (h_n)$ satisfies $h \to 0$, $\log n/(nh) \to 0$ and $nh^5/\log n = O(1)$ as $n \to \infty$. Then, for $x \in I$,

$$\{1 - \widehat{p}_h^c(x)\} - \{1 - p(x)\} = \{1 - p(x)\} \sum_{i=1}^{n} \widetilde{B}_{hi}(x)\zeta\big(T_i, \delta_i, \xi_i, \nu_i, \tau_0, x\big) + R_n(x),$$

where

$$\zeta\big(T_i, \delta_i, \xi_i, \nu_i, t, x\big) = \frac{\mathbf{1}(T_i \le t, \delta_i = 1)}{J(T_i^- \mid x)} - \int_0^t \big\{\mathbf{1}(T_i \ge v) + \mathbf{1}\big(T_i < v, \xi_i\nu_i = 1\big)\big\} \frac{dH^1\big(v \mid x\big)}{J^2\big(v^- \mid x\big)}, \tag{16}$$

$$\widetilde{B}_{hi}(x) = \frac{1}{m(x)} \frac{1}{nh} K\left(\frac{x - X_i}{h}\right) \tag{17}$$

and $R_n(x)$ satisfies

$$\sup_{x \in I} | R_n(x) | = O\Big(\big(\log n\big)^{3/4} (nh)^{-3/4}\Big) \ a.s. \tag{18}$$

The following corollary on the strong consistency of the estimator $1 - \widehat{p}_h^c(x)$ is deduced from Theorem 1.

**Corollary 1** (**Strong consistency**). Suppose that Assumptions 1 to 8 and condition (15) hold, and the bandwidth $h = (h_n)$ satisfies $h \to 0,\ \log n/(nh) \to 0$ and $nh^5/\log n = O(1)$ as $n \to \infty$. Then, for $x \in I$,

$$\sup_{x \in I} \mid \widehat{p}_h^c(x) - p(x) \mid = O\Big((nh)^{-1/2}(\log n)^{1/2}\Big) \ a.s.$$

The corollary is proved by considering the asymptotic representation in Theorem 1 and following similar arguments as in the proof of Corollary 1 in Safari et al.[12] for $t = T_{(n)}^1$, so it is omitted.

The proof of the next proposition is in Appendix B of the Supplemental Material.

**Proposition 3** (**Asymptotic bias and variance**). Suppose that Assumptions 1 to 8 and condition (15) hold and the bandwidth $h = (h_n)$ satisfies $h \to 0,\ \log n/(nh) \to 0$ and $nh^5/\log n = O(1)$ as $n \to \infty$. Then, the asymptotic bias and variance of the dominant term of $1 - \widehat{p}_h^c(x)$ are, respectively,

$$\mu_{h,c}(x) = h^2 B_c(x) + O\big(h^4\big) \text{ and } \sigma^2_{h,c}(x) = \frac{1}{nh} s_c^2(x) + O\Big(\frac{h}{n}\Big).$$

The function $B_c(x)$ in the dominant term of the bias is

$$B_c(x) = \big(c_{1,c}(x) + c_{2,c}(x)\big) d_K \tag{19}$$

with $d_K = \int v^2 K(v) dv$,

$$c_{1,c}(x) = \frac{2(1-p(x))' m'(x) + (1-p(x))'' m(x)}{2m(x)}, \tag{20}$$

and

$$c_{2,c}(x) = (1-p(x)) \int_0^{\tau_0} \frac{G'(v^- \mid x)}{1 - G(v^- \mid x)} \frac{d}{ds} \left( \frac{S'\big(s \mid x\big)}{S\big(s \mid x\big)} \right) \Big|_{s=v^-} dv. \tag{21}$$

Here $p'(x)$, $p''(x)$, $S'(t \mid x)$ and $G'(t \mid x)$ refer to the derivatives with respect to $x$. The function $s_c^2(x)$ in the dominant term of the variance is

$$s_c^2(x) = \frac{\big(1-p(x)\big)^2}{m(x)} \int_0^{\tau_0} \frac{dH^1\big(v^- \mid x\big)}{\big(1 - H\big(v^- \mid x\big) + H^{11}(v^- \mid x)\big)^2} c_K, \tag{22}$$

with $c_K = \int K^2(v) dv$.

The following theorem establishes the asymptotic normality of $1 - \widehat{p}_h^c(x)$. The proof is in Appendix B of the Supplemental Material.

**Theorem 2** (**Asymptotic normality**). Suppose that Assumptions 1 to 8 and (15) hold, and the bandwidth $h = (h_n)$ satisfies $h \to 0,\ \log n = (nh) \to 0$ and $nh^5/\log n = O(1)$ as $n \to \infty$. Then for $x \in I$ it follows that:

(i) If $nh^5 \to 0$ and $(\log n)^3/(nh) \to 0$, then

$$(nh)^{1/2}\big\{\widehat{p}_h^c(x) - p(x)\big\} \to N(0, s_c^2(x)) \text{ in distribution.}$$

(ii) If $nh^5 \to C$, where $C > 0$ is a constant then

$$(nh)^{1/2}\big\{\widehat{p}_h^c(x) - p(x)\big\} \to N(C^{5/2} B_c(x), s_c^2(x)) \text{ in distribution,}$$

where $B_c(x)$ is defined in (19) and $s_c^2(x)$ in (22).

## 2.4 Effect of ignoring the cure status information

The use of the information given by the cure status has an impact on both the bias and variance of the proposed estimator of the cure probability $1 - \widehat{p}_h^c(x)$. When the cure status is ignored and the observed times of the individuals known to be cured are considered as simple censored times, the asymptotic expressions of the bias and variance of the XP estimator, $1 - \widehat{p}_h(x)$, are

$$\mu_h(x) = h^2 B(x) + O\left(h^4\right) \text{ and } \sigma_h^2(x) = \frac{1}{nh} s^2(x) + O\left(\frac{h}{n}\right),$$

where $B(x) = (c_{1,c}(x) + c_2(x))d_K$, with $c_{1,c}(x)$ in (20),

$$c_2(x) = (1 - p(x)) \int_0^{\tau_0} \frac{G_0'(v^- \mid x)}{1 - G_0(v^- \mid x)} \frac{d}{ds} \left( \frac{S'\left(s \mid x\right)}{S\left(s \mid x\right)} \right) \Big|_{s=v^-} dv, \tag{23}$$

and

$$s^2(x) = \frac{\left(1 - p(x)\right)^2}{m(x)} \int_0^{\tau_0} \frac{dH^1\left(v^- \mid x\right)}{\left(1 - H\left(v^- \mid x\right)\right)^2} c_K.$$

The cure status information affects the bias of the proposed estimator only in the second term of $B_c(x)$ in (19). If the cure status information is ignored, the term $1 - \pi(x)$ in $G(t \mid x)$ disappears and results to $c_2(x)$ in (23). Therefore, in terms of bias, the gain of knowing the cure status is not straightforward as it depends on the derivatives of $(1 - \pi(x))$ and $G_0(t \mid x)$.

The effect of considering the cure status information on the variance is through the function $H^{11}(t \mid x)$ of $s_c^2(x)$ in (22). When the cure status information is ignored, then $H^{11}(t \mid x) = 0$ and $s_c^2(x) \leq s^2(x)$ for all $x$. So, when the known cure status is taken into account, the variance of the proposed estimator of the cure probability decreases asymptotically with respect to the XP estimator.

## 2.5 Bandwidth selection

The proposed estimator of the probability of cure $1 - \widehat{p}_h^c(x)$ in (6), like the XP estimator $1 - \widehat{p}_h(x)$, depends on the bandwidth $h$. Different approaches for bandwidth selection of kernel estimators are available in the literature.[7] Here, a bootstrap bandwidth selector is proposed to choose the smoothing parameter $h$ for the cure rate estimator $\widehat{p}_h^c(x)$. The principle is to select the bandwidth $h$ that minimizes the bootstrap version of the MSE approximated by Monte Carlo as:

$$\text{MSE}_x^*(h) \simeq \frac{1}{B} \sum_{b=1}^{B} \{\widehat{p}_h^{c,*b}(x) - \widehat{p}_g^c(x)\}^2, \tag{24}$$

where $1 - \widehat{p}_h^{c,*b}(x)$ is the proposed estimator computed with the $b$th bootstrap resample and bandwidth $h$, $1 - \widehat{p}_g^c(x)$ is the proposed estimator computed with the original sample and pilot bandwidth $g$. The algorithm to compute the bootstrap bandwidth for a fixed covariate value $x$ is

Step 1. With the original sample and the pilot bandwidth $g$, compute $1 - \widehat{p}_g^c(x)$ in (6).
Step 2. Choose a dense enough grid of $L$ bandwidths $\{h_1, \ldots, h_L\}$.
Step 3. Generate $B$ bootstrap resamples $\{(X_i^{*(b)}, T_i^{*(b)}, \delta_i^{*(b)}, \xi_i^{*(b)}\nu_i^{*(b)}) : i = 1, \ldots, n\}$, for $b = 1, \ldots, B$.
Step 4. With the $b$th bootstrap resample and the bandwidth $h_l$ compute $1 - \widehat{p}_{h_l}^{c,*b}(x)$, for $l = 1, \ldots, L$.
Step 5. For $h_l$, $l = 1, \ldots, L$, compute the Monte Carlo approximation $\text{MSE}_x^*(h_l)$ given by (24).
Step 6. The bootstrap bandwidth, $h_x^*$, is the bandwidth of the grid $\{h_1, \ldots, h_L\}$ that minimizes $\text{MSE}_x^*(h)$ in (24).

The bootstrap resamples in Step 3 are generated similarly as in Li and Datta,[38] using the simple weighted bootstrap or the obvious bootstrap resampling methods. Without ties in the observed times, both methods are equivalent.[39]

## 2.6 The simple weighted bootstrap

Generate $\{X_1^*, \ldots, X_n^*\}$ from the empirical distribution of $\{X_1, \ldots, X_n\}$. Next, for each $X_i^*$ generate $(T_i^*, \delta_i^*, \xi_i^*\nu_i^*)$ from the weighted empirical conditional distribution

$$\widehat{F}_g(t, d, z \mid X_i^*) = \sum_{j=1}^{n} B_{gj}(X_i^*) \mathbf{1}\left(T_j \leq t, \delta_j \leq d, \xi_j \nu_j \leq z\right)$$

where $B_{gj}(x)$ are the NW weights (5) with bandwidth $g$.

## 2.7 The obvious bootstrap

Consider the estimator of the survival function $1 - \widehat{F}_g^c(t \mid x)$ in (4) and the estimator for the censoring distribution:

$$1 - \widehat{G}_g^c(t \mid x) = \prod_{i=1}^{n} \left( 1 - \frac{(1 - \delta_{[i]}) \mathbf{1}(\xi_{[i]} \nu_{[i]} = 0) B_{g[i]}(x) \mathbf{1}(T_{(i)} \leq t)}{\sum_{j=i}^{n} B_{g[j]}(x) + \sum_{j=1}^{i-1} B_{g[j]}(x) \mathbf{1}(\xi_{[j]} \nu_{[j]} = 1)} \right),$$

computed with the same pilot bandwidth $g$. Simulate the bootstrap sample $\{(X_i^*, T_i^*, \delta_i^*, \xi_i^* \nu_i^*), i = 1, \ldots, n\}$ as follows.

Step 1. Generate $\{X_1^*, \ldots, X_n^*\}$ from the empirical distribution of $\{X_1, \ldots, X_n\}$.
Step 2. For $i = 1, \ldots, n$, set $Y_i^* = \infty$ with probability $1 - \widehat{F}_g^c(T_{(n)} \mid X_i^*) = 1 - \widehat{p}_g^c(X_i^*)$, and generate a finite survival time otherwise:

$$Y_i^* \sim \frac{1 - \widehat{F}_g^c(t \mid X_i^*) - (1 - \widehat{p}_g^c(X_i^*))}{\widehat{p}_g^c(X_i^*)}.$$

Generate $C_i^* = \infty$ with probability $1 - \widehat{G}_g^c(T_{(n)} \mid X_i^*) = \widehat{\pi}_g^c(X_i^*)$, and

$$C_i^* \sim \frac{1 - \widehat{G}_g^c(t \mid X_i^*) - \pi_g^c(X_i^*)}{1 - \widehat{\pi}_g^c(X_i^*)}, \text{ otherwise.}$$

Let $\widehat{G}_{0g}(t \mid x)$ be the kernel estimator of $G_0(t \mid x)$, the distribution function of the observed times of the individuals known to be cured:

$$\widehat{G}_{0g}(t \mid x) = \frac{\sum_{i=1}^{n} B_{gi}(x) \mathbf{1}(T_i \leq t, \xi_i \nu_i = 1)}{\sum_{i=1}^{n} B_{gi}(x) \mathbf{1}(\xi_i \nu_i = 1)}.$$

For each $i = 1, \ldots, n$, generate $C_{0i}^*$ from $\widehat{G}_{0g}(t \mid X_i^*)$. The bootstrap sample is $\{(T_i^*, \delta_i^*, \xi_i^* \nu_i^*), i = 1, \ldots, n\}$ where

$$\begin{aligned}
T_i^* &= \min(Y_i^*, C_i^*)[1 - \mathbf{1}(Y_i^* = \infty, C_i^* = \infty)] + C_{0i}^* \mathbf{1}(Y_i^* = \infty, C_i^* = \infty), \\
\delta_i^* &= \mathbf{1}(Y_i^* < C_i^*), \\
\xi_i^* \nu_i^* &= \mathbf{1}(Y_i^* = \infty, C_i^* = \infty).
\end{aligned}$$

For computational efficiency, see López-Cheda et al.[7,8] the bootstrap values of the covariate can be set to $X_i^* = X_i$ instead of resampling it randomly from $\{X_1, \ldots, X_n\}$.

López-Cheda et al.[7,8] and Safari et al.[12] proposed to use a local pilot bandwidth depending on the sample size and the distribution of the covariate:

$$g_x = \frac{d_k^+(x) + d_k^-(x)}{2} 100^{1/9} n^{-1/9},$$

where $d_k^+(x)$ and $d_k^-(x)$ are the distances from $x$ to the $k$th nearest neighbor on the right and left, and $k$ is the integer part of $n/4$. If there are not at least $k$ neighbors on the right (or left), we use $d_k^+(x) = d_k^-(x)$ (or $d_k^-(x) = d_k^+(x)$). The simulations in López-Cheda et al.[7] and Safari et al.[12] demonstrated that the choice of the pilot bandwidth has small effect on the bootstrap bandwidth.

## 3 Alternative estimators of the cure rate

In this section, we introduce alternative methods for estimating the cure rate with covariates.

### 3.1 Competing risks estimators

The competing risks model considers that an individual is exposed to $J$ types of failure or competing risks. For $j \in \{1, \ldots, J\}$, let $Y_j$ the time until the failure of type $j$ happens, and consider the random pair $(Y_F, D)$, where $Y_F = \min(Y_1, \ldots, Y_J)$ is the time until the first failure, and $D \in \{1, 2, \ldots, J\}$ indicates the type of failure. Let $C$ be a censoring variable. Under right random censoring, the observations $(Y_F, D)$ will be incomplete if follow-up ends before any failure occurs. In this situation only $(T, \Delta)$ is observed, where $T = \min(Y_F, C)$ is the possibly censored observed time, and $\Delta =$

$\mathbf{1}(Y_F < C)D$ is the type of event in the case a terminal event occurs and $\Delta = 0$ indicates that the failure type is unknown and the failure time is right-censored. The censoring mechanism is assumed to be non-informative.[40] This general competing risks model usually assumes that all patients will eventually experience one of the $J$ possible types of risks if there is sufficient follow-up and, therefore, does not consider the possibility of cure.

In the MCM with cured individuals randomly observed, Betensky and Schoenfeld[11] showed that the event of interest and cure can be regarded as two competing risks, in which cures are random and only the minimum between the cure and event times is observed. The probability of cure is then simply the limit as $t$ tends to infinity of the CIF of the cure, or just 1 minus the limit of the CIF of the event of interest. We adopt this perspective and introduce a competing risks model for the MCM when the cure status is partially observed in the presence of covariates. Let $\{Y_E, Y_c\}$ be the latent failure times of 2 type failures: the event of interest ($E$) and the classification of an individual as cured ($c$). Let $Y_F = \min(Y_E, Y_c)$ be the time of the first failure and $C$ the censoring time. For right censored competing risks data, let $T = \min(Y_E, Y_c, C)$ be the observed time, and the uncensoring indicator $\Delta = \mathbf{1}(Y_F < C)D$ where $D \in \{1, 2\}$ is the type of risk. In this context, the observed sample $\{(T_i, \delta_i, \xi_i\nu_i), i = 1, \ldots, n\}$ can be written as $\{(T_i, \Delta_i), i = 1, \ldots, n\}$, where

$$\Delta_i = \begin{cases} 0 & \text{if } \delta_i = 0,\ \xi_i\nu_i = 0 \text{ (censored)} \\ 1 & \text{if } \delta_i = 1,\ \xi_i\nu_i = 0 \text{ (event observed)} \\ 2 & \text{if } \delta_i = 0,\ \xi_i\nu_i = 1 \text{ (known to be cured).} \end{cases}$$

The CIF of the event of interest $E$ is the probability that a failure of type 1 occurs at or before time $t$:

$$F_1\left(t \mid x\right) = P\left(Y_F \le t, D = 1 \mid X = x\right).$$

The CIF of the second competing risk (individual known to be cured) is

$$F_2\left(t \mid x\right) = P\left(Y_F \le t, D = 2 \mid X = x\right).$$

The probability of cure is simply the CIF of the competing risk *cure* ($c$) evaluated at infinity or the complementary of the CIF of the event of interest ($E$) evaluated at infinity:

$$1 - p(x) = P\left(Y_E = \infty \mid X = x\right) = 1 - \lim_{t\to\infty} F_1(t \mid x).$$

Equivalently,

$$1 - p(x) = P\left(Y_c < \infty \mid X = x\right) = \lim_{t\to\infty} F_2(t \mid x).$$

The conditional version of the estimators of the CIFs in Klein and Moeschberger[41] are the following (also see Effraimidis and Dahl [42]):

$$\begin{aligned} \widehat{F}_{1,h}\left(t \mid x\right) &= \sum_{i=1}^{n} \frac{\delta_{[i]} B_{h[i]}(x) \mathbf{1}\left(T_{(i)} \le t\right)}{\sum_{j=i}^{n} B_{h[j]}(x)} \prod_{k=1}^{i-1} \left(1 - \frac{\left(\delta_{[k]} + \xi_{[k]}\nu_{[k]}\right) B_{h[k]}(x)}{\sum_{j=k}^{n} B_{h[j]}(x)}\right) \\ &= \sum_{i=1}^{n} \frac{\delta_{[i]} B_{h[i]}(x) \mathbf{1}\left(T_{(i)} \le t\right)}{\sum_{j=i}^{n} B_{h[j]}(x)} \widehat{S}_h\left(T_{(i)}^{-} \mid x\right), \end{aligned} \tag{25}$$

$$\begin{aligned} \widehat{F}_{2,h}\left(t \mid x\right) &= \sum_{i=1}^{n} \frac{\xi_{[i]}\nu_{[i]} B_{h[i]}(x) \mathbf{1}\left(T_{(i)} \le t\right)}{\sum_{j=i}^{n} B_{h[j]}(x)} \prod_{k=1}^{i-1} \left(1 - \frac{\left(\delta_{[k]} + \xi_{[k]}\nu_{[k]}\right) B_{h[k]}(x)}{\sum_{j=k}^{n} B_{h[j]}(x)}\right) \\ &= \sum_{i=1}^{n} \frac{\xi_{[i]}\nu_{[i]} B_{h[i]}(x) \mathbf{1}\left(T_{(i)} \le t\right)}{\sum_{j=i}^{n} B_{h[j]}(x)} \widehat{S}_h\left(T_{(i)}^{-} \mid x\right), \end{aligned} \tag{26}$$

where $\widehat{S}_h(t \mid x)$ is the Beran[43] estimator obtained by treating any of the competing risks as an event. The estimation of the conditional CIFs in (25) and (26) allows us to model the conditional cure probability.

**Proposition 4.** The conditional probability of cure $1 - p(x)$ can be estimated by

$$1 - \widehat{p}_{1,h}(x) = 1 - \lim_{t\to\infty} \widehat{F}_{1,h}(t \mid x) = 1 - \sum_{i=1}^{n} \frac{\delta_{[i]} B_{h[i]}(x)}{\sum_{j=i}^{n} B_{h[j]}(x)} \widehat{S}_h\left(T_{(i)}^{-} \mid x\right), \tag{27}$$

or by

$$1-\widehat{p}_{2,h}(x)=\lim_{t\to\infty}\widehat{F}_{2,h}\left(t \mid x\right)=\sum_{i=1}^{n}\frac{\xi_{[i]}\nu_{[i]}B_{h[i]}(x)}{\sum_{j=i}^{n}B_{h[j]}(x)}\widehat{S}_h\left(T_{(i)}^{-} \mid x\right). \tag{28}$$

If the last observation is an event or an observed cured individual, then the sum of both estimated CIFs equals 1 at the limit when $t$ tends to infinity, and $1-\widehat{p}_{1,h}(x)=1-\widehat{p}_{2,h}(x)$. If, however, the last observation is censored, then $1-\widehat{p}_{1,h}(x)$ and $1-\widehat{p}_{2,h}(x)$ are not equivalent. In this case, $1-\widehat{p}_{1,h}(x)$ is an upper bound for the cure rate and $1-\widehat{p}_{2,h}(x)$ is a lower bound. The differences in such a case are more apparent when the censoring rate is high.

Note that, in the absence of censoring,

$$\begin{aligned}\widehat{S}_h\left(T_{(i)}^{-} \mid x\right)&=\prod_{k=1}^{i-1}\left(1-\frac{\left(\delta_{[k]}+\xi_{[k]}\nu_{[k]}\right)B_{h[k]}(x)}{\sum_{j=k}^{n}B_{h[j]}(x)}\right)=\prod_{k=1}^{i-1}\frac{\sum_{j=k+1}^{n}B_{h[j]}(x)}{\sum_{j=k}^{n}B_{h[j]}(x)}\\&=\sum_{j=i}^{n}B_{h[j]}(x).\end{aligned}$$

So the cure probability estimator is

$$\begin{aligned}1-\widehat{p}_{1,h}(x)&=1-\sum_{i=1}^{n}\frac{\delta_{[i]}B_{h[i]}(x)}{\sum_{j=i}^{n}B_{h[j]}(x)}\widehat{S}_h\left(T_{(i)}^{-} \mid x\right)=1-\sum_{i=1}^{n}\frac{\delta_{[i]}B_{h[i]}(x)}{\sum_{j=i}^{n}B_{h[j]}(x)}\sum_{j=i}^{n}B_{h[j]}(x)\\&=1-\sum_{i=1}^{n}\delta_{[i]}B_{h[i]}(x)=\sum_{i=1}^{n}\xi_{[i]}\nu_{[i]}B_{h[i]}(x).\end{aligned}$$

Similarly,

$$1-\widehat{p}_{2,h}(x)=\sum_{i=1}^{n}\xi_{[i]}\nu_{[i]}B_{h[i]}(x).$$

Thus, the cure rate estimators reduce to the sum of the weights of the individuals known to be cured, which is the NW estimator of the cure probability in (8).

This approach to the estimation of the cure rate can be viewed as arising from a redistribution to the right algorithm.[44] In particular, the mass of $B_{hi}(x)$ initially assigned to the censored observations (neither event nor observed to be cured) is redistributed equally to all subjects at risk for the event and cure at the time of censoring. The cure rate is then simply the weighted sum of the mass attached to each subject that is cured.

## 3.2 Multiply imputed NW estimator

The proposed estimator in (6) is based on the relationship in (3) between the cure probability $1-p(x)$ and the survival function $S(t \mid x)$. So, to estimate the probability of cure it requires the observations $\{(T_i, \delta_i), i=1,\dots,n\}$. Nonetheless, the cure probability can also be written as $1-p(x)=E(\nu \mid X=x)$, that is, the conditional expectation of the cure status $\nu$, or equivalently $1-p=E(\nu)$ for an unconditional setting. An estimator based on this latter relationship would only require the observed values of the covariate $X$ and the cure status $\nu$, dismissing the observed values of $(T, \delta)$.

The NW estimator is one of the most frequently used estimators in nonparametric regression. So, the estimator in (8) might be considered for the estimation of the cure probability $1-p(x)=E(\nu \mid X=x)$. Similarly, the unconditional cure probability $1-p=E(\nu)$ might easily be estimated using the empirical estimator $1-\widehat{p}=\sum_{i=1}^{n}\nu_i/n$. These methods require that the cure status $\nu$ is completely observed. However, in the present setup, the cure status $\nu$ remains unknown for some of the censored observations. There has been extensive work dealing with regression modeling when the response variable is only partially observed.[28,45,46] Aerts et al.[20] developed a fully nonparametric local multiple imputation (MI) procedure to estimate the unconditional mean of a variable in the presence of missing response data. When the cure status is not completely observed because of censoring, but it is partially available, their methodology can be applied to the estimation of $1-p$. To the best of our knowledge, the MI methodology in Aerts et al.[20] has not been extended to estimate the conditional mean. We extend this methodology to estimate $1-p(x)$ in the presence of a covariate.

It is important to define the nature of the missingness mechanism, as it highly influences the performance of statistical techniques that deal with missing data. The MI estimator[20] requires the strongly ignorable missing at random (siMAR) assumption,[47] which implies the probability that the cure status is observed depends only on the covariate $X$ but not on the response variable $\nu$:

$$E(\xi \mid X) = E(\xi \mid X, \nu). \tag{29}$$

This is weaker than missingness completely at random (MCAR) from dependence on the observed variable $X$ is allowed. Note that this siMAR condition is not fulfilled under the MCM model with the cure status partially known if $Y$ and $C$ are conditionally independent given $X = x$, as the probability of observing the cure status is different for the cured ($\nu = 1$) and the susceptible ($\nu = 0$) individuals, and therefore it depends on the cure status:

$$\begin{aligned}
E(\xi \mid X, \nu = 1) &= P(\xi = 1 \mid X, Y = \infty) = P(C = \infty \mid X, Y = \infty) = \pi(X), \\
E(\xi \mid X, \nu = 0) &= P(\xi = 1 \mid X, Y < \infty) = P(Y < C \mid X, Y < \infty) \\
&= P(Y < C, C < \infty \mid X, Y < \infty) + P(C = \infty \mid X, Y < \infty) \\
&= P(Y < C \mid X, Y < \infty, C < \infty)(1 - \pi(X)) + \pi(X).
\end{aligned}$$

Unless $P(Y < C \mid X, Y < \infty, C < \infty) = 0$, which yields the time of all the susceptible individuals to be censored, the siMAR condition (29) cannot be assumed if $Y$ and $C$ are independent conditionally on $X = x$. Nonetheless, note that the higher the value of $\pi(X)$, which means the lower missingness rate in the cure status, the smaller the difference between $E(\xi \mid X, \nu = 1)$ and $E(\xi \mid X, \nu = 0)$, and the closer the siMAR assumption to hold.

The main idea in the approach of Aerts et al.[20] is to use the assumed regression relationship between $X$ and $\nu$ to impute locally the missing observations of $\nu$. This idea is extended to estimate the conditional expectation for a continuous covariate $X$. An outline of the algorithm is:

Step 1. (Resampling step) Fix an integer $M$, for $m = 1, \ldots, M$ perform a nonparametric resampling of the observed data. That is, for each observation $i = 1, \ldots, n$, if the cure status is observed ($\xi_i = 1$) generate $\nu_i^{*(m)}$ from the distribution $\mathcal{L}(X_i)$ with cumulative distribution function

$$\sum_{j=1}^{n} B_{g_1 j}^{\xi}(X_i) \mathbf{1}\left(\nu_j \leq u\right)$$

where $B_{g_1 j}^{\xi}(x)$ are the kernel weights with bandwidth $g_1$:

$$B_{g_1 j}^{\xi}(x) = \frac{\xi_j K_{g_1}(x - X_j)}{\sum_{i=1}^{n} \xi_i K_{g_1}(x - X_i)}.$$

Step 2. (Imputation step) Given the resampled data from Step 1., the missing values of $\nu$ are imputed using local resampling. More specifically, conditionally on the resampled data $\{(X_i, \nu_i^{*(m)}, \xi_i) : i = 1, \ldots, n\}$, a second distribution $\mathcal{L}^*(X_i)$ is constructed, with cumulative distribution function

$$\sum_{j=1}^{n} B_{g_2 j}^{\xi}(X_i) \mathbf{1}\left(\nu_j^{*(m)} \leq u\right)$$

where the kernel weights $B_{g_2 j}^{\xi}(x)$ are computed with a second bandwidth $g_2$. Then, if $\nu_i$ is missing, generate $\nu_i^{+,m}$ from $\mathcal{L}^*(X_i)$.

Step 3. (Computation of the final estimator) For $\widetilde{\nu}_i^m = \xi_i \nu_i + (1 - \xi_i)\nu_i^{+,m}$, let $1 - \widehat{p}_n^m = (1/n)\sum_{i=1}^{n} \widetilde{\nu}_i^m$ be the empirical estimator of the cure probability with the $m$th augmented dataset. The multiple imputation (MI) estimator for the cure probability $1 - p$ is

$$1 - \widehat{p}_n^{MI} = \frac{1}{M} \sum_{m=1}^{M} (1 - \widehat{p}_n^m). \tag{30}$$

Analogously, let us define $1 - \widehat{p}_h^m(x) = \sum_{i=1}^{n} B_{hi}(x) \widetilde{\nu}_i^m$ as the NW estimator in (8) computed with bandwidth $h$ and the $m$th augmented dataset. The final multiply imputed NW (MI-NW) estimator for the cure probability $1 - p(x)$ is

$$1 - \widehat{p}_h^{\text{MI-NW}}(x) = \frac{1}{M} \sum_{m=1}^{M} (1 - \widehat{p}_h^m(x)). \tag{31}$$

Note that Step 1 is needed to fully account for all uncertainty in predicting the missing values by adding extra variability into the multiply imputed values.[48] Under conditions similar to those in Cheng,[27] Aerts et al.[20] showed that the proposed estimator of $1-p$ in (30) is consistent, and provided asymptotic expressions for the bias and variance. Next, the asymptotic expressions of the bias and variance for the MI-NW estimator in (31), following the ideas in Aerts et al.[20] are derived. The proof is deferred to Appendix B of the Supplemental Material.

**Proposition 5.** Suppose that the siMAR condition and Assumptions 1 (i), 2 (i), 3 (i), 8 and 9 hold. Also, the bandwidths $h$, $g_1$, $g_2$ satisfy $h \to 0$, $g_1 \to 0$, $g_2 \to 0$, $nh \to \infty$, $ng_1 \to \infty$ and $ng_2 \to \infty$ as $n \to \infty$. The asymptotic bias of $1-\widehat{p}_h^{\text{MI-NW}}(x)$ is

$$\mu_{g_1,g_2,h}^{\text{MI-NW}}(x) = h^2 c_{1,c}(x) + \left(g_1^2 + g_2^2\right) c_{2,\text{MI-NW}}(x) + o\left((h^2 + g_1^2 + g_2^2)^2\right),$$

where $c_{1,c}(x)$ is defined in (20), and

$$c_{2,\text{MI-NW}}(x) = \frac{(1-\pi(x))\left[\pi(x)\left(1-p(x)\right)m(x)\right]''}{2m(x)\pi(x)} d_K. \tag{32}$$

If the bandwidths are $g_1/h \to C_1$ and $g_2/h \to C_2$, then the asymptotic variance is

$$\begin{aligned}
\sigma_{h,\text{MI-NW}}^2(x) = & \frac{1}{nh}\frac{1-p(x)}{m(x)}\left(\frac{c_K(1-\pi(x))p(x)}{M\pi(x)} + \left\{\pi(x)c_K + (1-\pi(x))\left[c_{K,C_1,C_2} + \frac{1-\pi(x)}{\pi(x)} d_{K,C_1,C_2}\right.\right.\right. \\
& \left.\left.\left. + (1-p(x))\left(c_K + 2c_{K,C_2} + \frac{1-\pi(x)}{\pi(x)}(c_{K,C_1,C_2} + 2d_{K,C_1,C_2})\right)\right]\right\}\right) \\
& + \frac{2}{ng_1}\left(1-p(x)\right)^2 \frac{1-\pi(x)}{\pi(x)} K(0) + o\left((Mnh)^{-1}\right) + o\left((nh)^{-1}\right) + o\left((ng_1)^{-1}\right),
\end{aligned} \tag{33}$$

where $c_{K,C} = \iint K(u)K(v)K(u+Cv)dudv$,

$$c_{K,C_1,C_2} = \iiint K(u)K(v)K(w)K(u + C_1 v + C_2 w)dudvdw$$

and

$$d_{K,C_1,C_2} = \iiint K(u)K(v)K(w)K(u + C_1 v + C_2(u+w))dudvdw.$$

The term $h^2 c_{1,c}(x)$ in the bias, which also appears in the bias $\mu_{h,c}(x)$ of the proposed estimator in (20) and in the bias $\mu_h(x)$ of the XP estimator, is the dominant term of the bias of the NW estimator, while $(g_1^2 + g_2^2)\ c_{2,\text{MI-NW}}(x)$ stems from the multiple imputation procedure in Steps 1 and 2 above. The comparison in terms of bias of the proposed estimator and the MI-NW estimator is a trade-off between the terms $c_{2,c}(x)$ in (21) and $c_{2,\text{MI-NW}}(x)$ in (32).

As for the variance, note that if $C_1 = C_2 = 0$ then $c_{K,C_2} = c_{K,C_1,C_2} = d_{K,C_1,C_2} = c_K$, whereas if $C_1 = \infty$ or $C_2 = \infty$, then $c_{K,C_2} = c_{K,C_1,C_2} = d_{K,C_1,C_2} = 0$. It should be noted that in terms of variance, the comparison between the proposed estimator and the MI-NW estimator is not straightforward. It is easy to prove that in the case of no missingness, the dominant term of the bias reduces to that of the NW estimator $c_{1,c}(x)$, whereas the leading term of the variance becomes $(1/nh)(\sigma^2(x) + \mu^2(x))/m(x)$, where $\sigma^2(x) = var(\nu \mid X = x) = p(x)(1-p(x))$ and $\mu(x) = E(\nu \mid X = x) = 1 - p(x)$.

## 4 Simulation study

A simulation study was conducted to assess the finite sample performance of the proposed estimator, $1-\widehat{p}_h^c(x)$, compared with:

(a) the competing risks estimators $1-\widehat{p}_{1,h}(x)$ in (27) and $1-\widehat{p}_{2,h}(x)$ in (28),
(b) the XP estimator $1-\widehat{p}_h(x)$ in (7),
(c) the MI-NW estimator $1-\widehat{p}_h^{\text{MI-NW}}(x)$ in (31) with $M = 5$ multiple imputations,
(d) and the semiparametric estimator[15] $1-p(x;\widehat{\boldsymbol{\gamma}})$, where $\widehat{\boldsymbol{\gamma}}$ are the parameter estimates.

Data were generated from the MCM in (1), where the latency part is modeled using a truncated exponential distribution:

$$S_0(t \mid x) = \begin{cases} \frac{\exp\{-\alpha(x)t\}-\exp\{-\alpha(x)4.605\}}{1-\exp\{-\alpha(x)4.605\}} & \text{if } 0 \le t \le 4.605 \\ 0 & \text{if } t > 4.605, \end{cases}$$

where $\alpha(x) = \exp\left(\frac{x+20}{40}\right)$. Six scenarios characterized by the cure probability function were considered. As Table 1 shows, the cure probability encompasses a wide range of functions, intended to cover any reasonable practical scenario.

The proportion of individuals identified as being cured was set to $\pi(x) = 0.2$ and 0.8. The censoring time $C$ was generated so that $C = \infty$ with probability $\pi(x)$, and with probability $1 - \pi(x)$, $C$ was generated from a Weibull distribution with shape parameter $\alpha = 2$, scale parameter $\beta = 2$, and density function

$$g(t; \beta, \alpha) = \beta\alpha^{-\beta}t^{\beta-1}\exp\left(-t/\alpha\right)^{\beta}, \text{ for } t > 0.$$

The covariate $X$ was uniformly distributed on the interval $[-20, 20]$. Note that $S_0(t \mid x)$ is truncated at $\tau_0 = 4.605$, so that the support for $C$ is larger than the support of $Y$ in order to fulfill condition (15). Depending on the scenario, the percentage of censored observations ranged from 22.6% (in Scenario 4 with $\pi(x) = 0.2$) to 80.8% (in Scenario 5 with $\pi(x) = 0.2$). For each scenario, 1000 datasets of sample sizes $n = 50$, 100 and 200 were generated.

Two different designs were considered. They differ with respect to the distribution of the observed times of the individuals known to be cured, represented by $H^{11}(t \mid x)$. In Design 1, the observed cure times were simulated to be falling within the largest censored times. This design was intended to reflect the pattern of the observed lifetimes of the patients known to be cured in the breast cancer data. No big differences are expected between the proposed cure rate estimator (equivalent to XP estimator computed with the observed cure times shifted to be arbitrarily large times) or ignoring the known cure status (XP estimator computed with the unmodified observed times). In Design 2, the distribution of the observed times of the known cured patients in COVID-19 data is mimicked. In this case the observed cure times were simply chosen at random among the censored times. Large differences are now expected the available cure status with the proposed estimator $1 - \widehat{p}_h^c(x)$ and XP estimator $1 - \widehat{p}_h(x)$. This section contains the results for $\pi(x) = 0.8$, $n = 100$ and for Design 2; additional results are in Appendix C of the Supplemental Material.

The first goal was to evaluate the small sample size performance of $1 - \widehat{p}_h^c(x)$ in terms of squared bias, variance and MSE when the optimal bandwidth is used. For nonparametric estimators, the search for the optimal bandwidth $h$ was performed in a grid of 21 values ranging from 1.5 to 100 and equispaced on a logarithmic scale. Besides, the pilot bandwidths $(g_1, g_2)$ required of the MI-NW estimator were searched in a grid of 11 bandwidths equispaced from 1.5 to 100 on a logarithmic scale. The Epanechnikov kernel was used.

The MSE of $1 - \widehat{p}_h^c(x)$, $1 - \widehat{p}_{1,h}(x)$, $1 - \widehat{p}_{2,h}(x)$, $1 - \widehat{p}_h(x)$ and $1 - \widehat{p}_h^{\text{MI-NW}}(x)$, all of them computed with the corresponding optimal bandwidths, and the MSE of $1 - p(x; \hat{\gamma})$ when $n = 100$, $\pi(x) = 0.8$ for Design 2 are illustrated in Figure 1. Note that while the performance of $1 - \widehat{p}_h^c(x)$, $1 - \widehat{p}_{1,h}(x)$, $1 - \widehat{p}_{2,h}(x)$, $1 - \widehat{p}_h^{\text{MI-NW}}(x)$ and $1 - p(x; \boldsymbol{\gamma})$ is affected by the design, that of $1 - \widehat{p}_h(x)$ is not affected as it ignores the information provided by the observations identified as cured. The estimator $1 - \widehat{p}_h^c(x)$ outdoes $1 - \widehat{p}_h(x)$ for most values of $X$. The differences in squared bias between $1 - \widehat{p}_h^c(x)$ and the competing risks estimators are quite apparent in Scenarios 1–3 and 6. As it can be seen, in all the scenarios $1 - \widehat{p}_h^c(x)$ outperforms $1 - \widehat{p}_{1,h}(x)$ and $1 - \widehat{p}_{2,h}(x)$. Table 2 collects the MSE, squared bias and variance of all the estimators. When the cured individuals are predominantly identified, the competing risks approaches $1 - \widehat{p}_{1,h}(x)$, $1 - \widehat{p}_{2,h}(x)$ and the proposed estimator $1 - \widehat{p}_h^c(x)$ provide similar results in terms of bias, variance as well as MSE when using the same bandwidth $h$ (see Table 2 for $\pi(x) = 0.8$). However, when a high percentage remains unidentified for the cured individuals, the proposed estimator is highly competitive, while $1 - \widehat{p}_{2,h}(x)$ seems to behave poorer, particularly when the cure probability is high

**Table 1.** Characteristics of the simulated scenarios.

| Scenario | | $1 - p(x)$ | % censoring $\pi(x) = 0.2$ | % censoring $\pi(x) = 0.8$ | % cured |
|---|---|---|---|---|---|
| 1 | Logistic function | $(1 + \exp(0.476 + 0.358x))^{-1}$ | 48.0 | 47.0 | 46.0 |
| 2 | Cubic function | $0.5 - x^3/16000$ | 52.0 | 51.0 | 50.0 |
| 3 | Linear function | $0.5 - 0.025x$ | 51.0 | 50.4 | 50.0 |
| 4 | Low constant | 0.2 | 22.6 | 20.5 | 20.0 |
| 5 | High constant | 0.8 | 80.8 | 80.3 | 80.0 |
| 6 | Convex function | $0.0025x^2$ | 36.0 | 34.0 | 33.3 |

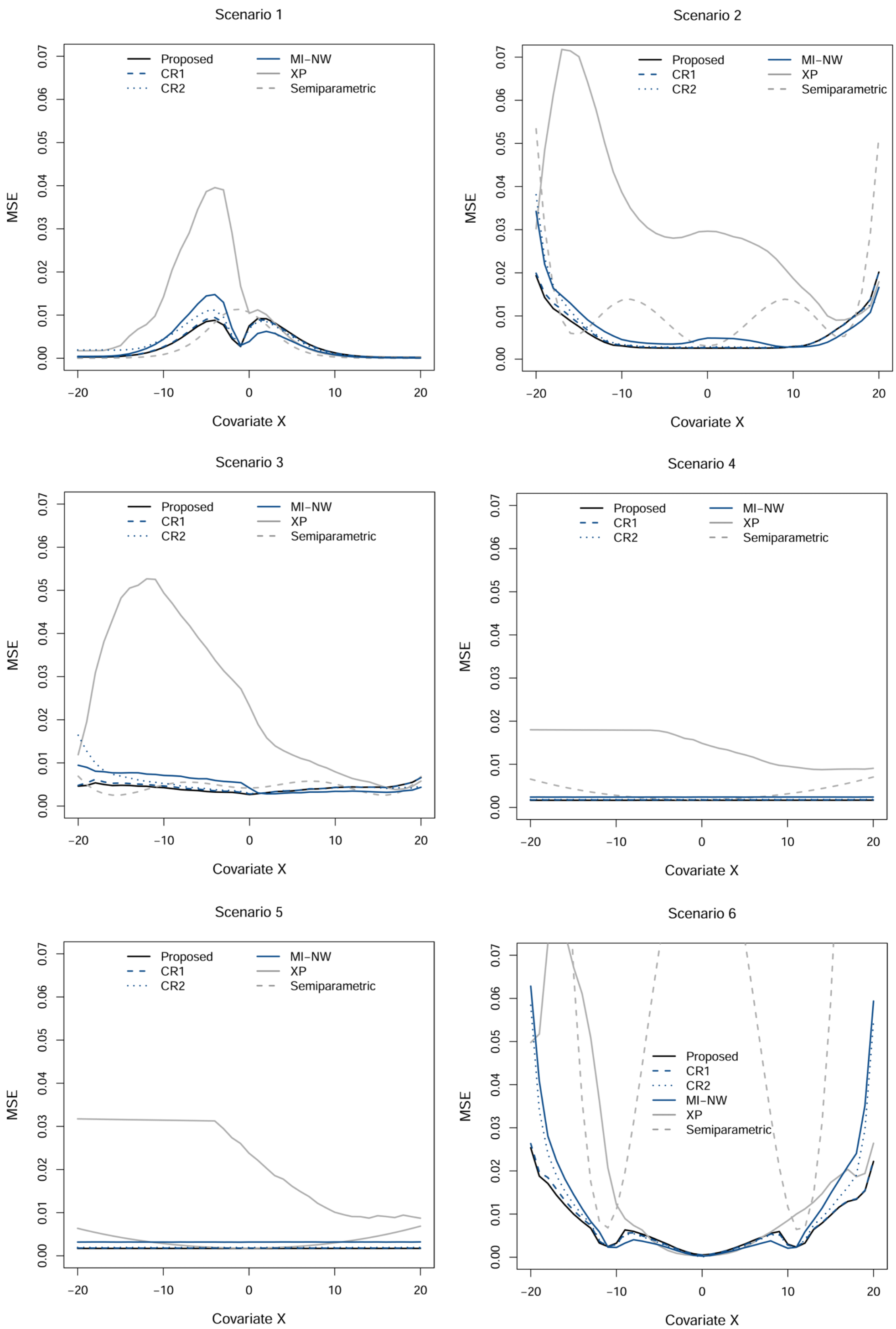


**Figure 1.** MSE of the proposed estimator $1-\widehat{p}_h^c(x)$, the competing risk estimators $1-\widehat{p}_{1,h}(x)$ (CR1) and $1-\widehat{p}_{2,h}(x)$ (CR2), the MI-NW estimator $1-\widehat{p}_h^{\text{MI-NW}}(x)$, the XP estimator $1-\widehat{p}_h(x)$ (all computed with the optimal bandwidth), and the semiparametric estimator $1-p(x;\hat{\gamma})$ in the simulated scenarios and under Design 2 for $\pi(x)=0.8$ and $n=100$.

(see Table S2 with $\pi(x)=0.2$ in Appendix C of the Supplemental Material). The estimators $1-\widehat{p}_h^c(x)$ and $1-\widehat{p}_h^{\text{MI-NW}}(x)$ show similar performances.

The simulation results for Design 2 are presented in Table 2 (as well as in Figure S1 in Appendix C of the Supplemental Material). In general the proposed estimator $1-\widehat{p}_h^c(x)$ has smaller MSE than XP estimator $1-\widehat{p}_h(x)$ for most values of $X$. This shows the loss of efficiency incurred in if the known cures are not incorporated in the estimation methodology. As

**Table 2.** Squared bias (Bias$^2$), variance (Var) and MSE of $1-\widehat{p}_h^c(x)$, $1-\widehat{p}_{1,h}(x)$, $1-\widehat{p}_{2,h}(x)$, $1-\widehat{p}_h(x)$, $1-\widehat{p}_h^{\text{MI-NW}}(x)$, all computed with the optimal MSE bandwidth, and $1-p(x;\hat{\gamma})$ in the simulated scenarios and under Designs 1 and 2, for $\pi(x)=0.8$ and $n=100$.

| | | $1-\widehat{p}_h^c(x)$ | | | | $1-\widehat{p}_{1,h}(x)$ | | | | $1-\widehat{p}_{2,h}(x)$ | | | | $1-\widehat{p}_h(x)$ | | | | $1-\widehat{p}_h^{MI-NW}(x)$ | | | | $1-p(x;\hat{\gamma})$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | $x$ | $h_x$ | Bias$^2$ $\times10^3$ | Var $\times10^3$ | MSE $\times10^3$ | $h_x$ | Bias$^2$ $\times10^3$ | Var $\times10^3$ | MSE $\times10^3$ | $h_x$ | Bias$^2$ $\times10^3$ | Var $\times10^3$ | MSE $\times10^3$ | $h_x$ | Bias$^2$ $\times10^3$ | Var $\times10^3$ | MSE $\times10^3$ | $h_x$ | Bias$^2$ $\times10^3$ | Var $\times10^3$ | MSE $\times10^3$ | Bias$^2$ $\times10^3$ | Var $\times10^3$ | MSE $\times10^3$ |
| **Design 1** | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | −10 | 5.288 | 0.302 | 2.579 | 2.882 | 5.288 | 0.302 | 2.579 | 2.882 | 5.288 | 0.302 | 2.579 | 2.882 | 2.283 | 0.317 | 13.885 | 14.202 | 2.283 | 1.101 | 2.873 | 3.974 | 0.001 | 0.998 | 0.999 |
| | 0 | 15.109 | 1.780 | 4.205 | 5.985 | 15.109 | 1.771 | 4.207 | 5.978 | 15.109 | 1.771 | 4.207 | 5.978 | 65.707 | 0.028 | 10.381 | 10.409 | 9.928 | 1.090 | 2.764 | 3.854 | 0.493 | 11.191 | 11.684 |
| | 10 | 6.523 | 0.098 | 0.966 | 1.064 | 6.523 | 0.095 | 0.957 | 1.052 | 6.523 | 0.081 | 0.938 | 1.019 | 8.047 | 0.003 | 0.678 | 0.681 | 2.816 | 0.157 | 0.964 | 1.121 | 0.000 | 0.253 | 0.253 |
| 2 | −10 | 28.368 | 0.171 | 3.193 | 3.364 | 28.368 | 0.175 | 3.194 | 3.370 | 28.368 | 0.175 | 3.194 | 3.370 | 28.368 | 17.188 | 21.477 | 38.665 | 9.928 | 0.480 | 3.473 | 3.953 | 6.471 | 4.930 | 11.401 |
| | 0 | 81.060 | 0.367 | 2.612 | 2.978 | 81.060 | 0.372 | 2.613 | 2.985 | 81.060 | 0.372 | 2.613 | 2.985 | 100 | 14.321 | 15.326 | 29.647 | 43.174 | 1.623 | 2.437 | 4.060 | 0.343 | 3.108 | 3.451 |
| | 10 | 34.996 | 0.009 | 2.686 | 2.695 | 34.996 | 0.008 | 2.687 | 2.695 | 34.996 | 0.008 | 2.687 | 2.695 | 100 | 3.640 | 15.059 | 18.699 | 22.995 | 0.005 | 2.394 | 2.399 | 12.906 | 4.213 | 17.119 |
| 3 | −10 | 12.247 | 0.539 | 4.571 | 5.111 | 12.247 | 0.543 | 4.576 | 5.119 | 12.247 | 0.543 | 4.576 | 5.119 | 12.247 | 15.379 | 33.983 | 49.362 | 5.288 | 1.581 | 4.843 | 6.424 | 0.559 | 4.055 | 4.613 |
| | 0 | 81.060 | 0.419 | 2.727 | 3.146 | 81.060 | 0.423 | 2.730 | 3.152 | 81.060 | 0.423 | 2.730 | 3.152 | 100 | 9.828 | 13.330 | 23.158 | 22.995 | 1.064 | 2.596 | 3.661 | 0.377 | 4.314 | 4.691 |
| | 10 | 15.109 | 0.094 | 3.692 | 3.785 | 15.109 | 0.091 | 3.695 | 3.786 | 15.109 | 0.091 | 3.695 | 3.786 | 18.640 | 0.369 | 7.561 | 7.930 | 6.523 | 0.003 | 3.315 | 3.318 | 2.419 | 3.598 | 6.017 |
| 4 | −10 | 81.060 | 0.229 | 1.483 | 1.712 | 81.060 | 0.249 | 1.483 | 1.732 | 81.060 | 0.250 | 1.484 | 1.734 | 100 | 14.109 | 3.821 | 17.930 | 34.996 | 0.262 | 1.311 | 1.573 | 0.202 | 2.583 | 2.785 |
| | 0 | 100 | 0.231 | 1.486 | 1.717 | 100 | 0.250 | 1.485 | 1.736 | 100 | 0.252 | 1.487 | 1.738 | 9.928 | 8.612 | 6.260 | 14.872 | 43.174 | 0.270 | 1.309 | 1.579 | 0.343 | 1.606 | 1.950 |
| | 10 | 100 | 0.232 | 1.491 | 1.723 | 100 | 0.251 | 1.490 | 1.742 | 100 | 0.252 | 1.492 | 1.744 | 12.247 | 4.237 | 5.315 | 9.552 | 100 | 0.278 | 1.307 | 1.585 | 0.298 | 2.936 | 3.234 |
| 5 | −10 | 100 | 0.086 | 1.862 | 1.948 | 100 | 0.087 | 1.863 | 1.949 | 100 | 0.087 | 1.863 | 1.949 | 100 | 7.357 | 24.080 | 31.437 | 22.995 | 0.878 | 2.098 | 2.976 | 0.070 | 3.181 | 3.251 |
| | 0 | 53.262 | 0.085 | 1.861 | 1.946 | 53.262 | 0.085 | 1.862 | 1.947 | 53.262 | 0.085 | 1.862 | 1.947 | 12.247 | 3.425 | 20.353 | 23.778 | 18.640 | 0.897 | 2.089 | 2.986 | 0.031 | 1.957 | 1.988 |
| | 10 | 100 | 0.085 | 1.862 | 1.947 | 100 | 0.086 | 1.862 | 1.948 | 100 | 0.086 | 1.862 | 1.948 | 9.928 | 0.836 | 9.263 | 10.099 | 65.707 | 0.917 | 2.085 | 3.001 | 0.062 | 3.224 | 3.285 |
| 6 | −10 | 28.368 | 0.053 | 2.559 | 2.613 | 28.368 | 0.045 | 2.562 | 2.607 | 28.368 | 0.045 | 2.563 | 2.608 | 100 | 3.003 | 9.191 | 12.194 | 15.109 | 0.275 | 2.004 | 2.279 | 3.501 | 5.615 | 9.117 |
| | 0 | 3.474 | 0.052 | 0.456 | 0.508 | 3.474 | 0.051 | 0.452 | 0.503 | 1.850 | 0.001 | 0.073 | 0.074 | 4.286 | 0.034 | 0.340 | 0.374 | 1.5 | 0.216 | 0.917 | 1.133 | 96.237 | 2.672 | 98.909 |
| | 10 | 28.368 | 0.042 | 2.317 | 2.359 | 28.368 | 0.038 | 2.317 | 2.355 | 28.368 | 0.037 | 2.317 | 2.354 | 12.247 | 0.313 | 8.190 | 8.503 | 15.109 | 0.217 | 1.881 | 2.098 | 4.559 | 5.019 | 9.578 |

(continued)

**Table 2.** Continued

| | | $1-\widehat{p}_h^c(x)$ | | | | $1-\widehat{p}_{1,h}(x)$ | | | | $1-\widehat{p}_{2,h}(x)$ | | | | $1-\widehat{p}_h(x)$ | | | | $1-\widehat{p}_h^{MI\text{-}NW}(x)$ | | | | $1-p(x;\hat{\gamma})$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | $x$ | $h_x$ | Bias$^2$ $\times 10^3$ | Var $\times 10^3$ | MSE $\times 10^3$ | $h_x$ | Bias$^2$ $\times 10^3$ | Var $\times 10^3$ | MSE $\times 10^3$ | $h_x$ | Bias$^2$ $\times 10^3$ | Var $\times 10^3$ | MSE $\times 10^3$ | $h_x$ | Bias$^2$ $\times 10^3$ | Var $\times 10^3$ | MSE $\times 10^3$ | $h_x$ | Bias$^2$ $\times 10^3$ | Var $\times 10^3$ | MSE $\times 10^3$ | Bias$^2$ $\times 10^3$ | Var $\times 10^3$ | MSE $\times 10^3$ |
| **Design 2** | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | −10 | 6.523 | 0.450 | 2.050 | 2.500 | 6.523 | 0.508 | 2.140 | 2.647 | 6.523 | 1.311 | 3.115 | 4.426 | 2.283 | 0.317 | 13.885 | 14.202 | 2.283 | 1.262 | 3.142 | 4.404 | 0.002 | 0.868 | 0.869 |
| | 0 | 12.247 | 2.569 | 4.920 | 7.489 | 12.247 | 2.156 | 5.046 | 7.202 | 12.247 | 1.631 | 5.238 | 6.870 | 65.707 | 0.028 | 10.381 | 10.409 | 9.928 | 0.763 | 3.090 | 3.853 | 0.027 | 10.650 | 10.677 |
| | 10 | 6.523 | 0.152 | 1.100 | 1.252 | 6.523 | 0.141 | 1.070 | 1.211 | 6.523 | 0.075 | 0.991 | 1.066 | 8.047 | 0.003 | 0.678 | 0.681 | 2.816 | 0.040 | 0.781 | 0.821 | 0.002 | 0.287 | 0.289 |
| | −10 | 28.368 | 0.020 | 3.054 | 3.074 | 28.368 | 0.001 | 3.198 | 3.199 | 28.368 | 0.028 | 3.304 | 3.331 | 28.368 | 17.188 | 21.477 | 38.665 | 9.928 | 0.336 | 4.188 | 4.524 | 9.033 | 4.628 | 13.661 |
| 2 | 0 | 65.707 | 0.002 | 2.539 | 2.541 | 65.707 | 0.039 | 2.629 | 2.668 | 65.707 | 0.098 | 2.727 | 2.825 | 100 | 14.321 | 15.326 | 29.647 | 43.174 | 2.051 | 2.816 | 4.867 | 0.002 | 3.026 | 3.028 |
| | 10 | 28.368 | 0.050 | 2.867 | 2.917 | 34.996 | 0.272 | 2.694 | 2.967 | 34.996 | 0.172 | 2.781 | 2.952 | 100 | 3.640 | 15.059 | 18.699 | 22.995 | 0.000 | 2.757 | 2.757 | 9.541 | 4.091 | 13.632 |
| 3 | −10 | 12.247 | 0.084 | 4.174 | 4.259 | 12.247 | 0.195 | 4.448 | 4.642 | 12.247 | 0.428 | 4.785 | 5.212 | 12.247 | 15.379 | 33.983 | 49.362 | 5.288 | 1.950 | 5.148 | 7.099 | 1.211 | 3.694 | 4.905 |
| | 0 | 65.707 | 0.008 | 2.653 | 2.660 | 81.06 | 0.044 | 2.752 | 2.796 | 65.707 | 0.112 | 2.827 | 2.939 | 100 | 9.828 | 13.330 | 23.158 | 43.174 | 1.372 | 2.790 | 4.162 | 0.003 | 4.172 | 4.175 |
| | 10 | 12.247 | 0.067 | 4.239 | 4.306 | 12.247 | 0.030 | 4.314 | 4.344 | 15.109 | 0.322 | 3.898 | 4.220 | 18.64 | 0.369 | 7.561 | 7.930 | 6.523 | 0.014 | 3.396 | 3.410 | 1.373 | 3.732 | 5.104 |
| 4 | −10 | 81.060 | 0.004 | 1.661 | 1.665 | 81.06 | 0.076 | 1.670 | 1.746 | 100 | 0.137 | 1.707 | 1.843 | 100 | 14.109 | 3.821 | 17.930 | 34.996 | 1.013 | 1.381 | 2.394 | 0.008 | 2.785 | 2.793 |
| | 0 | 100 | 0.004 | 1.665 | 1.669 | 100 | 0.076 | 1.674 | 1.749 | 100 | 0.136 | 1.709 | 1.846 | 9.928 | 8.612 | 6.260 | 14.872 | 100 | 0.995 | 1.413 | 2.408 | 0.043 | 1.762 | 1.805 |
| | 10 | 100 | 0.005 | 1.670 | 1.675 | 100 | 0.075 | 1.679 | 1.754 | 100 | 0.136 | 1.714 | 1.849 | 12.247 | 4.237 | 5.315 | 9.552 | 34.996 | 0.999 | 1.404 | 2.404 | 0.026 | 3.131 | 3.157 |
| 5 | −10 | 100 | 0.001 | 1.705 | 1.706 | 100 | 0.004 | 1.786 | 1.790 | 100 | 0.047 | 1.891 | 1.938 | 100 | 7.357 | 24.080 | 31.437 | 34.996 | 0.977 | 2.204 | 3.180 | 0.003 | 2.838 | 2.841 |
| | 0 | 65.707 | 0.001 | 1.704 | 1.705 | 53.262 | 0.003 | 1.781 | 1.785 | 43.174 | 0.045 | 1.886 | 1.931 | 12.247 | 3.425 | 20.353 | 23.778 | 34.996 | 0.966 | 2.227 | 3.193 | 0.011 | 1.763 | 1.774 |
| | 10 | 100 | 0.001 | 1.706 | 1.707 | 81.06 | 0.004 | 1.783 | 1.786 | 81.06 | 0.047 | 1.887 | 1.934 | 9.928 | 0.836 | 9.263 | 10.099 | 34.996 | 1.006 | 2.202 | 3.208 | 0.000 | 2.994 | 2.994 |
| 6 | −10 | 28.368 | 0.564 | 2.688 | 3.253 | 28.368 | 0.288 | 2.726 | 3.014 | 28.368 | 0.193 | 2.744 | 2.936 | 100 | 3.003 | 9.191 | 12.194 | 15.109 | 0.024 | 2.240 | 2.264 | 5.284 | 5.818 | 11.102 |
| | 0 | 3.474 | 0.058 | 0.475 | 0.532 | 3.474 | 0.056 | 0.468 | 0.525 | 1.5 | 0.000 | 0.062 | 0.062 | 4.286 | 0.034 | 0.340 | 0.374 | 1.5 | 0.027 | 0.227 | 0.255 | 105.037 | 2.844 | 107.881 |
| | 10 | 28.368 | 0.524 | 2.479 | 3.003 | 28.368 | 0.308 | 2.485 | 2.793 | 28.368 | 0.200 | 2.553 | 2.753 | 12.247 | 0.313 | 8.190 | 8.503 | 15.109 | 0.023 | 2.056 | 2.079 | 6.593 | 5.043 | 11.636 |

expected, in Scenario 1, the semiparametric estimator, which considers a logistic regression model to fit the probability of cure with an EM algorithm for estimating the regression parameter $\boldsymbol{\gamma}$, behaves well. However, the estimator $1-\widehat{p}_h^c(x)$ is competitive for a wide range of values close to $x=-20$ and $x=20$, and even beats $1-p(x;\hat{\gamma})$ for covariate values around $x=0$. The estimator $1-\widehat{p}_h^c(x)$ outperforms $1-p(x;\hat{\gamma})$ in Scenarios 2–6, where the underlying logistic model assumption for the cure probability in $1-p(x;\hat{\gamma})$ is not met. Finally, it must be noted that $1-\widehat{p}_h^c(x)$ is quite competitive with respect to $1-\widehat{p}_{1,h}(x)$, $1-\widehat{p}_{2,h}(x)$ and $1-\widehat{p}_h^{\text{MI-NW}}(x)$, behaving better in general.

For $n=200$ in Design 2 (see Figure S2 and Table S1 in Appendix C of the Supplemental Material), the differences in squared bias are much smaller for $1-\widehat{p}_h^c(x)$, $1-\widehat{p}_{1,h}(x)$, $1-\widehat{p}_{2,h}(x)$, $1-\widehat{p}_h^{\text{MI-NW}}(x)$ and $1-p(x;\hat{\gamma})$. Regarding the variance, $1-\widehat{p}_h^c(x)$ performs better in most scenarios.

Figure S2 (right) and Table S2 in Appendix C of the Supplemental Material show that our method is efficient when there is a small proportion of individuals identified as cured, when $n=100$ and $\pi(x)=0.2$. Interestingly, when $\pi(x)=0.2$, $1-\widehat{p}_h^c(x)$ is still efficient and even beats $1-\widehat{p}_h(x)$ for most values of $X$. Besides, $1-\widehat{p}_h^{\text{MI-NW}}(x)$ and $1-\widehat{p}_{2,h}(x)$ perform poorly due to a significant increase in both squared bias and variance leading to poor MSE results. This suggests that the use of $1-\widehat{p}_h^c(x)$ is advantageous even when only a few individuals are identified as cured.

Simulations were conducted to evaluate the performance of the bandwidth selector discussed in Section 2.5, using $B=1000$ resamples and a grid of bandwidths from 1.5 to 100. Figure S4 (in Appendix C of the Supplemental Material) shows the quartiles of the selected bootstrap bandwidth $h_x^*$ for Scenarios 1–6 under Design 2. The optimal bandwidth was also compared to $h_x^*$. The performance of $h_x^*$ varies depending on the scenario, but in general it performs well in all scenarios. Bandwidth choice seems to be more important in Scenarios 1–3 and 6, as different bandwidths result in sensibly different MSE. In Scenarios 4 and 5, different bandwidths yield approximately the same MSE. In this case, the bootstrap bandwidth being relatively far from the optimal bandwidth does not entail a significant loss of efficiency, see Figure S4 in Appendix C of the Supplemental Material.

## 5 Real data analysis

We applied all the estimators of the cure rate to the breast cancer and COVID-19 datasets described in Section 1.

### 5.1 Breast cancer data

When analyzing the survival of breast cancer patients, it is of great interest to study the effect of well-established clinicopathologic prognostic factors.[49,50] The aim of the analysis was to estimate the probability of not dying from breast cancer depending on cancer stage, number of positive lymph nodes, menopausal status, margin status, and age at diagnosis. In this dataset, only 42 (4.7%) patients died from cancer within the follow-up period. The observed times for the remaining patients were right-censored. In this censored group, 20 patients (2.2%) were cancer free for more than 10 years, suggesting they might be long-term survivors, in our model, known to be cured from the event "death from cancer." This results in a very high missingness rate, 93.1%, for the cure indicator, $\nu$. Maller and Zhou[3] test ($p$-value $<0.001$) provides evidence supporting condition (15), clearly endorsing the use of $1-\widehat{p}_h^c(x)$ and $1-\widehat{p}_n^c$ to estimate the probability of not dying from cancer.

The probability of cure as a function of the categorical covariates (cancer stage, number of positive lymph nodes, menopausal status, margin status) was estimated for the different groups of patients using:

(a) the proposed estimator $1-\widehat{p}_n^c$ in (9) with its $\widehat{se}_B(1-\widehat{p}_n^c)$;
(b) the empirical estimator $1-\widehat{p}=\sum_{i=1}^n \xi_i\nu_i(\sum_{i=1}^n \xi_i)^{-1}$ with its $\widehat{se}(1-\widehat{p})$;
(c) the unconditional competing risks estimators $1-\widehat{p}_{1,n}$ with its $\widehat{se}_B(1-\widehat{p}_{1,n})$, $1-\widehat{p}_{2,n}$ with its $\widehat{se}_B(1-\widehat{p}_{2,n})$;
(d) the MI estimator $1-\widehat{p}_n^{\text{MI}}$ in (30) with its $\widehat{se}_B(1-\widehat{p}_n^{\text{MI}})$ and $M=20$;
(e) the unconditional XP estimator $1-\widehat{p}_n$ in (10) with its $\widehat{se}(1-\widehat{p}_n)$.

The bootstrap procedure in Section 2.5 was used to estimate the standard errors $\widehat{se}_B(1-\widehat{p}_n^c)$, $\widehat{se}_B(1-\widehat{p}_{1,n})$, $\widehat{se}_B(1-\widehat{p}_{2,n})$, and $\widehat{se}_B(1-\widehat{p}_n^{\text{MI}})$, with $B=1000$ bootstrap resamples. The standard error $\widehat{se}(1-\widehat{p}_n)$ was computed with Greenwood's formula using the `R` package `survival`. Moreover, the standard error of $1-\widehat{p}$ was calculated using the standard error formula of the sample proportion $\widehat{se}(1-\widehat{p})=\sqrt{\widehat{p}(1-\widehat{p})(\sum_{i=1}^n \xi_i)^{-1}}$.

To the best of our knowledge, there is no any specifically tailored bandwidth selector for the pilot bandwidths $(g_1, g_2)$ required in the computation of $1-\widehat{p}_n^{\text{MI}}$. Thus, in this analysis $(g_1, g_2)$ were selected using the cross-validation selector of Bowman et al.,[51] available in the `R` package `kerdiest`.[52] The results are given in Table 3. The empirical estimator $1-\widehat{p}$

seems to underestimate the true $1-p$. Note that all the patients with unknown cure status are excluded, so the estimate is computed with a considerably reduced sample size. If the excluded patients are not MCAR, the reduced sample might not be representative. The MI estimator uses the complete sample, but it still appears to be performing poorly because 93.1% of patients have missing cure status. The unconditional XP estimator, $1-\widehat{p}_n$, does consider the censored observations, however, it dismisses the cure status information so it still underestimates the true cure probabilities. The estimators $1-\widehat{p}_n^c$, $1-\widehat{p}_{1,n}$ and $1-\widehat{p}_{2,n}$ make use of the available information of the cure status giving reasonably accurate estimates.

The estimated probability of not dying from breast cancer as a function of the continuous covariate age, is given in Figure 2. The estimator $1-\widehat{p}_h^c(x)$ is compared with the competing risks estimators $1-\widehat{p}_{1,h}(x)$, $1-\widehat{p}_{2,h}(x)$, and the XP estimator $1-\widehat{p}_h(x)$, all computed with the bootstrap bandwidth selector discussed in Section 2.5 using $B=1000$ resamples. It is also compared with the semiparametric estimator $1-p(x;\hat{\gamma})$ and the MI-NW estimator $1-\widehat{p}_h^{\text{MI-NW}}(x)$ computed with $M=20$. The bandwidth $h$ for the MI-NW estimator was chosen via an improved cross-validation bandwidth selector for the NW estimator,[53] using the R package np.[54] Figure 2 also shows the 95% confidence interval (CI) of the estimator $1-\widehat{p}_h^c(x)$. The CI was computed based on the asymptotic normality of the estimator $1-\widehat{p}_h^c(x)$ in which the bootstrap

**Table 3.** Demographic characteristics of breast cancer patients, and the numbers of dead patients (death), patients known to be cured (cured) and patients with unknown cure status (unknown). Also given are the estimated probability of being cured from breast cancer ($1-p$) estimated using $1-\widehat{p}_n^c$, $1-\widehat{p}$, $1-\widehat{p}_{1,n}$, $1-\widehat{p}_{2,n}$, $1-\widehat{p}_n^{MI}$, and $1-\widehat{p}_n$ with their respective estimated standard errors ($\widehat{se}$).

| | | Uncured | Censored | | Estimated probability of not dying from cancer | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Characteristic | Count (%) | Dead | Cured[b] | Unknown | $1-\widehat{p}_n^c$ ($\widehat{se}_B$) | $1-\widehat{p}$ ($\widehat{se}$) | $1-\widehat{p}_{1,n}$ ($\widehat{se}_B$) | $1-\widehat{p}_{2,n}$ ($\widehat{se}_B$) | $1-\widehat{p}_n^{MI}$ ($\widehat{se}_B$) | $1-\widehat{p}_n$ ($\widehat{se}$) |
| **Age** | | | | | | | | | | |
| <55 years | 371 (41.3) | 21 | 11 | 339 | 0.579 (0.084) | 0.344 (0.084) | 0.579 (0.084) | 0.579 (0.099) | 0.363 (0.093) | 0.561 (0.097) |
| ≥55 years | 527 (58.7) | 21 | 9 | 497 | 0.766 (0.089) | 0.300 (0.084) | 0.766 (0.089) | 0.766 (0.089) | 0.310 (0.095) | 0.766 (0.081) |
| **Stages**[a] | | | | | | | | | | |
| I | 164 (19.8) | 5 | 4 | 155 | 0.812 (0.099) | 0.444 (0.166) | 0.812 (0.099) | 0.812 (0.143) | 0.372 (0.191) | 0.812 (0.092) |
| II | 514 (62.0) | 20 | 10 | 484 | 0.630 (0.095) | 0.333 (0.086) | 0.631 (0.095) | 0.631 (0.103) | 0.352 (0.098) | 0.590 (0.118) |
| III | 151 (18.2) | 10 | 4 | 137 | 0.593 (0.148) | 0.286 (0.121) | 0.593 (0.148) | 0.474 (0.170) | 0.308 (0.136) | 0.593 (0.139) |
| **Menopausal** [a] **status**[a] | | | | | | | | | | |
| Pre | 185 (22.3) | 10 | 0 | 175 | 0.248 (0.244) | 0.000 (0.000) | 0.248 (0.244) | 0.000 (0.000) | 0.000 (0.000) | 0.248 (0.205) |
| Peri | 63 (7.6) | 7 | 11 | 45 | 0.725 (0.094) | 0.611 (0.115) | 0.725 (0.094) | 0.725 (0.105) | 0.596 (0.121) | 0.707 (0.105) |
| Post | 581 (70.1) | 17 | 9 | 555 | 0.803 (0.091) | 0.346 (0.093) | 0.803 (0.091) | 0.723 (0.109) | 0.397 (0.107) | 0.803 (0.083) |
| **No. of positive lymph nodes**[a] | | | | | | | | | | |
| 0 | 392 (50.5) | 12 | 11 | 369 | 0.754 (0.085) | 0.478 (0.104) | 0.754 (0.085) | 0.754 (0.085) | 0.484 (0.115) | 0.754 (0.087) |
| 1–3 | 299 (38.5) | 16 | 6 | 277 | 0.639 (0.118) | 0.273 (0.095) | 0.639 (0.118) | 0.548 (0.139) | 0.337 (0.111) | 0.639 (0.114) |
| > 3 | 86 (11.1) | 11 | 3 | 72 | 0.468 (0.152) | 0.214 (0.110) | 0.468 (0.152) | 0.351 (0.150) | 0.270 (0.124) | 0.439 (0.157) |
| **Margin status**[a] | | | | | | | | | | |
| Negative | 761 (92.5) | 21 | 14 | 726 | 0.744 (0.075) | 0.400 (0.083) | 0.744 (0.075) | 0.744 (0.090) | 0.406 (0.092) | 0.730 (0.083) |
| Positive | 62 (7.5) | 4 | 3 | 55 | 0.771 (0.131) | 0.429 (0.187) | 0.771 (0.131) | 0.771 (0.184) | 0.431 (0.216) | 0.771 (0.128) |

[a]Missing observations present.

[b]"Cured" from the event "death from breast cancer."

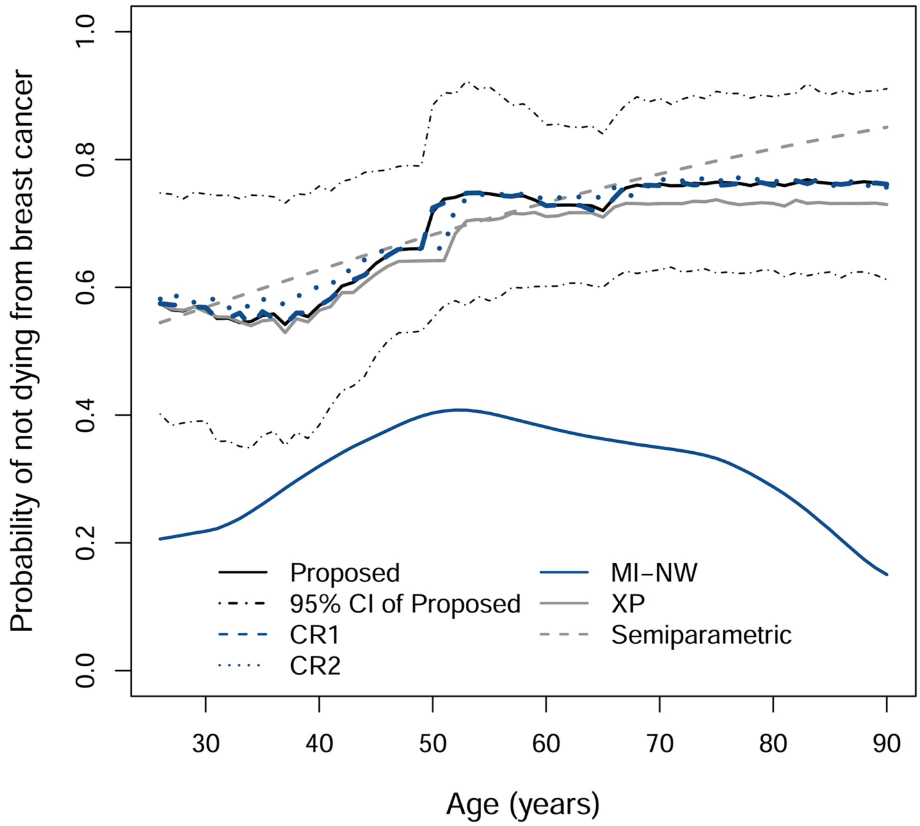


**Figure 2.** Estimation of the probability of not dying from breast cancer patients by using the proposed estimator $1 - \widehat{p}_h^c(x)$ and its 95% CI, the competing risks estimator $1 - \widehat{p}_{1,h}(x)$ (CR1) and $1 - \widehat{p}_{2,h}(x)$ (CR2), the XP estimator $1 - \widehat{p}_h(x)$ (all computed with the bootstrap bandwidth), the MI-NW estimator $1 - \widehat{p}_h^{\text{MI-NW}}(x)$ (computed using the cross-validation bandwidth), and the semiparametric estimator $1 - p(x; \hat{\gamma})$.

procedure was used to estimate $\widehat{se}_B(1 - \widehat{p}_h^c(x))$:

$$1 - \widehat{p}_h^c(x) \mp z_{1-\frac{\alpha}{2}} \widehat{se}_B\left(1 - \widehat{p}_h^c(x)\right),$$

where $z_\beta$ is the $\beta$th quantile of the standard normal.

Although $1 - p(x; \hat{\gamma})$ shows that the probability of not dying from breast cancer increases with age, the curves from the other estimators suggest that the logistic model assumed in the semiparametric estimator might not be appropriate. Specifically, they indicate an increment of that probability only for younger to middle age patients. The estimators $1 - \widehat{p}_h^c(x)$, $1 - \widehat{p}_{1,h}(x)$, $1 - \widehat{p}_{2,h}(x)$ and $1 - \widehat{p}_h(x)$ suggest no effect of the age on the probability for elderly patients, while $1 - \widehat{p}_h^{\text{MI-NW}}(x)$ implies that the probability decreases with age in older patients. Observe that the probability of not dying from breast cancer given by $1 - \widehat{p}_h(x)$, an estimator that disregards the available information about cure status, is equal or lower than the probability estimated with $1 - \widehat{p}_h^c(x)$. Nonetheless, the differences between $1 - \widehat{p}_h^c(x)$ and $1 - \widehat{p}_h(x)$ are subtle, as the proportion of the identified known cures is small. When the last observation is an event or known to be cured, the estimator $1 - \widehat{p}_h^c(x)$ produces the same estimate as the competing risks estimators $1 - \widehat{p}_{1,h}(x)$ and $1 - \widehat{p}_{2,h}(x)$.

Finally, the MI-NW estimator shows a similar trend as $1 - \widehat{p}_h^c(x)$ and $1 - \widehat{p}_h(x)$, although the estimated probabilities are substantially smaller. As pointed out before, the performance of $1 - \widehat{p}_h^{\text{MI-NW}}(x)$ worsens significantly because of the extremely high proportion of patients with missing cure status.

## 5.2 COVID-19 data

Since the COVID-19 pandemic started countries around the world are experiencing a large number of cases, with many patients requiring hospitalization wards. Although most infected people presented with mild disease, there were many severe cases that required long stays in ICU, overwhelming the healthcare systems with critical consequences on disease mortality. An accurate knowledge of the duration of hospitalization, and the prediction of the probability that a hospitalized inpatient would require a bed in ICU, are key for understanding the hospital demand for beds and crucial for decision-making and suitable planning.

As mentioned in Section 1, our second dataset contains the 10454 confirmed COVID-19 cases reported by the Galician Healthcare Service[32] between 6 March and 7 May 2020. The time of interest is the length of stay in hospital ward until admission to ICU, and the aim of this analysis was to estimate the probability of admission to ICU from hospital ward given age and sex as covariates of interest, see Table 4. For 2380 hospitalized patients for at least one day, 1063 (44.7%) were 75 years of age or above and 1262 (53%) were males. A total of 1638 (68.8%) patients were discharged alive before entering ICU, and 328 (13.8%) had died before entering ICU. None of them will require admission to ICU eventually, so all of them can be considered as "cured" from the event of interest, that is, admission to ICU. Note that "cure" means being free of experiencing admission to ICU, not cured in medical terms.

A total of 197 of the 2380 inpatients in hospital ward required admission to ICU, which gives an empirical estimated probability of admission to ICU of $\widehat{p} = 197/2380 = 0.083$. However, the true number of patients requiring ICU is expected to be larger than 197, as some of the 217 (9.1%) inpatients still in hospital bed at the end of the study might eventually need admission to the ICU. This shows that $\widehat{p} = 0.083$ might underestimate the probability of admission to ICU, motivating the use of alternative estimators than can handle censoring. It is assumed that condition (15) applies, as the result of the test of Maller and Zhou[3] suggests ($p$-value <0.001).

Table 4 shows the estimated probabilities of requiring ICU, given by the proposed estimator $\widehat{p}_n^c$, the unconditional competing risks estimators $\widehat{p}_{1,n}$, $\widehat{p}_{2,n}$, the empirical estimator $\widehat{p}$, the MI estimator $\widehat{p}_n^{\mathrm{MI}}$ with $M = 20$, and the unconditional XP estimator $\widehat{p}_n$. It should be noted that only 9.1% patients were still in a hospital bed at the end of the study, for whom eventual admission to ICU is unknown (missing cure status). Therefore, the proportion of individuals with observed cure status is high.

In this situation, the estimators $\widehat{p}_n^c$, $\widehat{p}_{1,n}$ and $\widehat{p}_n^{\mathrm{MI}}$ are expected to perform nicely, and the results for $\widehat{p}$ are likely to improve as the biased performance towards insufficient cure status information fades away. The estimator $\widehat{p}_{2,n}$ tends to overestimate probability of requiring ICU. In addition, XP estimator is expected to perform poorly since it dismisses the significant information given by the observed cured individuals.

As Table 4 shows, the estimated probabilities of admission to ICU given by the empirical estimator, the MI estimator and the proposed estimator are very similar. This suggests that it is possible that the missing data mechanism from this

**Table 4.** Demographic characteristics of COVID-19 patients, and the numbers of dead patients (death), patients known to be cured (cured) and patients with unknown cure status (unknown). Also given are the estimated probability of admission to ICU ($p$), when the probability of not requiring ICU ($1 - p$) is estimated using $1 - \widehat{p}_n^c$, $1 - \widehat{p}$, $1 - \widehat{p}_{1,n}$ , $1 - \widehat{p}_{2,n}$, $1 - \widehat{p}_n^{\mathrm{MI}}$ and $1 - \widehat{p}_n$ with their respective estimated standard errors ($\widehat{se}$).

| | | Uncured | Censored | | | Estimated probability of requiring ICU | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Variables | Count (%) | ICU admission | Dead[a] | Discharged[a] | Unknown | $\widehat{p}_n^c$ ($\widehat{se}_B$) | $\widehat{p}$ ($\widehat{se}$) | $\widehat{p}_{1,n}$ ($\widehat{se}_B$) | $\widehat{p}_{2,n}$ ($\widehat{se}_B$) | $\widehat{p}_n^{\mathrm{MI}}$ ($\widehat{se}_B$) | $\widehat{p}_n$ ($\widehat{se}$) |
| **Age** | | | | | | | | | | | |
| 0–24 years | 22 (0.9) | 0 | 0 | 17 | 5 | 0.000 (0.000) | 0.000 (0.000) | 0.000 (0.000) | 0.156 (0.087) | 0.000 (0.000) | 0.000 (0.000) |
| 25–54 years | 359 (15.1) | 25 | 5 | 316 | 13 | 0.071 (0.013) | 0.072 (0.014) | 0.071 (0.014) | 0.071 (0.014) | 0.077 (0.014) | 0.172 (0.056) |
| 55–64 years | 354 (14.9) | 41 | 24 | 271 | 18 | 0.114 (0.017) | 0.122 (0.018) | 0.117 (0.017) | 0.123 (0.018) | 0.121 (0.018) | 0.138 (0.022) |
| 65–74 years | 582 (24.5) | 96 | 35 | 411 | 40 | 0.165 (0.016) | 0.177 (0.016) | 0.169 (0.016) | 0.180 (0.018) | 0.176 (0.017) | 0.318 (0.076) |
| 75–84 years | 571 (24.0) | 34 | 101 | 377 | 59 | 0.059 (0.010) | 0.066 (0.011) | 0.061 (0.010) | 0.061 (0.011) | 0.067 (0.011) | 0.083 (0.016) |
| 85 years and over | 492 (20.7) | 1 | 163 | 246 | 82 | 0.002 (0.002) | 0.002 (0.002) | 0.002 (0.002) | 0.027 (0.012) | 0.003 (0.003) | 0.004 (0.004) |
| **Sex** | | | | | | | | | | | |
| Female | 1118 (47.0) | 55 | 136 | 822 | 105 | 0.049 (0.006) | 0.054 (0.007) | 0.051 (0.007) | 0.051 (0.007) | 0.053 (0.007) | 0.091 (0.022) |
| Male | 1262 (53.0) | 142 | 192 | 816 | 112 | 0.114 (0.009) | 0.124 (0.010) | 0.115 (0.009) | 0.131 (0.012) | 0.124 (0.010) | 0.175 (0.034) |

[a] "cured" from the event "admission to ICU."

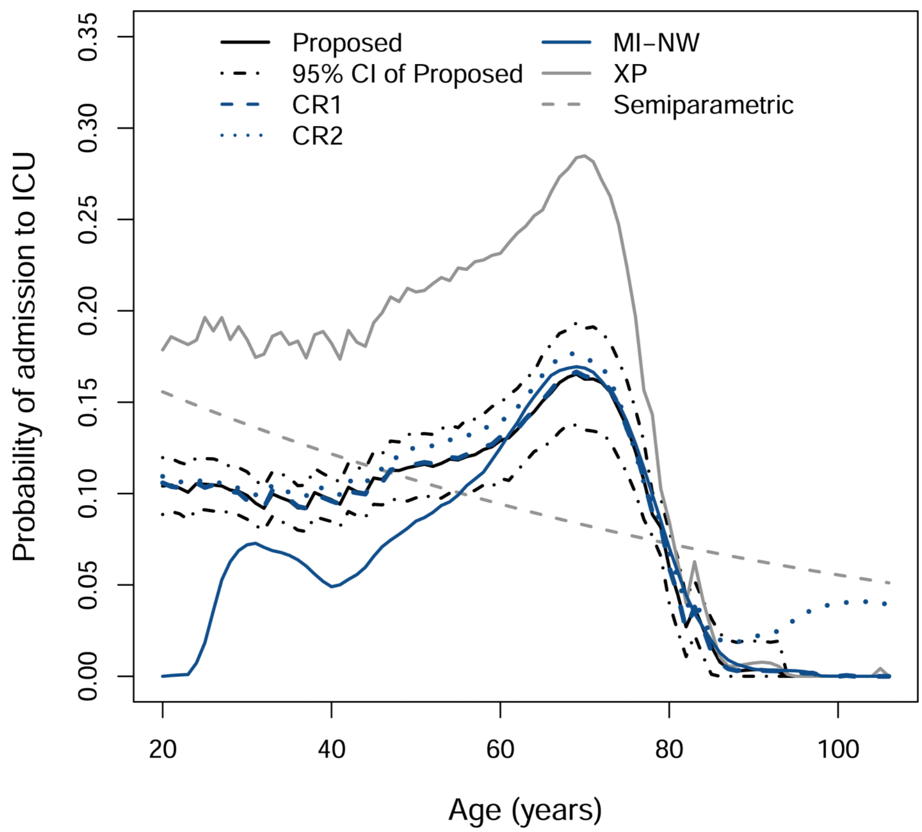


**Figure 3.** Estimation of the probability of admission to ICU for hospitalized COVID-19 patients estimated using the proposed estimator $1 - \widehat{p}_h^c(x)$ and its 95% CI, the competing risks estimators $1 - \widehat{p}_{1,h}(x)$ (CR1) and $1 - \widehat{p}_{2,h}(x)$ (CR2), the XP estimator $1 - \widehat{p}_h(x)$ (all computed with the bootstrap bandwidth), the MI-NW estimator $1 - \widehat{p}_h^{\text{MI-NW}}(x)$ (computed using the cross-validation bandwidth), and the semiparametric estimator $1 - p(x; \hat{\gamma})$.

dataset is close to siMAR assumption. On the other hand, the estimated probabilities given by the XP estimator $\widehat{p}_n$ seem to be too high.

Figure 3 shows the estimated probability of requiring admission to ICU depending on age, obtained using the estimator $1 - \widehat{p}_h^c(x)$, the competing risks estimators $1 - \widehat{p}_{1,h}(x)$, $1 - \widehat{p}_{2,h}(x)$, the XP estimator $1 - \widehat{p}_h(x)$, all computed using the bootstrap bandwidth selector as in the breast cancer example, the semiparametric estimator $1 - p(x; \hat{\gamma})$, and the MI-NW estimator $1 - \widehat{p}_h^{\text{MI-NW}}(x)$ computed using the same bandwidth selectors as in the breast cancer example. Although the semiparametric estimator suggests a uniformly decreasing effect of age on the probability of admission to the ICU, the other three estimators indicate that the logistic assumption for the cure probability might not be acceptable, as the curve patterns are characterized by a constant to a slightly increasing probability of admission to the ICU for younger patients (below 55 years), a sharp increase of the probability for middle age patients (from 55 to 69 years) and a decrease for elderly patients (70 years or older). For the aforementioned reasons, the XP estimator seems to overestimate the probability of ICU admission. Regarding the MI-NW estimator, the pattern of the estimated probability is consistent with that of the proposed estimator. However, it seems to underestimate the probability of admission to ICU for young-to-middle age patients. This is due to the low percentage of observed admissions to ICU for patients of those ages, resulting in an estimation with a high percentage of missing response.

## 6 Discussion

A novel nonparametric estimator of the conditional probability of cure is proposed for the MCM when some censored individuals can be observed to be cured from the event. It reduces to well-known estimators in the literature for the unconditional setting, when there is a cure threshold, if there are no observed cured individuals, and when there is no censoring. In contrast to regression based estimators, the proposed estimator is based on the MCM. It uses the available information of the observed times, and therefore can lead to substantial gain in efficiency.

When compared with the XP estimator, also based on the MCM but disregarding the information given by the cure status, it has been demonstrated to have smaller asymptotic variance. The advantage in terms of bias is not guaranteed, as it depends on the conditional probability of knowing the cure status information and the censoring distribution. On the other hand, simulations and the analysis of two real data examples show that our estimator yields significant improvement compared to XP estimator, which ignores cure status information.

The cure rate can alternatively be estimated using a competing risks model. The main disadvantage of this approach is that if the last observation is not an event nor an observed cured individual, then the estimator of the cure rate is not unique, and only upper and lower bounds are provided. In that case, if the censoring rate is high both estimates can be quite different, and deciding which one is preferable is not straightforward. However, when a high percentage remains unidentified for the cured individuals, the proposed estimator is highly competitive, while the CR2 estimator behaves poorly, particularly when the cure probability is high.

The MI-NW estimator performs well when the cure status is highly observed, but it performs poorly when few censored individuals are known to be cured as the siMAR assumption is violated. Besides, it is computationally quite expensive, particularly when the sample size is large, and it requires the selection of three different bandwidths. The semiparametric estimator[15] is somewhat affected when the logistic assumption is violated and it might be challenging to obtain stable estimates for the model parameters if the sample size is small. Moreover, the empirical estimator of the unconditional cure probability clearly underestimates the true probability and it cannot handle continuous covariates.

As discussed in Section 3.2, the probability of cure can be estimated as a regression function with the cure indicator as the response. We are aware of the existence of other ways of dealing with missingness in the response. To name one of them, the problem can also be addressed using IPW method.[55,24,56] However, it has not been considered because the performance is expected to be similar to that of the MI-NW estimator, and the efficiency of this method is expected to depend largely on the level of missingness.

In summary, the proposed estimator performs well when there are individuals known to be “cured” from the considered event, and it is efficient under high proportion of missingness in the cure status. Moreover, it does not require any parametric assumption for the cure probability. Also, it can be applied when the cure identification does not rely necessarily on the observed time-to-event being larger than a cure threshold, and when there is not any individual known to be cured or without censoring.

The proposed estimator of the cure rate requires the assumption of independent censoring. Besides, it is focused on ordinary right censored time-to-event data. However, in practical studies, lifetimes might be correlated to the censoring distribution, and observations may suffer from other types of censoring or truncation. Approaches that handle these complexities deserve further investigation.

Finally, all estimators in the present paper are specifically designed to estimate the probability of an event as a function of a one-dimensional covariate. We note that, to avoid the curse of dimensionality when estimating the probability for multiple covariates, alternative techniques such as product kernels[57] or a single-index model[58] can be considered. Such developments warrant future research.

## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research/work has been supported by MICINN grant PID2020-113578RB-I00, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020/14 and Centro de Investigación del Sistema universitario de Galicia ED431G 2019/01), all of them through the ERDF.


## Supplemental material

Supplementary material for this article is available online.

## References


1. Cognetti DM, Weber RS and Lai SY. Head and neck cancer: an evolving treatment paradigm. *Cancer* 2008; **113**: 1911–1932.
2. Peng Y and Yu B. *Cure models: methods, applications, and implementation*. Boca Raton, Florida: Chapman and Hall/CRC, 2021.

3. Maller RA and Zhou S. Estimating the proportion of immunes in a censored sample. *Biometrika* 1992; **79**: 731–739.
4. Patilea V and Van Keilegom I. A general approach for cure models in survival analysis. *Ann Stat* 2020; **48**: 2323–2346.
5. Amico M and Van Keilegom I. Cure models in survival analysis. *Annu Rev Stat Appl* 2018; **5**: 311–342.
6. Xu J and Peng Y. Nonparametric cure rate estimation with covariates. *Can J Stat* 2014; **42**: 1–17.
7. López-Cheda A, Cao R, Jácome MA et al. Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Comput Stat Data Anal* 2017; **105**: 144–165.
8. López-Cheda A, Jácome MA and Cao R. Nonparametric latency estimation for mixture cure models. *TEST* 2017; **26**: 353–376.
9. Taylor J. Semi-parametric estimation in failure time mixture models. *Biometrics* 1995; **51**: 899–907.
10. Laska EM and Meisner MJ. Nonparametric estimation and testing in a cure model. *Biometrics* 1992; **48**: 1223–1234.
11. Betensky R and Schoenfeld D. Nonparametric estimation in a cure model with random cure times. *Biometrics* 2001; **57**: 282–286.
12. Safari WC, López-de-Ullibarri I and Jácome MA. A product-limit estimator of the conditional survival function when cure status is partially known. *Biom J* 2021; **63**: 984–1005.
13. Nieto-Barajas LE and Yin G. Bayesian semiparametric cure rate model with an unknown threshold. *Scandinavian J Stat* 2008; **35**: 540–556.
14. Wu Y, Lin Y, Lu S et al. Extension of a Cox proportional hazards cure model when cure information is partially known. *Biostatistics* 2014; **15**: 540–554.
15. Bernhardt P. A flexible cure rate model with dependent censoring and a known cure threshold. *Stat Med* 2016; **35**: 4607–4623.
16. Nicolaie MA, Taylor JM and Legrand C. Vertical modeling: analysis of competing risks data with a cure fraction. *Lifetime Data Anal* 2019; **25**: 1–25.
17. Horvitz DG and Thompson DJ. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc* 1952; **47**: 663–685.
18. Robins JM, Rotnitzky A and Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc* 1994; **89**: 846–866.
19. Lipsitz SR, Zhao LP and Molenberghs G. A semiparametric method of multiple imputation. *J R Stat Soc: Ser B (Statistical Methodology)* 1998; **60**: 127–144.
20. Aerts M, Claeskens G, Hens N et al. Local multiple imputation. *Biometrika* 2002; **89**: 375–388.
21. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons, 2004.
22. Carpenter J and Kenward M. *Multiple Imputation and its Application*. New York, NY: John Wiley & Sons, 2012.
23. Wei Y, Ma Y and Carroll RJ. Multiple imputation in quantile regression. *Biometrika* 2012; **99**: 423–438.
24. Seaman SR, White IR, Copas AJ et al. Combining multiple imputation and inverse-probability weighting. *Biometrics* 2012; **68**: 129–137.
25. Wang Q, Linton O and Härdle W. Semiparametric regression analysis with missing response at random. *J Am Stat Assoc* 2004; **99**: 334–345.
26. Andridge RR and Little RJ. A review of hot deck imputation for survey non-response. *Int Stat Rev* 2010; **78**: 40–64.
27. Cheng P. Nonparametric estimation of mean functionals with data missing at random. *J Am Stat Assoc* 1994; **89**: 81–87.
28. Hsu C, He Y, Li Y et al. Doubly robust multiple imputation using kernel-based techniques. *Biommetrical J* 2016; **58**: 588–606.
29. The Cancer Genome Atlas. TCGA Research Network. https://tcgadata.nci.nih.gov/publications/tcga, 2021. Online accessed: 04-April-2022.
30. Pan H, Gray R, Braybrooke J et al. 20-year risks of breast-cancer recurrence after stopping endocrine therapy at 5 years. *N Engl J Med* 2017; **377**: 1836–1846.
31. Barnadas A, Algara M, Cordoba O et al. Recommendations for the follow-up care of female breast cancer survivors: a guideline of the Spanish Society of Medical Oncology (SEOM), Spanish Society of General Medicine (SEMERGEN), Spanish Society for Family and Community Medicine (SEMFYC), Spanish Society for General and Family Physicians (SEMG), Spanish Society of Obstetrics and Gynecology (SEGO), Spanish Society of Radiation Oncology (SEOR), Spanish Society of Senology and Breast Pathology (SESPM), and Spanish Society of Cardiology (SEC). *Clin Transl Oncol* 2018; **20**: 687–694.
32. Galician Healthcare Service. Dirección Xeral de Saúde Pública. https://www.sergas.es/Saude-publica, 2021. Online accessed: 04-April-2022.
33. Li C, Taylor JM and Sy J. Identifiability of cure models. *Stat Probab Lett* 2001; **54**: 389–395.
34. Hanin L and Huang L. Identifiability of cure models revisited. *J Multivar Anal* 2014; **130**: 261–274.
35. Nadaraya E. On estimating regression. *Theory Probab Appl* 1964; **9**: 141–142.
36. Watson G. Smooth regression analysis. *Indian J Stat, Ser A* 1964; **26**: 359–372.
37. Maller RA and Zhou S. Testing for sufficient follow-up and outliers in survival data. *J Am Stat Assoc* 1994; **89**: 1499–1506.
38. Li G and Datta S. A bootstrap approach to nonparametric regression for right censored data. *Ann Inst Stat Math* 2001; **53**: 708–729.
39. Safari WC. Nonparametric inference for the mixture cure model when the cure status is partially known. PhD Thesis, University of A Coruña, Spain, 2022.
40. Lagakos SW. General right censoring and its impact on the analysis of survival data. *Biometrics* 1979; **35**: 139–156.
41. Klein JP and Moeschberger ML. *Survival analysis: techniques for censored and truncated data*. 1230. New York: Springer-Verlag, 2003.
42. Effraimidis G and Dahl CM. Nonparametric estimation of cumulative incidence functions for competing risks data with missing cause of failure. *Stat Probab Lett* 2014; **89**: 1–7.

43. Beran R. *Nonparametric regression with randomly censored survival data*. Berkeley: University of California, Technical report. 1981.
44. Efron B. The two sample problem with censored data. *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability* 1967; 4 (University of California Press, Berkeley, CA): 831–853.
45. Verhasselt A, Flórez A, Van Keilegom I et al. The impact of incomplete data on quantile regression for longitudinal data. *FEB Research Report KBI_1906* 2019.
46. Vakulenko-Lagun B, Mandel M and Betensky RA. Inverse probability weighting methods for Cox regression with right-truncated data. *Biometrics* 2020; **76**: 484–495.
47. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
48. Efron B. Missing data, imputation, and the bootstrap. *J Am Stat Assoc* 1994; **89**: 463–475.
49. Duffy M, Harbeck N, Nap M et al. Clinical use of biomarkers in breast cancer: updated guidelines from the European group on tumor markers (EGTM). *Eur J Cancer* 2017; **75**: 284–298.
50. Colomer R, Aranda-López I, Albanell J et al. Biomarkers in breast cancer: a consensus statement by the Spanish Society of Medical Oncology and the Spanish Society of Pathology. *Clin Transl Oncol* 2018; **20**: 815–826.
51. Bowman A, Hall P and Prvan T. Bandwidth selection for the smoothing of distribution functions. *Biometrika* 1998; **85**: 799–808.
52. Quintela-del-Río A and Estévez-Pérez G. Nonparametric kernel distribution function estimation with kerdiest: an R package for bandwidth choice and applications. *J Stat Softw* 2012; **50**: 1–21.
53. Hurvich C, Simonoff J and Tsai C. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J R Stat Soc Ser B (Statistical Methodology)* 1998; **60**: 271–293.
54. Tristen H and Jeffrey R. Nonparametric econometrics: the np package. *J Stat Softw* 2008; **27**: 1–32.
55. Wang L, Rotnitzky A and Lin X. Nonparametric regression with missing outcomes using weighted kernel estimating equations. *J Am Stat Assoc* 2010; **105**: 1135–1146.
56. Seaman SR and White I. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res* 2013; **22**: 278–295.
57. Li Q and Racine JS. Nonparametric estimation of conditional distribution and quantile functions with mixed categorical and continuous data. *J Bus Econ Stat* 2008; **26**: 423–434.
58. Amico M, Van Keilegom I and Legrand C. The single-index/Cox mixture cure model. *Biometrics* 2019; **75**: 452–462.