Open access • Journal Article • DOI:10.1111/BIOM.12999

The single-index/Cox mixture cure model. — Source link

Maïlis Amico, Maïlis Amico, Ingrid Van Keilegom, Ingrid Van Keilegom ...+1 more authors

Institutions: Université catholique de Louvain, Katholieke Universiteit Leuven

Published on: 29 Mar 2019 - Biometrics (Biometrics)

Topics: Survival function, Proportional hazards model, Population, Logistic regression and Kernel smoother

Biometrics 63, 1–21 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2007

The Single-Index/Cox Mixture Cure Model

Maïlis Amico[1,2,*], Ingrid Van Keilegom[1,2,], and Catherine Legrand[2,***]**

[1]Research Center for Operations Research and Business Statistics, KU Leuven, Leuven, Belgium

[2]Institute of Statistics, Biostatistics and Actuarial Sciences, UCLouvain, Louvain-la-Neuve, Belgium

*_email:_ mailis.amico@kuleuven.be

**_email:_ ingrid.vankeilegom@kuleuven.be

***_email:_ catherine.legrand@uclouvain.be

Summary: In survival analysis it often happens that a certain fraction of the subjects under study never experience the event of interest, i.e. they are considered 'cured'. In the presence of covariates, a common model for this type of data is the mixture cure model, which assumes that the population consists of two subpopulations, namely the cured and the non-cured ones, and it writes the survival function of the whole population given a set of covariates as a mixture of the survival function of the cured subjects (which equals one), and the survival function of the non-cured ones. In the literature one usually assumes that the mixing probabilities follow a logistic model. This is however a strong modeling assumption, which might not be met in practice. Therefore, in order to have a flexible model which at the same time does not suffer from curse-of-dimensionality problems, we propose in this paper a single-index model for the mixing probabilities. For the survival function of the non-cured subjects we assume a Cox proportional hazards model. We estimate this model using a maximum likelihood approach. We also carry out a simulation study, in which we compare the estimators under the single-index model and under the logistic model for various model settings, and we apply the new model and estimation method on a breast cancer data set.

Key words: Cure models; EM algorithm; kernel smoothing; logistic model; proportional hazards model; survival analysis.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

In survival analysis one usually assumes that all subjects under study will eventually experience the event of interest. However, there are various situations for which this assumption is not realistic. For instance, when the event of interest is the time until a patient progresses or relapses from a certain disease, then patients who are cured from the disease will never experience the event. Likewise, when interest lies in the time until someone finds a new job or the time until an electronic component fails, the event of interest may never happen for some observations. Those observations will be considered as 'long-term survivors' or as 'cured', and their survival time will be set to infinity. Since it is not possible in practice to follow all individuals until they all have experienced the event of interest, a typical feature of survival data is (right-)censoring, meaning that for some individuals only a lower bound of the survival time is known. When survival data contain a fraction of long-term survivors, censored observations include uncured individuals for whom the event could not be observed, but also cured ones who will never experience the event. As a result, the cure status is unknown and survival data contain then a mix of cured and uncured individuals that are not distinguishable a priori. Cure models are survival models that have been developed to take this feature into account.

A typical field in which cure models are used is cancer studies. In clinical studies on breast cancer for example, we know that some patients will never relapse from the disease. Beside the contextual evidence for the presence of a cure fraction, the observation of a long plateau, with a non-negligible fraction of censored observations, that levels-off at a value greater than 0 in the Kaplan and Meier (1958) estimator of the survival function, is an indicator and a prerequisite for cure models, as explained by Sy and Taylor (2000). Web Figure 1, which represents this estimator for the time to distant metastasis for patients that experienced

a lymph-node negative breast cancer (Wang et al., 2005), gives a good example of cured survival data.

When information on covariates is present, a commonly used cure regression model is the mixture cure model. Originally proposed by Boag (1949) and Farewell (1982), it assumes that the survival function $S(t|\boldsymbol{x},\boldsymbol{z}) = P(T > t|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z})$ of the survival time T given a set of covariates $(\boldsymbol{X}^t, \boldsymbol{Z}^t)^t$ is given by

$$S(t|\boldsymbol{x},\boldsymbol{z}) = 1 - p(\boldsymbol{x}) + p(\boldsymbol{x})S_u(t|\boldsymbol{z}), \qquad t \geqslant 0, \tag{1}$$

where $p(\boldsymbol{x}) = P(B = 1|\boldsymbol{X} = \boldsymbol{x})$ is the conditional probability of being uncured (often referred to as the 'incidence') with $B = I(T < \infty)$ the latent uncured status, and $I(\cdot)$ the indicator function; and $S_u(t|\boldsymbol{z}) = P(T > t|B = 1, \boldsymbol{Z} = \boldsymbol{z})$ is the conditional survival function for the uncured subjects (often referred to as the 'latency'). Here, the vectors $\boldsymbol{X}$ and $\boldsymbol{Z}$ can contain (partially) the same covariates, but they can also be completely different.

In the literature various models have been considered for the latency. Parametric models are, for example, given in Boag (1949) and Farewell (1982). A semiparametric approach based on a Cox (1972) proportional hazards (PH) model is provided by e.g. Kuk and Chen (1992), Sy and Taylor (2000), and Lu (2008), whereas a completely nonparametric estimation approach for $S_u(t|\boldsymbol{z})$ is given in Patilea and Van Keilegom (2018). On the other hand, much less attention has been paid to the incidence and the modeling and estimation of the cure rate $1 - p(\boldsymbol{x})$. In fact, it is common practice to assume a logistic model, i.e. $p(\boldsymbol{x}) = \exp(\gamma_0 + \boldsymbol{\gamma}^t\boldsymbol{x})/\{1 + \exp(\gamma_0 + \boldsymbol{\gamma}^t\boldsymbol{x})\}$ for some parameter vector $\boldsymbol{\gamma}$ and an intercept γ_0. The logistic model has certainly a number of important qualities. For instance, it is easy to interpret, it is easy to estimate, and it is available in various statistical software packages. However, it also has an important drawback. Indeed, the logistic function $\exp(u)/\{1+\exp(u)\}$ is of a fixed known form, while there is no reason to constraint the cure rate $1-p(\boldsymbol{x})$ to have a S-shape. Furthermore, even if $p(\boldsymbol{x})$ is often monotone in practical applications, we might

think of situations where $p(\boldsymbol{x})$ would e.g. be increasing in $\boldsymbol{\gamma}^t\boldsymbol{x}$ up to some threshold, after which it would become decreasing. In order to accommodate these concerns, we propose in this paper a single-index model for $p(\boldsymbol{x})$, i.e. we assume that there exists an unknown link function $g(\cdot)$ (monotone or non-monotone) such that

$$p(\boldsymbol{x}) = g(\boldsymbol{\gamma}^t\boldsymbol{x}). \tag{2}$$

The link function g can be any (smooth) function with values between 0 and 1, and will be estimated non parametrically using kernel methods. As noted by a referee, the single-index model is a special case (when $M = 1$) of Projection Pursuit Regression (Friedman and Stuetzle, 1981), where $E(B|\boldsymbol{X} = \boldsymbol{x}) = \sum_{m=1}^{M} g_m(\omega_m^t\boldsymbol{x})$, with g_m, $m = 1, \ldots, M$, some unknown functions, and ω_m, $m = 1, \ldots, M$, some vectors of unknown parameters associates with $\boldsymbol{X}$. Considering the popular Cox PH model for the latency, and a single-index model for the incidence, we will refer to our model as the Single-Index/Cox (SIC) cure model.

Single-index models have been used in various contexts and have several advantages. First of all, as explained above, they are much more flexible than purely parametric models. Second, contrary to the completely nonparametric models for $p(\boldsymbol{x})$, they do not suffer from curse-of-dimensionality problems, since they summarize the covariate vector $\boldsymbol{X}$ into one single score $\boldsymbol{\gamma}^t\boldsymbol{X}$. Third, despite their nonparametric nature, they remain quite easy to interpret. Indeed, to compare the relative importance of one covariate with respect to another, it suffices to standardize the covariates and to compare the absolute value of the corresponding $\boldsymbol{\gamma}$-coefficients. We refer to Ichimura (1993) and Klein and Spady (1993) for some early references, to the book by Horowitz (2009) for a nice overview of existing results, and to Lu and Burke (2005), Wang et al. (2007), Lopez (2009), and Lopez et al. (2013) for papers which have used the single-index model in survival analysis.

This paper is organized as follows. In Section 2 we detail our proposed model, i.e. the SIC cure model, we provide results about the identifiability of the model and we define a

maximum likelihood based estimation procedure. The proof of the identifiability of the model is given in the Supplementary Materials. The finite sample performance of the proposed estimators is investigated through a numerical study in Section 3. It also contains a discussion on the issue of bandwidth selection for the kernel based non-parametric estimator of g. The breast cancer data introduced previously is analyzed with the SIC cure model in Section 4, and we end the paper with some concluding remarks in Section 5.

2. The Model and its Estimation

The survival time T is subject to random right censoring, i.e. instead of observing T we observe the couple (Y, Δ), where $Y = \min(T, C)$ is the follow-up time, $\Delta = I(T \leqslant C)$ is the censoring indicator, and C is the censoring time. As often, we assume that T and C are independent given the covariates $(\boldsymbol{X}^t, \boldsymbol{Z}^t)^t$. Let $(Y_i, \Delta_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$, $i = 1, \ldots, n$, be i.i.d. realizations of $(Y, \Delta, \boldsymbol{X}, \boldsymbol{Z})$.

Consider the mixture cure model (1) and assume a single-index structure (2) for $p(\boldsymbol{x})$, where $\boldsymbol{\gamma}^t = (\gamma_1, \ldots, \gamma_d)$ and d is the dimension of $\boldsymbol{X}$. For identifiability reasons we do not include an intercept in the model. For the latency, a Cox PH model is considered with survival function

$$S_u(t|\boldsymbol{z}) = S_0(t)^{\exp(\boldsymbol{\beta}^t \boldsymbol{z})}, \tag{3}$$

where $S_0(t) = P(T > t|B = 1)$ is the baseline conditional survival function, which is totally unspecified and which corresponds to the conditional survival function for $\boldsymbol{Z} = \boldsymbol{0}$, and $\boldsymbol{\beta}^t = (\beta_1, \ldots, \beta_q)$ is a vector of parameters associated with $\boldsymbol{Z}$ that does not include an intercept (with $q = \dim(\boldsymbol{Z})$). The conditional hazard function is given by $\lambda_u(t|\boldsymbol{z}) = \lambda_0(t) \exp(\boldsymbol{\beta}^t \boldsymbol{z})$, where $\lambda_0(t) = f_0(t)/S_0(t)$ is the baseline conditional hazard function, with $f_0(t) = -(d/dt)S_0(t)$. Note that $S_u(t|\boldsymbol{z})$ is a proper survival function, i.e. $S_u(t|\boldsymbol{z}) \to 0$ when $t \to \infty$, implying that the survival function is such that $S(t|\boldsymbol{x}, \boldsymbol{z}) \to 1 - p(\boldsymbol{x})$ when $t \to \infty$.

Note also that even if the conditional hazard function $\lambda_u(t|\boldsymbol{z})$ satisfies the proportional hazard property, this is not the case for the hazard function of the entire population.

2.1 *Identifiability of the Model*

Under the data generating process described before, and assuming non-informative censoring, the likelihood of an observation $(y, \delta, \boldsymbol{x}, \boldsymbol{z})$ is given by

$$\mathrm{L}(y, \delta, \boldsymbol{x}, \boldsymbol{z}) = \{g(\boldsymbol{\gamma}^t\boldsymbol{x}) f_u(y|\boldsymbol{z})\}^{\delta}\{1 - g(\boldsymbol{\gamma}^t\boldsymbol{x}) + g(\boldsymbol{\gamma}^t\boldsymbol{x}) S_u(y|\boldsymbol{z})\}^{1-\delta},$$

where $f_u(t|\boldsymbol{z}) = -(d/dt)S_u(t|\boldsymbol{z})$.

In practice, the censoring time C is bounded. This prevents us from observing cured subjects in the data. A way around this problem is to assume the existence of a so-called 'cure threshold' $\tau < \infty$, such that $T > \tau$ implies that $T = +\infty$ (Taylor, 1995). This assumption is commonly accepted and often used in cure models literature. As a consequence, whenever Y is observed to be greater than τ, the individual is assumed to be cured.

We first derive our identifiability result for the case where $\boldsymbol{X}$ is only composed of continuous random variables. Additional assumptions are necessary when $\boldsymbol{X}$ is composed of both continuous and discrete random variables. We give them at the end of this subsection.

For a continuous random vector $\boldsymbol{X}$, our identifiability result is derived under the following set of assumptions:

(A1) (i) The function g is differentiable and not constant on the support of $\boldsymbol{\gamma}^t\boldsymbol{X}$.

(ii) The components of $\boldsymbol{X}$ are continuous random variables that have a joint probability density function.

(iii) The support of $\boldsymbol{X}$ is not contained in any proper linear subspace of R^p.

(iv) $\boldsymbol{\gamma}^t\boldsymbol{X}$ does not contain an intercept.

(v) Either $\gamma_1 = 1$, or $\|\boldsymbol{\gamma}\| = 1$ and the sign of γ_1 is fixed, with $\|\cdot\|$ the Euclidean norm.

(A2) (i) $\boldsymbol{\beta}^t\boldsymbol{Z}$ does not contain an intercept.

(ii) The matrix Var($\boldsymbol{Z}$) has full rank.

(A3) (i)There exists a $\tau < \infty$ (cure threshold) such that $T > \tau \iff T = \infty$, and $P(C > \tau|\boldsymbol{X}, \boldsymbol{Z}) > 0$ for almost all $\boldsymbol{X}$ and $\boldsymbol{Z}$.

(ii) For all $\boldsymbol{x}$, $0 < p(\boldsymbol{x}) < 1$.

Assumption (A1) is required to identify the single-index model (see Horowitz (2009), Theorem 2.1, p. 14), whereas (A2) is needed to make sure that the Cox model is identifiable, and (A3) is a typical assumption needed to identify the mixture cure model (see e.g. Taylor (1995)).

PROPOSITION 2.1: *Under (A1)-(A3), the model given by (1), (2) and (3) is identifiable.*

The proof of this Proposition is given in the Supplementary Materials.

The identifiability result stated above only considers that $\boldsymbol{X}$ is a vector of continuous covariates. When $\boldsymbol{X}$ is a mixture of continuous and discrete variables, two additional conditions, as described in Horowitz (2009), are necessary:

(A4) The support of $\boldsymbol{\gamma}^t\boldsymbol{X}$ must not be divided in disjoint subsets when the values of the discrete components vary.

(A5) The function g is not periodic.

2.2 *Maximum Likelihood Estimation*

In the context of mixture cure models, the likelihood function takes the form $\mathrm{L} = \prod_{i=1}^{n}\{p(\boldsymbol{X}_i) f_u(Y_i|\boldsymbol{Z}_i)\}^{\Delta_i}[\{1 - p(\boldsymbol{X}_i)\} + p(\boldsymbol{X}_i)S_u(Y_i|\boldsymbol{Z}_i)]^{1-\Delta_i}$, with only two different types of contributions: from censored and from uncensored observations. When the latency is modeled as a Cox PH model, a particular feature of this model is that the baseline hazard is left unspecified (Cox, 1972). In classical survival analysis, the Cox PH model is usually fit using a profile likelihood approach (Murphy and Van der Vaart, 2000), which consists in first maximizing the likelihood function with respect to λ_0 for fixed $\boldsymbol{\beta}$, and then plugging-in the solution

in the likelihood function. The so-obtained 'partial likelihood' does not depend on λ_0, and hence the likelihood is to be maximized only with respect to $\boldsymbol{\beta}$. However in mixture cure models, one can not eliminate $\lambda_0(t)$ from the likelihood L. Indeed $\lambda_0(t)$ is related to the incidence because it is conditional on B. It then contains information about the uncure status. Furthermore, cured and uncured censored observations have the same contribution to the likelihood function and no distinction is made between the two. By eliminating $\lambda_0(t)$, one would lose part of the information about the incidence. As a solution, Sy and Taylor (2000) propose to use the Expectation-Maximization algorithm (Dempster et al., 1977) in order to handle both the fact that the cure status is partially unobserved and the fact that $\lambda_0(t)$ is unspecified. The method is based on a complete-data likelihood, that is, the likelihood function that would be obtained if the uncure status B would be observed for all individuals : $\mathrm{L}_c = \prod_{i=1}^{n} \{p(\boldsymbol{X}_i)\lambda_u(Y_i|\boldsymbol{Z}_i)S_u(Y_i|\boldsymbol{Z}_i)\}^{B_i\Delta_i} \left[\{p(\boldsymbol{X}_i)S_u(Y_i|\boldsymbol{Z}_i)\}^{B_i} \times \{1-p(\boldsymbol{X}_i)\}^{1-B_i}\right]^{1-\Delta_i}$. As it can be seen, cured and uncured censored subjects have a different contribution to this likelihood, and one can therefore use a profile likelihood-type approach to estimate the survival function for uncured observations without losing information about the incidence.

The EM algorithm consists in maximizing the complete-data likelihood by alternating between an E(xpectation) and a M(aximization) step until convergence. At the m^{th} iteration of the algorithm, the E-step consists in computing the expectation of the logarithm of the complete-data likelihood with respect to the latent variable (the uncure status), given the observed data and the current parameter values. As the log-complete-data likelihood is linear in B, it is the same as computing $E\left(B_i|O,\theta^{(m-1)}\right) = \Delta_i \ [1 \times P(B_i = 1|Y = Y_i, \Delta_i = 1, \boldsymbol{X} = \boldsymbol{X}_i, \boldsymbol{Z} = \boldsymbol{Z}_i, \theta^{(m-1)})] + \ (1-\Delta_i) \ [1 \times P(B_i = 1|Y = Y_i, \Delta_i = 0, \boldsymbol{X} = \boldsymbol{X}_i, \boldsymbol{Z} = \boldsymbol{Z}_i, \theta^{(m-1)})]$, where $O = \{(Y_i, \Delta_i, \boldsymbol{X}_i, \boldsymbol{Z}_i), i = 1, ..., n\}$, are the observed data, $\theta = (\boldsymbol{\gamma}, \boldsymbol{\beta}, S_0, g)$ is the vector of parameters, and $\theta^{(m-1)}$ denotes the set of parameter values at the $(m-1)^{th}$ iteration. Given that $P(B_i = 1|Y = Y_i, \Delta_i = 1, \boldsymbol{X} = \boldsymbol{X}_i, \boldsymbol{Z} =$

$\boldsymbol{Z}_i, \theta^{(m-1)}) = 1$, and that $P(B_i = 1|Y = Y_i, \Delta_i = 0, \boldsymbol{X} = \boldsymbol{X}_i, \boldsymbol{Z} = \boldsymbol{Z}_i, \theta^{(m-1)}) = \{P(B_i = 1, T > Y_i|\boldsymbol{X} = \boldsymbol{X}_i, \boldsymbol{Z} = \boldsymbol{Z}_i, \theta^{(m-1)})\}/\{P(T > Y_i|\boldsymbol{X} = \boldsymbol{X}_i, \boldsymbol{Z} = \boldsymbol{Z}_i, \theta^{(m-1)})\}$, it follows that $E\left(B_i|O, \theta^{(m-1)}\right) = \Delta_i + (1 - \Delta_i)\{p^{(m-1)}(\boldsymbol{X}_i)S_u^{(m-1)}(Y_i|\boldsymbol{Z}_i)\}/\{1 - p^{(m-1)}(\boldsymbol{X}_i) + p^{(m-1)}(\boldsymbol{X}_i)S_u^{(m-1)}(Y_i|\boldsymbol{Z}_i)\} = W_i^{(m)}$. The substitution of $W_i^{(m)}$ in the logarithm of L_c gives the expected log-complete-data likelihood or equivalently the expected complete-data likelihood $\tilde{\mathrm{L}}_c = \prod_{i=1}^n \{p(\boldsymbol{X}_i)\lambda_u(Y_i|\boldsymbol{Z}_i)S_u(Y_i|\boldsymbol{Z}_i)\}^{W_i^{(m)}\Delta_i} \times \prod_{i=1}^n [\{p(\boldsymbol{X}_i)S_u(Y_i|\boldsymbol{Z}_i)\}^{W_i^{(m)}} \times \{1 - p(\boldsymbol{X}_i)\}^{1-W_i^{(m)}}]^{1-\Delta_i}$. In the M-step, $\tilde{\mathrm{L}}_c$ is maximized with respect to the parameters of the model. Since it can be re-written as the product of two factors, each of them containing the parameters of one part of the model, $\tilde{\mathrm{L}}_c = \prod_{i=1}^n [p(\boldsymbol{X}_i)^{W_i^{(m)}}\{1 - p(\boldsymbol{X}_i)\}^{1-W_i^{(m)}}] \times \prod_{i=1}^n \{\lambda_u(Y_i|\boldsymbol{Z}_i)^{\Delta_i} S_u(Y_i|\boldsymbol{Z}_i)\}^{W_i^{(m)}} = \tilde{\mathrm{L}}_1 \times \tilde{\mathrm{L}}_2$, it may be maximized separately for the two parts of the model.

For the incidence, we follow the maximum likelihood approach of Klein and Spady (1993) to estimate a single-index model with a binary outcome. The likelihood function given by $\tilde{\mathrm{L}}_1 = \prod_{i=1}^n g(\boldsymbol{\gamma}^t\boldsymbol{X}_i)^{W_i^{(m)}}\{1 - g(\boldsymbol{\gamma}^t\boldsymbol{X}_i)\}^{1-W_i^{(m)}}$ is the same as for a logistic regression model, except that the link function $g(\cdot)$ is totally unknown. So there is first a need to estimate the link function $g(\cdot)$. The following (unfeasible) leave-one-out kernel estimator of $g(\boldsymbol{\gamma}^t\boldsymbol{X}_i)$ based on Nadaraya-Watson weights:

$$\sum_{j\neq i}^{n} \frac{K\left(\frac{\boldsymbol{\gamma}^t\boldsymbol{X}_i - \boldsymbol{\gamma}^t\boldsymbol{X}_j}{h}\right)}{\sum_{l\neq i}^n K\left(\frac{\boldsymbol{\gamma}^t\boldsymbol{X}_i - \boldsymbol{\gamma}^t\boldsymbol{X}_l}{h}\right)} B_j,$$

where h is a one-dimensional bandwidth, can not be used in practice as B is latent. Alternatively, B can be replaced by its expectation $W^{(m)}$ obtained in the E-step of the algorithm. The Nadaraya-Watson estimator becomes

$$\tilde{g}_{-i}^{(m)}(\boldsymbol{\gamma}^t\boldsymbol{X}_i) = \sum_{j\neq i}^{n} \frac{K\left(\frac{\boldsymbol{\gamma}^t\boldsymbol{X}_i - \boldsymbol{\gamma}^t\boldsymbol{X}_j}{h}\right)}{\sum_{l\neq i}^n K\left(\frac{\boldsymbol{\gamma}^t\boldsymbol{X}_i - \boldsymbol{\gamma}^t\boldsymbol{X}_l}{h}\right)} W_j^{(m)}. \tag{4}$$

The kernel estimator (4) is substitued in $\tilde{\mathrm{L}}_1$, and $\boldsymbol{\gamma}$ is estimated by maximizing the likelihood function with numerical techniques such as the Newton-Raphson algorithm. The resulting estimator of $\boldsymbol{\gamma}$ is denoted by $\hat{\boldsymbol{\gamma}}^{(m)}$. Once $\hat{\boldsymbol{\gamma}}^{(m)}$ is computed, $g(\boldsymbol{\gamma}^t\boldsymbol{X}_i)$ is estimated

by the estimator given in (4), but with $\boldsymbol{\gamma}$ replaced by $\hat{\boldsymbol{\gamma}}^{(m)}$. The estimator is denoted by $\hat{g}_{-i}^{(m)}\{(\hat{\boldsymbol{\gamma}}^{(m)})^t \boldsymbol{X}_i\}$. Note that when $d = 1$, as $\gamma_1 = 1$, the estimator (4) reduces to a non-parametric estimator.

For the latency, the likelihood function given by $\tilde{\mathrm{L}}_2 = \prod_{i=1}^{n}[\{\lambda_0(Y_i)\exp(\boldsymbol{\beta}^t \boldsymbol{Z}_i)\}^{\Delta_i} \exp\{-\Lambda_0(Y_i)$ $\exp(\boldsymbol{\beta}^t \boldsymbol{Z}_i)\}]^{W_i^{(m)}}$ is similar to the likelihood function for the classical Cox PH model except that the baseline cumulative hazard $\Lambda_0(t) = \int_0^t \lambda_0(u)\,du$ and $\lambda_0(t)$ are conditional on $B = 1$. Sy and Taylor (2000) propose a profile likelihood approach to estimate $\boldsymbol{\beta}$ based on the work of Breslow (1974). In a first step, the value of $\boldsymbol{\beta}$ is fixed and $\Lambda_0(t)$ is estimated non-parametrically by

$$\sum_{j:Y_{(j)}\leqslant t} \frac{D_j}{\sum_{k\in R_j} W_k^{(m)} \exp(\boldsymbol{\beta}^t \boldsymbol{Z}_k)}, \tag{5}$$

where $Y_{(1)} < \ldots < Y_{(r)}$ are the ordered uncensored event times with r the number of distinct event times, D_j is the number of events at time $Y_{(j)}$, and R_j is the set of observations at risk just before $Y_{(j)}$. In a second step, (5) is plugged in $\tilde{\mathrm{L}}_2$, obtaining the partial likelihood

$$\breve{\mathrm{L}}_2 = \prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{\beta}^t \boldsymbol{Z}_i)}{\sum_{k\in R_i} W_k^{(m)} \exp(\boldsymbol{\beta}^t \boldsymbol{Z}_k)} \right\}^{\Delta_i}. \tag{6}$$

The maximum likelihood estimator of $\boldsymbol{\beta}$ is obtained by maximizing (6), and is denoted by $\hat{\boldsymbol{\beta}}^{(m)}$. $\hat{\Lambda}_0^{(m)}(t)$ denotes the estimator of $\Lambda_0(t)$ given in (5) but with $\boldsymbol{\beta}$ replaced by $\hat{\boldsymbol{\beta}}^{(m)}$. The EM algorithm iterates until the difference between two consecutive values of the estimates of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $S_0(\cdot)$ is smaller than a certain value a priori defined. The final estimators are denoted by $\hat{\boldsymbol{\gamma}}$, $\hat{g}(\cdot)$, $\hat{\boldsymbol{\beta}}$ and $\hat{\Lambda}_0(\cdot)$.

When the latency is modeled non- or semi-parametrically, little information is available to distinguish cured from uncured individuals among censored subjects. The tail of the conditional survival function $\hat{S}_u(t|\boldsymbol{z})$ may be hard to estimate and we often observe that $\hat{S}_u(t|\boldsymbol{z}) > 0$ when $t > Y_{(r)}$, implying identifiability issues. To solve this problem, Taylor (1995) proposes to consider $Y_{(r)}$ as a cure threshold τ, and to force $\hat{S}_u(t|\boldsymbol{z})$ to be equal to zero beyond $Y_{(r)}$ by setting, in the M-step of the EM algorithm, $W_i^{(m)}$ equal to 0 when an

observation i is such that $\Delta_i = 0$ and $Y_i > Y_{(r)}$. This proposal is equivalent to considering that such an observation i is cured. To understand the intuition behind this proposal, we have to recall some elements introduced previously. When cured observations are present in survival data, $\lim_{t\to\infty} S(t|\boldsymbol{x}, \boldsymbol{z}) > 0$. In practice, this assumption translates into a Kaplan-Meier estimator of the survival function with a large plateau, and leveling-off to a value greater than 0 for $t > Y_{(r)}$. It corresponds to a situation where a certain number of censored individuals have a follow-up time greater than $Y_{(r)}$. If the follow-up is sufficiently long and if the number of observations in the plateau is sufficiently large, it reasonably makes sense to consider these censored individuals as cured, as Taylor (1995) proposes. Under the assumed Cox PH model for the latency, this constraint is also applied to the SIC cure model.

3. Numerical Study

The objective of this numerical study is to compare the fit of the SIC cure model with the fit of the classical logistic/Cox (LC) cure model in various settings. The two models differ in their incidence (a single-index versus a logistic structure), while they both assume a Cox PH model in the latency. The impact of a single-index structure on both the estimates of $p(\boldsymbol{x})$ and $S_u(t|\boldsymbol{z})$ are investigated as well as the impact of different censoring rates.

Data for the incidence is generated according to three scenarios, each of them corresponding to a different link function. The first scenario assumes a logistic link function of the form $g(u) = \exp(\gamma_0 + u)/\{1 + \exp(\gamma_0 + u)\}$, where γ_0 is an intercept term. The second scenario is an adaptation of a model coming from Müller and Schmitt (1988), namely $g(u) = \exp[0.75\ \Phi\{(\gamma_0 + u) + 0.5\} + 0.25\ \Phi\{0.5(\gamma_0 + u)^3\}]/(1 + \exp[0.75\ \Phi\{(\gamma_0 + u) + 0.5\} + 0.25\ \Phi\{0.5(\gamma_0 + u)^3\}])$, with $\Phi(\cdot)$ the standard normal cumulative distribution, leading to a non-logistic but monotone link function. The third scenario assumes a non-monotone link function of the form $g(u) = \exp[0.4\{(\gamma_0+u)^3-(\gamma_0+u)^2-0.9(\gamma_0+u)+1\}]/(1+\exp[0.4\{(\gamma_0+u)^3 - (\gamma_0 + u)^2 - 0.9(\gamma_0 + u) + 1\}])$. Figure 1 shows the three link functions. Note that the

last two link functions were chosen sufficiently different from a logistic link function in order to assess the performance of a single-index and a logistic model in such situations. Note also that we included a logistic transformation in the two last link functions in order to restrict the probability to the interval $[0, 1]$.

[Figure 1 about here.]

For all scenarios, we consider four independent covariates: X_1 and X_2 have a standard normal distribution, and X_3 and X_4 have a Bernoulli distribution with parameters 0.3 and 0.6 respectively. The parameters $\gamma_0, \ldots, \gamma_4$ are chosen so that a sufficient cure proportion is obtained and so that the identification condition (A4) is fulfilled. Table 1 summarizes the parameter values and the cure proportions. Note that the intercept γ_0 is only estimated under the logistic model. Under the single-index model it is incorporated into the link function g.

For the latency, we consider one binary covariate Z that is independent of $\boldsymbol{X} = (X_1, X_2, X_3, X_4)^t$ and that has a Bernoulli distribution with parameter 0.6. The survival times for the uncured observations are generated according to a Weibull PH model, i.e. $S_u(t|z) = \exp(-\lambda t^\rho)^{\exp(\beta z)}$ for given choices of λ, ρ and β. For the three scenarios, the scale parameter λ equals 1.5, the shape parameter ρ equals 1.2, and $\beta = 1.5$. The censoring time C is independent of the vector $(\boldsymbol{X}, Z, T)$, and is generated from an exponential distribution with density function $f(t) = \lambda_c \exp(-\lambda_c t)$. Three different values for the rate λ_c, denoted by level 1, level 2 and level 3, are considered so that the difference between the censoring rate and the cure rate increases. An increment of 5% between each level has been considered. The values for each scenario are given in Table 1.

[Table 1 about here.]

For both the survival and the censoring times, the parameter values were chosen in such a way that a certain number of the censored observations have a follow-up time larger than $Y_{(r)}$, and such that these observations are cured. In such a way we mimic the type of real

survival data on which cure models are typically used. Additionally by increasing λ_c our goal is to assess the impact of larger censoring rates on the model estimates. Note that when the censoring rate increases, the proportion of observations with a follow-up time larger than $Y_{(r)}$ tends to decrease. As a consequence, very large censoring rates are not considered in order to still have a non-negligible number of observations in the plateau. Table 1 gives the censoring rates and the percentage of observations in the plateau for the different values of λ_c.

For each setting, we consider samples of size $n = 250$ and 500. This leads to a total of 18 settings (3 model scenarios, 3 censoring rates and 2 sample sizes). For each setting we generate 500 datasets. For each dataset, we fit a LC cure model and a SIC cure model. For the single-index model, we use an Epanechnikov kernel, $K(u) = 3(1-u^2/5)/(4\sqrt{5})I(u^2 \leqslant 5)$. The bandwidth is selected according to a likelihood cross-validation procedure at the beginning of the M-step, prior to the incidence estimation. It is computed at each iteration of the algorithm. The cross-validation criterion is given by minus the logarithm of $\tilde{\mathrm{L}}_1$ evaluated at $\hat{\boldsymbol{\gamma}}^{(m-1)}$, i.e. $CV^{(m)}(h) = -\sum_{i=1}^{n} W_i^{(m)} \log \hat{g}_{h,-i}^{(m-1)}\{(\hat{\boldsymbol{\gamma}}^{(m-1)})^t X_i\} - \sum_{i=1}^{n}(1 - W_i^{(m)}) \log[1 - \hat{g}_{h,-i}^{(m-1)}\{(\hat{\boldsymbol{\gamma}}^{(m-1)})^t X_i\}]$, where $\hat{g}_{h,-i}^{(m-1)}(\cdot)$ is the leave-one-out estimator (4) with $\boldsymbol{\gamma}$ replaced by $\hat{\boldsymbol{\gamma}}^{(m-1)}$ and based on a bandwidth h. The selected bandwidth is the minimizer on the interval [0.4,1] of the cross-validation criterion. It is given by $h_{CV}^{(m)} = \arg\min_h CV^{(m)}(h)$. The interval [0.4,1] is chosen based on visual inspection of what are reasonable bandwidths in this setting. The bandwidth obtained at the last iteration of the algorithm is the final bandwidth.

In order to make both models identifiable, the conditional survival function is forced to be equal to 0 beyond $Y_{(r)}$ as proposed by Taylor (1995). For the SIC cure model, we apply the identification conditions mentioned previously. For condition (A1)-(v), we set $\hat{\gamma}_1 = 1$ or $\hat{\gamma}_1 = -1$, depending on the sign of $\hat{\gamma}_1$ for the LC cure model. These identification conditions will be assumed for the remaining of the article. For the two models, the initial values of

the EM algorithm are the parameters coming from a logistic regression model taking the censoring indicator as response variable for the incidence and coming from a Cox PH model fitted on the uncensored observations only for the latency. We consider that the algorithm converges when the difference between two consecutive values of the parameters is smaller than 10^{-5}.

In order to evaluate the performance of the SIC cure model for the incidence, we compute the Average Squared Error (ASE) of $\hat{p}(\boldsymbol{x})$ for each model, $ASE(\hat{p}) = M^{-1}\sum_{j=1}^{M}\left\{\hat{g}(\hat{\boldsymbol{\gamma}}^t\boldsymbol{x}_j) - g(\boldsymbol{\gamma}^t\boldsymbol{x}_j)\right\}^2$. The ASE is computed on a grid $(\boldsymbol{x}_j)_j = (\{x_{j1}, x_{j2}, x_{j3}, x_{j4}\})_j$, $j = 1, ..., M$, of values of the vector (X_1, X_2, X_3, X_4). For X_1 and X_2 we take grid points on $[-1.5, 1.5]$ with step size 0.01, while X_3 and X_4 take values in $\{0, 1\}$. For the latency, we compute the bias, the variance and the mean squared error (MSE) of $\hat{\beta}$ for each model.

Table 2 shows the bias, variance, and MSE of $\hat{\beta}$ for the two models. As can be seen from the table, the bias of $\hat{\beta}$ is small for the two models, and we observe very small differences between them. These small differences however tend to increase when λ_c increases. This situation occurs because when λ_c increases, the censoring rate increases, and the number of observations in the plateau becomes substantially smaller than the number of cured observations. In such a situation, the number of censored observations such that $Y \leqslant Y_{(r)}$ gets larger. The weight $W_i^{(m)}$ for these observations will be equal to $\{p^{(m-1)}(\boldsymbol{X}_i)S_u^{(m-1)}(Y_i|Z_i)\}/\{1 - p^{(m-1)}(\boldsymbol{X}_i) + p^{(m-1)}(\boldsymbol{X}_i)S_u^{(m-1)}(Y_i|Z_i)\}$, that is, it will rely more on $\hat{p}(\boldsymbol{x})$ than when an observation is uncensored ($W_i^{(m)} = 1$) or when an observation is censored after $Y_{(r)}$ ($W_i^{(m)} = 0$). Depending on the estimate for $p(\boldsymbol{x})$, these observations censored before $Y_{(r)}$ will have a different contribution to the likelihood function for the latency. Larger differences in the estimation of β are observed between the two models. Despite this effect, the bias is still small for the two models and it seems that choosing between a logistic regression and a single-index model does not have a large impact on the latency parameter estimate when the latency modeling

is the same. Furthermore, the variance and MSE of $\hat{\beta}$ are very similar for the two models. Nevertheless, when λ_c increases, $\hat{\beta}$ becomes more variable, and the MSE increases for the same reason as above.

[Table 2 about here.]

For the incidence, Figure 2 shows the boxplots of the ASE of $\hat{p}(\boldsymbol{x})$ for both models. As can be expected, the LC cure model performs better than the SIC cure model when the true model is a logistic regression, regardless of the sample size (see the boxplots in Figure 2, scenario 1). For scenarios 2 and 3 on the contrary, when the true link function is different from a logistic one, the SIC cure model performs better for all censoring rates and sample sizes considered. As is the case for the latency, the censoring rate has an impact on the quality and the precision of the estimates of $p(\boldsymbol{x})$, with the ASE taking on higher values and being more variable when λ_c increases. The explanation is the same as above. When the censoring rate gets much larger than the cure proportion, more observations have a weight $W_i^{(m)}$ taking values between 0 and 1. Because $W_i^{(m)}$ is involved in the likelihood function for estimating $p(\boldsymbol{x})$, the uncertainty in the estimation of $p(\boldsymbol{x})$ increases, and larger and more variable values of ASE are observed. On the contrary, as can be expected, when the sample size increases, a decrease in the value of the ASE is observed, meaning a better fit to the data.

[Figure 2 about here.]

Estimating a SIC cure model is computationally more demanding. However, although a difference by a factor of around 85 has been observed in our simulations, the required time remains reasonable, going from 0.10 seconds for the LC cure model to 8.4 seconds for the SIC cure model when n=250, and from 0.18 seconds to 15 seconds when n=500 (2.7 GH processor and 16 Go of memory computer). With a more complex model, for e.g. with a higher dimension of $\boldsymbol{X}$ or a more complex link function, the computation time increases

for both models, mainly because the EM algorithm gets slower. In term of EM algorithm convergence, between 99.2 % and 100% of the datasets converge when estimating a SIC cure model.

4. Real Data Application

Our application concerns a breast cancer study which consists of 286 patients that experienced a lymph-node-negative breast cancer between 1980 to 1995 (Wang et al., 2005). The event of interest is distant metastasis, and the associated survival time is the time to distant metastasis (DM). Among the 286 patients, 107 experienced a distant recurrence from breast cancer. As can be seen from Web Figure 1, the Kaplan-Meier estimator of the survival function shows a large plateau at about 60%, and 88% of the censored observations are in the plateau. A cure model seems therefore appropriate for these data. Four covariates are considered : age of the patient (ranges from 26 to 83 with a median of 52 years), estrogen receptor (ER) status (0 = ER−: less than 10 fmol per mg protein - 77 patients, 1 = ER+: 10 fmol per mg protein or more - 209 patients), size of the tumor (ranges from 1 to 4 with a median of 1), and menopausal status (0 = pre-menopausal - 129 patients, 1 = post-menopausal - 157 patients) which has been obtained by dichotomizing age (pre-menopausal: age $\leqslant 50$ - post-menopausal: age > 50). The original data are split in a training and a test set following the 2/3 - 1/3 recommendations of Hastie et al. (2009). For our application, it corresponds to 190 and 96 observations, respectively. The training set is used to estimate and to interpret the model. The test set is used to compute the prediction error of the incidence, given by $PE = -\sum_{j=1}^{96} \log[\hat{p}(\boldsymbol{x}_j^{test})^{\hat{w}_j}\{1-\hat{p}(\boldsymbol{x}_j^{test})\}^{1-\hat{w}_j}]$, where $\hat{p}(\boldsymbol{x}_j^{test})$ and $\hat{w}_j$ are the predicted uncure probability and the predicted weight for the j^{th} observation in the test set, respectively, computed based on the parameter estimates (and the link function for the single-index) obtained from the training set. On the training set, a LC and a SIC cure model including the four covariates in both parts of the two models are fitted. For interpretation

convenience, a standardized version of all covariates, continuous and discrete, is considered in the incidence. For the single-index model, we proceed as in Section 3 for model fitting, bandwidth selection, and model identification, and the parameters are further normalized so that $\|\hat{\boldsymbol{\gamma}}^{SI}\| = 1$, with $\hat{\boldsymbol{\gamma}}^{SI}$ the vector of parameter estimates under the single-index model, in order to ease the interpretation. The standard errors of the parameter estimates for both models have been computed using a bootstrap approach, based on 250 bootstrap samples.

Table 3 gives the results jointly with p-values for the Wald test of the parameters. For both models, the effects for the four covariates on the latency have the same direction, and the estimates are very close. Moreover, only the ER status affects significantly the survival of uncured patients. A positive ER status implies a longer time to DM.

[Table 3 about here.]

For the incidence, we first compute the prediction error for both models to compare their fits. The prediction error for the SIC cure model equals to 67.71, while it is equal to 71.57 for the LC cure model, meaning that the SIC cure model performs better than the LC cure model in predicting the uncure status. To understand why the single-index model performs better, we compare the estimates of the two models. Beside different parameter estimates, the link functions, given in Figure 3 (a)-(b), are also different, the single-index link function showing a non-monotone trend which is quite different from the logistic one. Nevertheless, as already pointed out by Li and Duan (1989) in a broader context, even if two models are estimated with different link functions, the relative importance of the parameter estimates is similar, permitting to evaluate which variable impacts the most the outcome. Here, age and menopausal status are the most influential covariates on the uncure probability because of their larger estimates in comparison with the other covariates. However, to interpret covariate effects on $p(\boldsymbol{x})$, the link function and the parameters have to be considered jointly as the sign of the effect depends on the shape of the link function. If the link function is monotone

increasing (resp. decreasing), the sign of the effect will be the same as (resp. the opposite) the sign of the estimate.

[Figure 3 about here.]

For this dataset, the parameter estimates show, at a 10% level, that age and menopausal status have a significant effect on the uncure probability for the single-index model, whereas none of them are significant for the logistic model. As the link function is not monotone, we can not interpret the parameter estimates directly. But we can instead plot the uncure probability against covariates to better understand their effects. Figure 3 (c)-(d) shows the effect of age on $p(\boldsymbol{x})$ for the two models. For both of them, two different groups appear, corresponding to pre-menopausal and post-menopausal women. Contrary to the logistic model where the relationship seems to be almost equivalent and linear in the two groups, it seems that the trend for pre-menopausal in the single-index model is flat between roughly 35 and 50, and much higher in the 25-35 range, while for post-menopausal it is roughly flat from 50 to 70, then much lower for ages greater than 70.

5. Conclusion

We proposed a new semi-parametric modeling approach for the mixture cure model by considering a single-index structure for the uncure probability. We proved the identifiability of the proposed model, parameter estimators are obtained via maximum likelihood using the EM algorithm, and for the unknown link function of the single-index a Nadaraya-Watson estimator is proposed with a likelihood cross-validation method for the bandwidth parameter. Based on a numerical study, we showed the good performance of the method for the estimation of the uncure probability, and its ability to outperform the logistic model when the true link function is not the logistic one. The application of the method to a breast cancer dataset illustrates its practical use and the usefulness of a single-index model. The

analysis uncovered a non-monotone link function, and both the parameter estimates and the graphical representation indicate a more complex effect of age not observed with the logistic model. To further investigate age effect on $p(\boldsymbol{x})$, an alternative modeling would be to consider splines in the incidence as age seems to have a non-linear effect. Li and Taylor (2002) propose such modeling by considering a generalized additive model for the incidence with cubic B-splines.

A single-index model is a flexible alternative to the logistic regression, and the SIC cure model may be considered as a diagnostic tool to investigate misspecification of the incidence of the mixture cure model. A possible extension of our work will be to develop a test for the parametric form of $p(\cdot)$. Müller and Van Keilegom (2018) propose such a test based on the non-parametric estimator of the cure rate proposed by Xu and Peng (2014) for the case where $\boldsymbol{X}$ is one-dimensional. By avoiding curse of dimensionality problems, our estimator $\hat{g}(\hat{\boldsymbol{\gamma}}^t\boldsymbol{X})$ may constitute a natural extension of their work when $\boldsymbol{X}$ is multi-dimensional. Furthermore, with its flexibility, our proposal also permits to model a wide range of link functions, monotone or not. However in certain contexts it is natural to assume that the link function is monotone. A possible adaptation of this work would be to restrict the link function to the monotone case.

Acknowledgements

The authors like to thank Roberto d'Ambrosio and Professor Pierre Dupont for providing the breast cancer data set, Professor Raymond J. Carroll for his helpful comments, as well as the associate editor and two anonymous referees for their constructive comments and interesting suggestions. This work was supported by the Interuniversity Attraction Pole Research Network of the Belgian government (Belgian Science Policy) P7/06. Maïlis Amico and Ingrid Van Keilegom were also supported by the European Research Council 2016-2021, Horizon 2020 / ERC grant, agreement No. 694409. The computational resources and services

used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government, department EWI.

Supplementary Materials

Web Figure 1, referenced in Sections 1 and 4, the proof of Proposition 2.1, as well as R codes implementing the single-index/Cox mixture cure model are available with this paper at the Biometrics website on Wiley Online Library.

References

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

Breslow, N.E. (1974). Covariates analysis of censored survival data. *Biometrics* **30,** 89–99.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Dempster A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via EM Algorithm. *Journal of the Royal Statistical Society, Series B* **39,** 1–38.

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

Friedman, J.H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association* **76,** 817–823.

Hastie, T., Tibshirani, R., and Friedman, R. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd Edition. New York: Springer.

Horowitz, J.L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. New York: Springer.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* **58,** 71–120.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53,** 457–481.

Klein, R.W. and Spady, R.H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61,** 387–421.

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

Li, K.C. and Duan, N. (1989). Regression Analysis under Link Violation. *The Annals of Statistics* **17,** 1009–1052.

Li, C.S. and Taylor, J.M.G. (2002). Smoothing covariate effects in cure models. *Communications in Statistics - Theory and Methods,* **31,** 477–493.

Lopez, O. (2009). Single-index regression models with right-censored responses. *Journal of Statistical Planning and Inference* **139,** 1082–1097.

Lopez, O., Patilea, V., and Van Keilegom, I. (2013). Single index regression models in the presence of censoring depending on the covariates. *Bernoulli* **19,** 721–747.

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

Lu, X. and Burke, M.D. (2005). Censored multiple regression by the method of average derivatives. *Journal of Multivariate Analysis* **95,** 182–205.

Müller, H.-G. and Schmitt, T. (1988). Kernel and probit estimates in quantal bioassay. *Journal of the American Statistical Society* **83,** 750–759.

Müller, U.U. and Van Keilegom, I. (2018). Goodness-of-fit tests for the cure rate in a mixture cure model. *Biometrika*, in press.

Murphy, S.A. and Van der Vaart, A.W. (2000). On profile likelihood. *Journal of the American Statistical Association* **95,** 449–465.

Patilea, V. and Van Keilegom, I. (2018). A general approach for cure models in survival

analysis. *Annals of Statistics*, under revision.

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

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

Wang, Y., He, S., Zhu, L., and Yuen, K.C. (2007). Asymptotics for a censored generalized linear model with unknown link function. *Probability Theory and Related Fields* **138,** 235–267.

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

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

Received April 2007. *Revised April* 2007. *Accepted April* 2007.

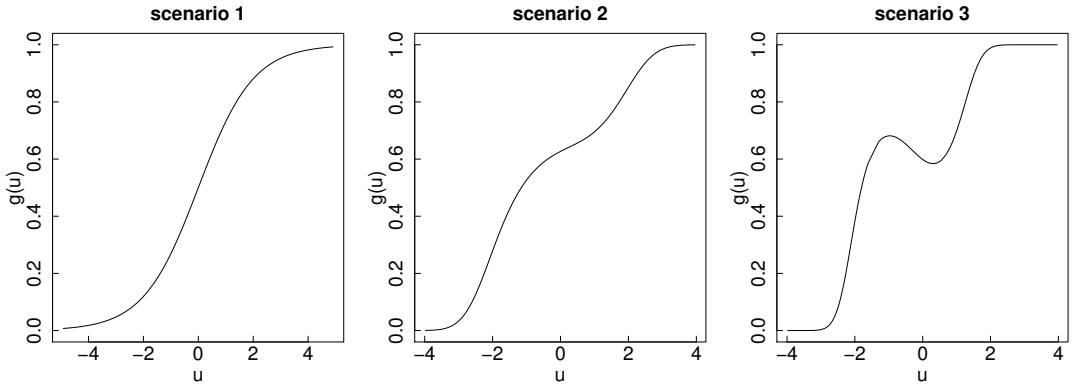

Figure 1. Link functions considered for the incidence in the data generation process when $\gamma_0 = 0$.

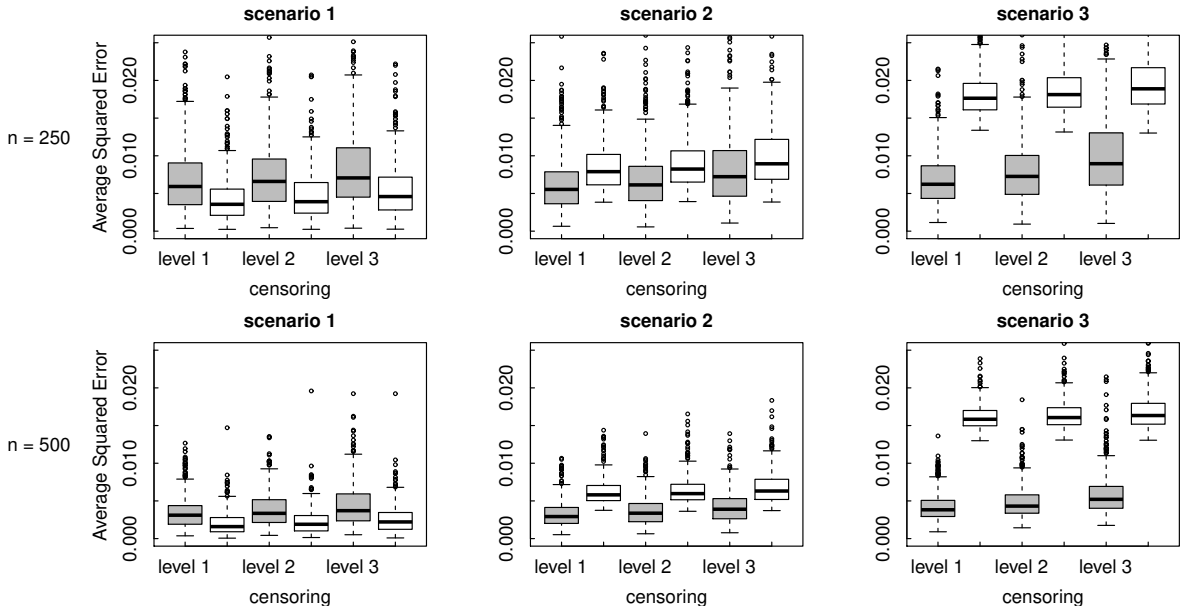

Figure 2. Boxplots of the Average Squared Error (ASE) for the single-index model (grey boxplots) and the logistic model (white boxplots).

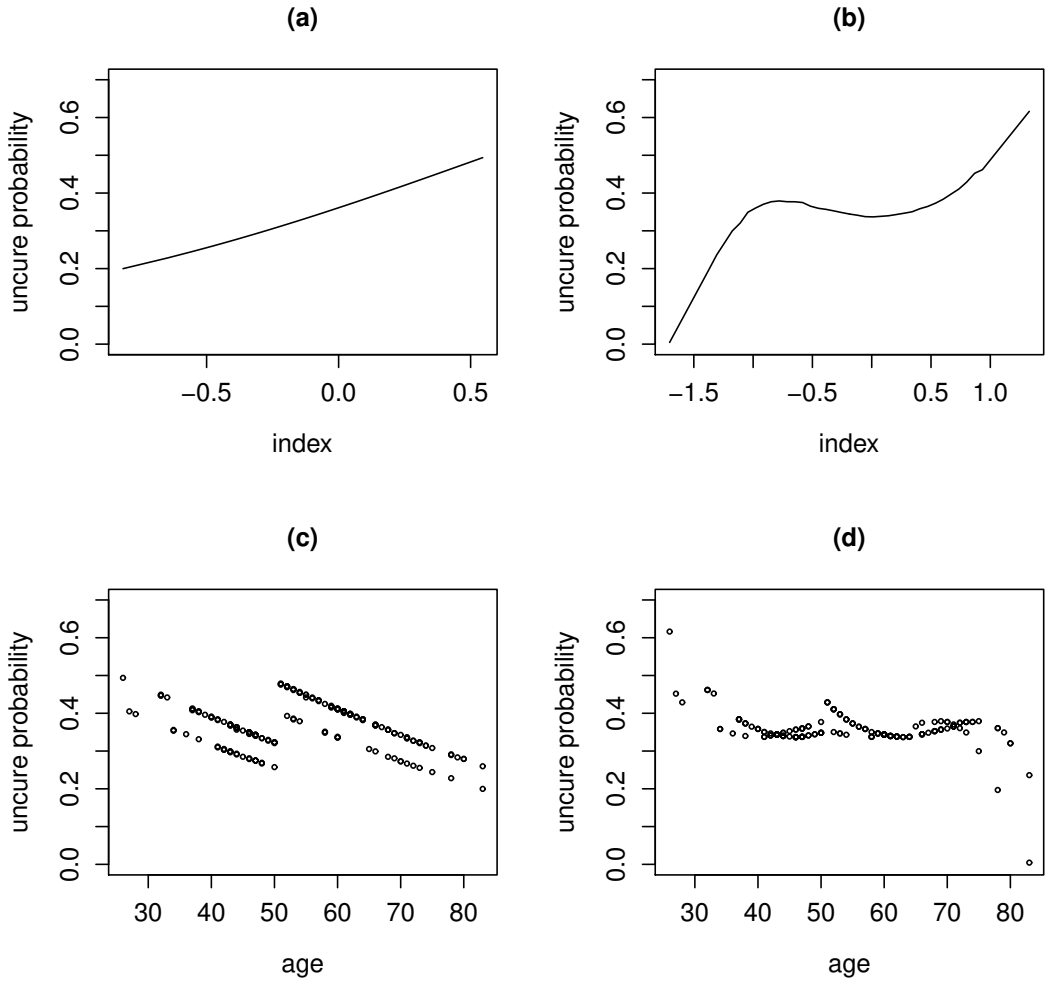

Figure 3. Estimated link function for (a) the logistic model, (b) the single-index model, and plot of the effect of age on the uncure probability for (c) the logistic model and (d) the single-index model.

Table 1
The parameter values for the incidence, the cure proportions, the censoring rates, and the proportion of observations in the plateau for each scenario.

	incidence parameters						censoring scheme			
scen.	γ_0	γ_1	γ_2	γ_3	γ_4	rate	level	λ_c	rate	plateau
1	1.4	-1.5	0.5	2.3	-1.3	18.6%	level 1	0.05	20.1%	16.3%
						18.6%	level 2	0.25	25.7%	10.3%
						18.6%	level 3	0.5	31.4%	6.7%
2	-0.3	1.5	-0.9	2	-1	34.4%	level 1	0.05	35.6%	30.3%
						34.4%	level 2	0.25	40.1%	19.5%
						34.4%	level 3	0.55	45.6%	12.0%
3	1	1.3	-2	-2	1	37.9%	level 1	0.1	40.2%	29.7%
						37.9%	level 2	0.35	45.2%	18.2%
						37.9%	level 3	0.7	50.7%	11.2%

Table 2
Bias, variance and mean squared error (MSE) of $\hat{\beta}$ for the SIC cure model and for the LC cure model.

			censoring schemes								
			level 1			level 2			level 3		
n	scen.		bias	var.	MSE	bias	var.	MSE	bias	var.	MSE
250	1	SIC	0.0154	0.0343	0.0345	0.0248	0.0372	0.0378	0.0271	0.0394	0.0401
		LC	0.0149	0.0343	0.0346	0.0230	0.0370	0.0375	0.0233	0.0390	0.0396
	2	SIC	0.0206	0.0450	0.0454	0.0278	0.0480	0.0487	0.0285	0.0528	0.0536
		LC	0.0208	0.0450	0.0454	0.0289	0.0480	0.0489	0.0297	0.0524	0.0533
	3	SIC	0.0155	0.0473	0.0475	0.0165	0.0547	0.0549	0.0170	0.0665	0.0668
		LC	0.0160	0.0468	0.0471	0.0185	0.0551	0.0555	0.0189	0.0655	0.0659
500	1	SIC	0.0227	0.0169	0.0174	0.0192	0.0180	0.0184	0.0198	0.0199	0.0203
		LC	0.0224	0.0169	0.0174	0.0176	0.0180	0.0183	0.0174	0.0199	0.0202
	2	SIC	0.0117	0.0210	0.0211	0.0092	0.0224	0.0225	0.0047	0.0262	0.0262
		LC	0.0120	0.0210	0.0212	0.0102	0.0225	0.0227	0.0062	0.0262	0.0262
	3	SIC	0.0168	0.0220	0.0223	0.0123	0.0242	0.0244	0.0151	0.0292	0.0294
		LC	0.0178	0.0220	0.0223	0.0150	0.0244	0.0247	0.0176	0.0296	0.0299

Table 3
Parameter estimates and standard errors for the SIC cure model and for the LC cure model.

	SIC cure model			*LC cure model*		
Incidence	Estimate	Std.Error	p-value	Estimate	Std.Error	p-value
(intercept)	-	-	-	-0.5707	0.1514	0.0002
age	-0.8111	0.2394	0.0007	-0.3668	0.2593	0.1573
ER status [ER+ vs. ER-]	0.1758	0.2923	0.5475	0.1422	0.1805	0.4308
size of the tumour	-0.0003	0.3496	0.9992	-0.0058	0.1563	0.9705
menopausal [post vs. pre]	0.5577	0.3293	0.0894	0.3413	0.2678	0.2026
bandwidth	*0.3245*	-	-	-	-	-
Latency	Estimate	Std.Error	p-value	Estimate	Std.Error	p-value
age	-0.0113	0.0212	0.5953	-0.0107	0.0211	0.6122
ER status [ER+ vs. ER-]	-1.3639	0.3855	0.0004	-1.3640	0.3886	0.0004
size of the tumour	0.4527	0.2977	0.1284	0.4524	0.3027	0.1351
menopausal [post vs. pre]	-0.2519	0.6178	0.6835	-0.2644	0.6166	0.6682