A 2-step estimation procedure for semiparametric mixture cure models

Musta, E.; Patilea, V.; Van Keilegom, I.

Publication date
2022

Document Version
Final published version

License
CC BY

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
A 2-step estimation procedure for semiparametric mixture cure models

Eni Musta¹, Valentin Patilea² and Ingrid Van Keilegom³

¹Korteweg de Vries Institute for Mathematics, University of Amsterdam, Netherlands, e-mail: e.musta@uva.nl
²CREST, Ensai, France, e-mail: valentin.patilea@ensai.fr
³ORSTAT, KU Leuven, Belgium, e-mail: ingrid.vankeilegom@kuleuven.be

Abstract: Cure models have been developed as an alternative modelling approach to conventional survival analysis in order to account for the presence of cured subjects that will never experience the event of interest. Mixture cure models, which model separately the cure probability and the survival of uncured subjects depending on a set of covariates, are particularly useful for distinguishing curative from life-prolonging effects. In practice, it is common to assume a parametric model for the cure probability and a semiparametric model for the survival of the susceptibles. Because of the latent cure status, maximum likelihood estimation is performed by means of the iterative EM algorithm. Here, we focus on the cure probabilities and propose a two-step procedure to improve upon the performance of the maximum likelihood estimator when the sample size is not large. The new method is based on the idea of presmoothing by first constructing a non-parametric estimator and then projecting it into the desired parametric class. We investigate the theoretical properties of the resulting estimator and show through an extensive simulation study for the logistic-Cox model that it outperforms the existing method. Practical use of the method is illustrated through two melanoma datasets.

MSC2020 subject classifications: Primary 62N02.

Keywords and phrases: cure model, logistic model, presmoothing, survival analysis.

1. Introduction

Cure models are used to analyze time until occurrence of an event of interest when a proportion of the study population is immune to that event (cured). They are recently becoming increasingly popular in oncology as curative treatments are now a possibility, meaning that some patients will not experience cancer relapse/death (see for example [15, 25]). More broadly, cure models find applications in studies of fertility ([37]), hospitalization of COVID-19 patients ([28]), equipment failure in engineering ([22]), credit scoring in economics ([9, 8]), etc.. What makes statistical modeling and estimation challenging when not all subjects are susceptible to the event of interest, is the unobserved cure status. As a consequence of a limited follow-up period, all cured subject are observed as censored, hence mixed with the uncured ones.
There are two main families of cure models: promotion time models and mixture cure models (see [2] and [32] for an overview). The latter ones are particularly attractive in practice because, by separately modeling the uncure probability (incidence) and the survival of the susceptibles (latency) given possibly different sets of covariates, they are able to distinguish a curative from a life-prolonging effect. Early works on mixture cure models were fully parametric approaches ([11, 41, 14]), while more recently semi-parametric ([35, 31, 16, 43]) and non-parametric ([40, 27, 3, 18, 19]) models have been proposed. Among them, the semiparametric models are often used in practice as a reasonable compromise between flexibility and simplicity. These models assume a parametric form of the incidence and a semiparametric form for the latency, with the most common choice being the mixture of the logistic with the Cox proportional hazards model (e.g. [42, 33, 39]).

Estimation in the logistic-Cox or in general semi-parametric mixture cure models is mostly carried out via the Expectation-Maximization algorithm because of the latent cure status. Such estimators were proposed in [31] and [35] for the logistic-Cox model; in [16, 43, 21] for the logistic-accelerated failure time model. The procedure is implemented in the R-package smcure ([5]). However, for limited sample sizes which are common in practice, such iterative procedures are characterized by large mean-squared-error (MSE), convergence problems and instability of the estimators for the incidence component depending on which variables are included in the latency model (see for e.g. [24]). This might lead to incorrect conclusions regarding significant effects.

Here we propose a new second stage estimator based on presmoothing with the aim of improving upon an initially available estimator, that can for example be the smcure estimator, for small and moderate sample sizes. The initial estimator is used to construct a one-dimensional covariate, conditional on which we compute a nonparametric estimator of the cure probabilities. Afterwards, the nonparametric estimator is projected on the desired parametric class (for example logistic). This allows for direct estimation of the parametric incidence component despite the latent cure status. We focus on the cure fraction, but once that is estimated, one can also fit a semiparametric model to the latency component. Compared to the method proposed in [24], this approach does not restrict us to a one-dimensional covariate and does not require multidimensional smoothing, which is essential for practical purposes. Apart from the cure model setting, the idea of constructing a parametric estimator by nonparametric estimation has been previously proposed in the context of linear regression, variable selection and functional linear regression ([7, 1, 12]). The novelty of our method lies in using presmoothing as a second stage estimator where a preliminary available estimator is used to reduce the covariate dimension to one. In this way we only need to choose one bandwidth independently of the number of covariates and still profit from the advantages of presmoothing: lower MSE and more stable estimators.

The paper is organized as follows. In Sections 2 and 3, we describe the model and the estimation procedure. In Section 4 we show that the resulting estimator is consistent and square-root-n convergent with a Gaussian limit distribution,
provided that the initial estimator is consistent. As a particular case, we focus on the logistic/Cox mixture cure model in Section 5 and illustrate through an extensive simulation study that the proposed estimator outperforms the \textsc{smcure} estimator by significantly reducing its mean squared error. In addition, the second step using presmoothing makes the estimator more stable towards misspecifications in the latency model. Finally, in Section 6, we apply the method to two medical datasets and show that in practice it can lead to different conclusions compared to the \textsc{smcure} estimator.

2. The semiparametric mixture cure model

Suppose we are interested in the time $T$ until a certain event happens for a mixed population of cured ($T = \infty$) and uncured ($T = T_0 < \infty$) subjects. Let $B$ be a 0-1 random variable indicating the uncured status: $B = 1$ for susceptible individuals and $B = 0$ otherwise. Due to the limited follow-up period, we cannot actually observe $T$ and $B$. Instead we observe a finite follow-up time $Y = \min(T, C)$ and a censoring indicator $\Delta = I_{\{T \leq C\}}$, where $C$ denotes the censoring time. As a result, for all the censored observations, the cure status is unknown. In the mixture cure model, the survival function of $T$ given two covariate vectors $X \in \mathbb{R}^p$, $Z \in \mathbb{R}^q$, is given by

$$S(t|x, z) = P(T > t|X = x, Z = z) = \pi_0(x) + (1 - \pi_0(x)) S_u(t|z),$$

where $S_u(t|z) = P(T > t|Z = z, B = 1)$ is the survival function of the susceptibles and $\pi_0(x) = P(B = 0|X = x)$ denotes the cure probability. Using two covariate vectors $X$ and $Z$ for modeling the incidence and the latency allows the cure probability and the survival of the uncured to be affected by different variables. However it does not exclude situations in which the two vectors $X$ and $Z$ are exactly the same or share some components.

In the context of mixture cure models, the classical survival analysis assumption of independent censoring, means that $T_0 \perp (C, X)|Z$ and $B \perp (C, T_0, Z)|X$, which imply that

$$T \perp C|(X, Z),$$

(see Lemma 1 in the supplementary material of [24]). As a result, we also have

$$P(T = \infty|X, Z) = P(T = \infty|X) \quad \text{and} \quad P(T_0 \leq t|X, Z) = P(T_0 \leq t|Z).$$

Among various modeling approaches for the incidence and the latency, the most common choice is a parametric (logistic) model for the incidence and a semiparametric (Cox or accelerated failure time) model for the latency ([27, 4, 15, 42, 33, 39]). The popularity of such a choice is due to the simplicity and ease of interpretation. We focus on this type of model and assume that

$$\pi_0(x) = 1 - \phi(\gamma_0^T x),$$

where $\phi: \mathbb{R} \rightarrow [0, 1]$ is a known function, $\gamma_0 \in \mathbb{R}^{p+1}$ and $\gamma_0^T$ denotes the transpose of the vector $\gamma_0$. Here the first component of $x$ is taken to be equal to...
1 and the first component of $\gamma$ corresponds to the intercept. In particular, for
the logistic model, we have

$$\phi(u) = \frac{e^u}{1 + e^u}.$$  \hspace{1cm} (3)

To check the fit of this model in practice, one can compare the prediction error with that of a more flexible single-index model as done in [3] or use the

test proposed in [23] which is currently developed only for a one-dimensional
covariate.

For the latency, we assume a semiparametric model $S_u(t|z) = S_u(t|z; \beta_0, \Lambda_0)$
depending on a finite-dimensional parameter $\beta_0 \in \mathbb{R}^q$, and an infinite-dimensional
parameter $\Lambda_0$. The main examples we keep in mind are the Cox proportional
hazards (PH) model

$$S_u(t|z) = \exp\{-\Lambda_0(t) \exp(\beta_0^T z)\}, \hspace{1cm} (4)$$

and the accelerated failure time model (AFT)

$$S_u(t|z) = \exp \{-\Lambda_0 \left( \exp (\beta_0^T z) t \right) \},$$

where $\Lambda_0$ is the baseline cumulative hazard.

The goal is to estimate the true parameters $\gamma_0$, $\beta_0$ and $\Lambda_0$ on the basis of $n$
i.i.d. observations $(Y_1, \Delta_1, X_1, Z_1), \ldots, (Y_n, \Delta_n, X_n, Z_n)$. The general conditions
under which the semiparametric mixture cure model is identifiable, meaning
that different parameter values lead to different distributions of the observed
variables $(Y, \Delta, X, Z)$, were derived in [26] and are the following:

(I1) if $\phi(\gamma, X) = \phi(\tilde{\gamma}, X)$ almost surely, then $\gamma = \tilde{\gamma}$,

(I2) the function $S_u(\cdot|z)$ has support $[0, \tau(z)]$,

(I3) $P(C > \tau(Z)|X, Z) > 0$ for almost all $X$ and $Z$,

(I4) if, for all $t \geq 0$, we have $S_u(t|Z; \Lambda, \beta) = S_u(t|Z; \tilde{\Lambda}, \tilde{\beta})$ almost surely, then

$\Lambda = \tilde{\Lambda}$ and $\beta = \tilde{\beta}$.

In the particular case of the logistic-Cox model the conditions become:

(I1’) for all $x$, $0 < \phi(\gamma_0^T x) < 1$,

(I2’) the function $S_u$ has support $[0; \tau_0]$ for some $\tau_0 < \infty$,

(I3’) $P(C > \tau_0|X; Z) > 0$ for almost all $X$ and $Z$,

(I4’) the matrices $Var(X)$ and $Var(Z)$ are positive definite,

(see Proposition 1 and 2 in [26]). Conditions I3 and I3’ are of particular
importance in the context of mixture cure models and essentially tell us that, in
order to correctly identify the cure proportion, we need sufficiently long follow-
up beyond the time when the events occur. In practice, this can be evaluated
based on the plateau of the Kaplan-Meier estimator and the expert (medical)
knowledge.
3. The 2-step estimation procedure

Estimation in semiparametric mixture cure models is usually performed via the expectation maximization algorithm because of the latent cure status. Such method has been proposed by \[31, 35\] for the logistic-Cox mixture cure model, and by \[16, 43\] for the logistic-AFT model. The procedure is implemented in the R package smcure \([5]\). Despite the simplicity of the method, simultaneous computation of \(\gamma_0\), \(\beta_0\) and \(L_0\) through an iterative procedure leads to several problems for finite, not large sample sizes which are commonly encountered in practice. This has been previously reported and illustrated in \[24, 4, 13\]. The main concerns are the large MSE, convergence problems and instability of the estimator for the incidence component depending on which variables are included in the latency model. In particular, if the latency model is misspecified, even the estimators of the incidence parameters suffer from induced bias. To alleviate these problems, we propose the following 2-step estimation procedure that makes use of presmoothing.

We start with some preliminary estimator \(\tilde{\gamma}_n\) of \(\gamma_0\). This can be any estimator that satisfies the conditions described in Section 4 and in particular for the logistic-Cox or logistic-AFT model we can use the smcure estimator. We use this preliminary estimator to construct the one-dimensional index \(V = \tilde{\gamma}_n X\) estimating \(V = \gamma_0 X\). Based on this new one-dimensional covariate, we compute a nonparametric estimator of the cure probability for each subject defined as follows

\[
\hat{\pi}_n(x) = \prod_{t \in \mathbb{R}} \left( 1 - \frac{\hat{H}_{1, \tilde{\gamma}_n} (dt|\tilde{\gamma}_n T x)}{H_{\tilde{\gamma}_n} ([t, \infty)|\tilde{\gamma}_n T x)} \right),
\]

(5)

where \(\hat{H}_{\gamma_n} ([t, \infty)|u) = \hat{H}_{1, \gamma_n} ([t, \infty)|u) + \hat{H}_{0, \gamma_n} ([t, \infty)|u), \hat{H}_{1, \gamma_n} (dt|u) = \hat{H}_{1, \gamma_n} ((t-\Delta) dt, t)|u)\) for small \(\Delta\) and

\[
\hat{H}_{l, \gamma_n} ([t, \infty)|u) = \sum_{i=1}^{n} \frac{k_b (\tilde{\gamma}_n T X_i - u)}{\sum_{j=1}^{n} k_b (\tilde{\gamma}_n T X_j - u)} \mathbb{I}(Y_i \geq t, \Delta_i = l), \quad l = 0, 1,
\]

are estimators of

\[
H_{l} ([t, \infty)|u) = \mathbb{P} (Y \geq t, \Delta = l|\gamma_0 T X = u),
\]

(6)

and \(H([t, \infty)|u) = H_1([t, \infty)|u) + H_0([t, \infty)|u)\). Here \(k\) is a one-dimensional kernel function, \(b = b_n\) is a bandwidth sequence and \(k_b(\cdot) = k(\cdot/b)/b\).

The estimator \(\hat{\pi}_n(x)\) coincides with the Beran estimator of the conditional survival function \(S\) at the largest observed event time \(Y_{(m)}\) and does not require any specification of \(\tau_0\). Since \(\hat{H}_{1, \tilde{\gamma}_n} (dt|\tilde{\gamma}_n T x)\) is different from zero only at the observed event times, computation of \(\hat{\pi}_n(x)\) requires only a product over \(t\) in the set of the observed event times. Afterwards, we consider the logistic likelihood

\[
\hat{L}_{n,1}(\gamma) = \prod_{i=1}^{n} \phi(\gamma T X_i)^{1-\hat{\pi}_n(X_i)}[1 - \phi(\gamma T X_i)]^{\hat{\pi}_n(X_i)},
\]
and define \( \hat{\gamma}_n \) as the maximizer of

\[
\log \hat{L}_{n,1}(\gamma) = \sum_{i=1}^{n} \left\{ 1 - \hat{\pi}_n(X_i) \log \phi(\gamma^T X_i) + \hat{\pi}_n(X_i) \log \left[ 1 - \phi(\gamma^T X_i) \right] \right\} \tau(X_i).
\]

(7)

We introduce a trimming function \( \tau(\cdot) \geq 0 \) to avoid regions where the density function of the index \( \gamma^T X \), for \( \gamma \) in a neighborhood of \( \gamma_0 \), approaches zero (as done for example in [17]). We discuss possible choices of \( \tau(\cdot) \) in Section 4. Existence and uniqueness of \( \hat{\gamma}_n \) hold under the same conditions as for the maximum likelihood estimator in the binary outcome regression model where \( 1 - \hat{\pi}_n(X_i) \) is replaced by the outcome \( B_i \). For example, in the logistic model, it is required that \( p < n \) and the matrix of the variables \( X \) has full rank. Estimation of the latency component can then be performed by maximizing the likelihood of the mixture model

\[
\prod_{i=1}^{n} \left\{ \phi(\gamma^T X_i) f_u(Y_i|Z_i; \beta, \Lambda) \right\}^{\Delta_i} \left\{ 1 - \phi(\gamma^T X_i) + \phi(\gamma^T X_i) S_u(Y_i|Z_i; \beta, \Lambda) \right\}^{1-\Delta_i},
\]

with respect to \( \beta \) and \( \Lambda \) for \( \gamma = \hat{\gamma}_n \). Here \( f_u(t|Z; \beta, \Lambda) = -\frac{d}{dt} S_u(t|Z; \beta, \Lambda) \). In practice this would mean performing the EM algorithm (as in the smcure package) only on the latency component, i.e. keeping \( \gamma = \hat{\gamma}_n \) fixed and updating \( \beta \) and \( \Lambda \) in each iteration.

We call this a 2-step estimator because it relies on a preliminary estimator \( \hat{\gamma}_n \), which is used to construct the one-dimensional covariate \( \hat{\gamma}_n^T X \). In this way, independently of the dimension of \( X \), the kernel estimator requires only one bandwidth parameter. The idea of a single-index structure is used in several papers to avoid multidimensional regression (e.g. [34, 17]), but has not previously been exploited in the context of cure models and presmoothing. A nonparametric estimator for the cure probability in (5) could be obtained using any nonparametric estimator of the conditional survival function as done for example in [30, 29] for estimation of the default probability. Here we use the Beran estimator since it is easy to compute and exhibits a good behavior.

### 3.1. Rationale behind the new approach

By definition we have

\[
\pi_0(x) = \mathbb{P}(T = \infty \mid X = x) = \mathbb{P}(T = \infty \mid \gamma_0^T X = \gamma_0^T x).
\]

Moreover, since from our model it follows that \( T \perp C \mid \gamma_0^T X, \beta_0^T Z \), we have

\[
H_1 \left( dt|\gamma_0^T x, \beta_0^T z \right) = F_C \left( [t, \infty) \mid \gamma_0^T x, \beta_0^T z \right) F_T \left( dt|\gamma_0^T x, \beta_0^T z \right),
\]

\[
H \left( [t, \infty) \mid \gamma_0^T x, \beta_0^T z \right) = F_T \left( [t, \infty) \mid \gamma_0^T x, \beta_0^T z \right) F_C \left( [t, \infty) \mid \gamma_0^T x, \beta_0^T z \right),
\]

which yields

\[
\frac{F_T \left( dt|\gamma_0^T x, \beta_0^T z \right)}{F_T \left( [t, \infty) \mid \gamma_0^T x, \beta_0^T z \right)} = \frac{H_1 \left( dt|\gamma_0^T x, \beta_0^T z \right)}{H \left( [t, \infty) \mid \gamma_0^T x, \beta_0^T z \right)}.
\]
As in [27], we obtain that
\[
P(T = \infty \mid \gamma T X = \gamma T_0 X, \beta T Z = \beta T_0 Z) = \prod_{t \in \mathbb{R}} \left\{ 1 - \frac{H_1(dT|\gamma T_0 X, \beta T_0 Z)}{H([t, \infty)|\gamma T_0 X, \beta T_0 Z)} \right\},
\]
where \(\prod_{t \in \mathbb{R}}\) denotes the product integral. By the first part of (2), the product integral is also equal to \(P(T = \infty \mid \gamma T_0 X = \gamma T_0 x)\). Similarly, under the slightly stronger assumption that \(T \perp C \mid \gamma T_0 X\), we obtain
\[
P(T = \infty \mid \gamma T_0 X = \gamma T_0 x) = \prod_{t \in \mathbb{R}} \left\{ 1 - \frac{H_1(dT|\gamma T_0 X)}{H([t, \infty)|\gamma T_0 x)} \right\}.
\]
This justifies the definition of our estimator in (5). This assumption is satisfied if \(C \perp (X, Z) \mid \gamma T_0 X\). We restrict ourselves to this case for simplicity in order to have conditioning on only one index. However, the method can be used in general conditioning on both \(\gamma T_0 X\) and \(\beta T_0 Z\). We illustrate this through one of the simulation settings in Section 5.

4. Asymptotic results

In this section we focus on models where the survival function of the susceptibles \(S_u\) has fixed support \([0, \tau_0]\) such that
\[
\inf_{\gamma \in G} \mathbb{P}(C > \tau_0 \mid \gamma T X) > 0 \ \text{almost surely.} \tag{8}
\]
This is the case when the latency follows a Cox regression model and the identifiability assumption (I3') is satisfied. Moreover, we assume that \(T \perp C \mid \gamma T_0 X\) but the results can be generalized as mentioned in the previous section.

Let us sketch the arguments we will use to obtain asymptotic properties of \(\hat{\gamma}_n\). Note that \(\hat{\gamma}_n\) is the maximizer of the Bernoulli-type log-likelihood in (7). Hence, the main issue is dealing with the nonparametric estimator \(\hat{\pi}_n\), which replaces the latent binary outcome. By definition we have
\[
\hat{\pi}_n(x) = \hat{g}_{\hat{\gamma}_n}(\hat{\gamma}_n T x),
\]
where
\[
\hat{g}_{\gamma}(u) = 1 - \tilde{F}_n(\tau_0 \mid \gamma T X = u) = \prod_{t \in \mathbb{R}} \left( 1 - \frac{\hat{H}_{1,\gamma}(dt|u)}{H_{\gamma}([t, \infty)|u)} \right). \tag{9}
\]
Note that the product integral actually over \(t \in \mathbb{R}\) is the same as over \(t \leq \tau_0\) because \(\hat{H}_{1,\gamma}(dt|u) = 0\) for \(t > Y_{(m)}\), where \(Y_{(m)} \leq \tau_0\) is the last observed event time. We can also write
\[
\pi_0(x) = g_{\gamma_0}(\gamma_0 T x),
\]
with
\[
g_{\gamma}(u) = 1 - F_T(\tau_0 \mid \gamma T X = u) = \prod_{t \in \mathbb{R}} \left( 1 - \frac{H_1(dt|u)}{H([t, \infty)|u)} \right). \tag{10}
\]
Hence
\[
\hat{\pi}_n(x) - \pi_0(x) = \hat{g}_{\hat{\gamma}_n} \left( \hat{\gamma}_n^T x \right) - g_{\gamma_0} \left( \gamma_0^T x \right) \\
= \left\{ \hat{g}_{\gamma_0} \left( \gamma_0^T x \right) - g_{\gamma_0} \left( \gamma_0^T x \right) \right\} + \left\{ \hat{g}_{\hat{\gamma}_n} \left( \hat{\gamma}_n^T x \right) - \hat{g}_{\gamma_0} \left( \gamma_0^T x \right) \right\} \\
= \left\{ F_n \left( \tau_0 | \gamma_0^T x = \gamma_0^T x \right) - F_T \left( \gamma_0^T x = \gamma_0^T x \right) \right\} \\
+ \left\{ \hat{g}_{\hat{\gamma}_n} \left( \hat{\gamma}_n^T x \right) - \hat{g}_{\gamma_0} \left( \gamma_0^T x \right) \right\}. \tag{11}
\]

The first term on the right hand side of the equation, can be dealt with as usual being the difference between $F_n$ and $F$ conditionally on a one dimensional covariate $\gamma_0^T X$. The second term results from using $\hat{\gamma}_n$ instead of $\gamma_0$ when constructing the one-dimensional covariate $\hat{\gamma}_n^T X$. The behaviour of this term depends on the properties of the preliminary estimator $\hat{\gamma}_n$. We first formulate the results for a general preliminary estimator $\hat{\gamma}_n$ and a general parametric function $\phi$. We then show that, for the logistic-Cox model, the maximum likelihood estimator satisfies the required conditions.

The following assumptions are needed for consistency of $\hat{\gamma}_n$.

(C1) The preliminary estimator is consistent, i.e. $\hat{\gamma}_n - \gamma_0 = o_P(1)$.

(C2) The parameter $\gamma_0$ lies in the interior of a compact set $G \subset \mathbb{R}^p$.

(C3) There exist some constants $a > 0$, $c > 0$ such that

\[
|\phi(\gamma_1^T x) - \phi(\gamma_2^T x)| \leq c \|\gamma_1 - \gamma_2\|^a, \quad \forall \gamma_1, \gamma_2 \in G, \forall x \in X,
\]

where $\| \cdot \|$ denotes the Euclidean norm and $X \subset \mathbb{R}^p$ is the support of $X$.

(C4) $\inf_{\gamma \in G} \inf_{x \in X} \phi(\gamma^T x) > 0$ and $\sup_{\gamma \in G} \sup_{x \in X} \phi(\gamma^T x) < 1$.

(C5) For any $\gamma \in G$, the support $X_\gamma$ of $\gamma^T X$ is a bounded convex subset of $\mathbb{R}$.

(C6) The density $f_{X,\gamma}(\cdot)$ of $X_\gamma$ is twice differentiable with a bounded second derivative.

(C7) The bandwidth $b$ is such that $nb^4 \to 0$ and $nb^{3+\xi}/(\log b^{-1}) \to \infty$ for some $\xi > 0$.

(C8) (i) The functions $H([0, t]|u)$, $H_1([0, t]|u)$ defined in (6) are twice differentiable with respect to $u$, with uniformly bounded derivatives for all $t \leq \tau_0$, $u \in X_{\tau_0}$. Moreover, there exist continuous nondecreasing functions $L_1$, $L_2$, $L_3$ such that $L_1(0) = 0$, $L_2(\tau_0) < \infty$ and for all $t, s \in [0, \tau_0]$, $u \in X_{\tau_0}$,

\[
|H_c(t|u) - H_c(s|u)| \leq |L_1(t) - L_1(s)|,
\]

\[
|H_{1c}(t|u) - H_{1c}(s|u)| \leq |L_1(t) - L_1(s)|,
\]

\[
\left| \frac{\partial H_c(t|u)}{\partial u} - \frac{\partial H_c(s|u)}{\partial u} \right| \leq |L_2(t) - L_2(s)|,
\]

\[
\left| \frac{\partial H_{1c}(t|u)}{\partial u} - \frac{\partial H_{1c}(s|u)}{\partial u} \right| \leq |L_3(t) - L_3(s)|,
\]

where the subscript $c$ denotes the continuous part of a function.
(ii) The number of jump points for the distribution function $F_C(t|u)$ of the censoring times given the index $\gamma_0^T X = u$, are finite and the same for all $u$. The partial derivative of $F_C(t|u)$ with respect to $u$ exists and is uniformly bounded for all $t \leq \tau_0$, $u \in \mathcal{X}_{\gamma_0}$. Moreover, the partial derivative with respect to $u$ of $F_T(t|u)$ (distribution function of the survival times $T$ given $\gamma_0^T X = u$) exists and is uniformly bounded for all $t \leq \tau_0$, $u \in \mathcal{X}_{\gamma_0}$.

(C9) The function $(x, \gamma) \mapsto g_{\gamma}(\gamma^T x)$ is continuously differentiable with respect to $\gamma$ and the vector $\nabla_{\gamma} g_{\gamma}(\gamma^T x)$ is continuous with respect to $(x, \gamma)$.

Assumptions (C2)-(C4), (C6)-(C8) are standard assumptions (see for example [27, 24, 38]). Assumptions (C5) and (C9) are needed because we compute the nonparametric estimator using the index $\tilde{\gamma}_n^T X$ instead of $\gamma_0^T X$. Such assumptions appear for example in [17].

A possible choice of the trimming function $\tau(\cdot)$ in (7) could be $\tau(x) = 1_{\{f_{X\gamma}(\gamma_0^T X) \geq c\}}$ for some $c > 0$, which is asymptotically equivalent to the previous proposal. In practice, we can use $\tau(x) = 1_{\{\hat{f}_{X\gamma}(\tilde{\gamma}_n^T X) \geq c\}}$ based on the preliminary estimator $\tilde{\gamma}_n$.

**Theorem 1.** Assume that conditions (C1)-(C9) are satisfied. Then

$$\hat{\gamma}_n - \gamma_0 = o_P(1).$$

In order to obtain asymptotic normality of $\hat{\gamma}_n$ at rate $\sqrt{n}$, we need the following additional assumptions.

(N1) For each $x \in \mathcal{X}$, the function $\gamma \mapsto \phi(\gamma^T x)$ is twice continuously differentiable with uniformly bounded derivatives in $G \times X$.

(N2) The matrix $E \left[ \phi'(\gamma_0^T X)^2 X X^T \right]$ is positive definite.

(N3) The preliminary estimator $\tilde{\gamma}_n$ is $\sqrt{n}$ consistent and such that there exists a function $\zeta$ such that

$$\tilde{\gamma}_n - \gamma_0 = \frac{1}{n} \sum_{i=1}^{n} \zeta(T_i, \Delta_i, X_i, Z_i) + R_n,$$

with $\|R_n\| = o_P(n^{-1/2})$ and $E[\zeta(T, \Delta, X, Z)] = 0$.

Again (N1)-(N2) are standard assumptions, while (N3) arises from the use of the index $\tilde{\gamma}_n^T X$ instead of $\gamma_0^T X$. As a result the asymptotic variance of $\hat{\gamma}_n$ will also depend on the asymptotic variance of the preliminary estimator $\tilde{\gamma}_n$.

**Theorem 2.** Assume that conditions (C1)-(C9), (N1)-(N3) are satisfied. Then

$$\sqrt{n}(\hat{\gamma}_n - \gamma_0) \overset{d}{\to} N(0, \Sigma_{\gamma}),$$

with covariance matrix $\Sigma_{\gamma}$ defined in (27).
Given the complicated form of the covariance matrix $\Sigma_\gamma$, we suggest using a bootstrap procedure for estimating the standard errors as also done for the maximum likelihood estimator of a semi-parametric mixture cure model.

If we consider the particular case of a logistic-Cox mixture cure model and take the maximum likelihood estimator as a preliminary estimator $\hat{\gamma}_n$, then assumptions (C2)-(C4), (C9), (N1)-(N2) are obviously satisfied for the logistic model. Moreover (C1) and (N2) are satisfied if the cumulative baseline function $\Lambda_0$ is strictly increasing and continuously differentiable under the condition

$$\inf_z \mathbb{P}(T_0 \geq \tau_0 | Z = z) > 0,$$

(see Theorem 2 and Theorem 3 in [20]). Then, from Theorems 1 and 2 it follows that the 2-step estimator is also consistent and $\sqrt{n}$-convergent. If we continue estimating the latency sub-model using this estimator of $\gamma_0$, then the resulting estimator of $\beta_0$ and $\Lambda_0$ have the desired asymptotic behavior as in Theorems 2 and 4 in [24]. The proof remains the same given that they only use consistency and the asymptotic i.i.d. expression of the estimator as in assumption (N3).

5. Simulation study

In this section we investigate the finite-sample behaviour of the 2-step approach in the logistic/Cox mixture cure model and compare it with the maximum likelihood estimator implemented in the R package smcure. We use the smcure estimator as preliminary estimator $\hat{\gamma}_n$ for the new method.

We make some standard and common choices when computing the nonparametric estimator in (5). The kernel function $k$ is taken to be the Epanechnikov kernel $k(u) = (3/4)(1-u^2)1_{|u| \leq 1}$. Using the preliminary estimator $\hat{\gamma}_n$, we compute the smoothing bandwidth by cross-validation as implemented in the R package np for kernel estimators of conditional distribution functions, in our case for the estimation of $H = H_0 + H_1$ given $\hat{\gamma}^T X$. In addition, we restrict ourselves to the interval $[0, \hat{Y}(m)]$, where $\hat{Y}(m)$ is the last observed event time since the estimator of the cure probability $\hat{\pi}$ in (5) is essentially a product over values of $t$ that are equal to the observed event times. This means that we use the cross-validation bandwidth for estimating the conditional distribution $H(t | \hat{\gamma}^T x)$ for $t \leq \hat{Y}(m)$.

We could use a trimming function

$$\tau(x) = 1_{(f_{X_n}(\hat{\gamma}_n^T x) \geq c)},$$

for some small value of $c$ as proposed in Section 4. However, we observe that in practice this does not affect the results since $c$ can be chosen as arbitrarily small. Hence, we do not do any trimming so that we do not have to worry about the choice of the trimming constant. The trimming is mainly introduced for the asymptotic study in order to avoid the assumption that the density of the index is bounded from below by a positive constant.
We consider four different models and for each of them, three scenarios, covering a wide range of settings with different number and choice of covariates (continuous and discrete), different cure and censoring rate, and different censoring mechanisms (independent of covariates, depending on the same index as the incidence model, depending on both indexes of the incidence and latency). The models are as follows.

Model 1. Both incidence and latency depend on two independent continuous covariates $X_1 = Z_1 \sim N(0,1)$ and $X_2 = Z_2 \sim \text{Unif}(-1,1)$. We generate the cure status $B$ as a Bernoulli random variable with success probability $\phi(\gamma_T X)$ where $\phi$ is the logistic function in (3) and $\gamma = (\gamma_0, 1.5, 1.5)$. The survival times for the uncured observations are generated according to a Weibull proportional hazards model

$$S_u(t|z) = \exp \left(-\mu t^\rho \exp \left(\beta^T z\right)\right),$$

and are truncated at $\tau_0 = 15$ for $\rho = 0.75$, $\mu = 1.5$ and $\beta = (0.5, 0.3)$. The censoring times are independent from $X$ and $T$. They are generated from the exponential distribution with parameter $\lambda_C$ and are truncated at $\tau = 17$.

Model 2. Both incidence and latency depend on three independent covariates $X_1 = Z_1 \sim N(0,1)$, $X_2 = Z_2 \sim \text{Bernoulli}(0.3)$ and $X_3 = Z_3 \sim \text{Bernoulli}(0.7)$. The cure status and the survival times for the uncured observations are generated as in Model 1 for $\gamma = (\gamma_0, -1, 1, -0.3)$, $\beta = (-0.8, 1.5, -0.5)$, $\rho = 0.75$, $\mu = 1.5$, and $\tau_0 = 7$. The censoring times are generated according to a Weibull proportional hazards model

$$S_C(t|x) = \exp \left(-\lambda_C \mu t^\rho \exp \left(\gamma^T x\right)\right),$$

for various choices of $\lambda_C$ and are truncated at $\tau = 9$.

Model 3. For the incidence we consider four independent covariates: $X_1 \sim N(0,1)$, $X_2 \sim \text{Unif}(-1,1)$, $X_3$ and $X_4$ are Bernoulli random variables with parameters 0.4 and 0.6 respectively. The latency depends on three covariates: $Z_1 \sim N(0,1)$, $Z_2 = X_2$ and $Z_3 = X_4$. The cure status and the survival times for the uncured observations are generated as in Model 1 for $\gamma = (\gamma_0, -0.3, 0.8, 0.5, -1)$, $\rho = 0.75$, $\mu = 1.5$, $\beta = (0.1, 0.4, -0.2)$ and $\tau_0 = 10$. The censoring times are generated according to a Weibull proportional hazards model

$$S_C(t|x) = \exp \left(-\lambda_C \mu t^\rho \exp \left(0.4\gamma^T x + 0.5\beta^T z\right)\right),$$

for various choices of $\lambda_C$ and are truncated at $\tau = 12$.

Model 4. For the incidence we consider five independent covariates: $X_1 \sim N(0,1)$, $X_2 \sim \text{Unif}(-1,1)$, $X_3$ is Binomial with parameters 2 and 0.5, $X_4$ and $X_5$ are Bernoulli random variables with parameters 0.4 and 0.6 respectively. The latency depends on three covariates: $Z_1 \sim N(0,1)$, $Z_2 = X_3$ and $Z_3 = X_4$. The cure status and the survival times for the uncured observations are generated as in Model 1 for $\gamma = (\gamma_0, -0.8, 0.3, -0.4, 0.5, 0.6)$, $\rho = 0.75$, $\mu = 1.5$, $\beta = (0.2, -0.5, 0.3)$ and $\tau_0 = 7$. The censoring times are independent from $X$, $Z$ and $T$. They are generated from the exponential distribution with parameter $\lambda_C$ and are truncated at $\tau = 9$.
For the four models we choose the values of the unspecified parameters $\gamma_0$ and $\lambda_C$ in such a way that the cure rate is around 20%, 40% or 60% and the difference between the cure and the censoring rate is around 5%, 10% or 15%. The specification of the parameters and the corresponding censoring and cure rates are given Table 1. The truncation of the survival and censoring times on $[0, \tau_0]$ and $[0, \tau]$ is made in such a way that $\tau_0 < \tau$ and condition (13) is satisfied but in practice it is unlikely to observe event times at $\tau_0$. In this way, we try to find a compromise between theoretical assumptions and real-life scenarios. Model 3 illustrates the behavior of the method when the censoring times depends on both the indexes of the incidence and latency models (which was assumed for simplicity in the theoretical study).

We consider samples of size $n = 200$ and $n = 400$ since we aim to provide a method that improves upon the maximum likelihood estimator for small and moderate sample size. For each configuration, 1000 datasets were generated and the estimators of $\beta$ and $\gamma$ were computed through smcure and the proposed 2-step approach. We report the bias, variance and mean squared error (MSE) of the estimators, computed over the iterations for which the smcure procedure converges, in Tables 2-4. In some scenarios, mainly corresponding to the ones with 15% additional censoring compared to the cure rate and smaller sample size, the iterative procedure of the EM algorithm in smcure does not converge. The most problematic setting in this regard is Model 2 scenario 3, for which $73/1000$ iterations do not converge for $n = 200$ and $36/1000$ for $n = 400$. The boxplots of the estimators for both methods in these non-convergent iterations are shown in Figure 1. In the other settings, only around 2% or less of the iterations do not converge.

Simulations show that the 2-step approach improves considerably upon smcure for estimation of $\gamma$ when $n = 200$ and the censoring rate among the uncured observations is higher. In almost all scenarios the 2-step approach has a smaller variance, which is expected due to presmoothing, but it also often exhibits a lower bias. As the sample size increases or the censoring rate decreases, we see less difference between the two methods but still, the 2-step approach is usually
Table 2
Bias, variance and MSE of $\hat{\gamma}$ and $\hat{\beta}$ for secure (second rows) and the 2-step approach (first rows) in Models 1 and 2.

<table>
<thead>
<tr>
<th>Mod.</th>
<th>n</th>
<th>Par.</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bias</td>
<td>Var.</td>
<td>MSE</td>
</tr>
<tr>
<td>1</td>
<td>200</td>
<td>$\gamma_1$</td>
<td>-0.007 0.246 0.246</td>
<td>-0.013 0.062 0.062</td>
<td>0.088 0.019 0.057</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_2$</td>
<td>-0.009 0.217 0.217</td>
<td>-0.035 0.116 0.117</td>
<td>-0.043 0.089 0.091</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_3$</td>
<td>-0.050 0.406 0.408</td>
<td>-0.028 0.205 0.206</td>
<td>-0.031 0.152 0.153</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_1$</td>
<td>-0.006 0.015 0.015</td>
<td>-0.005 0.024 0.024</td>
<td>0.015 0.034 0.034</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_2$</td>
<td>0.003 0.036 0.026</td>
<td>-0.002 0.052 0.052</td>
<td>0.002 0.073 0.073</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>$\gamma_1$</td>
<td>0.049 0.248 0.250</td>
<td>0.011 0.114 0.114</td>
<td>0.000 0.244 0.244</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_2$</td>
<td>0.017 0.073 0.074</td>
<td>0.010 0.047 0.047</td>
<td>0.090 0.091 0.099</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_3$</td>
<td>0.017 0.947 1.923</td>
<td>0.035 0.177 0.178</td>
<td>0.084 0.444 0.451</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_1$</td>
<td>-0.021 0.303 0.303</td>
<td>-0.011 0.162 0.162</td>
<td>-0.040 0.459 0.461</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_2$</td>
<td>0.029 0.050 0.051</td>
<td>0.043 0.069 0.071</td>
<td>0.099 0.236 0.245</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_3$</td>
<td>-0.009 0.044 0.044</td>
<td>-0.014 0.052 0.052</td>
<td>-0.034 0.220 0.221</td>
</tr>
<tr>
<td>400</td>
<td></td>
<td>$\gamma_1$</td>
<td>0.029 0.111 0.112</td>
<td>0.010 0.055 0.055</td>
<td>0.017 0.107 0.107</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_2$</td>
<td>0.016 0.040 0.040</td>
<td>0.006 0.022 0.022</td>
<td>0.061 0.046 0.049</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_3$</td>
<td>-0.030 0.155 0.156</td>
<td>-0.017 0.079 0.080</td>
<td>-0.111 0.142 0.154</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_4$</td>
<td>0.010 0.114 0.114</td>
<td>0.015 0.066 0.067</td>
<td>0.030 0.128 0.129</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_1$</td>
<td>0.004 0.007 0.007</td>
<td>-0.001 0.008 0.008</td>
<td>-0.037 0.030 0.032</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_2$</td>
<td>0.010 0.024 0.024</td>
<td>0.018 0.029 0.029</td>
<td>0.063 0.104 0.108</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_3$</td>
<td>0.001 0.021 0.021</td>
<td>-0.005 0.026 0.026</td>
<td>-0.015 0.100 0.100</td>
</tr>
</tbody>
</table>
Table 3
Bias, variance and MSE of $\hat{\gamma}$ and $\hat{\beta}$ for smcure (second rows) and the 2-step approach (first rows) in Models 3-4 for $n = 200$.

<table>
<thead>
<tr>
<th>Mod.</th>
<th>n</th>
<th>Par.</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bias</td>
<td>Var.</td>
<td>MSE</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>$\gamma_1$</td>
<td>-0.065</td>
<td>0.267</td>
<td>0.290</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_2$</td>
<td>0.083</td>
<td>0.041</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_3$</td>
<td>-0.029</td>
<td>0.071</td>
<td>0.072</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_4$</td>
<td>0.200</td>
<td>0.216</td>
<td>0.222</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_1$</td>
<td>0.159</td>
<td>1.233</td>
<td>1.259</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_2$</td>
<td>0.124</td>
<td>0.325</td>
<td>0.340</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_3$</td>
<td>-0.359</td>
<td>4.817</td>
<td>4.946</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_4$</td>
<td>0.003</td>
<td>0.011</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_5$</td>
<td>0.003</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_6$</td>
<td>0.022</td>
<td>0.034</td>
<td>0.034</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_7$</td>
<td>0.010</td>
<td>0.036</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_8$</td>
<td>-0.048</td>
<td>0.044</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_9$</td>
<td>-0.032</td>
<td>0.048</td>
<td>0.049</td>
</tr>
<tr>
<td>4</td>
<td>200</td>
<td>$\gamma_1$</td>
<td>0.086</td>
<td>0.247</td>
<td>0.245</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_2$</td>
<td>0.110</td>
<td>0.273</td>
<td>0.285</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_3$</td>
<td>0.010</td>
<td>0.058</td>
<td>0.058</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_4$</td>
<td>-0.069</td>
<td>0.064</td>
<td>0.069</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_5$</td>
<td>-0.025</td>
<td>0.131</td>
<td>0.132</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_6$</td>
<td>0.000</td>
<td>0.153</td>
<td>0.153</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_7$</td>
<td>-0.025</td>
<td>0.110</td>
<td>0.111</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_8$</td>
<td>-0.026</td>
<td>0.212</td>
<td>0.212</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_9$</td>
<td>0.019</td>
<td>0.245</td>
<td>0.245</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_1$</td>
<td>0.031</td>
<td>0.203</td>
<td>0.204</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_2$</td>
<td>0.031</td>
<td>0.203</td>
<td>0.204</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_3$</td>
<td>0.000</td>
<td>0.033</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_4$</td>
<td>-0.001</td>
<td>0.033</td>
<td>0.033</td>
</tr>
</tbody>
</table>

Fig 1: Boxplots of estimates of the four $\gamma$ parameters in Model 2, scenario 3 for the iterations in which smcure does not converge. The horizontal lines correspond to the true values of the parameters. Left panel: for $n = 200$. Right panel: for $n = 400$. 
Bias, variance and MSE of $\hat{\gamma}$ and $\hat{\beta}$ for single (second rows) and the 2-step approach (first rows) in Models 3-4 for $n = 400$.

<table>
<thead>
<tr>
<th>Mod</th>
<th>n</th>
<th>Par</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bias Var. MSE</td>
<td>Bias Var. MSE</td>
<td>Bias Var. MSE</td>
</tr>
<tr>
<td>3</td>
<td>400</td>
<td>$\gamma_1$</td>
<td>-0.045 0.149 0.151</td>
<td>-0.009 0.069 0.070</td>
<td>-0.020 0.046 0.047</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_2$</td>
<td>0.124 0.213 0.228</td>
<td>0.043 0.071 0.073</td>
<td>-0.003 0.045 0.045</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_3$</td>
<td>0.063 0.021 0.025</td>
<td>0.034 0.016 0.017</td>
<td>0.023 0.015 0.015</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_4$</td>
<td>-0.018 0.030 0.030</td>
<td>-0.008 0.018 0.018</td>
<td>-0.003 0.016 0.016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_5$</td>
<td>-0.041 0.098 0.099</td>
<td>-0.025 0.058 0.059</td>
<td>0.001 0.050 0.050</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_1$</td>
<td>0.052 0.116 0.119</td>
<td>0.019 0.061 0.062</td>
<td>0.026 0.052 0.052</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_2$</td>
<td>-0.088 0.111 0.119</td>
<td>-0.052 0.069 0.071</td>
<td>-0.020 0.059 0.059</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_3$</td>
<td>0.029 0.149 0.149</td>
<td>0.007 0.076 0.076</td>
<td>0.019 0.060 0.060</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_4$</td>
<td>0.103 0.165 0.175</td>
<td>0.047 0.083 0.085</td>
<td>0.033 0.068 0.069</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_5$</td>
<td>-0.079 0.226 0.232</td>
<td>-0.030 0.085 0.086</td>
<td>-0.013 0.066 0.067</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_6$</td>
<td>-0.001 0.005 0.005</td>
<td>0.001 0.007 0.007</td>
<td>0.003 0.010 0.010</td>
</tr>
<tr>
<td>4</td>
<td>400</td>
<td>$\gamma_1$</td>
<td>0.042 0.115 0.116</td>
<td>0.016 0.105 0.105</td>
<td>0.003 0.118 0.118</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_2$</td>
<td>0.056 0.124 0.127</td>
<td>0.036 0.119 0.121</td>
<td>0.007 0.125 0.125</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_3$</td>
<td>0.025 0.027 0.027</td>
<td>0.045 0.036 0.038</td>
<td>0.047 0.036 0.039</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_4$</td>
<td>0.029 0.028 0.029</td>
<td>-0.040 0.032 0.034</td>
<td>-0.038 0.032 0.034</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_5$</td>
<td>0.002 0.064 0.064</td>
<td>-0.022 0.062 0.062</td>
<td>-0.018 0.063 0.063</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_6$</td>
<td>0.024 0.072 0.072</td>
<td>0.009 0.073 0.073</td>
<td>0.012 0.074 0.074</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_1$</td>
<td>0.010 0.045 0.045</td>
<td>0.016 0.054 0.054</td>
<td>0.018 0.061 0.061</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_2$</td>
<td>-0.014 0.049 0.049</td>
<td>-0.019 0.058 0.059</td>
<td>-0.018 0.069 0.070</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_3$</td>
<td>0.000 0.101 0.101</td>
<td>-0.019 0.097 0.097</td>
<td>-0.012 0.107 0.108</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_4$</td>
<td>0.034 0.111 0.112</td>
<td>0.033 0.111 0.112</td>
<td>0.038 0.119 0.120</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_5$</td>
<td>-0.039 0.084 0.086</td>
<td>-0.045 0.081 0.084</td>
<td>-0.041 0.099 0.099</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_6$</td>
<td>-0.002 0.094 0.094</td>
<td>0.015 0.091 0.092</td>
<td>0.018 0.098 0.098</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_7$</td>
<td>0.000 0.004 0.004</td>
<td>0.003 0.009 0.009</td>
<td>-0.002 0.015 0.015</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_8$</td>
<td>0.000 0.004 0.004</td>
<td>0.005 0.009 0.009</td>
<td>0.001 0.015 0.015</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_9$</td>
<td>-0.002 0.009 0.009</td>
<td>-0.007 0.023 0.023</td>
<td>-0.011 0.045 0.045</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_{10}$</td>
<td>0.005 0.016 0.017</td>
<td>0.012 0.039 0.039</td>
<td>0.015 0.072 0.072</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_{11}$</td>
<td>0.004 0.017 0.017</td>
<td>0.006 0.040 0.040</td>
<td>0.006 0.074 0.074</td>
</tr>
</tbody>
</table>
better. In terms of $\beta$ estimators, both approaches give very similar results. In addition, the boxplots in Figure 1, indicate that, even when smcure does not converge, the 2-step approach still gives more reasonable estimates.

Since the second step of the new method does not depend on the latency model, we expect it to be more stable than smcure with respect to misspecifications of the latency model. We investigate this issue by considering two additional settings: one corresponding to a non-Cox latency model (Model 5 below) for which we still apply the two methods as if the Cox model was true and one corresponding to a logistic-Cox model but in which we don’t use the correct covariates. For the latter, we use Model 4, scenario 2 described above but fit a latency model with covariates $X_1, \ldots, X_5$ instead of $Z_1, Z_2, Z_3$. In particular, this means that we are including covariates $X_1, X_2, X_5$ that actually do not have any effect and are excluding $Z_1$ which affects the survival of the uncured.

Model 5. Both incidence and latency depend on three independent covariates $X_1 = Z_1 \sim \text{Unif}(-1,1)$, $X_2 = Z_2 \sim \text{Bernoulli}(0.4)$ and $X_3 = Z_3 \sim \text{Bernoulli}(0.6)$. The cure status and the survival times for the uncured observations are generated as in Model 1 for $\gamma = (1.4, 2.1, -1)$, $\rho$ depending on $z$, $\rho(z) = 0.75 + \exp(\beta^T z)$, $\mu = 1.5$, $\beta = (1, 0.4, -0.6)$. In particular this means that the latency model does not satisfy the proportional hazards assumption. For an observation with covariate $z$, the event time is truncated at $\tau_0(z)$ equal to the 97% quantile of the Weibull distribution with parameters $\rho(z)$ and $\mu^{-1/\rho(z)}$. The censoring times are generated according to a Weibull proportional hazards model

$$S_C(t|x) = \exp \left( -\frac{1}{22} \mu t \tilde{\rho} \exp \left( \gamma^T x \right) \right),$$

with $\mu = 1.5$ and $\tilde{\rho} = 2.5$, truncated at $\tau = \max_z \tau_0(z) + 2$. This scenario corresponds to a cure rate of 30% and a censoring rate of 45%.

Results for sample size 200 and 400, reported in Table 5, show that when the true latency model is not a Cox proportional hazards model, even $\gamma$ estimates are biased. However, the 2-step approach has lower bias and MSE, hence suffers less from the misspecification of the latency. On the other hand, misspecification of the latency covariates when the model is still Cox, seems to be less critical. It leads to a slight increase in bias and variance compared to the results in Table 3-4 but again the 2-step approach performs better.

6. Application

In this section we illustrate the practical use of the method through two medical datasets for melanoma cancer patients and compare the results with those provided by the smcure package. Melanoma is a common skin cancer type for which nowadays it is expected that a considerable fraction of the patients get cured as a consequence of medical advances in diagnostics and treatment. Therefore, it is important to account for the presence of cured patients in the statistical analysis of melanoma survival data and to evaluate new treatments focusing on cure and not only survival prolongation.
Table 5
Bias, variance and MSE of $\hat{\gamma}$ and $\hat{\beta}$ for the cure (second rows) and the 2-step approach (first rows) in Model 5 and Model 4 with misspecification.

<table>
<thead>
<tr>
<th>Mod.</th>
<th>Par.</th>
<th>Bias</th>
<th>Var.</th>
<th>MSE</th>
<th>Bias</th>
<th>Var.</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>$\gamma_1$</td>
<td>-0.154</td>
<td>0.187</td>
<td>0.211</td>
<td>-0.087</td>
<td>0.100</td>
<td>0.107</td>
</tr>
<tr>
<td></td>
<td>$\gamma_2$</td>
<td>0.514</td>
<td>0.303</td>
<td>0.457</td>
<td>0.437</td>
<td>0.152</td>
<td>0.343</td>
</tr>
<tr>
<td></td>
<td>$\gamma_3$</td>
<td>-0.304</td>
<td>0.277</td>
<td>0.370</td>
<td>0.213</td>
<td>0.144</td>
<td>0.189</td>
</tr>
<tr>
<td></td>
<td>$\gamma_4$</td>
<td>0.623</td>
<td>0.464</td>
<td>0.652</td>
<td>0.507</td>
<td>0.184</td>
<td>0.441</td>
</tr>
<tr>
<td></td>
<td>$\gamma_5$</td>
<td>-0.232</td>
<td>0.203</td>
<td>0.257</td>
<td>-0.166</td>
<td>0.108</td>
<td>0.136</td>
</tr>
<tr>
<td></td>
<td>$\gamma_6$</td>
<td>0.297</td>
<td>0.655</td>
<td>0.744</td>
<td>0.255</td>
<td>0.176</td>
<td>0.231</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>0.207</td>
<td>0.188</td>
<td>0.231</td>
<td>0.132</td>
<td>0.112</td>
<td>0.130</td>
</tr>
<tr>
<td></td>
<td>$\beta_2$</td>
<td>-0.326</td>
<td>0.377</td>
<td>0.483</td>
<td>-0.281</td>
<td>0.166</td>
<td>0.245</td>
</tr>
<tr>
<td></td>
<td>$\beta_3$</td>
<td>-0.281</td>
<td>0.051</td>
<td>0.130</td>
<td>-0.296</td>
<td>0.025</td>
<td>0.112</td>
</tr>
<tr>
<td></td>
<td>$\beta_4$</td>
<td>-0.354</td>
<td>0.056</td>
<td>0.181</td>
<td>-0.353</td>
<td>0.027</td>
<td>0.151</td>
</tr>
<tr>
<td></td>
<td>$\beta_5$</td>
<td>-0.137</td>
<td>0.043</td>
<td>0.062</td>
<td>-0.144</td>
<td>0.022</td>
<td>0.043</td>
</tr>
<tr>
<td></td>
<td>$\beta_6$</td>
<td>-0.178</td>
<td>0.046</td>
<td>0.076</td>
<td>-0.169</td>
<td>0.023</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>$\beta_7$</td>
<td>0.221</td>
<td>0.051</td>
<td>0.099</td>
<td>0.210</td>
<td>0.023</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td>$\beta_8$</td>
<td>0.257</td>
<td>0.054</td>
<td>0.120</td>
<td>0.235</td>
<td>0.024</td>
<td>0.079</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n = 200</th>
<th>n = 400</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_1$</td>
<td>0.036</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>0.097</td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>-0.089</td>
</tr>
<tr>
<td>$\gamma_4$</td>
<td>0.029</td>
</tr>
<tr>
<td>$\gamma_5$</td>
<td>0.025</td>
</tr>
<tr>
<td>$\gamma_6$</td>
<td>0.005</td>
</tr>
<tr>
<td>$\gamma_7$</td>
<td>-0.059</td>
</tr>
<tr>
<td>$\gamma_8$</td>
<td>-0.057</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.022</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.029</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.075</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>-0.025</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-0.084</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>0.017</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>0.010</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>0.004</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>0.011</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>0.013</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>0.004</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>-0.001</td>
</tr>
<tr>
<td>$\beta_{13}$</td>
<td>-0.019</td>
</tr>
</tbody>
</table>
6.1. Eastern Cooperative Oncology Group (ECOG) Data

The ECOG phase III clinical trial e1684 aimed at evaluating the effect of treatment (high dose interferon alpha-2b regimen) as the postoperative adjuvant therapy for melanoma patients. The corresponding dataset, consisting of 284 observations (after deleting missing data), is available in the smcure package [5]. The event time is the time from initial treatment to recurrence of melanoma and three covariates have been considered: age (continuous variable centered to the mean), gender (0=male and 1=female) and treatment (0=control and 1=treatment). Around 30% of the observations are censored. The Kaplan-Meier curve is shown in Figure 2.

We fit a logistic-Cox mixture cure model by using the maximum likelihood principle (smcure package) and the proposed 2-step approach. For our method we use the smcure estimator as a preliminary estimator. In both cases, standard errors are computed through 500 naive bootstrap samples. The resulting parameter estimates, standard errors and corresponding p-values for the Wald test are reported in Table 6.

We observe that, despite exhibiting the same effect directions for all covariates, the two approaches give quite different results in terms of treatment effect. Age and treatment are both found to have a significant effect on the cure frac-
Fig 3: Boxplot of the difference between the PE of the new method and the PE of \texttt{smcure}, over these 1000 iterations for the ECOG data.

Fig 3: Boxplot of the difference between the PE of the new method and the PE of \texttt{smcure}, over these 1000 iterations for the ECOG data.

Fig 3: Boxplot of the difference between the PE of the new method and the PE of \texttt{smcure}, over these 1000 iterations for the ECOG data.

The prediction error (PE) is calculated as follows:

\[
PE = -\sum_{j \in \text{test set}} \log \left[ \phi \left( \hat{\gamma}_n^T X_j \right) \hat{w}_j \left( 1 - \phi \left( \hat{\gamma}_n^T X_j \right) \right) \right],
\]

where \( \hat{\gamma}_n \) are the parameter estimates from the training set and \( \hat{w}_j \) are the predicted uncure probabilities given the observations, i.e.

\[
\hat{w}_j = \Delta_j + (1 - \Delta_j) \frac{\phi(\hat{\gamma}_n^T X_j) \hat{S}_u(Y_j | Z_j)}{1 - \phi(\hat{\gamma}_n^T X_j) + \phi(\hat{\gamma}_n^T X_j) \hat{S}_u(Y_j | Z_j)},
\]

We repeat this procedure 1000 times, for random selection of the train and test set. The boxplot of the difference between the PE of the new method and the PE of \texttt{smcure}, over these 1000 iteration, is given in Figure 3. We observe that the 2-step approach leads to lower PE (negative difference) in more than 50% of the cases and the improvement in PE for the new method is usually larger compared to the cases in which \texttt{smcure} does better.

In addition, we expect the new approach to be more stable with respect to the latency model since that does not influence the second step of the estimation. To illustrate this point, we also fit a cure model with only gender as covariate for the survival of uncured patients (see Table 7) and see that in that case, \texttt{smcure} also detects the effect of the treatment to be significant.

### 6.2. Surveillance, Epidemiology and End Results database

Here we consider melanoma data extracted from the SEER database to illustrate the performance of the method for more than one continuous covariate. The SEER database collects cancer incidence data from population-based cancer
Table 7
Results for the incidence (logistic component) and the latency (Cox PH component) from the ECOG data.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>smcure package Estimates</th>
<th>SE</th>
<th>p-value</th>
<th>2-step approach Estimates</th>
<th>SE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>1.4000</td>
<td>0.2791</td>
<td>5·10^{-4}</td>
<td>1.9073</td>
<td>0.5225</td>
<td>0.0002</td>
</tr>
<tr>
<td>Age</td>
<td>0.0165</td>
<td>0.0121</td>
<td>0.1709</td>
<td>0.0357</td>
<td>0.0174</td>
<td>0.0399</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.0538</td>
<td>0.3101</td>
<td>0.8623</td>
<td>-0.0429</td>
<td>0.3979</td>
<td>0.9141</td>
</tr>
<tr>
<td>Treatment</td>
<td>-0.6765</td>
<td>0.3118</td>
<td>0.0300</td>
<td>1.2590</td>
<td>0.5809</td>
<td>0.0302</td>
</tr>
</tbody>
</table>

| Latency    |                           |    |         |                           |    |         |
| Gender     | 0.0637                    | 0.1935 | 0.7421  | 0.0235                   | 0.1924 | 0.9028 |

Fig 4: Kaplan-Meier survival curve for the SEER data.

registries in US. We select the database ‘Incidence - SEER 18 Regs Research Data’ and, in order to have a reasonable sample size, we extract the melanoma cancer data for the county of San Francisco in California during the period 2005 – 2010. We consider only patients with known follow-up time and tumor size (in the range 1-90 mm) and restrict the study to white people because of the very small number of cases from other races. The event of interest is death because of melanoma. This cohort consists of 384 observations out of which 228 are male. The age ranges from 23 to 101 years old, the follow-up from 1 to 143 months with no events observed after 108 months. Because of the high expected cure rate, 89% of the observations are censored. We consider as covariates in the model: gender (0=male, 1=female), age and tumor size (continuous). The use of cure models is justified from the presence of a long plateau containing around 25% of the observations (see the Kaplan-Meier curve in Figure 4).

As in the previous section, we compute parameter estimates, standard errors and corresponding p-values for both methods (see Table 8). We observe that both methods agree on the directions of the effects and give similar parameter estimates. Note that the sample size in this case is larger than in the previous data example. However, smcure only finds age to be significant while the 2-step approach also detects the tumor size. For the latency, none of the covariates is found significant with both methods. Also in this case, most of the time the new method leads to an improvement in terms of prediction errors, computed according to the procedure described in the previous subsection (see Figure 5). In addition, we also observe that if we remove the covariate tumor
Table 8
Results for the incidence (logistic component) and the latency (Cox PH component) from the SEER data.

<table>
<thead>
<tr>
<th></th>
<th>smcure package</th>
<th></th>
<th></th>
<th>2-step approach</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Covariates</td>
<td>Estimates</td>
<td>SE</td>
<td>p-value</td>
<td>Estimates</td>
<td>SE</td>
</tr>
<tr>
<td>incidence</td>
<td>Intercept</td>
<td>−4.4952</td>
<td>0.9670</td>
<td>3 \times 10^{-6}</td>
<td>−4.9634</td>
<td>1.1269</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>0.0411</td>
<td>0.0142</td>
<td>0.0037</td>
<td>0.0455</td>
<td>0.0167</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>−0.6805</td>
<td>0.4551</td>
<td>0.1349</td>
<td>−0.7902</td>
<td>0.5162</td>
</tr>
<tr>
<td></td>
<td>Tumor size</td>
<td>0.0214</td>
<td>0.0136</td>
<td>0.1147</td>
<td>0.0295</td>
<td>0.0131</td>
</tr>
<tr>
<td>latency</td>
<td>Age</td>
<td>−0.0161</td>
<td>0.0152</td>
<td>0.2874</td>
<td>−0.0171</td>
<td>0.0151</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>−0.3599</td>
<td>0.4217</td>
<td>0.3934</td>
<td>−0.3125</td>
<td>0.4184</td>
</tr>
<tr>
<td></td>
<td>Tumor size</td>
<td>0.0308</td>
<td>0.0224</td>
<td>0.1695</td>
<td>0.0282</td>
<td>0.0217</td>
</tr>
</tbody>
</table>

Fig 5: Boxplot of the difference between the PE of the new method and the PE of smcure, over these 1000 iterations for the SEER data.

From the latency model, the 2-step approach gives similar results while this time smcure also detects tumor size as significant for the incidence component (see Table 9). Once more, this behavior reflects the strong dependence of the incidence estimates on the latency model for smcure.

Table 9
Results for the incidence (logistic component) and the latency (Cox PH component) from the SEER data.

<table>
<thead>
<tr>
<th></th>
<th>smcure package</th>
<th></th>
<th></th>
<th>2-step approach</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Covariates</td>
<td>Estimates</td>
<td>SE</td>
<td>p-value</td>
<td>Estimates</td>
<td>SE</td>
</tr>
<tr>
<td>incidence</td>
<td>Intercept</td>
<td>−4.7297</td>
<td>0.9211</td>
<td>2 \times 10^{-5}</td>
<td>−4.5477</td>
<td>1.0829</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>0.0423</td>
<td>0.0137</td>
<td>0.0029</td>
<td>0.0392</td>
<td>0.0159</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>−0.6393</td>
<td>0.4579</td>
<td>0.1627</td>
<td>−0.7021</td>
<td>0.4430</td>
</tr>
<tr>
<td></td>
<td>Tumor size</td>
<td>0.0288</td>
<td>0.0129</td>
<td>0.0251</td>
<td>0.0250</td>
<td>0.0129</td>
</tr>
<tr>
<td>latency</td>
<td>Age</td>
<td>−0.0165</td>
<td>0.0147</td>
<td>0.2626</td>
<td>−0.0145</td>
<td>0.0146</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>−0.3208</td>
<td>0.4282</td>
<td>0.4538</td>
<td>−0.2559</td>
<td>0.4092</td>
</tr>
</tbody>
</table>
Appendix

Appendix B: Proofs

Proof of Theorem 1. We start by showing that, when $\hat{\gamma}_n - \gamma_0 = o_P(1)$, we have

$$\sup_{x \in \mathcal{X}} |\hat{\pi}(x) - \pi_0(x)| = o_P(1).$$

(14)

Note that, independently of the choice of the trimming function,

$$E \left[ \left\{ (1 - \pi_0(X)) \log \phi(\gamma^T X) + \pi_0(X) \log (1 - \phi(\gamma^T X)) \right\} \tau(X) \right],$$

is maximized at $\gamma = \gamma_0$ because of condition (I1) and the fact that, for any $x \in \mathcal{X}$, the function

$$g_z(x) = \left\{ \phi(\gamma_0^T x) \log \frac{z}{\phi(\gamma_0^T x)} + \left( 1 - \phi(\gamma_0^T x) \right) \log \frac{1 - z}{1 - \phi(\gamma_0^T x)} \right\} \tau(x), \quad z \in (0, 1),$$

is strictly negative for $z \neq \phi(\gamma_0^T x)$ and $g_z(\phi(\gamma_0^T x)) = 0$.

Next, we show uniform consistency of $\hat{\pi}_n$ using the decomposition (11) and restricting to $\mathcal{X}$ as in (12). We have

$$||\hat{\pi}_n - \pi_0||_\infty = \sup_{x \in \mathcal{X}} |\hat{g}_{\hat{\gamma}_n}(\gamma_n^T x) - g_{\gamma_0}(\gamma_0^T x)|$$

$$= \sup_{x \in \mathcal{X}} |\hat{g}_{\gamma_0}(\gamma_0^T x) - g_{\gamma_0}(\gamma_0^T x)| + \sup_{x \in \mathcal{X}} |\hat{g}_{\hat{\gamma}_n}(\gamma_n^T x) - \hat{g}_{\gamma_0}(\gamma_0^T x)|$$

$$\leq \sup_{x \in \mathcal{X}} \sup_{t \leq \tau_0} ||\hat{F}_n(t \mid \gamma_0^T X = \gamma_0^T x) - F_t(t \mid \gamma_0^T X = \gamma_0^T x)|| + \|\gamma_n^T - \gamma_0\| \sup_{x \in \mathcal{X}} \gamma \in G \sup_{x \in \mathcal{X}} |\nabla_{\gamma} \hat{g}_{\gamma}(\gamma^T x)|.$$

(15)

The first term on the right-hand side of the equation converges to zero by Theorem 4.1 in [38]. The second term converges to zero because of assumption (C1) and the fact that

$$\sup_{x \in \mathcal{X}} \sup_{\gamma \in G} |\nabla_{\gamma} \hat{g}_{\gamma}(\gamma^T x)| = O_P(1).$$

(16)

Indeed, (16) follows from assumption (C9) and

$$\sup_{x \in \mathcal{X}} \sup_{\gamma \in G} |\nabla_{\gamma} \hat{g}_{\gamma}(\gamma^T x) - \nabla_{\gamma} g_{\gamma}(\gamma^T x)| = o_P(1),$$

(17)

which can be proved as in Lemma A.2 in [17]. Their estimator $\hat{G}_n(t \mid \lambda(\theta, x))$ is the same as our $\hat{F}_n(t \mid \gamma^T X = \gamma^T x)$ if we replace $\theta$ by $\gamma$, consider $\lambda(\theta, x) = \theta^T x$ and exchange $T$ with $C$ (they are interested in the conditional distribution of $C$ given $\lambda(\theta, X)$). This concludes the proof of (14).

The consistency of $\hat{\gamma}_n$ then follows similarly to Theorem 1 in [24]. There it is required that $\sup_{x \in \mathcal{X}} |\hat{\pi}(x) - \pi_0(x)| \to 0$ almost surely in order to obtain strong consistency of $\hat{\gamma}_n$. Here we restrict to convergence in probability and conclude via Theorem 1 in [6] formulated for M-estimators.

\[\square\]
Proof of Theorem 2. The result follows from Theorem 2 in [24] once we show that the assumptions (AN1)-(AN4) of that paper hold. Note that the introduction of the trimming function \( \tau(\cdot) \) would not change anything in the proof. It just allows us to restrict ourselves to the set \( \tilde{X} \) where the density of the index is bounded from below, in order to apply standard results from the literature. The assumptions (AN1) and (AN3) of [24] are the same as assumptions (C2), (N1) and (N2) here. It remains to verify assumptions (AN2) and (AN4) which for completeness we state below:

(AN2) \( \pi_0(\cdot) \) belongs to a class of functions \( \Pi \) such that

\[
\int_0^\infty \sqrt{\log N(\epsilon, \Pi, \| \cdot \|_\infty)} \, d\epsilon < \infty,
\]

where \( N(\epsilon, \Pi, \| \cdot \|_\infty) \) denotes the \( \epsilon \)-covering number of the space \( \Pi \) with respect to \( \| \pi \|_\infty = \sup_{x \in \tilde{X}} |\pi(x)| \).

(AN4) The estimator \( \hat{\pi}(\cdot) \) satisfies the following properties:

(i) \( \mathbb{P}(\hat{\pi}(\cdot) \in \Pi) \to 1 \).

(ii) \( \| \hat{\pi}(x) - \pi_0(x) \|_\infty = o_P(n^{-1/4}) \).

(iii) There exists a function \( \Psi \) such that

\[
\mathbb{E}^* \left[ \left( \hat{\pi}(X) - \pi_0(X) \right) \left( \frac{1}{\phi(\gamma_0^T X)} + \frac{1}{1 - \phi(\gamma_0^T X)} \right) \phi'(\gamma_0^T X) X \tau(X) \right] = \frac{1}{n} \sum_{i=1}^n \Psi(Y_i, \Delta_i, X_i, Z_i) + R_n,
\]

where \( \mathbb{E}^* \) denotes the conditional expectation given the sample, taken with respect to the generic variable \( X \), \( \mathbb{E}[\Psi(Y, \Delta, X, Z)] = 0 \) and \( \| R_n \| = o_P(n^{-1/2}) \).

Step 1. For (AN2) we use the class of functions

\[
\Pi = \left\{ f : \mathcal{X} \to [0, 1], f(x) = g(\gamma^T x) \text{ for some } \gamma \in G \text{ and } g \in \tilde{\Pi}_g \right\},
\]

where \( \tilde{\Pi}_g \) is the space of continuously differentiable functions \( g \) from \( \mathcal{X} \) to \( [0, 1] \) such that \( \sup_{u \in \mathcal{X}_\gamma} |g'(u)| \leq M \) and

\[
\sup_{u_1 \neq u_2 \in \mathcal{X}_\gamma} \frac{|g'(u_1) - g'(u_2)|}{|u_1 - u_2|^\xi} \leq M,
\]

for some \( M > 0 \) and \( \xi \in (0, 1] \) independent of \( \gamma \). The norm that we consider on \( \Pi \) is the sup norm. By assumption (A2) we have \( \pi_0 \in \Pi \). Next we compute the \( \epsilon \)-covering number of \( \Pi \). Let \( \gamma_1, \ldots, \gamma_{N_1} \) be a \( \delta \)-covering of the compact \( G \in \mathbb{R}^p \) with respect to the \( l_1 \) norm. We have \( N_1 \leq \frac{K_1}{\delta^p} \) for some constant
Consider the class $\hat{\Pi}$ of continuously differentiable functions $g$ from $\mathcal{U} = \{\gamma^T x \mid \gamma \in \mathcal{G}, x \in \mathcal{X}\}$ to $[0, 1]$ such that $\sup_{u \in \mathcal{U}} |g'(u)| \leq M$ and
\[
\sup_{u_1 \neq u_2} \frac{|g'(u_1) - g'(u_2)|}{|u_1 - u_2|^\xi} \leq M.
\]

We have in particular that if $\rho$ can be extended to a function $\bar{\rho}$ on $\tilde{\Pi}$, then the restriction of $\tilde{\Pi}$ to any $\gamma \in \mathcal{G}$. Let $g_1, \ldots, g_{N_2}$ be a $\rho$-covering of $\Pi$ with respect to the sup norm. Since $\mathcal{U}$ is bounded and convex ($\mathcal{G}$ can be chosen to be a small neighborhood of $\gamma_0$), from Theorem 2.7.1 in [36], we have
\[
\log N_2 \leq K_2 \left(\frac{1}{\rho}\right)^{1/(1+\xi)},
\]
for some constant $K_2 > 0$. For assumption (AN4)(i), note that $\hat{\gamma}_n = \hat{g}_n(\gamma_n^T x)$ with $\hat{g}_n \gamma$ as in (9). From consistency of $\hat{\gamma}_n$, it follows that $\hat{\gamma}_n \in \mathcal{G}$. Let $\hat{g}_n'$ denote the derivative of the function $u \mapsto \hat{g}_n(u)$. Let $g_n$, be as in (10). Next we show that
\[
\sup_{x \in \mathcal{X}} |\hat{g}_n'(\gamma_n^T x) - g'_\gamma(\gamma_0^T x)| = o_P(1),
\]
and
\[
\sup_{\gamma_n^T x_1 \neq \gamma_n^T x_2} \frac{|\hat{g}_n'(\gamma_n^T x_1) - g'_\gamma(\gamma_0^T x_1) - \hat{g}_n'(\gamma_n^T x_2) + g'_\gamma(\gamma_0^T x_2)|}{|\gamma_n^T x_1 - \gamma_n^T x_2|^\delta} = o_P(1),
\]
from which we can derive that (AN4)(i) is satisfied since \( \sup_{x \in \mathcal{X}} |g'_{\gamma_0} (\gamma_0^T x)| < \infty \) and
\[
\sup_{x_1, x_2 \in \mathcal{X}, \gamma_0^T x_1 \neq \gamma_0^T x_2} \left| \frac{g'_{\gamma_0} (\gamma_0^T x_2) - g'_{\gamma_0} (\gamma_0^T x_1)}{\gamma_0^T x_1 - \gamma_0^T x_2} \right| < \infty.
\]
From Theorems 4.1 and 4.2 in [38] we have
\[
\sup_{x \in \mathcal{X}} |\hat{g}'_{\gamma_0} (\gamma_0^T x) - g'_{\gamma_0} (\gamma_0^T x)| = o_P(1), \tag{20}
\]
and
\[
\sup_{x_1, x_2 \in \mathcal{X}, \gamma_0^T x_1 \neq \gamma_0^T x_2} \left| \frac{\hat{g}'_{\gamma_0} (\gamma_0^T x_1) - \hat{g}'_{\gamma_0} (\gamma_0^T x_2)}{\gamma_0^T x_1 - \gamma_0^T x_2} \right| = o_P(1), \tag{21}
\]
by conditioning on the variable \( \gamma_0^T X \). Next, we need to deal with the fact that we are using \( \tilde{\gamma}_0^T X \) instead of \( \gamma_0^T X \).

We can write
\[
\log \hat{g}_{\gamma_0} (u) = \sum_{i=1}^{n} \Delta_i \log (1 - W_i(u; \tilde{\gamma}_n)),
\]
where
\[
W_i(u; \tilde{\gamma}_n) = k \left( \tilde{\gamma}_n^T X_i - u \right) / \left( \sum_{j=1}^{n} k \left( \tilde{\gamma}_n^T X_j - u \right) \right).
\]
Hence
\[
\hat{g}'_i(u) = \hat{g}_{\tilde{\gamma}_n}(u) \sum_{i=1}^{n} \Delta_i \frac{W'_i(u; \tilde{\gamma}_n)}{1 - W_i(u; \tilde{\gamma}_n)}. \tag{22}
\]
Using the triangular inequality we obtain
\[
\sup_{x \in \mathcal{X}} |\hat{g}'_{\gamma_0} (\gamma_0^T x) - \hat{g}'_{\gamma_0} (\gamma_0^T x)|
\leq \sup_{x \in \mathcal{X}} |\hat{g}_{\gamma_0} (\gamma_0^T x) - \hat{g}_0 (\gamma_0^T x)| \sup_{x \in \mathcal{X}} \left| \sum_{i=1}^{n} \Delta_i \frac{W'_i(\gamma_0^T x; \tilde{\gamma}_n)}{1 - W_i(\gamma_0^T x; \tilde{\gamma}_n)} \right|
\tag{23}
+ \sup_{x \in \mathcal{X}} |\hat{g}'_{\gamma_0} (\gamma_0^T x)| \sup_{x \in \mathcal{X}} \left| \sum_{i=1}^{n} \Delta_i \left\{ \frac{W'_i(\gamma_0^T x; \tilde{\gamma}_n)}{1 - W_i(\gamma_0^T x; \tilde{\gamma}_n)} - \frac{W'_i(\gamma_0^T x; \gamma_0)}{1 - W_i(\gamma_0^T x; \gamma_0)} \right\} \right|.
\]
As is (15), we have
\[
\sup_{x \in \mathcal{X}} |\hat{g}_{\gamma_0} (\gamma_0^T x) - \hat{g}_0 (\gamma_0^T x)| = O_P(n^{-1/2}).
\]
Moreover, \( \sup_{x \in X} |g_0'(\gamma_0^T x)| = O_P(1) \) and, as in the proof of Proposition 4.1 in [38], it can be seen that

\[
\sup_{x \in X} \left| \sum_{i=1}^{n} \Delta_i \frac{W_i'((\gamma_0^T x; \gamma_0))}{1 - W_i((\gamma_0^T x; \gamma_0))} \right| = O_P(1),
\]

and

\[
\sup_{x \in X} \left| \sum_{i=1}^{n} \Delta_i \nabla_\gamma \left( \frac{W_i'((\gamma^T x; \gamma))}{1 - W_i((\gamma^T x; \gamma))} \right) \right| = O_P(1).
\]

From a Taylor expansion and \( \hat{\gamma}_n - \gamma_0 = O(n^{-1/2}) \), we conclude that the left hand side of (23) converges to zero. Together with (20) we obtain (18).

Next we show that

\[
\sup_{\tilde{x}_n \neq x_2} \frac{|\hat{g}'_n (\tilde{\gamma}_n^T x_1) - \hat{g}'_n (\gamma_0^T x_1) - \hat{g}'_n (\tilde{\gamma}_n^T x_2) + \hat{g}'_n (\gamma_0^T x_2)|}{|\tilde{\gamma}_n^T x_1 - \gamma_0^T x_2|^2} = o_p(1).
\]

Note that by (22) and the fact that \( \sup_{x \in X} |g_\gamma'(\gamma^T x)| = O_P(1) \) for both \( \gamma = \gamma_0 \) and \( \gamma = \hat{\gamma}_n \), it is sufficient to consider

\[
\sup_{\tilde{x}_n \neq x_2} \frac{\left| V_{\gamma_n} (\tilde{\gamma}_n^T x_1) - V_{\gamma_0} (\gamma_0^T x_1) - V_{\gamma_n} (\tilde{\gamma}_n^T x_2) + V_{\gamma_0} (\gamma_0^T x_2) \right|}{|\tilde{\gamma}_n^T x_1 - \gamma_0^T x_2|^2}, \tag{24}
\]

where

\[
V_\gamma(u) = \sum_{i=1}^{n} \Delta_i \frac{W_i'(u; \gamma)}{1 - W_i(u; \gamma)}.
\]

In addition we can also restrict the supremum over \( x_1, x_2 \) such that \( \tilde{\gamma}_n^T x_1 \neq \gamma_0^T x_2 \) and \( \gamma_0^T x_1 \neq \gamma_0^T x_2 \) because otherwise we would have \( |\tilde{\gamma}_n^T x_1 - \gamma_0^T x_2| \leq cn^{-1/2} \) and (24) would obviously hold for \( \xi \) sufficiently small by using a Taylor expansion and \( \sup u V_n'(u) = O_P(1/b) \). Note also that

\[
\sup_{\tilde{x}_n \neq x_2} \frac{|\gamma_0^T x_1 - \gamma_0^T x_2|}{|\tilde{\gamma}_n^T x_1 - \gamma_0^T x_2|^2} = O_P(1),
\]

hence the denominator can be replaced by \( |\gamma_0^T x_1 - \gamma_0^T x_2|^2 \). By Taylor expansion we obtain that the largest order term in (24) is

\[
(\hat{\gamma}_n - \gamma_0) \sup_{\tilde{x}_n \neq x_2} \left| \sum_{i=1}^{n} \Delta_i \nabla_\gamma \left( \frac{W_i'((\gamma^T x_1; \gamma))}{1 - W_i((\gamma^T x_1; \gamma))} \right) \right| \frac{|\gamma_0^T x_1 - \gamma_0^T x_2|^2}{|\tilde{\gamma}_n^T x_1 - \gamma_0^T x_2|^2}.
\]
Using Lemma A.1. in [38] and that for a smooth function \( l \) (in our case \( l \) is the kernel function or its derivatives) we have
\[
|l\left(\frac{x-z}{b}\right) - l\left(\frac{y-z}{b}\right)|/|x-y|^\xi = O(b^{-\xi}),
\]
uniformly over \( x, y, z \), we obtain that the expression in the previous equation is of the order \( O_P(n^{-1/2}b^{-1-\xi}) = o_P(1) \). Together with (21) this yields (19).

**Step 3.** Assumption (AN4)(ii) can be checked using (15). The first term on the right-hand side of that equation is of order \( o_P(n^{-1/4}) \) because of Theorem 4.1 in [38] and the assumption (C6). The second term on the right-hand side of (15) is also of order \( o_P(n^{-1/4}) \) because of assumptions (N3), (C9) and (17).

**Step 4.** For (AN4)(iii), using the decomposition (11) we have
\[
E^* \left[ \left( \hat{\pi}(X) - \pi_0(X) \right) \left( \frac{1}{\phi(\gamma_0^T X)} + \frac{1}{1 - \phi(\gamma_0^T X)} \right) \phi'(\gamma_0^T X) X \tau(X) \right] = E^* \left[ \frac{\hat{F}_n(\tau_0 | \gamma_0^T X) - F_T(\tau_0 | \gamma_0^T X)}{\phi(\gamma_0^T X) \{1 - \phi(\gamma_0^T X)\}} \phi'(\gamma_0^T X) X \tau(X) \right] + E^* \left[ \frac{\hat{g}_{\gamma_n}(\gamma_0^T X) - \hat{g}_{\gamma_0}(\gamma_0^T X)}{\phi(\gamma_0^T X) \{1 - \phi(\gamma_0^T X)\}} \phi'(\gamma_0^T X) X \tau(X) \right].
\]

Since \( \sup_{x\in\mathcal{X}} H(|t, \infty)|\gamma_0^T x| < 1 \) by condition (8), from Theorem 3.2 in [10], it follows as in [24] (see proof of Theorem 5) that the first term on the right-hand side of (25) is equal to \( \frac{1}{n} \sum_{i=1}^n \psi(Y_i, \Delta_i, X_i) + R_n \) with \( ||R_n|| = o_P(n^{-1/2}) \) and
\[
\psi(Y, \Delta, X) = \begin{cases} \Delta \mathbb{1}_{\{Y \leq \tau_0\}} H(|Y, \infty)|\gamma_0^T X| & - \int_0^{Y \wedge \tau_0} \frac{H_1(ds|\gamma_0^T X)}{H^2([s, \infty)|\gamma_0^T X)} \phi'\left(\gamma_0^T X\right) d\tau(X). 
\end{cases}
\]  

For the second term in (25), by the mean value theorem, we write
\[
\hat{g}_{\gamma_n}(\gamma_0^T x) - \hat{g}_{\gamma_0}(\gamma_0^T x) = (\hat{\gamma}_n - \gamma_0)^T \nabla_{\gamma} \hat{g}_{\gamma}(\gamma T x) |_{\gamma^*} = (\hat{\gamma}_n - \gamma_0)^T \nabla_{\gamma} \hat{g}_{\gamma}(\gamma T x) |_{\gamma_n},
\]
for some \( \gamma^* \) such that \( ||\gamma^* - \gamma_0|| \leq ||\hat{\gamma}_n - \gamma_0|| \). Using (17) and assumption (C9), we obtain
\[
\hat{g}_{\gamma_n}(\gamma_0^T x) - \hat{g}_{\gamma_0}(\gamma_0^T x) = (\hat{\gamma}_n - \gamma_0)^T \nabla_{\gamma} g_{\gamma}(\gamma T x) |_{\gamma_0} + o_P(n^{-1/2})
\]
where the \( o_P(n^{-1/2}) \) term is uniform with respect to \( x \in \mathcal{X} \). It follows that
\[
E^* \left[ \frac{\hat{g}_{\gamma_n}(\gamma_0^T X) - \hat{g}_{\gamma_0}(\gamma_0^T X)}{\phi(\gamma_0^T X) \{1 - \phi(\gamma_0^T X)\}} \phi'(\gamma_0^T X) X \tau(X) \right] = E \left[ \nabla_{\gamma} g_{\gamma}(\gamma^T X) |_{\gamma_0} \phi'(\gamma_0^T X) X \tau(X) \right] (\hat{\gamma}_n - \gamma_0) + o_P(n^{-1/2}),
\]
where also follows that the covariance matrix is given by
\[
\Sigma_n = (\Gamma_1 \Gamma_1')^{-1}(\Gamma_1' \Gamma_1)^{-1} = \Gamma_1^{-1}V \Gamma_1^{-1},
\]
(27)
where \(V = \text{Var}(\Psi(Y, \Delta, X, Z))\) and
\[
\Gamma_1 = -\mathbb{E} \left[ \left( \frac{1}{\phi(\gamma_0 X)} + \frac{1}{1 - \phi(\gamma_0 X)} \right) \phi'(\gamma_0 X)^2 X^T \tau(X) \right].
\]
Acknowledgments

I. Van Keilegom acknowledges financial support from the European Research Council (2016-2021, Horizon 2020 and grant agreement 694409). V. Patilea gratefully acknowledges support from the Joint Research Initiative ‘Models and mathematical processing of very large data’ under the aegis of Risk Foundation, in partnership with MEDIAMETRIE and GENES, France. For the simulations we used the Lisa cluster of the Dutch national Supercomputer. The data that were analysed in Section 5 have been provided by the SEER programme (www.seer.cancer.gov) research data (1973–2014), National Cancer Institute, Division of Cancer Control and Population Sciences, Surveillance Research Program.

References