A presmoothing approach for estimation in the semiparametric Cox mixture cure model

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

DOI
10.3150/21-BEJ1434

Publication date
2022

Document Version
Final published version

Published in
Bernoulli

Bernoulli **28**(4), 2022, 2689–2715
https://doi.org/10.3150/21-BEJ1434

A presmoothing approach for estimation in the semiparametric Cox mixture cure model

ENI MUSTA[1,2,b], VALENTIN PATILEA[3,c] and INGRID VAN KEILEGOM[1,a]

[1]*ORSTAT, KU Leuven, Belgium.* [a]*ingrid.vankeilegom@kuleuven.be*
[2]*Korteweg de Vries Institute for Mathematics, University of Amsterdam, Netherlands.* [b]*e.musta@uva.nl*
[3]*CREST, Ensai, France.* [c]*valentin.patilea@ensai.fr*

A challenge when dealing with survival analysis data is accounting for a cure fraction, meaning that some subjects will never experience the event of interest. Mixture cure models have been frequently used to estimate both the probability of being cured and the time to event for the susceptible subjects, by usually assuming a parametric (logistic) form of the incidence. We propose a new estimation procedure for a parametric cure rate that relies on a preliminary smooth estimator and is independent of the model assumed for the latency. On a second stage one can assume a semiparametric model for the latency and estimate also the survival distribution of the uncured subject. For the particular case of the logistic/Cox model, we investigate the theoretical properties of the estimators and show through simulations that presmoothing leads to more accurate results compared to the maximum likelihood estimator. To illustrate the practical use, we apply the new estimation procedure to two studies of melanoma survival data.

Keywords: Cure models; kernel smoothing; logistic model; survival analysis

1. Introduction

There are many situations in survival analysis problems where some of the subjects will never experience the event of interest. For instance, as significant progress is being made for treatment of different types of cancers, many of the patients get cured of the disease and do not experience recurrence or cancer-related death. Other examples include study of time to natural conception, time to default in finance and risk management, time to early failure of integrated circuits in engineering, time to find a job after a layoff. However, because of the finite duration of the studies and censoring, the cured subjects (for which the event never takes place) cannot be distinguished from the 'susceptible' ones. We can just get an indication of the presence of a cure fraction from the context of the study and a long plateau (containing many censored observations) with height greater than zero in the Kaplan-Meier estimator of the survival function. Predicting the probability of being cured given a set of characteristics is often of particular interest in order to make better decisions in terms of treatment, management strategies or public policies. This lead to the development of mixture cure models.

Mixture cure models were first proposed by [5] and [4]. They assume that the population is a mixture of two groups: the cured and the susceptible subjects. Within this very wide class of models, various approaches have been considered in the literature for modelling and estimating the incidence (probability of being uncured) and the latency (survival function of the uncured subjects). Initially, fully parametric models with a logistic regression form of the incidence and various parametric distributions for the latency were used in [11,15,31]. Later on, more flexible semi-parametric approaches were proposed for the latency based on the Cox proportional hazards model [22,25] or accelerated failure time models [17,32]. However, they still maintain the logistic regression model for the incidence. More recently, nonparametric methods have been developed for both or one of the model components in [2,21,30]. In this wide range of models, probably the most commonly used one in practice is the logistic/Cox mixture cure model [16,24,29].

There have been different proposals for estimation in the logistic/Cox mixture cure model. The presence of a latent variable (the unknown cure status), does not allow for a 'direct' approach as in the classical Cox proportional hazards model. [15] adapted a marginal likelihood approach computed through Monte Carlo approximations, whereas [22] and [25] computed the maximum likelihood estimator via the Expectation-Maximization algorithm. Asymptotic properties of the latter estimators are investigated in [18], while the procedure is implemented in the package `smcure` [7]. One concern about the previous estimators is that they are obtained by iterative procedures which could be unstable in practice. In particular, when the sample size is small there are situations in which the EM algorithm fails to converge (even though the `smcure` package can still provide without error the estimates obtained when the maximum number of iterations is reached). Such problems are for example reported in [14]. In addition, the maximum likelihood estimator for the incidence component depends on which variables are included in the latency model (see for example the illustration in Section 7) and this instability might in practice lead to unobserved effects (when the effect is not very strong). In particular, if the latency model is misspecified, even the estimators of the incidence parameters suffer from induced bias (see for example [6]).

In this paper, we introduce an alternative estimation method which applies very broadly and, in particular, for the logistic/Cox mixture cure model. Our approach focuses on direct estimation of the cure probability without using distributional assumptions on the latency and iterative algorithms. It relies on a preliminary nonparametric estimator for the incidence which is then 'projected' on a parametric class of functions (like logistic functions). The idea of constructing a parametric estimator by nonparametric estimation has been previously proposed for the classical linear regression by [9]. Later on it was shown to be effective also in the context of variable selection and functional linear regression [1,12]. However, its extension to nonlinear setups has been very little investigated. Here we show that in the context of mixture cure models, even when a parametric form is assumed for the incidence, the use of a presmoothed estimator as an intermediate step for obtaining the parameter estimates often leads to more accurate results. Once the cure fraction is estimated, we estimate the survival distribution of the uncured subjects. In the case of the logistic/Cox cure model, this is done by maximizing the Cox component of the likelihood. In this step, an iterative algorithm is used to compute the estimators of the baseline cumulative hazard and the regression parameters. This new approach is of practical relevance given the popularity of the semiparametric logistic/Cox mixture cure model. However, the method can be applied more in general to a mixture cure model with a parametric form of the incidence and other type of models for the uncured subjects, such as the semiparametric proportional odds model or the semiparametric AFT model. Our findings suggest that presmoothing has potential to improve parameter estimation for small and moderate sample size.

The paper is organized as follows. In Sections 2 and 3 we describe the model and the estimation procedure. Section 4 focuses on the estimation method in the case of the logistic/Cox mixture cure model. Consistency and asymptotic normality of the estimators are shown in Section 5. Thanks to the presmoothing, we are able to present theoretical results under more reasonable assumptions and thus we contribute to fill a gap between unrealistic technical conditions and applications. The finite sample performance of the method is investigated through a simulation study and results are reported in Section 6. For practical purposes, we propose to make simple and commonly used choices for the bandwidth and the kernel function in the presmoothing step, and we show that these choices provide satisfactory results. The proposed estimation procedure is applied to two medical datasets about studies of patients with melanoma cancer (see Section 7). We conclude in Section 8 with some discussion and ideas for further research. Finally, some of the proofs can be found in Section 8, while the remaining proofs and additional simulation results are collected in the online Supplementary Material [20].

2. Model description

In the mixture cure model the survival time T can be decomposed as

$$T = BT_0 + (1 - B)\infty,$$

where T_0 represents the finite survival time for an uncured individual and B is an unobserved 0-1 random variable giving the uncured status: $B = 1$ for uncured individuals and $B = 0$ otherwise. By convention $0 \cdot \infty = 0$. Let C be the censoring time and $(X', Z')'$ a $(p + q)$-dimensional vector of covariates, where x' denotes the transpose of the vector x. Let $\mathcal{X}$ and $\mathcal{Z}$ be the supports of X and Z respectively. Observations consist of n i.i.d. realizations of (Y, Δ, X, Z), where $Y = \min(T, C)$ is the finite follow-up time and $\Delta = \mathbb{1}_{\{T \leq C\}}$ is the censoring indicator. Since Y is finite, then necessarily $\mathbb{P}(C < \infty) = 1$, that means the censoring times are finite (which makes sense given the limited duration of the studies). As a result, censored survival times of the uncured subjects cannot be distinguished from the cured ones.

The covariates included in X are those used to model the cure rate, while the ones in Z affect the survival conditional on the uncured status. This allows in general to use different variables for modelling the incidence and the latency but does not exclude situations in which the two vectors X and Z share some components or are exactly the same. Apart from the standard assumption in survival analysis that $T_0 \perp (C, X)|Z$, here we also need

$$B \perp (C, T_0, Z)|X. \tag{1}$$

This implies in particular that

$$T \perp C|(X, Z) \tag{2}$$

(see Lemma 1 in the Supplementary Material [20]). Moreover, (1) implies

$$\mathbb{P}(T = \infty|X, Z) = \mathbb{P}(T = \infty|X). \tag{3}$$

In addition, in the cure model context we need that the event time T_0 has support $[0, \tau_0]$, i.e. $\{T > \tau_0\} = \{T = \infty\}$, such that

$$\inf_x \mathbb{P}(C > \tau_0|X = x) > 0. \tag{4}$$

(If the support of T_0 given $Z = z$ depends on z, then we let $\tau_0 = \sup \tau_0(z)$, where $\tau_0(z)$ is the right endpoint of this support.) This condition tells us that all the observations with $Y > \tau_0$ are cured. Even if it might seem restrictive, it is reasonable when a cure model is justified by a 'good' follow-up beyond the time when most of the events occur and it is commonly accepted in the cure model literature in order for the mixture cure model to be identifiable and not to overestimate the cure rate. Since $T_0 \perp X|Z$, we have

$$\mathbb{P}(T_0 \leq t|X, Z) = \mathbb{P}(T_0 \leq t|Z), \quad \forall t \in [0, \tau_0].$$

We assume a parametric model for the cure rate and we denote by $\pi_0(x)$ the cure probability of a subject with covariate x, i.e

$$\pi_0(x) = \mathbb{P}(T = \infty|X = x) = 1 - \phi(\gamma_0, x),$$

for some parametric model $\{\phi(\gamma, x) : \gamma \in G\}$ and $\gamma_0 \in G$. The first component of X is equal to one and the first component of γ corresponds to the intercept. In order for γ to be identifiable we need the following condition

$$\mathbb{P}\left(\phi(\gamma, X) = \phi(\tilde{\gamma}, X)\right) = 1 \qquad \text{implies that} \qquad \gamma = \tilde{\gamma}. \tag{5}$$

Choosing a parametric model for the incidence seems quite standard in the literature of mixture cure models ([6,21,23]) because of its simplicity and ease of interpretability (particularly for multiple covariates). 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 [2] and for our real data application in Section 7. It is also possible to test whether this assumption is reasonable using the test proposed in [19], but this is currently developed only for one covariate. Among the parametric models for the incidence component, the most common example is the logistic model, where

$$\phi(\gamma, x) = 1/(1 + \exp(-\gamma' x)). \tag{6}$$

We state the results in Section 5 for a general parametric model for the incidence, but then we focus on the logistic function in the simulation study in Section 6 since it is more of interest in practice. For the uncured subjects, we can consider a general semiparametric model defined through the survival function

$$S_u(t|z) = S_u(t|z; \beta, \Lambda) = \mathbb{P}(T_0 > t|Z = z, B = 1) \quad \text{and} \quad S_u(\tau_0|z) = 0, \tag{7}$$

where the conditional survival function S_u is allowed to depend on a finite-dimensional parameter, denoted by $\beta \in \mathcal{B}$, and/or an infinite-dimensional parameter, denoted by $\Lambda \in \mathcal{H}$, with $\mathcal{B}$ and $\mathcal{H}$ the respective parameter sets. Let $\beta_0 \in \mathcal{B}$ and $\Lambda_0 \in \mathcal{H}$ be the true values of these parameters. As a result, the conditional survival function corresponding to T is then

$$S(t|x, z) = \mathbb{P}(T > t|X = x, Z = z) = 1 - \phi(\gamma_0, x) + \phi(\gamma_0, x) S_u(t|z).$$

The main example we keep in mind is the Cox proportional hazards (PH) model where Λ_0 is the baseline cumulative hazard. In this case

$$S_u(t|z) = S_0(t)^{\exp(\beta_0' z)} = \exp(-\Lambda_0(t) \exp(\beta_0' z)), \tag{8}$$

where S_0 is the baseline survival and β_0 does not contain an intercept.

3. Presmoothing estimation approach

The estimation method we propose is based on a two step procedure. We first estimate nonparametrically the cure probability for each observation and then compute an estimator of γ as the maximizer of the logistic likelihood, ignoring the model for the uncured subjects. In the second step, we plug-in this estimator of γ in the full likelihood of the mixture cure model and fit the latency model using maximum likelihood estimation. In what follows, we describe in more details these two steps.

Step 1. Even though a parametric model is assumed for the incidence, we start by computing a nonparametric estimator of the cure probability for each subject. One possibility is to use the method followed by [21] (see also [30]), but other estimators are possible as well, as long as the conditions given in Section 5 are satisfied. The estimator of [21] is defined as follows:

$$\hat{\pi}(x) = \prod_{t \in \mathbb{R}} \left(1 - \frac{\hat{H}_1(dt|x)}{\hat{H}([t, \infty)|x)}\right), \tag{9}$$

where $\hat{H}([t,\infty)|x) = \hat{H}_1([t,\infty)|x) + \hat{H}_0([t,\infty)|x)$, $\hat{H}_1(dt|x) = \hat{H}_1((t - dt, t]|x)$ for small dt and

$$\hat{H}_k([t,\infty)|x) = \sum_{i=1}^{n} \frac{\tilde{K}_b(X_i - x)}{\sum_{j=1}^{n} \tilde{K}_b(X_j - x)} \mathbb{1}_{\{Y_i \geq t, \Delta_i = k\}}, \quad k = 0, 1,$$

are estimators of

$$H_k([t,\infty)|x) = \mathbb{P}(Y \geq t, \Delta = k | X = x),$$

$H([t,\infty)|x) = H_1([t,\infty)|x) + H_0([t,\infty)|x)$. Here $\tilde{K}_b$ is a multidimensional kernel function defined in the following way. If X is composed of continuous and discrete components, $X = (X_c, X_d) \in \mathcal{X}_c \times \mathcal{X}_d \subset \mathbb{R}^{p_c} \times \mathbb{R}^{p_d}$ with $p_c + p_d = p$, then

$$\tilde{K}_b(X_i - x) = K_b(X_{c,i} - x_c)\mathbb{1}_{\{X_{d,i} = x_d\}},$$

where $b = b_n$ is a bandwidth sequence, $K_b(\cdot) = K(\cdot/b)/b^{p_c}$ and $K(u) = \prod_{j=1}^{p_c} k(u_j)$, with k a kernel. Note that, one can compute this estimator with any covariate but here we only use X because of our assumption (3). The estimator $\hat{\pi}(x)$ 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 τ_0. Since $\hat{H}_1(dt|x)$ is different from zero only at the observed event times, computation of $\hat{\pi}(x)$ requires only a product over t in the set of the observed event times. Afterwards, we consider the logistic likelihood

$$\hat{L}_{n,1}(\gamma) = \prod_{i=1}^{n} \phi(\gamma, X_i)^{1-\hat{\pi}(X_i)} (1 - \phi(\gamma, X_i))^{\hat{\pi}(X_i)},$$

and define $\hat{\gamma}_n$ as the maximizer of

$$\log \hat{L}_{n,1}(\gamma) = \sum_{i=1}^{n} \Big\{ [1 - \hat{\pi}(X_i)] \log \phi(\gamma, X_i) + \hat{\pi}(X_i) \log [1 - \phi(\gamma, X_i)] \Big\}. \tag{10}$$

Existence and uniqueness of $\hat{\gamma}_n$ holds under the same conditions as for the maximum likelihood estimator in the binary outcome regression model where $1 - \hat{\pi}(X_i)$ 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.

Step 2. Now we consider the likelihood of the mixture cure model. Let

$$f_u(t|z;\beta,\Lambda) = -(\partial/\partial t) S_u(t|z;\beta,\Lambda)$$

with $S_u(t|z;\beta,\Lambda)$ as defined in (7), denote an element in the model for the conditional density of T_0 given $Z = z$, which is supposed to exist and belong to the model. Assuming non informative censoring and that the distribution of the covariates does not carry information on the parameters β, Λ, the likelihood criterion is then

$$L_{n,2}(\beta,\Lambda,\gamma) = \prod_{i=1}^{n} \{\phi(\gamma, X_i) f_u(Y_i|Z_i;\beta,\Lambda)\}^{\Delta_i} \{1 - \phi(\gamma, X_i) + \phi(\gamma, X_i) S_u(Y_i|Z_i;\beta,\Lambda)\}^{1-\Delta_i}, \tag{11}$$

and we maximize it w.r.t. β and Λ for $\gamma = \hat{\gamma}_n$, i.e. $(\hat{\beta}_n, \hat{\Lambda}_n)$ are the maximizers of

$$\hat{l}_n(\beta,\Lambda,\hat{\gamma}_n) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \Delta_i, X_i, Z_i; \beta, \Lambda, \hat{\gamma}_n), \tag{12}$$

over a set of possible values for β and Λ, where

$$\begin{aligned} \ell(Y_i, \Delta_i, X_i, Z_i; \beta, \Lambda, \gamma) = {} & \Delta_i \log f_u(Y_i|Z_i;\beta,\Lambda) \\ & + (1 - \Delta_i) \log \{1 - \phi(\gamma, X_i) + \phi(\gamma, X_i) S_u(Y_i|Z_i;\beta,\Lambda)\}. \end{aligned} \tag{13}$$

4. Presmoothing estimation for the parametric/Cox mixture cure model

In the sequel we focus on the case of a Cox PH model defined in (8) for the conditional law of T_0. The criterion defined in (12) becomes

$$\hat{l}_n(\beta,\Lambda,\hat{\gamma}_n) = \frac{1}{n}\sum_{i=1}^{n}\Delta_i\left\{\mathbb{1}_{\{Y_i<\tau_0\}}[\log\Delta\Lambda(Y_i)+\beta' Z_i]-\Lambda(Y_i)e^{\beta' Z_i}\right\} + \frac{1}{n}\sum_{i=1}^{n}(1-\Delta_i)\log\left\{1-\phi(\hat{\gamma}_n,X_i)+\phi(\hat{\gamma}_n,X_i)\exp\left(-\Lambda(Y_i)e^{\beta' Z_i}\right)\right\}, \quad (14)$$

and has to be maximized with respect to β and Λ in the class of step functions Λ defined on $[0,\tau_0]$ (thus by definition $\Lambda(t)=\infty$ if $t>\tau_0$), with jumps of size $\Delta\Lambda$ at the event times. The indicator of the event $\{Y_i<\tau_0\}$ in the first term is needed in case the distribution of the event times has a jump at τ_0 meaning that $\mathbb{P}(T_0=\tau_0|Z)>0$. In such a case $f_u(\tau_0|Z;\beta,\Lambda)=\exp(-\Lambda(\tau_0)e^{\beta' Z})$ where $\Lambda(\tau_0)=\lim_{t\uparrow\tau_0}\Lambda(t)$. Otherwise, if $\mathbb{P}(T_0=\tau_0|Z)=0$, then for all uncensored observations we have $\mathbb{1}_{\{Y<\tau_0\}}=1$ with probability one. Thus, the presence of the indicator function can be neglected. As in [18], it can be shown that

$$(\hat{\beta}_n,\hat{\Lambda}_n) = \arg\max_{\beta,\Lambda}\hat{l}_n(\beta,\Lambda,\hat{\gamma}_n) \quad (15)$$

exists and it is finite. Moreover, for any given β and γ, the $\Lambda_{n,\beta,\gamma}$ which maximizes $\hat{l}_n(\beta,\Lambda,\gamma)$ in (14), with respect to Λ with jumps at the event times, can be characterized as

$$\Lambda_{n,\beta,\gamma}(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i\mathbb{1}_{\{Y_i\le t,Y_i<\tau_0\}}}{\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{Y_i\le Y_j\le\tau_0\}}\exp(\beta' Z_j)\left\{\Delta_j+(1-\Delta_j)g_j(Y_j,\Lambda_{n,\beta},\beta,\gamma)\right\}}, \quad (16)$$

where

$$g_j(t,\Lambda,\beta,\gamma) = \frac{\phi(\gamma,X_j)\exp\left(-\Lambda(t)\exp\left(\beta' Z_j\right)\right)}{1-\phi(\gamma,X_j)+\phi(\gamma,X_j)\exp\left(-\Lambda(t)\exp\left(\beta' Z_j\right)\right)}. \quad (17)$$

Next, we could define

$$\hat{\beta}_n = \arg\max_{\beta}\hat{l}_n(\beta,\Lambda_{n,\beta,\hat{\gamma}_n},\hat{\gamma}_n) \quad \text{and} \quad \hat{\Lambda}_n = \Lambda_{n,\hat{\beta}_n,\hat{\gamma}_n}.$$

To compute $(\hat{\beta}_n,\hat{\Lambda}_n)$ we use an iterative algorithm based on profiling. To be precise, we start with initial values which are the maximum partial likelihood estimator and the Breslow estimator (as if there was no cure fraction) and we iterate between the next two steps until convergence:

a) Compute the weights

$$w_j^{(m)} = \Delta_j + (1-\Delta_j)\frac{\phi(\hat{\gamma}_n,X_j)\hat{S}_u^{(m)}(Y_j|Z_j)}{1-\phi(\hat{\gamma}_n,X_j)+\phi(\hat{\gamma}_n,X_j)\hat{S}_u^{(m)}(Y_j|Z_j)},$$

where

$$\hat{S}_u^{(m)}(Y_j|Z_j) = \exp\left(-\hat{\Lambda}_n^{(m)}(Y_j)\exp\left(\hat{\beta}_n^{(m)'}Z_j\right)\right),$$

using the estimators $\hat{\Lambda}_n^{(m)}$, $\hat{\beta}_n^{(m)}$ of the previous step.

b) Using the previous weights, update the estimators for Λ and β, i.e. $\hat{\beta}_n^{(m+1)}$ is the maximizer of

$$\prod_{i=1}^{n}\left\{\frac{e^{\beta' Z_i}}{\sum_{Y_k \geq Y_i} w_k^{(m)} e^{\beta' Z_k}}\right\}^{\Delta_i}$$

and

$$\hat{\Lambda}_n^{(m+1)}(t) = \sum_{i=1}^{n} \frac{\Delta_i \mathbb{1}_{\{Y_i \leq t, Y_i < \tau_0\}}}{\sum_{j=1}^{n} \mathbb{1}_{\{Y_i \leq Y_j \leq \tau_0\}} w_j^{(m)} \exp\left(\hat{\beta}_n^{(m+1)'} Z_j\right)}. \tag{18}$$

The update of Λ an β in Step (b) coincides with the maximization step of the EM algorithm and the weights $w^{(m)}$ correspond to the expectation of the latent variable B given the observed data and the current parameter values. However, unlike the maximum likelihood estimation [25], we are keeping $\hat{\gamma}_n$ fixed while performing this iterative algorithm. The estimator $\hat{\Lambda}_n$ seems to depend on the unknown τ_0. However, with data at hand, one could easily proceed without knowing τ_0. Indeed, if there are ties at the last uncensored observation, then τ_0 is revealed by the data. On the other hand, if there are no ties, all uncensored observations will be smaller than τ_0, hence no need to know τ_0.

As suggested in [25,26], we impose the zero-tail constraint, meaning that $\hat{S}_u^{(m)}$ is forced to be equal to zero beyond the last event. In this way, all censored observations in the plateau are assigned to the cured group.

5. Asymptotic results for the parametric/Cox mixture cure model

We first explain why presmoothing allows for more realistic asymptotic results in semiparametric mixture cure models. Next, we show consistency and asymptotic normality of the proposed estimators $\hat{\gamma}_n$, $\hat{\beta}_n$ and $\hat{\Lambda}_n$ for the parametric/Cox mixture cure model when, in Step 1, we use a general nonparametric estimator $\hat{\pi}$ of π_0 that satisfies certain assumptions. Afterwards, we verify these conditions for the particular estimator $\hat{\pi}$ in (9). Some of the proofs can be found in Section 8 and the rest in the online Supplementary Material [20]. The assumptions mentioned in Section 2 are assumed to be satisfied throughout this section. In addition $Var(Z)$ is supposed to have full rank.

5.1. A challenge with mixture cure models

To derive asymptotic results, in most of the existing literature it has been assumed that

$$\inf_z \mathbb{P}(T_0 \geq \tau_0 | Z = z) > 0, \tag{19}$$

[18,21]. In nonparametric approaches such a condition keeps the denominators away from zero. In the parametric/Cox mixture cure model, it guarantees that the baseline distribution stays bounded on the compact support $[0, \tau_0]$. However, condition (19) implies that $\inf_z \mathbb{P}(Y = \tau_0, \Delta = 1 | Z = z) > 0$, a condition which is not frequently satisfied in real-data applications.

One could imagine that, instead of imposing condition (19), it could be possible to proceed as follows: first restrict to events on $[0, \tau^*]$ for some $\tau^* < \tau_0$ such that

$$\inf_z \mathbb{P}(Y \geq \tau^*, \Delta = 1 | Z = z) > 0, \tag{20}$$

next derive the asymptotics, and finally let τ^* tend to τ_0. This idea is used, for instance, in Cox PH model, see [13] chapter 8, or [3]. However, this idea does not seem to work for mixture cure models without suitable adaptation. This is because it implicitly requires that β_0 and Λ_0 are identifiable from the restricted data. Here, identifiability means that the true values β_0 and Λ_0 of the parameters maximize the expectation of the criterion maximized to obtain the estimators. Two aspects have to be taken into account when analyzing this identifiability. The first aspect is related to the parameter identifiability in the semiparametric model for T_0 when the events are restricted to $[0, \tau^*]$. This property is satisfied in the common models, in particular it holds true in the Cox PH model as soon as $Var(Z)$ has full rank. The second aspect is the additional complexity induced by the mixture with a cure fraction. If the cure fraction is unknown and one decides to restrict to events on $[0, \tau^*]$, the parameter identifiability is likely lost because the events $\{T_0 \in (\tau^*, \tau_0]\}$ and $\{T = \infty\}$ are not distinguishable. The usual remedy for this is to impose (19), so that τ^* could be taken equal to τ_0.

Presmoothing allows to avoid condition (19) and thus to fill the gap between the technical conditions and the reality of the data. This is possible because, when using the presmoothing, the conditional probability of the event $\{T = \infty\}$ is identified by other means. We are thus able to prove the consistency of $\hat{\beta}$ and $\hat{\Lambda}$ without imposing (19). Deriving the asymptotic normality without (19) remains an open problem which will be addressed elsewhere.

5.2. Consistency

We first prove consistency of $\hat{\gamma}_n$ and then use that result to obtain consistency of $\hat{\Lambda}_n$ and $\hat{\beta}_n$. In order to proceed with our results, the following conditions will be used.

(AC1) $\sup_{x \in \mathcal{X}} |\hat{\pi}(x) - \pi_0(x)| \to 0$ almost surely.
(AC2) The parameters β_0 and γ_0 lie in the interior of compact sets $B \subset \mathbb{R}^q$, $G \subset \mathbb{R}^p$.
(AC3) There exist some constants $a > 0$, $c > 0$ such that

$$|\phi(\gamma_1, x) - \phi(\gamma_2, x)| \leq c\|\gamma_1 - \gamma_2\|^a, \qquad \forall \gamma_1, \gamma_2 \in G, \forall x \in \mathcal{X},$$

where $\|\cdot\|$ denotes the Euclidean distance.
(AC4) $\inf_{\gamma \in G} \inf_{x \in \mathcal{X}} \phi(\gamma, x) > 0$ and $\inf_{\gamma \in G} \inf_{x \in \mathcal{X}} \phi(\gamma, x) < 1$.
(AC5) The covariates are bounded: $\mathbb{P}(\|Z\| < m \text{ and } \|X\| < m) = 1$ for some $m > 0$.
(AC6) The baseline hazard function $\lambda_0(t) = \Lambda_0'(t)$ is strictly positive and continuous on $[0, \tau_0)$.
(AC7) With probability one, the conditional distribution function of the censoring times $F_C(t|x, z)$ is continuous in t on $[0, \tau_0]$ and there exists a constant $C > 0$ such that

$$\inf_{0 \leq t_1 < t_2 \leq \tau_0} \inf_{x,z} \frac{F_C(t_2|x, z) - F_C(t_1|x, z)}{t_2 - t_1} > C.$$

(AC1) is a minimal assumption given that we want to match $\phi(\gamma, \cdot)$ to $\hat{\pi}(\cdot)$. (AC2) to (AC4) are mild conditions satisfied by usual binary regression models, like for instance the logistic one, and (AC5) is always satisfied in practice for large m.

Theorem 1. *Let the estimator $\hat{\gamma}_n$ be defined as in (10). Assume that (AC1)-(AC4) hold. Then, $\hat{\gamma}_n \to \gamma_0$ almost surely.*

Theorem 2. *Let the estimators $\hat{\beta}_n$ and $\hat{\Lambda}_n$ be defined as in Section 4. Assume that (AC1)-(AC7) hold. Then, with probability one, $\|\hat{\beta}_n - \beta_0\| \to 0$, where $\|\cdot\|$ denotes the Euclidean distance. Moreover, for*

any $\tau^* \leq \tau_0$ *satisfying* (20)*, with probability one,*

$$\sup_{t\in[0,\tau^*]} \left|\hat{\Lambda}_n(t) - \Lambda_0(t)\right| \to 0.$$

When condition (19) is satisfied and $\tau^* = \tau_0$ in the previous Theorem, we are referring to the continuous version of Λ_0, i.e. $\Lambda_0(\tau_0) = \lim_{t\uparrow\tau_0} \Lambda_0(t)$. Note that, by definition, we also have $\hat{\Lambda}_n(\tau_0) = \lim_{t\uparrow\tau_0} \hat{\Lambda}_n(t)$.

5.3. Asymptotic normality

We first derive asymptotic normality of $\hat{\gamma}_n$ following the approach in [8]. Theorem 2 in that paper provides sufficient conditions for the $\sqrt{n}$ normality of parametric estimators obtained by minimizing an objective function that depends on a preliminary infinite dimensional estimator $\hat{\pi}$. In our case, since $\hat{\gamma}_n$ solves

$$\frac{1}{n}\nabla_\gamma \log \hat{L}_{n,1}(\gamma) = 0,$$

where ∇_γ denotes the vector-valued partial differentiation operator with respect to the components of γ, it follows that $\hat{\gamma}_n$ minimizes the function

$$\left\|\frac{1}{n}\nabla_\gamma \log \hat{L}_{n,1}(\gamma)\right\| = \left\|\frac{1}{n}\sum_{i=1}^{n} m(X_i;\gamma,\hat{\pi})\right\|,$$

where

$$m(x;\gamma,\pi) = \left[\frac{1-\pi(x)}{\phi(\gamma,x)} - \frac{\pi(x)}{1-\phi(\gamma,x)}\right]\nabla_\gamma\phi(\gamma,x). \tag{21}$$

Hence, we only need to check that the conditions of Theorem 2 in [8] are satisfied. To do that, we need the following assumptions which are stronger than the previous (AC1)-(AC4).

(AN1) The parameter γ_0 lies in the interior of a compact set $G \subset \mathbb{R}^p$ and, for each $x \in \mathcal{X}$, the function $\gamma \mapsto \phi(\gamma,x)$ is twice continuously differentiable with uniformly bounded derivatives in $G \times \mathcal{X}$ and satisfies (AC4).

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

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

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

(AN3) The matrix $\mathbb{E}\left[\nabla_\gamma\phi(\gamma_0,X)\nabla_\gamma\phi(\gamma_0,X)'\right]$ is positive definite.

(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 Ψ such that

$$\mathbb{E}^*\left[(\hat{\pi}(X)-\pi_0(X))\left(\frac{1}{\phi(\gamma_0,X)}+\frac{1}{1-\phi(\gamma_0,X)}\right)\nabla_\gamma\phi(\gamma_0,X)\right] = \frac{1}{n}\sum_{i=1}^{n}\Psi(Y_i,\Delta_i,X_i)+R_n,$$

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

Theorem 3. *Let the estimator $\hat{\gamma}_n$ be defined as in (10). Assume that (AN1)-(AN4) hold. Then,*

$$n^{1/2}(\hat{\gamma}_n-\gamma_0)\xrightarrow{d} N(0,\Sigma_\gamma)$$

with covariance matrix Σ_γ defined in (A28) of [20].

For deriving the asymptotic distribution of $\hat{\beta}_n$ and $\hat{\Lambda}_n$ we assume, for simplicity, that condition (19) is satisfied. In such case, in Theorem 2 we can take $\tau^*=\tau_0$ and obtain uniform strong consistency of $\hat{\Lambda}_n$ on the whole support $[0,\tau_0]$. We believe that, at the price of additional technicalities, asymptotic distributional theory can be obtained also without imposing (19), as we did for the consistency in Theorem 2. This conjecture is supported by simulations but we leave the problem to be addressed by future research.

Theorem 4. *Let the estimators $\hat{\beta}_n$ and $\hat{\Lambda}_n$ be defined as in Section 4. Assume that condition (19), (AN1)-(AN4) and (AC2), (AC5)-(AC7) hold. Then,*

$$\left\langle \sqrt{n}\left(\hat{\Lambda}_n-\Lambda_0\right), \sqrt{n}\left(\hat{\beta}_n-\beta_0\right)\right\rangle \to G$$

weakly in $l^\infty(\mathcal{H}_{\mathfrak{m}})$, where $\mathcal{H}_{\mathfrak{m}}$ is a functional space defined in Section A.1, $l^\infty(\mathcal{H}_{\mathfrak{m}})$ denotes the space of bounded real-valued functions on $\mathcal{H}_{\mathfrak{m}}$, G is a tight Gaussian process in $l^\infty(\mathcal{H}_{\mathfrak{m}})$ with mean zero and covariance process given in (A39) of [20] and for $h=(h_1,h_2)\in\mathcal{H}_{\mathfrak{m}}$

$$\langle\Lambda,\beta\rangle(h)=\int_0^{\tau_0}h_1(t)\,\mathrm{d}\Lambda(t)+h_2'\beta.$$

The asymptotic variances of each component of $\hat{\beta}_n$ and of $\hat{\Lambda}_n(t)$ can be obtained from the covariance process in (A39) by taking $h_1(t)=0$ for all t and $h_2=e_i$ (the ith unit vector) or $h_2=0$ and $h_1(s)=\mathbb{1}_{\{s\le t\}}$. We leave the details about these covariance matrices in the Supplementary Material because they have quite complicated expressions that require definitions of several other quantities. Even though it could be possible in principle to estimate the asymptotic standard errors through plug-in estimators and numerical inverse, we think that this is not feasible in practice and we do not intend to exploit it further. Instead, we use a bootstrap procedure for estimation of the standard errors in the application discussed in Section 7. However, the maximum likelihood estimators are not more favorable in this regard. For example, in the logistic/Cox model, the proposed estimators of the asymptotic variance in [18] also involve solving numerically complicated nonlinear equalitons. For this reason, bootstrap is used in practice to estimate the standard errors even for the maximum likelihood estimators.

By considering a two-step procedure, where estimation of the incidence parameters is performed independently of the latency model, we expect to loose efficiency of the estimators. However, this

does not cause major concern because our purpose is to provide an alternative estimation method that performs better than the maximum likelihood estimation with sample sizes usually encountered in practice. Efficiency is a key concept for the asymptotics of the estimators, and in general there is no particular need for another method since the MLE would be the best choice. However, in many nonlinear models, like the mixture cure models, the asymptotic approximation is poor and the efficiency becomes a less relevant purpose for real data sample sizes. Hence, we choose to trade efficiency for better performance in a wider range of applications.

5.4. Verification of assumptions for $\hat{\pi}$

Next we show that our assumptions (AN1)-(AN4) of the asymptotic theory are satisfied for the nonparametric estimator $\hat{\pi}$ defined in (9) and the logistic model in (6). For reasons of simplicity, since we use results available in the literature only for a one-dimensional covariate, we consider only cases with one continuous covariate. In order for assumption (AN4) to be satisfied we need the following conditions:

(C1) The bandwidth b is such that $nb^4 \to 0$ and $nb^{3+\xi}/(\log b^{-1}) \to \infty$ for some $\xi > 0$.

(C2) The support $\mathcal{X}$ of X is a compact subset of $\mathbb{R}$. The density $f_X(\cdot)$ of X is bounded away from zero and twice differentiable with bounded second derivative.

(C3) The kernel k is a twice continuously differentiable, symmetric probability density function with compact support and $\int uk(u)\,\mathrm{d}u = 0$.

(C4) (i) The functions $H([0,t]|x)$, $H_1([0,t]|x)$ are twice differentiable with respect to x, with uniformly bounded derivatives for all $t \le \tau_0$, $x \in \mathcal{X}$. Moreover, there exist continuous nondecreasing functions L_1, L_2, L_3 such that $L_i(0) = 0$, $L_i(\tau_0) < \infty$ and for all $t, s \in [0, \tau_0]$, $x \in \mathcal{X}$,

$$|H_c(t|x) - H_c(s|x)| \le |L_1(t) - L_1(s)|, \quad |H_{1c}(t|x) - H_{1c}(s|x)| \le |L_1(t) - L_1(s)|$$

$$\left|\frac{\partial H_c(t|x)}{\partial x} - \frac{\partial H_c(s|x)}{\partial x}\right| \le |L_2(t) - L_2(s)|$$

$$\left|\frac{\partial H_{1c}(t|x)}{\partial x} - \frac{\partial H_{1c}(s|x)}{\partial x}\right| \le |L_3(t) - L_3(s)|,$$

where the subscript c denotes the continuous part of a function.

(ii) The jump points for the distribution function $G(t|x)$ of the censoring times given the covariate, are finite and the same for all x. The partial derivative of $G(t|x)$ with respect to x exists and is uniformly bounded for all $t \le \tau_0$, $x \in \mathcal{X}$. Moreover, the partial derivative with respect to x of $F(t|x)$ (distribution function of the survival times T given $X = x$) exists and is uniformly bounded for all $t \le \tau_0$, $x \in \mathcal{X}$.

(C5) The survival time T and the censoring time C are independent given X.

(C1) to (C5) are conditions guaranteeing the rates of convergence and the i.i.d. representation [10]. In case of discrete covariates we also need to have only a finite number of atoms. Assumption (C5) is needed because we are dealing with the distribution of T conditional only on the covariate X (since the cure rate depends only on X).

Theorem 5. *Under the conditions (C1)-(C5), the assumptions (AN1)-(AN4) hold true for the logistic model and the estimator $\hat{\pi}(x)$ defined in (9).*

6. Simulation study

In this section we focus on the logistic/Cox mixture cure model and evaluate the finite sample performance of the proposed method. Comparison is made with the maximum likelihood estimator implemented in the package `smcure`.

We first illustrate through a brief example the convergence problems of the smcure estimator. We consider a model where the incidence depends on four independent covariates: $X_1 \sim N(0,2)$, $X_2 \sim$ Uniform$(-1,1)$, $X_3 \sim$ Bernoulli(0.8), $X_4 \sim$ Bernoulli(0.2). The latency depends on $Z_1 = X_1$, $Z_2 = X_3$ and $Z_3 = X_4$. We generate the cure status B as a Bernoulli random variable with success probability $\phi(\gamma, X)$ where ϕ is the logistic function and $\gamma = (0.6, -1, 1, 2.5, 1.2)$. The survival times for the uncured observations are generated according to a Weibull proportional hazards model

$$S_u(t|z) = \exp\left(-\mu t^{\rho} \exp(\beta' z)\right),$$

and are truncated at $\tau_0 = 14$ for $\rho = 1.75$, $\mu = 1.5$, $\beta = (-0.8, 0.9, 0.5)$. The censoring times are independent from X and T. They are generated from the exponential distribution with parameter $\lambda_C = 0.22$ and are truncated at $\tau = 16$. We generate 1000 datasets according to this model with sample size $n = 100$, and we observe that smcure fails to converge in 43% of the cases. Convergence fails mainly in the γ parameter, with only 17% of the cases failing to converge also for the β parameter (because of the unreasonable γ estimators). On the other hand, there was no convergence problem in the second step of the presmoothing approach. In addition, even among the cases where smcure converged, the presmoothing approach showed significantly better behavior, as can be seen in Table 1.

Hence, in the cases in which smcure exhibits very poor behavior, the presmoothing is obviously superior. Next, we focus on models for which smcure behaves reasonable (there are convergence problems in less then 3% of the cases) and show that, even in such scenarios presmoothing can lead to more accurate results.

We consider four different models and for each of them various choices of the parameters in order to cover a wide range of scenarios. The models are as follows.

Model 1. Both incidence and latency depend on one covariate X, which is uniform on $(-1,1)$. We generate the cure status B as a Bernoulli random variable with success probability $\phi(\gamma, X)$ where ϕ is the logistic function. The survival times for the uncured observations are generated according to a Weibull proportional hazards model

$$S_u(t|x) = \exp\left(-\mu t^{\rho} \exp(\beta x)\right),$$

Table 1. Bias, variance and MSE of $\hat{\gamma}$ and $\hat{\beta}$ for `smcure` and our approach among the iterations that converged for smcure.

	presmoothing			smcure		
Par.	Bias	Var.	MSE	Bias	Var.	MSE
γ_1	−0.113	0.620	0.633	0.200	8.318	8.358
γ_2	−0.073	0.156	0.162	−0.388	3.085	3.236
γ_3	−0.071	0.546	0.551	0.280	1.957	2.035
γ_4	0.037	1.326	1.327	0.704	14.395	14.891
γ_5	−0.250	8.398	8.461	1.621	36.450	36.945
β_1	−0.014	0.011	0.012	−0.017	0.012	0.012
β_2	0.024	0.064	0.065	0.026	0.065	0.065
β_3	−0.053	0.165	0.168	−0.053	0.166	0.169

and are truncated at τ_0 for $\rho = 1.75$, $\mu = 1.5$, $\beta = 1$ and $\tau_0 = 4$. The censoring times are independent from X and T. They are generated from the exponential distribution with parameter λ_C and are truncated at $\tau = 6$.

Model 2. Both incidence and latency depend on one covariate X with standard normal distribution. The cure status and the survival times for the uncured observations are generated as in Model 1 for $\rho = 1.75$, $\mu = 1.5$, $\beta = 1$ and $\tau_0 = 10$. The censoring times are generated according to a Weibull proportional hazards model

$$S_C(t|x) = \exp\left(-\nu\mu t^{\rho}\exp(\beta_C x)\right),$$

for $\beta_C = 1$ and various choices of ν and are truncated at $\tau = 15$.

Model 3. For the incidence we consider three independent covariates: X_1 is normal with mean zero and standard deviation 2, X_2 and X_3 are Bernoulli random variables with parameters 0.6 and 0.4 respectively. The latency also depends on three covariates: $Z_1 = X_1$, Z_2 is a uniform random variable on $(-3,3)$ independent of the previous ones and $Z_3 = X_2$. The cure status and the survival times for the uncured observations are generated as in Model 1 for $\rho = 1.75$, $\mu = 1.5$ and different choices of the other parameters. The censoring times are generated independently of the previous variables from an exponential distribution with parameter λ_C and are truncated at τ, for given choices of λ_C and τ.

Model 4. This setting is obtained by adding an additional continuous covariate to the incidence component of Model 3. To be precise, X_1 is normal with mean zero and standard deviation 2, X_2 is uniform on $(-1,1)$ independent of the other variables, X_3 and X_4 are Bernoulli random variables with parameters 0.6 and 0.4 respectively. As in Model 3, $Z_1 = X_1$, Z_2 is a uniform random variable on $(-3,3)$ independent of the previous ones and $Z_3 = X_3$. The event and censoring times are generated as in the previous model.

For the four models we choose the values of the unspecified parameters in such a way that the cure rate is around 20%, 30%, 50% (corresponding respectively to scenarios 1, 2 and 3) and the censoring rate corresponds to three levels (with a difference of 5% between each other). The specification of the parameters and the corresponding censoring rate and percentage of the observations in the plateau are given Table 2. Note that, within each scenario, the fraction of the observations in the plateau decreases as the censoring rate increases because more cured observations are censored earlier and as a result are not observed in the plateau. This makes the estimation of the cure rate more difficult. 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 (19) is satisfied but in practice it is unlikely to observe event times at τ_0. In this way, we try to find a compromise between theoretical assumptions and real-life scenarios.

For each setting we consider samples of size $n = 200, 400, 1000$. This leads to a total of 108 settings (4 models, 3 scenarios for the cure rate, 3 censoring levels and 3 sample sizes). In this way, we hope to address a number of issues such as the effect of the cure proportion, the sample size, amount and type of censoring, covariates (number, relation between X and Z and their distribution). For each configuration 1020 datasets were generated and the estimators of β_0 and γ_0 were computed through `smcure` and our method. We report the bias, variance and mean squared error (MSE) of the estimators, computed after omitting the lowest and the highest 1% of the estimators (for stability of the reported results) and rounded to three decimals. Tables 3-5 show some of the results, while the rest can be found in the online Supplementary Material [20]. We aim to provide a ready-to-use method that works well in practice without needing to think about how to choose the kernel function or the bandwidth. Hence, we illustrate the performance of the method for some standard and commonly used choices. The kernel function k is taken to be the Epanechnikov kernel $k(u) = (3/4)(1-u^2)\mathbb{1}_{\{|u|\leq 1\}}$. We use the cross-validation bandwidth (implemented in the R package `np`) for kernel estimators of conditional distribution functions, in our case for estimation of $H = H_0 + H_1$ given the continuous covariates (affecting the incidence). In

Table 2. Parameter values and model characteristics for each scenario.

Model	Parameters	Scenario	Cens. level	Cens. parameters	Cens. rate	Plateau
1	$\gamma = (1.75, 2)$	1	1	$\lambda_C = 0.1$	25%	15%
			2	$\lambda_C = 0.2$	30%	11%
			3	$\lambda_C = 0.3$	35%	9%
	$\gamma = (1, 1.5)$	2	1	$\lambda_C = 0.1$	34%	22%
			2	$\lambda_C = 0.25$	40%	15%
			3	$\lambda_C = 0.4$	46%	10%
	$\gamma = (0.1, 5)$	3	1	$\lambda_C = 0.2$	54%	32%
			2	$\lambda_C = 0.4$	59%	23%
			3	$\lambda_C = 0.7$	65%	15%
2	$\gamma = (1.5, 0.5)$	1	1	$\nu = 1/15$	25%	7%
			2	$\nu = 1/7$	30%	4%
			3	$\nu = 1/4$	35%	2%
	$\gamma = (1, 1)$	2	1	$\nu = 1/13$	35%	14%
			2	$\nu = 1/10$	40%	9%
			3	$\nu = 5/18$	45%	6%
	$\gamma = (-0.1, 5)$	3	1	$\nu = 1/9$	56%	38%
			2	$\nu = 1/4$	60%	30%
			3	$\nu = 2/5$	65%	25%
3	$\gamma = (0.5, -1, 2.5, 1.2)$, $\beta = (-1, 0.5, 1.5)$, $\tau_0 = 30$, $\tau = 35$	1	1	$\lambda_C = 0.12$	25%	10%
			2	$\lambda_C = 0.25$	30%	6%
			3	$\lambda_C = 0.45$	35%	4%
	$\gamma = (1, 2, 1.8, 0.5)$, $\beta = (1, 0.5, 2)$, $\tau_0 = 6$, $\tau = 8$	2	1	$\lambda_C = 0.2$	35%	16%
			2	$\lambda_C = 0.5$	40%	9%
			3	$\lambda_C = 0.8$	45%	6%
	$\gamma = (-0.8, 1.3, 1.5, -0.2)$, $\beta = (1, -0.1, 0.8)$, $\tau_0 = 5$, $\tau = 7$	3	1	$\lambda_C = 0.3$	55%	24%
			2	$\lambda_C = 0.7$	59%	14%
			3	$\lambda_C = 1.3$	65%	8%
4	$\gamma = (0.6, -1, 1, 2.5, 1.2)$, $\beta = (-0.8, 0.3, 0.5)$, $\tau_0 = 14$, $\tau = 16$	1	1	$\lambda_C = 0.1$	25%	11%
			2	$\lambda_C = 0.22$	30%	7%
			3	$\lambda_C = 0.35$	35%	5%
	$\gamma = (0.45, 0.5, 2, 1, 0.5)$, $\beta = (1, 0.5, 2)$, $\tau_0 = 18$, $\tau = 20$	2	1	$\lambda_C = 0.15$	35%	11%
			2	$\lambda_C = 0.35$	40%	7%
			3	$\lambda_C = 0.6$	45%	5%
	$\gamma = (-0.22, 0.3, -0.4, 0.5, -0.2)$, $\beta = (0.4, -0.1, 0.5)$, $\tau_0 = 6$, $\tau = 8$	3	1	$\lambda_C = 0.2$	55%	30%
			2	$\lambda_C = 0.4$	59%	20%
			3	$\lambda_C = 0.7$	65%	12%

addition, we restrict to the interval $[0, Y_{(m)}]$, where $Y_{(m)}$ is the last observed event time since the estimator of the cure probability $\hat{\pi}$ in (9) 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 estimation of the conditional distribution $H(t|x)$ for $t \leq Y_{(m)}$. This choice of bandwidth improves significantly the performance of

Table 3. Bias, variance and MSE of $\hat{\gamma}$ and $\hat{\beta}$ for `smcure` (second rows) and our approach (first rows) in Model 1 and 2.

				Cens. level 1			Cens. level 2			Cens. level 3		
Mod.	n	scen.	Par.	Bias	Var.	MSE	Bias	Var.	MSE	Bias	Var.	MSE
1	200	1	γ_1	0.001	0.060	0.060	0.020	0.065	0.065	0.005	0.078	0.078
				0.021	0.063	0.063	0.050	0.068	0.071	0.044	0.084	0.086
			γ_2	−0.034	0.164	0.165	−0.014	0.202	0.202	−0.051	0.209	0.212
				0.026	0.173	0.173	0.067	0.222	0.226	0.044	0.229	0.230
			β	0.008	0.028	0.028	0.015	0.029	0.029	0.013	0.034	0.035
				0.007	0.028	0.028	0.012	0.029	0.029	0.009	0.035	0.035
		3	γ_1	−0.001	0.059	0.059	0.009	0.065	0.065	−0.014	0.091	0.092
				0.010	0.064	0.064	0.029	0.074	0.075	0.037	0.113	0.115
			γ_2	−0.034	0.536	0.537	−0.111	0.595	0.608	−0.085	0.809	0.816
				0.201	0.649	0.689	0.218	0.768	0.816	0.400	1.146	1.306
			β	0.011	0.090	0.090	0.024	0.109	0.110	0.014	0.128	0.128
				0.007	0.091	0.091	0.014	0.110	0.110	−0.001	0.129	0.129
	400	1	γ_1	0.001	0.028	0.028	0.007	0.032	0.032	0.001	0.037	0.037
				0.015	0.029	0.030	0.027	0.033	0.034	0.024	0.039	0.039
			γ_2	−0.024	0.083	0.084	−0.004	0.088	0.088	−0.018	0.107	0.107
				0.021	0.087	0.087	0.049	0.093	0.095	0.041	0.111	0.113
			β	0.003	0.013	0.013	0.007	0.015	0.015	0.002	0.016	0.016
				0.002	0.013	0.013	0.005	0.015	0.015	0.000	0.016	0.016
		3	γ_1	−0.004	0.029	0.029	−0.004	0.030	0.030	−0.007	0.048	0.048
				0.002	0.030	0.030	0.009	0.033	0.033	0.015	0.053	0.053
			γ_2	−0.050	0.237	0.239	−0.080	0.312	0.318	−0.134	0.432	0.450
				0.111	0.260	0.273	0.142	0.361	0.381	0.167	0.491	0.519
			β	−0.003	0.039	0.039	0.024	0.051	0.052	0.013	0.071	0.071
				−0.007	0.039	0.039	0.017	0.051	0.052	0.000	0.071	0.071
2	200	1	γ_1	0.004	0.040	0.040	0.020	0.045	0.045	−0.016	0.060	0.060
				0.017	0.040	0.040	0.058	0.047	0.050	0.083	0.079	0.086
			γ_2	0.001	0.039	0.039	−0.022	0.042	0.043	−0.027	0.055	0.056
				0.016	0.040	0.040	0.008	0.047	0.047	0.029	0.072	0.073
			β	0.006	0.011	0.011	0.000	0.014	0.014	0.011	0.015	0.015
				0.005	0.011	0.011	−0.002	0.014	0.014	0.004	0.016	0.016
		3	γ_1	−0.016	0.071	0.071	−0.057	0.065	0.068	−0.139	0.083	0.102
				0.029	0.092	0.092	0.051	0.119	0.121	0.024	0.175	0.176
			γ_2	−0.468	0.723	0.942	−0.943	0.823	1.713	−1.348	0.829	2.646
				0.364	0.926	1.058	0.495	1.453	1.698	0.596	2.128	2.482
			β	0.017	0.035	0.035	0.022	0.039	0.039	0.036	0.052	0.054
				0.014	0.035	0.035	0.017	0.040	0.040	0.025	0.053	0.054
	400	1	γ_1	0.011	0.019	0.019	0.019	0.023	0.023	0.002	0.032	0.032
				0.018	0.019	0.019	0.037	0.023	0.025	0.047	0.034	0.036
			γ_2	−0.002	0.018	0.018	−0.010	0.023	0.023	−0.019	0.027	0.028
				0.009	0.018	0.018	0.007	0.025	0.025	0.008	0.032	0.032
			β	0.000	0.006	0.006	0.004	0.006	0.006	0.003	0.008	0.008
				0.000	0.006	0.006	0.002	0.006	0.006	0.000	0.008	0.008
		3	γ_1	−0.015	0.031	0.031	−0.071	0.034	0.039	−0.086	0.041	0.048
				0.014	0.037	0.038	0.001	0.050	0.050	0.047	0.072	0.074
			γ_2	−0.444	0.330	0.527	−0.802	0.410	1.053	−1.191	0.463	1.881
				0.149	0.364	0.386	0.244	0.557	0.616	0.325	0.739	0.845
			β	0.007	0.016	0.016	0.015	0.019	0.020	0.017	0.024	0.024
				0.004	0.016	0.016	0.010	0.019	0.020	0.010	0.024	0.024

Table 4. Bias, variance and MSE of $\hat{\gamma}$ for `smcure` (second rows) and our approach (first rows) in Model 3.

				Cens. level 1			Cens. level 2			Cens. level 3		
Mod.	n	scen.	Par.	Bias	Var.	MSE	Bias	Var.	MSE	Bias	Var.	MSE
3	200	1	γ_1	0.025	0.147	0.147	0.010	0.192	0.192	−0.008	0.243	0.243
				0.034	0.147	0.148	0.034	0.191	0.192	0.062	0.249	0.253
			γ_2	−0.045	0.042	0.044	−0.078	0.049	0.055	−0.085	0.059	0.066
				−0.077	0.050	0.056	−0.122	0.065	0.080	−0.148	0.092	0.144
			γ_3	0.081	0.366	0.373	0.074	0.485	0.491	0.029	0.536	0.537
				0.174	0.397	0.427	0.266	0.574	0.644	0.309	0.799	0.895
			γ_4	−0.046	0.326	0.373	−0.160	0.412	0.437	−0.289	0.453	0.537
				0.087	0.366	0.374	0.089	0.528	0.535	0.186	0.908	0.943
		3	γ_1	−0.059	0.161	0.164	−0.091	0.258	0.266	−0.223	0.419	0.468
				−0.053	0.163	0.166	−0.071	0.261	0.266	−0.138	0.524	0.543
			γ_2	0.018	0.046	0.046	0.026	0.063	0.064	0.086	0.088	0.096
				0.080	0.052	0.058	0.121	0.080	0.095	0.252	0.170	0.233
			γ_3	0.060	0.235	0.238	0.076	0.366	0.372	0.135	0.517	0.535
				0.091	0.242	0.251	0.135	0.375	0.393	0.228	0.642	0.694
			γ_4	−0.030	0.202	0.203	−0.040	0.292	0.293	−0.081	0.479	0.486
				−0.027	0.205	0.205	−0.017	0.277	0.277	−0.037	0.534	0.535
	400	1	γ_1	0.016	0.074	0.074	0.021	0.091	0.092	0.003	0.128	0.128
				0.017	0.072	0.073	0.022	0.082	0.082	0.023	0.108	0.108
			γ_2	−0.026	0.019	0.019	−0.039	0.023	0.025	−0.070	0.032	0.037
				−0.042	0.020	0.021	−0.049	0.025	0.027	−0.081	0.035	0.041
			γ_3	0.039	0.194	0.195	0.028	0.219	0.220	0.026	0.298	0.298
				0.093	0.190	0.198	0.097	0.206	0.215	0.158	0.297	0.322
			γ_4	0.010	0.171	0.171	−0.091	0.193	0.201	−0.178	0.276	0.307
				0.070	0.177	0.182	0.038	0.198	0.200	0.088	0.289	0.297
		3	γ_1	−0.023	0.089	0.089	−0.051	0.118	0.121	−0.124	0.212	0.228
				−0.029	0.092	0.093	−0.032	0.112	0.113	−0.062	0.200	0.204
			γ_2	0.003	0.023	0.023	0.006	0.033	0.033	0.042	0.048	0.050
				0.042	0.023	0.025	0.057	0.034	0.037	0.104	0.055	0.066
			γ_3	0.010	0.113	0.113	0.042	0.166	0.168	0.090	0.276	0.284
				0.039	0.111	0.111	0.060	0.152	0.156	0.108	0.250	0.262
			γ_4	0.012	0.110	0.110	−0.021	0.131	0.131	−0.047	0.220	0.223
				0.014	0.111	0.111	−0.018	0.117	0.118	−0.020	0.183	0.183

the estimators, compared to the cross-validation bandwidth on the whole interval $[0,\tau]$, in situations with a large percentage of observations in the plateau, while it leads to little difference otherwise.

Simulations show that, for not large sample size, the new method performs better than `smcure` for estimation of γ_0, mostly because of a smaller variance. As the sample size increases, they tend to behave quite similarly. On the other hand, both methods give almost the same estimates for β_0 and Λ. The most favorable situation for our method is when there is little censoring among uncured observations and the censored uncured observations are in the region of covariates that corresponds to higher cure rate. This comes from the fact that the nonparametric estimator in (9) takes larger values when the product has more terms equal to one. This should not be a problem when we expect that subjects with high probability of being cured correspond to longer survival times, meaning that it is more probable for them to be censored compared to those with small cure probability and shorter survival times.

This is indeed the case in Model 1 and we observe that our approach outperforms `smcure` in all the scenarios. The difference between the two is more marked when n is small and the absolute value of the γ coefficient is larger. In Model 2, the situation is more difficult because censoring depends on the covariate in such a way that, the non-cured subjects have the same probability of being censored independently of their cure probability. However, for the first two scenarios the new method is still

Table 5. Bias, variance and MSE of $\hat{\gamma}$ for `smcure` (second rows) and our approach (first rows) in Model 4.

				Cens. level 1			Cens. level 2			Cens. level 3		
Mod.	n	scen.	Par.	Bias	Var.	MSE	Bias	Var.	MSE	Bias	Var.	MSE
4	200	1	γ_1	0.041	0.157	0.159	0.016	0.187	0.188	−0.010	0.210	0.210
				0.077	0.178	0.184	0.096	0.228	0.238	0.127	0.285	0.301
			γ_2	−0.017	0.039	0.039	−0.019	0.042	0.042	−0.015	0.049	0.049
				−0.090	0.052	0.060	−0.125	0.069	0.085	−0.164	0.108	0.135
			γ_3	−0.245	0.159	0.219	−0.281	0.165	0.244	−0.355	0.179	0.305
				0.064	0.244	0.249	0.084	0.304	0.311	0.114	0.395	0.408
			γ_4	−0.068	0.331	0.336	−0.162	0.385	0.411	−0.285	0.443	0.524
				0.171	0.401	0.430	0.241	0.561	0.619	0.314	0.842	0.941
			γ_5	−0.095	0.301	0.310	−0.234	0.349	0.404	−0.366	0.371	0.505
				0.106	0.363	0.375	0.143	0.509	0.529	0.177	0.693	0.724
		3	γ_1	−0.044	0.079	0.081	−0.079	0.095	0.101	−0.148	0.132	0.154
				0.000	0.079	0.079	0.003	0.096	0.096	0.009	0.141	0.141
			γ_2	0.018	0.007	0.008	0.024	0.008	0.009	0.041	0.010	0.012
				0.015	0.008	0.008	0.017	0.009	0.010	0.028	0.013	0.014
			γ_3	0.034	0.066	0.067	0.046	0.073	0.075	0.067	0.087	0.091
				−0.025	0.080	0.080	−0.033	0.091	0.092	−0.041	0.120	0.122
			γ_4	0.041	0.102	0.104	0.054	0.125	0.128	0.082	0.166	0.173
				0.022	0.100	0.101	0.023	0.126	0.126	0.026	0.179	0.180
			γ_5	−0.031	0.103	0.104	−0.034	0.120	0.121	−0.054	0.159	0.162
				−0.016	0.099	0.099	−0.013	0.115	0.115	−0.018	0.150	0.151
	400	1	γ_1	0.013	0.067	0.067	0.015	0.079	0.080	0.005	0.097	0.097
				0.024	0.067	0.068	0.037	0.080	0.082	0.043	0.101	0.103
			γ_2	−0.001	0.017	0.017	−0.003	0.020	0.020	−0.007	0.023	0.023
				−0.042	0.020	0.021	−0.055	0.025	0.028	−0.079	0.034	0.041
			γ_3	−0.229	0.089	0.141	−0.207	0.090	0.133	−0.275	0.102	0.178
				0.046	0.107	0.109	0.061	0.137	0.141	0.066	0.162	0.166
			γ_4	−0.063	0.161	0.165	−0.143	0.175	0.196	−0.222	0.215	0.265
				0.085	0.176	0.183	0.107	0.222	0.234	0.145	0.318	0.339
			γ_5	−0.075	0.146	0.151	−0.192	0.177	0.214	−0.299	0.199	0.289
				0.043	0.157	0.159	0.049	0.194	0.196	0.060	0.253	0.257
		3	γ_1	−0.024	0.038	0.039	−0.038	0.047	0.048	−0.092	0.064	0.073
				−0.003	0.036	0.036	0.002	0.044	0.044	0.004	0.060	0.060
			γ_2	0.006	0.003	0.003	0.010	0.004	0.004	0.019	0.005	0.006
				0.005	0.004	0.004	0.005	0.004	0.004	0.006	0.006	0.006
			γ_3	0.028	0.033	0.034	0.045	0.038	0.040	0.065	0.045	0.049
				−0.010	0.036	0.036	−0.008	0.042	0.042	−0.008	0.052	0.052
			γ_4	0.021	0.052	0.053	0.032	0.062	0.063	0.060	0.088	0.092
				0.015	0.050	0.050	0.016	0.060	0.060	0.019	0.083	0.083
			γ_5	−0.024	0.049	0.050	−0.041	0.059	0.061	−0.048	0.077	0.079
				−0.017	0.049	0.049	−0.022	0.056	0.057	−0.022	0.069	0.069

superior. The third scenario is more problematic because the cure probability drops very fast from almost one to almost zero, resulting in a large fraction of uncured observations with almost zero cure probability. The presence of censoring in this region leads to overestimation of the cure rate. If we would take $\beta_C = 0.1$ (meaning larger probability of being censored for higher cure rate), then the new approach is significantly superior (see Table 6 for $n = 400$ and scenario 3). In Model 3, complications arise because of the presence of different covariates for the incidence and latency. Hence, subjects with higher cure rate might correspond to shorter survival times. As a result, the previous problem might still happen and its effects are more visible for large sample size and large censoring rate. Finally, Model 4 suggests that, even though the assumptions in Section 5 were shown to be satisfied only for

Table 6. Bias, variance and MSE of $\hat{\gamma}$ and $\hat{\beta}$ for `smcure` and our approach in Model 2, scenario 3 when $\beta_C = 0.1$ and $n = 400$.

	`smcure` package			Our approach		
Parameter	Bias	Var	MSE	Bias	Var	MSE
γ_1	0.014	0.123	0.123	−0.058	0.103	0.106
γ_2	0.418	1.243	1.418	−0.535	0.652	0.937
β	0.001	0.025	0.025	0.001	0.027	0.027

one continuous covariate, the method could be applied in more general cases. We noticed that, when a continuous covariate affects only the incidence and not the latency, the bandwidth selected by the `np` package is often very large, meaning that it fails to capture the effect of this covariate on the conditional distribution function. In those cases, we truncate the selected bandwidth from above at 2. Note that the bandwidth is chosen for standardized covariates so the truncation level can be fixed regardless of the distribution of the covariate. We decided to truncate at 2 since it seems to be a kind of boundary for a 'reasonable' bandwidth with standardized covariates (we do not want to externally affect chosen bandwidths smaller than 2 but we only replace extremely large values by 2). However, even when reasonable, the `np` bandwidth for X_2 seems to be larger than it should, resulting in more bias in the estimator of γ_3. Nevertheless, in terms of mean squared error, the method performs well for not large sample size. If X_2 would affect also the latency, the selected bandwidth would be more adequate and there would be no bias problems.

To conclude, the new approach seems to perform significantly better than `smcure` when the sample size is not large and the fraction of censored observations is not much higher than the expected cure proportion. In other situations, both methods are comparable. However, one has to be more careful when there is no reason to expect that the censored subjects correspond to higher cure probabilities.

In the previous settings, we truncated the event times at τ_0 in such a way that condition (19) is satisfied but in practice it is unlikely to observe event times at τ_0. Next, we consider one additional model for which condition (19) is not satisfied. The covariates and the parameters are as in Model 3 described above, but the event times are generated from a Weibull distribution on $[0, \tau_0]$ with $\tau_0 = 15$, i.e.

$$S_u(t|z) = \frac{\exp\{-\mu t^\rho \exp(\beta' z)\} - \exp\{-\mu \tau_0^\rho \exp(\beta' z)\}}{1 - \exp\{-\mu \tau_0^\rho \exp(\beta' z)\}}$$

The censoring times are exponentially distributed as in Model 3 and truncated at $\tau = 20$. Results for sample size $n = 200$ and three censoring levels are shown in Table 7. Compared to Model 3 above, we observe that, when condition (19) is not satisfied, presmoothing is even more superior than the smcure estimator.

Finally we conclude with a remark about the computational aspect. The proposed approach is computationally more intensive than the MLE mainly because of the bandwidth selection through a cross-validation procedure. For example, for one iteration in Model 3 with sample size 200 and 400, `smcure` computes the estimates in 0.7 and 0.8 seconds respectively, while the new approach requires 4.1 and 23.5 seconds (with a Core i7-8665U CPU desktop). However, this seems still reasonable since the method is not meant for much larger sample sizes.

Table 7. Bias, variance and MSE of $\hat{\gamma}$ for `smcure` (second rows) and our approach (first rows) in Model 3 without condition (19).

	Cens. level 1			Cens. level 2			Cens. level 3		
Par.	Bias	Var.	MSE	Bias	Var.	MSE	Bias	Var.	MSE
γ_1	0.015	0.152	0.152	0.000	0.196	0.196	−0.032	0.246	0.247
	0.017	0.150	0.151	0.027	0.193	0.194	0.035	0.260	0.262
γ_2	−0.054	0.044	0.047	−0.077	0.052	0.058	−0.109	0.064	0.076
	−0.085	0.050	0.057	−0.119	0.069	0.083	−0.171	0.101	0.130
γ_3	0.087	0.379	0.386	0.073	0.450	0.456	0.045	0.578	0.580
	0.197	0.423	0.462	0.249	0.561	0.623	0.343	0.885	1.002
γ_4	−0.010	0.339	0.339	−0.106	0.373	0.385	−0.228	0.498	0.550
	0.125	0.364	0.380	0.156	0.523	0.548	0.260	1.513	1.581

7. Application: Melanoma study

To illustrate the practical performance, we apply the proposed estimation procedure to two medical datasets for patients with melanoma and compare the results with `smcure`. Melanoma is the third most common skin cancer type with overall incidence rate 21.8 out of 100,000 people in the US (Cancer statistics from the Center for Disease Control and Prevention) and according to the American Cancer Society, 6850 people are expected to die of melanoma in 2020. However, in the recent years, the chances of survival for melanoma patients have increased due to earlier diagnosis and improvement of treatment and surgical techniques. The 5-year survival rates based on the stage of the cancer when it was first diagnosed are 92% for localized, 65% for regional and 25% for distant stage. It is also known that this disease is more common among white people and the death rate is higher for men than women. Even though most melanoma patients are cured by their initial treatment, it is not possible to distinguish them from the uncured patients. Hence, accurately estimating the probability of being cured is important in order to plan further treatment and prevent recurrence of uncured patients.

7.1. Eastern cooperative oncology group (ECOG) data

We use the melanoma data (ECOG phase III clinical trial e1684) from the `smcure` package [7] in order to compare our results with those of `smcure`. The purpose of this study was to evaluate the effect of treatment (high dose interferon alpha-2b regimen) as the postoperative adjuvant therapy. 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). The data consists of 284 observations (after deleting missing data) out of which 196 had recurrence of the melanoma cancer (around 30% censoring). The Kaplan-Meier curve is shown in Figure 1. The parameter estimates, standard errors and corresponding p-values for the Wald test using our method and the `smcure` package are given in Table 8. Standard errors are computed through 500 naive bootstrap samples.

We observe that, for both methods, the effects of the covariates have the same direction. Only the intercept was found significant for the incidence with `smcure`, while our method concludes that also age and treatment are significant. In particular, the probability of recurring melanoma is higher for the control group compared to the treatment group. This seems to be indeed the case if we look at the Kaplan Meier survival curves for the two groups in Figure 1. On the other hand, both methods agree that none of the covariates is significant for the latency.

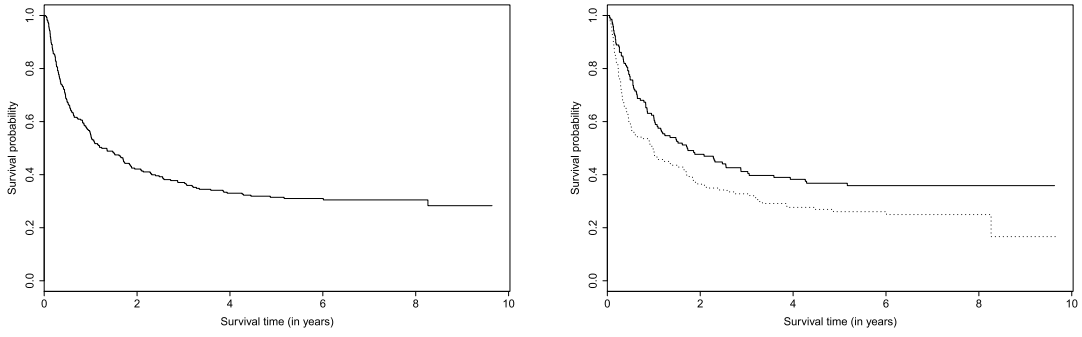

Figure 1. Left panel: Kaplan-Meier survival curve for ECOG data. Right panel: Kaplan-Meier survival curves for the treatment group (solid) and control group (dotted) in the ECOG data.

To illustrate another advantage of the new approach, we also compute the maximum likelihood estimator with the `smcure` package for different choices of the latency model. We see in Table 9 that the estimators of the incidence component (and their significance) change depending on which variables are included in the latency. On the other hand, the new method does not suffer from this problem because it estimates the incidence independently of the latency.

7.2. Surveillance, epidemiology and end results (SEER) database

The SEER database collects cancer incidence data from population-based cancer registries in US. These data consist of patient demographic characteristics, primary tumor site, tumor morphology, stage at diagnosis, length of follow up and vital status. We select the database 'Incidence—SEER 18 Regs Research Data' and extract the melanoma cancer data for the county of San Francisco in California during the period 2004 – 2015. We consider only patients with stage at diagnosis: localized, regional and distant and exclude those with unknown or zero follow-up time and restrict the study to white people because of the very small number of cases from other races. The event time is death because of melanoma. This cohort consists of 1445 melanoma cases out of which 596 are female and 849 male. The age ranges from 11 to 101 years old, the follow-up from 1 to 155 months. For most of the patients the cancer has been diagnosed at early stage (localized), while for 101 of them the stage at diagnosis is 'regional' and only for 42 it is 'distant'. We aim at evaluating how age, gender and stage at diagnosis affect the survival of melanoma patients in this cohort.

Table 8. Results for the incidence (logistic component) and the latency (Cox PH component) from the ECOG data.

		`smcure` package			Our approach		
	Covariates	Estimates	SE	p-value	Estimates	SE	p-value
incidence	Intercept	1.3649	0.3457	$8 \cdot 10^{-5}$	1.6697	0.3415	10^{-6}
	Age	0.0203	0.0159	0.2029	0.0220	0.0104	0.0344
	Gender	−0.0869	0.3347	0.7949	−0.3039	0.3448	0.3493
	Treatment	−0.5884	0.3706	0.1123	−0.9345	0.3603	0.0095
latency	Age	−0.0077	0.0069	0.2663	−0.0079	0.0060	0.1861
	Gender	0.0994	0.1932	0.6067	0.1240	0.1653	0.4534
	Treatment	−0.1535	0.1715	0.3707	−0.0947	0.1692	0.5756

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

		Model 1			Model 2			Model 3		
	Covariates	Estimates	SE	p-value	Estimates	SE	p-value	Estimates	SE	p-value
incidence	Intercept	1.3507	0.3001	$7 \cdot 10^{-6}$	1.4148	0.3213	10^{-5}	1.4181	0.3073	$4 \cdot 10^{-6}$
	Age	0.0164	0.0125	0.1905	0.0205	0.0154	0.1803	0.0209	0.0146	0.1528
	Gender	−0.0265	0.3113	0.9320	−0.0673	0.3352	0.8407	−0.0222	0.3130	0.9432
	Treatment	−0.6060	0.3509	0.0842	−0.6773	0.3223	0.0415	−0.6913	0.3439	0.0444
latency	Age				−0.0074	0.0066	0.2568	−0.0073	0.0064	0.2579
	Gender				0.0789	0.1863	0.6719			
	Treatment	−0.1324	0.1561	0.3963						

The use of cure models is justified from the presence of a long plateau containing around 20% of the observations (see the Kaplan-Meier curve in Figure 2). Moreover, the Kaplan-Meier curves depending on gender and stage at diagnosis in Figure 2 confirm that gender and stage affect the cure rate.

We checked the fit of the logistic model by comparing it with the single-index mixture cure model proposed in [2] through the prediction error of the incidence. More precisely, as in [2], we divide the data into a training set and a test set of size 964 and 481 respectively. Using the training set, we estimate

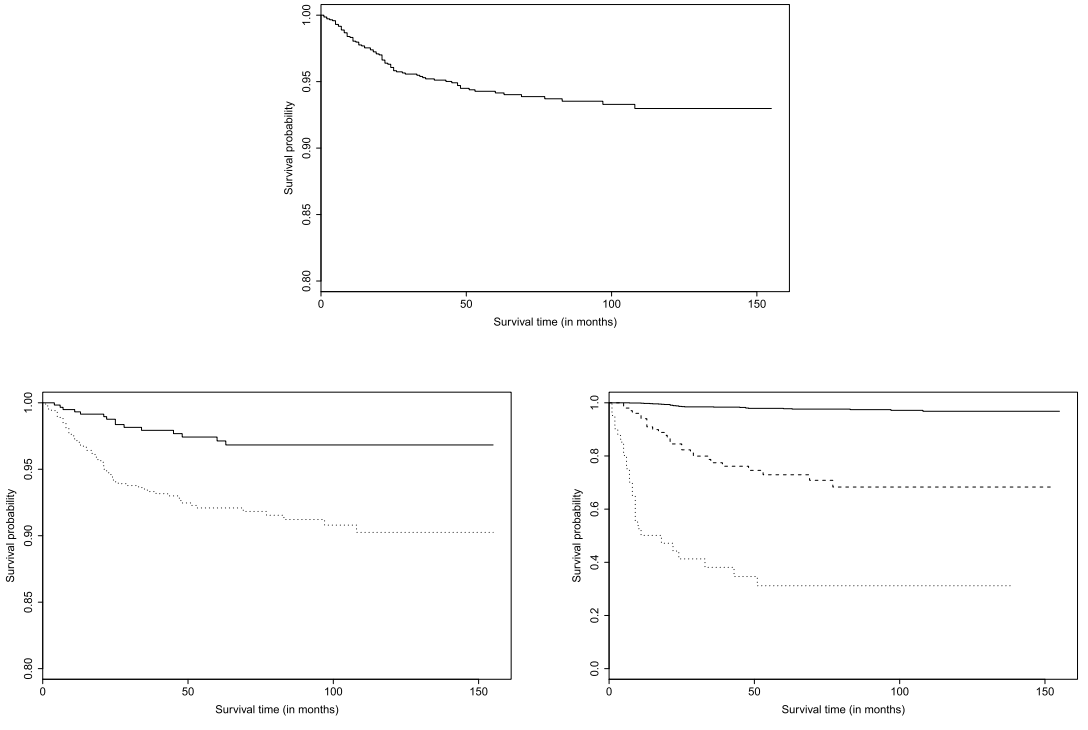

Figure 2. Upper panel: Kaplan-Meier survival curves for SEER data. Left panel: group division based on gender, females (solid) and males (dotted). Right panel: group division based on cancer stage at diagnosis, localized (solid), regional (dashed) and distant (dotted).

Table 10. Results for the incidence (logistic component) and the latency (Cox PH component) from the SEER data.

		`smcure` package			Our approach		
	Covariates	Estimates	SE	p-value	Estimates	SE	p-value
incidence	Intercept	−4.2071	0.3817	0	−4.2436	0.3980	0
	Age	0.0304	0.0122	0.0124	0.0328	0.0172	0.0565
	Gender	1.1318	0.4211	0.0072	1.2341	0.4792	0.010
	S_1	2.6738	0.3702	$5 \cdot 10^{-13}$	2.4474	0.4247	$8 \cdot 10^{-9}$
	S_2	4.0763	0.5067	$8 \cdot 10^{-16}$	3.9426	0.4536	0
latency	Age	−0.0139	0.0098	0.1577	−0.0143	0.0106	0.1756
	Gender	−0.0549	0.4065	0.8925	−0.0871	0.3687	0.8131
	S_1	0.5176	0.3993	0.1949	0.6130	0.3971	0.1226
	S_2	1.8039	0.4529	$7 \cdot 10^{-5}$	1.8623	0.5072	0.0002

the logistic/Cox model and the single-index/Cox model. Afterwards, we compute the prediction error in the test set given by

$$PE = -\sum_{j=1}^{481} \left\{ \hat{W}_j \log[1 - \hat{\pi}(X_j^{\text{test}})] + (1 - \hat{W}_j) \log \hat{\pi}(X_j^{\text{test}}) \right\}$$

where $\hat{\pi}(X_j^{\text{test}})$ and $\hat{W}_j$ are the predicted cure probability and the predicted weight for the jth observation in the test set, computed based on the parameter estimates (and the link function for the single-index model) in the training set. More precisely, for the logistic/Cox model we have $\hat{\pi}(X_j^{\text{test}}) = \phi(\hat{\gamma}_n, X_j^{\text{test}})$ and

$$\hat{W}_j = \Delta_j^{\text{test}} + (1 - \Delta_j^{\text{test}}) \frac{\hat{\pi}(X_j^{\text{test}}) \exp\left(-\hat{\Lambda}_n(Y_j^{\text{test}}) e^{\hat{\beta}_n' Z_j^{\text{test}}}\right)}{1 - \hat{\pi}(X_j^{\text{test}}) + \hat{\pi}(X_j^{\text{test}}) \exp\left(-\hat{\Lambda}_n(Y_j^{\text{test}}) e^{\hat{\beta}_n' Z_j^{\text{test}}}\right)}$$

where $\hat{\gamma}_n$, $\hat{\beta}_n$ and $\hat{\Lambda}_n$ are the estimated parameters and the estimated hazard function in the training set. For the single-index/Cox model, the only difference is that $\hat{\pi}(X_j^{\text{test}}) = \hat{g}_n(\hat{\gamma}_n, X_j^{\text{test}})$ where $\hat{g}_n$ is the estimated link function as in [2]. The weights $\hat{W}_j$ correspond to the conditional expectation of the cure status B given the observations. We find that the prediction error for the logistic model is 98.53, whereas for the single-index model it is 156.55. This means that the logistic model performs better.

We also performed the test for the logistic model proposed in [19], which is based on the distance between a parametric and a nonparametric estimator of the cure probability. The p-value, estimated through a bootstrap procedure as described in [19], was 0.91. The bandwidth was chosen using the np package but even for a range of bandwidths around that value we obtain p-values larger than 0.6. Hence the logistic assumption for the incidence model seems reasonable.

The parameter estimates, standard errors and corresponding p-values for the Wald test using our method and the `smcure` package are given in Table 10. Standard errors are computed with 500 naive bootstrap samples. The covariate stage is classified using two dummy Bernoulli variables S_1 and S_2, where $S_1 = 1$ indicates the regional stage and $S_2 = 1$ indicates the distant stage. The gender variable is equal to zero for females and one for males. We observe that both methods agree that all the considered covariates are significant for the incidence (with age being a borderline case for our approach). For the latency, only being in the distant stage is found significant with both methods. Moreover, again the effects of all the covariates on the latency and incidence have the same direction for both methods.

8. Discussion

In this paper we proposed a new estimation procedure for the mixture cure model with a parametric form of the incidence (for example logistic) and any semiparametric model for the latency. We investigated more in detail the logistic/Cox model given its practical relevance. Instead of using an iterative algorithm for dealing with the unknown cure status, this method relies on a preliminary nonparametric estimator of the cure probabilities. We showed through simulations that the new approach improves upon the classical maximum likelihood estimator implemented in the package `smcure`, mainly for smaller sample sizes. For the latency, both methods behave similarly. Hence, it is of particular interest in situations in which the focus is on the estimation of cure probabilities. The real data application on the ECOG clinical trial also showed that the improvement in estimation can be meaningful in practice and help detecting significant effects.

The proposed method has the advantage of direct estimation of the incidence component, without relying on the latency, which makes it robust to latency model misspecification. On the contrary, the `smcure` estimator strongly depends on the choice of the variables for the latency and could be biased for a misspecified Cox model. Hence, for practical reason, confronting the estimators obtained with the two methods is valuable for confirming the results or obtaining new insights. From the theoretical point of view, unlike the standard maximum likelihood estimation, presmoothing allows us to obtain consistency and asymptotic normality without requiring the 'unrealistic' assumption that the distribution of uncured subjects has a positive mass at the end point of the support.

It might be argued that since the proposed method relies on smoothing, it is more complex and the results can be affected by the choice of the kernel function or the bandwidth. Our purpose was to show that the user doesn't have to think about this because the standard choices proposed in this paper perform well in practice. In addition, since the final estimator is a parametric one and the kernel estimator is only a preliminary step of the procedure, the results would anyway be more stable with respect to these choices than in a nonparametric setting. The main challenge this method faces is extension to many continuous covariates for the incidence. We did not deeply investigate such situations since, in that case, multiple bandwidths have to be chosen, which can be more problematic and computationally intensive. However, our approach based on presmoothing allows to efficiently handle these situations if the estimator $\hat{\pi}$ is constructed in a more adequate way. One possibility would be to construct the estimator assuming a single-index model for the latency, which is reasonable since the final goal is a parametric estimator. With this approach one can avoid the choice of multiple bandwidths and perform the estimation as in the one dimensional case. However, this problem will be addressed by future research. In this regard, even though considering only one continuous covariate might seem restrictive in practice, the proposed procedure constitutes the basis for further developments of new estimators for general dimension scenarios that do not require multidimensional smoothing.

Appendix

A.1. Proof of Theorem 4

We obtain the asymptotic normality of $\hat{\Lambda}_n$, $\hat{\beta}_n$ following the proof of Theorem 3 in [18]. In order to work with a one-dimensional submodel, for d in a neighbourhood of the origin, let $\Lambda_d(t) = \int_0^t \{1 + dh_1(s)\}\mathrm{d}\hat{\Lambda}_n(s)$ and $\beta_d = dh_2 + \hat{\beta}_n$, where h_1 is a function of bounded variation on $[0, \tau_0]$ and h_2 is a q-dimensional real vector. Let $\dot{S}_n(\hat{\Lambda}_n, \hat{\beta}_n)(h_1, h_2)$ denote the derivative of $\hat{l}_n(\Lambda_d, \beta_d)$ (defined in (14))

with respect to d and evaluated at $d = 0$. We have

$$\hat{S}_n(\hat{\Lambda}_n,\hat{\beta}_n)(h_1,h_2) = \frac{1}{n}\sum_{i=1}^{n}\Delta_i \mathbb{1}_{\{Y_i<\tau_0\}}\left[h_1(Y_i)+h_2'Z_i\right] - \frac{1}{n}\sum_{i=1}^{n}\left\{\Delta_i+(1-\Delta_i)\mathbb{1}_{\{Y_i\le\tau_0\}}g_i(Y_i,\hat{\Lambda}_n,\hat{\beta}_n,\hat{\gamma}_n)\right\} \times\left\{e^{\hat{\beta}_n'Z_i}\int_0^{Y_i}h_1(s)\mathrm{d}\hat{\Lambda}_n(s)+e^{\hat{\beta}_n'Z_i}\hat{\Lambda}_n(Y_i)h_2'Z_i\right\},$$

where g_j is defined in (17) and $\hat{\gamma}_n$ is the maximizer of (10). Let $\Upsilon_n = (\hat{\Lambda}_n,\hat{\beta}_n)$ and $\Upsilon_0 = (\Lambda_0,\beta_0)$. Furthermore, denote by S the asymptotic version of $\hat{S}_n$:

$$S(\Lambda,\beta)(h_1,h_2) = \mathbb{E}\Bigg[\Delta\mathbb{1}_{\{Y<\tau_0\}}\{h_1(Y)+h_2'Z\} - \left\{\Delta+(1-\Delta)\mathbb{1}_{\{Y\le\tau_0\}}g(Y,\Lambda,\beta,\gamma_0)\right\} \times\left\{e^{\beta'Z}\int_0^{Y}h_1(s)\mathrm{d}\Lambda(s)+e^{\beta'Z}\Lambda(Y)h_2'Z\right\}\Bigg].$$

We have $\hat{S}_n(\Upsilon_n) = 0$ and $S(\Upsilon_0) = 0$. The score function S_n and S are respectively a random and a deterministic map from Ξ to $l^\infty(\mathcal{H}_\mathfrak{m})$ (the space of bounded real-valued functions on $\mathcal{H}_\mathfrak{m}$), where

$$\Xi = \left\{(\Lambda,\beta) : \sup_{h\in\mathcal{H}_\mathfrak{m}}\left|\int_0^{\tau_0}h_1(s)\mathrm{d}\Lambda(s)+h_2'\beta\right| < \infty\right\}$$

and $\mathcal{H}_\mathfrak{m} = \{h\in\mathcal{H} : \|h\|_H \le \mathfrak{m}\}$. Here $\|h\|_H = \|h_1\|_v + \|h_2\|_{L_1}$, $\|h_2\|_{L_1} = \sum_{j=1}^{q}|h_{2,j}|$, $\|h_1\|_v = |h_1(0)| + V_0^{\tau_0}(h_1)$ and $V_0^{\tau_0}(h_1)$ denotes the total variation of h_1 on $[0,\tau_0]$. This means that S_n is a random variable defined in the abstract probability space $(\Omega,\mathcal{F},\mathbb{P})$ (where the random vector (B,T_0,C,X,Z) is defined) with values in the space of bounded functions $\Xi\mapsto l^\infty(\mathcal{H}_\mathfrak{m})$ with respect to the supremum norm. The latter one is a Banach space equipped with the Borel σ-field.

We need to show that conditions 1-4 of Theorem 4 in [18] (or Theorem 3.3.1 in [28]) are satisfied. The main difference of the function S from the one in [18] is that here $\gamma = \gamma_0$ fixed. We are only considering variation with respect to β and not γ, so the components of h that correspond to γ are set to zero. However, conditions 2 and 3 of Theorem 4 in [18] for S can be shown in the same way as in [18]. Details about conditions 1 and 4 can be found in the online Supplementary Material [20]. □

A.2. Proof of Theorem 5

The logistic model for the cure probability obviously satisfies assumptions (AN1) and (AN3). Let Π be the space of continuously differentiable functions f from $\mathcal{X}$ to $[0,1]$ such that $\sup_{x\in\mathcal{X}}|f'(x)| \le M$ and

$$\sup_{x_1,x_2\in\mathcal{X}}\frac{|f'(x_1)-f'(x_2)|}{|x_1-x_2|^\xi} \le M$$

for some $M > 0$ and $\xi\in(0,1]$. If such space is equipped with the supremum norm, the covering numbers satisfy

$$\log N(\epsilon,\Pi,\|\cdot\|_\infty) \le K\frac{1}{\epsilon^{1/(1+\xi)}}$$

for some constant $K > 0$ independent of ϵ (see Theorem 2.7.1 in [28]). Obviously, for $\epsilon > 1$, $\log N(\epsilon, \Pi, \|\cdot\|_\infty) = 0$. Hence, assumption (AN2) is satisfied. It remains to check (AN4). Recall that the estimator of the cure probability $\hat{\pi}(x)$ is the value at time τ_0 of the Beran estimator $\hat{S}(t|x)$, while $\pi_0(x) = S(\tau_0|x)$. Moreover, by assumption (4), we have $\inf_x H((\tau_0, \infty)|x) > 0$. From Proposition 4.1 and 4.2 in [27] it follows that

$$\sup_x |\hat{\pi}(x) - \pi_0(x)| = O\left((nb)^{-1/2}(\log b^{-1})^{1/2}\right) \quad a.s.,$$

$$\sup_x \left|\hat{\pi}'(x) - \pi_0'(x)\right| = O\left((nb^3)^{-1/2}(\log b^{-1})^{1/2}\right) \quad a.s.$$

and

$$\sup_{x_1, x_2 \in \mathcal{X}} \frac{|\hat{\pi}'(x_1) - \pi_0'(x_1) - \hat{\pi}'(x_2) + \pi_0'(x_2)|}{|x_1 - x_2|^{\xi/2}} = O\left(\left(nb^{3+\xi}\right)^{-1/2}(\log b^{-1})^{1/2}\right) \quad a.s.,$$

where ξ is as in assumption (C1). Since π_0 is twice continuously differentiable, from assumption (C1) it follows that $\hat{\pi}$ satisfies (i,ii) of (AN4). From Theorem 3.2 of [10] (with $T = \tau_0$) we have $\hat{\pi}(x) - \pi_0(x) = \frac{1}{n}\sum_{i=1}^n A_i(x) + R_n(x)$, where

$$A_i(x) = -\frac{1 - \phi(\gamma_0, x)}{f_X(x)} \frac{1}{b} k\left(\frac{x - X_i}{b}\right) \left\{ \frac{\Delta_i \mathbb{1}_{\{Y_i \le \tau_0\}}}{H([Y_i, \infty)|x)} - \int_0^{Y_i \wedge \tau_0} \frac{H_1(ds|x)}{H^2([s, \infty)|x)} \right\} \tag{22}$$

and $\sup_x |R_n(x)| = O\left((nb)^{-3/4}(\log n)^{3/4}\right)$ a.s.. Hence

$$\begin{aligned}
&\mathbb{E}^*\left[(\hat{\pi}(X) - \pi_0(X))\left(\frac{1}{\phi(\gamma_0, X)} + \frac{1}{1 - \phi(\gamma_0, X)}\right)\nabla_\gamma \phi(\gamma_0, X)\right] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}^*\left[A_i(x)\left(\frac{1}{\phi(\gamma_0, X)} + \frac{1}{1 - \phi(\gamma_0, X)}\right)\nabla_\gamma \phi(\gamma_0, X)\right] \\
&\quad + \mathbb{E}^*\left[R_n(X)\left(\frac{1}{\phi(\gamma_0, X)} + \frac{1}{1 - \phi(\gamma_0, X)}\right)\nabla_\gamma \phi(\gamma_0, X)\right].
\end{aligned}$$

The second term on the right hand side of the previous display is bounded by $c \sup_x |R_n(x)| = o(n^{-1/2})$ for some $c > 0$ because of assumptions (C1) and (AN1). Furthermore, from (AN1) and (AC4) and a Taylor expansion, it follows that the generic element of the sum in the first term is equal to

$$\begin{aligned}
&-\int_{\mathcal{X}} \frac{1}{b} k\left(\frac{x - X_i}{b}\right)\left\{\frac{\Delta_i \mathbb{1}_{\{Y_i \le \tau_0\}}}{H([Y_i, \infty)|x)} - \int_0^{Y_i \wedge \tau_0} \frac{H_1(ds|x)}{H^2([s, \infty)|x)}\right\} \frac{1}{\phi(\gamma_0, x)} \nabla_\gamma \phi(\gamma_0, x)\, \mathrm{d}x \\
&= -\left\{\frac{\Delta_i \mathbb{1}_{\{Y_i \le \tau_0\}}}{H([Y_i, \infty)|X_i)} - \int_0^{Y_i \wedge \tau_0} \frac{H_1(ds|X_i)}{H^2([s, \infty)|X_i)}\right\} \frac{1}{\phi(\gamma_0, X_i)} \nabla_\gamma \phi(\gamma_0, X_i) + O(b^2).
\end{aligned}$$

Since because of (C1) we have $O(b^2) = o(n^{-1/2})$, (AN4-iii) holds with

$$\Psi(Y, \Delta, X) = -\left\{\frac{\Delta \mathbb{1}_{\{Y \le \tau_0\}}}{H([Y, \infty)|X)} - \int_0^{Y \wedge \tau_0} \frac{H_1(ds|X)}{H^2([s, \infty)|X)}\right\} \frac{1}{\phi(\gamma_0, X)} \nabla_\gamma \phi(\gamma_0, X).$$

□

Acknowledgements

I. Van Keilegom and E. Musta acknowledge financial support from the European Research Council (2016-2021, Horizon 2020 and grant agreement 694409). For the simulations we used the infrastructure of the Flemish Supercomputer Center (VSC).

Supplementary Material

Supplement to "A presmoothing approach for estimation in the semiparametric Cox mixture cure model" (DOI: 10.3150/21-BEJ1434SUPP; .pdf). It contains the proofs of Theorems 1, 2 and 3 in Section 5 and additional simulation results.

References

[1] Aerts, M., Hens, N. and Simonoff, J.S. (2010). Model selection in regression based on pre-smoothing. *J. Appl. Stat.* **37** 1455–1472. MR2758660 https://doi.org/10.1080/02664760903046086

[2] Amico, M., Van Keilegom, I. and Legrand, C. (2019). The single-index/Cox mixture cure model. *Biometrics* **75** 452–462. MR3999169 https://doi.org/10.1111/biom.12999

[3] Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10** 1100–1120. MR0673646

[4] Berkson, J. and Gage, R.P. (1952). Survival curve for cancer patients following treatment. *J. Amer. Statist. Assoc.* **47** 501–515.

[5] Boag, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. Roy. Statist. Soc. B* **11** 15–53.

[6] Burke, K. and Patilea, V. (2021). A likelihood-based approach for cure regression models. *TEST* **30** 693–712. MR4297274 https://doi.org/10.1007/s11749-020-00738-8

[7] Cai, C., Zou, Y., Peng, Y. and Zhang, J. (2012). smcure: An R-Package for estimating semiparametric mixture cure models. *Comput. Methods Programs Biomed.* **108** 1255–1260.

[8] Chen, X., Linton, O. and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* **71** 1591–1608. MR2000259 https://doi.org/10.1111/1468-0262.00461

[9] Cristóbal Cristóbal, J.A., Faraldo Roca, P. and González Manteiga, W. (1987). A class of linear regression parameter estimators constructed by nonparametric estimation. *Ann. Statist.* **15** 603–609. MR0888428 https://doi.org/10.1214/aos/1176350363

[10] Du, Y. and Akritas, M.G. (2002). Uniform strong representation of the conditional Kaplan-Meier process. *Math. Methods Statist.* **11** 152–182. MR1941314

[11] Farewell, V.T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 1041–1046.

[12] Ferraty, F., González-Manteiga, W., Martínez-Calvo, A. and Vieu, P. (2012). Presmoothing in functional linear regression. *Statist. Sinica* **22** 69–94. MR2933168 https://doi.org/10.5705/ss.2010.085

[13] Fleming, T.R. and Harrington, D.P. (2011). *Counting Processes and Survival Analysis* **169**. New York: Wiley.

[14] Han, X. (2017). *Statistical Methods for Analysis of Genetic and Survival Data with Latent Heterogeneity*. Ann Arbor, MI: ProQuest LLC. Thesis (Ph.D.)—New York University. MR3755013

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

[16] Lee, T.E., Fisher, D.O., Blomberg, S.P. and Wintle, B.A. (2017). Extinct or still out there? Disentangling influences on extinction and rediscovery helps to clarify the fate of species on the edge. *Glob. Change Biol.* **23** 621–634.

[17] Li, C.-S. and Taylor, J.M. (2002). A semi-parametric accelerated failure time cure model. *Stat. Med.* **21** 3235–3247.

[18] Lu, W. (2008). Maximum likelihood estimation in the proportional hazards cure model. *Ann. Inst. Statist. Math.* **60** 545–574. MR2434411 https://doi.org/10.1007/s10463-007-0120-x
[19] Müller, U.U. and Van Keilegom, I. (2019). Goodness-of-fit tests for the cure rate in a mixture cure model. *Biometrika* **106** 211–227. MR3912392 https://doi.org/10.1093/biomet/asy058
[20] Musta, E., Patilea, V. and Van Keilegom, I. (2022). Supplement to "A presmoothing approach for estimation in the semiparametric Cox mixture cure model." https://doi.org/10.3150/21-BEJ1434SUPP
[21] Patilea, V. and Van Keilegom, I. (2020). A general approach for cure models in survival analysis. *Ann. Statist.* **48** 2323–2346. MR4134797 https://doi.org/10.1214/19-AOS1889
[22] Peng, Y. and Dear, K.B. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics* **56** 237–243.
[23] Sposto, R. (2002). Cure model analysis in cancer: An application to data from the Children's Cancer Group. *Stat. Med.* **21** 293–312. https://doi.org/10.1002/sim.987
[24] Stringer, S., Denys, D., Kahn, R.S. and Derks, E.M. (2016). What cure models can teach us about genome-wide survival analysis. *Behav. Genet.* **46** 269–280.
[25] Sy, J.P. and Taylor, J.M.G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics* **56** 227–236. MR1767631 https://doi.org/10.1111/j.0006-341X.2000.00227.x
[26] Taylor, J.M. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics* **51** 899–907.
[27] Van Keilegom, I. and Akritas, M.G. (1999). Transfer of tail information in censored regression models. *Ann. Statist.* **27** 1745–1784. MR1742508 https://doi.org/10.1214/aos/1017939150
[28] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes, with Applications to Statistics*. *Springer Series in Statistics*. New York: Springer. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2
[29] Wycinka, E. and Jurkiewicz, T. (2017). Mixture cure models in prediction of time to default: Comparison with logit and Cox models. In *Contemporary Trends and Challenges in Finance* 221–231. Springer.
[30] Xu, J. and Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Canad. J. Statist.* **42** 1–17. MR3181580 https://doi.org/10.1002/cjs.11197
[31] Yamaguchi, K. (1992). Accelerated failure-time regression models with a regression model of surviving fraction: An application to the analysis of "permanent employment" in Japan. *J. Amer. Statist. Assoc.* **87** 284–292.
[32] Zhang, J. and Peng, Y. (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Stat. Med.* **26** 3157–3171. MR2380509 https://doi.org/10.1002/sim.2748

Received December 2020 and revised October 2021