Biometrika (2017), **103**, 1, *pp.* 1–17
Printed in Great Britain Advance Access publication on 31 July 2016

Goodness-of-fit tests for the cure rate in a mixture cure model

BY U.U. MÜLLER
Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.
uschi@stat.tamu.edu

AND I. VAN KEILEGOM
Research Centre for Operations Research and Business Statistics, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium
ingrid.vankeilegom@kuleuven.be

SUMMARY

We consider models for time-to-event data that allow that an event, e.g., a relapse of a disease, never occurs for a certain percentage p of the population, called the cure rate. We suppose that these data are subject to random right censoring and we model the data using a mixture cure model, in which the survival function of the uncured subjects is left unspecified. The aim is to test whether the cure rate p, as a function of the covariates, satisfies a certain parametric model. To do so, we propose a test statistic that is inspired by a goodness-of-fit test for a regression function by Härdle and Mammen. We show that the statistic is asymptotically normally distributed under the null hypothesis that the model is correctly specified and under local alternatives. A bootstrap procedure is proposed to implement the test. The good performance of the approach is confirmed with simulations. For illustration we apply the test to data on the times between first and second birth.

Some key words: Beran estimator; bootstrap; censoring; cure fraction; Kaplan–Meier estimator; logistic model; kernel estimator.

1. INTRODUCTION

The classical approach to time-to-event data assumes that an event will occur eventually. This is appropriate in many situations but not all: the events that a person marries or becomes a parent never occur for a certain percentage of the population. The same is true of the event that an unemployed person finds a job, since some people always stay unemployed. The proportion of subjects for which the event does not happen is often called cure rate and is of particular interest in health sciences, especially if there are covariates that are likely to affect it. In this case it would be useful to have a parametric model for the cure rate as a function of the covariates that is easy to interpret. We want to find out whether such a model fits the data well.

Let T denote the time until an event of interest occurs. The random variable T is typically called the survival time or failure time. Write $S(t \mid x) = \mathrm{pr}(T > t \mid X = x)$ $(t \geq 0)$ for the conditional survival function given that the covariate X equals x. We set $T = \infty$ if a subject is cured. Then $S_0(t \mid x) = \mathrm{pr}(T > t \mid T < \infty, X = x)$ is the conditional survival function given the covariate x and given that the subject is not cured.

The presence of censoring is typical for survival data, e.g., when patients leave a study before an event occurs. This makes estimation of the cure rate difficult because it is not clear whether they were cured or not when they left. In this paper we consider cure models for survival time data with right censoring. Two types of cure models are typically considered in the literature, the mixture cure model,

$$S(t \mid x) = \mathrm{pr}(T > t \mid X = x) = p(x) + \{1 - p(x)\}S_0(t \mid x), \quad t \geq 0, \tag{1}$$

where $p(\cdot)$ denotes the cure rate, and the promotion time cure model,

$$S(t \mid x) = \exp\{-\theta(x)F_0(t)\}, \quad t \geq 0,$$

where F_0 is a baseline distribution function, and usually $\theta(x) = \exp(\beta_0 + \beta_1 x)$. Parametric estimation in the latter model is, for example, discussed in Yakovlev et al. (1994); for semiparametric methods see, e.g., Tsodikov (1998, 2003).

A number of articles study the estimation of the mixture cure model (1). Most assume a logistic model for the cure rate $p(x)$, while various model assumptions have been made on the survival function $S_0(t \mid x)$. Parametric estimators are, for example, given in Boag (1949) and Farewell (1982). A semiparametric approach based on a proportional hazards model is provided by Kuk and Chen (1992), Sy and Taylor (2000), and Lu (2008), just to name a few, whereas nonparametric estimation of $S_0(t \mid x)$ is studied in a 2018 preprint on cure models in survival analysis by Patilea and Van Keilegom. The literature on cure models that do not assume a logistic model for $p(x)$ is much more sparse. Maller and Zhou (1992) propose a nonparametric estimator of the cure rate in a model that does not involve covariates; Laska and Meisner (1992) discuss nonparametric inference with discrete covariates, and Xu and Peng (2014) propose a conditional Kaplan–Meier estimator which involves one continuous covariate, and which we will use here to construct goodness-of-fit tests of the form of the cure rate.

Most papers in the literature assume either that a cure rate does not exist, or that it does exist and follows a logistic model. However, neither assumption may be correct. First, the existence of a cure rate has been disputed in the literature on cure models. Second, there is no reason to believe that the cure rate is always monotone in x, let alone that it is logistic. Hence we consider tests for the parametric form of the cure rate function. Tests for the logistic model and for the existence of a cure rate will be special cases. Tests for the existence of a cure rate have been proposed by Maller and Zhou (1994) and by Li et al. (2007), but neither incorporates covariates. To the best of our knowledge, the literature does not contain any test when covariates are present. Hence, our test is the first to allow the cure rate to depend on a covariate. The study of our test is also a first hurdle to be taken before the case of multi-dimensional covariates can be considered. The latter case requires a semi- or nonparametric estimator of $p(x)$ when X is multi-dimensional, however, which is yet to be developed in the literature.

Our test statistic is an adaptation of the statistic proposed by Härdle and Mammen (1993), who test for the parametric form of a regression function. We describe it in the next section, which also contains our asymptotic result, Theorem 1, which states the limiting normality of the test statistic. For the implementation of the test we recommend a bootstrap procedure.

2. MODEL AND THEORETICAL RESULTS

Let T be a survival time and C a censoring time. The observations are independent copies (Y_i, δ_i, X_i) $(i = 1, \ldots, n)$ of

$$Y = \min(T, C), \quad \delta = 1(T \leq C), \quad X,$$

where X is a covariate. We consider the mixture cure model (1), i.e.

$$S(t \mid x) = p(x) + \{1 - p(x)\}S_0(t \mid x), \quad t \geq 0.$$

The usual model for the cure rate p is the logistic model, i.e. $p(x) = p_\theta(x) = 1/\{1 + \exp(\theta_0 + \theta_1 x)\}$, with parameter vector $\theta = (\theta_0, \theta_1)$. We want to test whether the logistic model, or any other parametric model, is appropriate. The hypotheses are $H_0 : p = p_\theta$ for some $\theta \in \Theta$, and $H_1 : p \neq p_\theta$ for all $\theta \in \Theta$, where Θ is a finite-dimensional parameter space and the function p_θ is known up to a parameter vector $\theta \in \Theta$. The test statistic we propose is a weighted L_2 distance between a nonparametric estimator of $p(x)$ and a parametric estimator obtained under H_0. This is in the same spirit as the test statistic proposed by Härdle and Mammen (1993) in the context of tests for the parametric form of a regression function. More precisely, we define

$$\mathcal{T}_n = nh^{1/2} \int \{\hat{p}(x) - p_{\hat{\theta}}(x)\}^2 \pi(x)\, dx, \tag{2}$$

where $\hat{p}$ is a nonparametric estimator with bandwidth h, $\hat{\theta}$ estimates θ, and $\pi(\cdot)$ is a given density function. For practical purposes we recommend an empirical version $\tilde{\mathcal{T}}_n$ of the special case of $\mathcal{T}_n$ where $\pi(\cdot)$ equals the covariate density $f(\cdot)$ (Cao and González-Manteiga, 2008),

$$\tilde{\mathcal{T}}_n = nh^{1/2} \frac{1}{n} \sum_{i=1}^{n} \{\hat{p}(X_i) - p_{\hat{\theta}}(X_i)\}^2. \tag{3}$$

This statistic naturally puts more weight on covariates that are more likely.

Both $\mathcal{T}_n$ and $\tilde{\mathcal{T}}_n$ require a nonparametric estimator $\hat{p}$ and a parametric estimator $p_{\hat{\theta}}$ of the cure rate. The estimator $p_{\hat{\theta}}$ is obtained by plugging in an estimator $\hat{\theta}$ of θ. Estimation of θ is addressed in Remark 1 below. Our nonparametric estimator $\hat{p}$ is the estimator introduced in Xu and Peng (2014), which equals the conditional Kaplan–Meier estimator $\hat{S}$, evaluated at the largest uncensored survival time. The estimator $\hat{S}$ was introduced in 1981 by Beran in a Technical Report of the University of California at Berkeley entitled Nonparametric regression with randomly censored survival data. More precisely, we use

$$\hat{p}(x) = \hat{S}(Y^1_{(n)} \mid x), \quad Y^1_{(n)} = \max_{i:\delta_i=1} Y_i \tag{4}$$

and

$$\hat{S}(t \mid x) = 1(t \leq Y_{(n)}) \prod_{j=1}^{n} \Big\{ \frac{1 - \sum_{i=1}^n 1(Y_i \leq Y_j) w_{hi}(x)}{1 - \sum_{i=1}^n 1(Y_i < Y_j) w_{hi}(x)} \Big\}^{1(Y_j \leq t; \delta_j = 1)},$$

where $w_{h1}(x), \ldots, w_{hn}(x)$ are non-negative weights which add up to one, and $Y_{(i)}$ is the i-th order statistic. For simplicity we use the Nadaraya–Watson weights,

$$w_{hi}(x) = \frac{K_h(x - X_i)}{\sum_{j=1}^n K_h(x - X_j)}, \quad K_h(\cdot) = \frac{1}{h} K\Big(\frac{\cdot}{h}\Big),$$

where h is a bandwidth and K a kernel function. Note that $\hat{p}(x) = \hat{S}(\infty \mid x)$, since $\hat{S}(\cdot \mid x)$ is constant for values to the right of $Y^1_{(n)}$.

To ensure that $\hat{p}(x)$ is consistent for $p(x)$, we assume that

$$\tau_0(x) = \inf\{t : S_0(t \mid x) = 0\} \;<\; \inf\{t : G(t \mid x) = 0\} \text{ for every } x, \tag{5}$$

where G is the survival function of the censoring time C. The assumption is slightly weaker than the corresponding condition by Xu and Peng, who consider the supremum with respect to

x on the left-hand side. It implies that subjects censored after the time point $\tau_0(x)$ are known to be cured, although the cure state of censored subjects before $\tau_0(x)$ is unknown. The quantity $\tau_0(x)$ does not need to be estimated in practice. In fact, the assumption in (5) is used implicitly in the way we estimate the cure rate $p(x)$, which is done using the Beran estimator $\hat{S}(Y^1_{(n)}|x)$; see equation (4). Xu and Peng (2014) show that this is a consistent estimator of $p(x) = S\{\tau_0(x)|x\}$. They also prove that $(nh)^{1/2}\{S(Y^1_{(n)} \mid x) - p(x)\} = o_p(1)$, so asymptotically the replacement of $p(x)$ by $S(Y^1_{(n)}|x)$ has no effect on the variance. Whether assumption (5) is reasonable will depend on the application. However, this assumption, or a version of it, is generally accepted in the literature, and is needed to identify the model parameters. We refer to e.g., Taylor (1995) for a general discussion on the identifiability of mixture cure models.

Before we state our main result, the limiting normality of $\mathcal{T}_n$, we provide some notation and a list of technical assumptions. In the following we will write

$$H(t \mid x) = \mathrm{pr}(Y \le t \mid X = x), \quad H_1(t \mid x) = \mathrm{pr}(Y \le t, \delta = 1 \mid X = x)$$

and

$$\zeta_i(t \mid x) = \frac{1(Y_i \le t, \delta_i = 1)}{1 - H(Y_i \mid x)} - \int_0^t \frac{1(Y_i \ge s)}{\{1 - H(s \mid x)\}^2}\, dH_1(s \mid x), \quad i = 1, \ldots, n. \tag{6}$$

Let $\zeta(t \mid x)$ denote the version of $\zeta_i(t \mid x)$ that involves the base observations (Y, δ) instead of (Y_i, δ_i). Further set

$$\mu_{xz}(t) = E[\zeta\{\tau_0(x) \mid x\}\zeta\{\tau_0(z) \mid z\} \mid X = t]$$

and note that $\mu_{xx}(t) = E[\zeta^2\{\tau_0(x) \mid x\} \mid X = t]$.

To prove the limit theorem we will need the following assumptions.

(A1) The random variable X is quasi-uniform, i.e., X has a density f with support $[0, 1]$ and f is bounded and bounded away from zero on $[0, 1]$. The distribution function F has bounded second and third derivatives f' and f''.

(A2) The kernel function K is a symmetric probability density of order four with compact support and two bounded derivatives.

(A3) (i) There is a continuous nondecreasing and bounded function L_1 with $L_1(0) = 0$ and

$$|H(t_1 \mid x) - H(t_2 \mid x)| \le |L_1(t_1) - L_1(t_2)|, \quad x \in \mathbb{R},\, t_1, t_2 \ge 0.$$

(ii) The first three derivatives of $H(t \mid x)$ and $H_1(t \mid x)$ with respect to x exist and are bounded uniformly for all $x \in \mathbb{R}$ and $t \ge 0$.

(iii) There are continuous nondecreasing and bounded functions L_2 and L_3 with $L_2(0) = L_3(0) = 0$ and

$$\begin{aligned}
\Big|\frac{\partial H(t_1 \mid x)}{\partial x} - \frac{\partial H(t_2 \mid x)}{\partial x}\Big| &\le |L_2(t_1) - L_2(t_2)|, \quad x \in \mathbb{R},\, t_1, t_2 \ge 0, \\
\Big|\frac{\partial H_1(t_1 \mid x)}{\partial x} - \frac{\partial H_1(t_2 \mid x)}{\partial x}\Big| &\le |L_3(t_1) - L_3(t_2)|, \quad x \in \mathbb{R},\, t_1, t_2 \ge 0.
\end{aligned}$$

(A4) The functions $H(t \mid x)$, $S_0(t \mid x)$ and $G(t \mid x)$ are continuously differentiable in t, for any value of x, and have bounded second derivatives with respect to x, for any given value t.

(A5) The function $\mu_{xz}(z)$ is continuous in z for any value of x.

(A6) The cure rate $p(\cdot)$ is continuously differentiable; the vector of partial derivatives with respect to θ, $\dot{p}_\theta(\cdot)$, is Lipschitz in θ uniformly in x, i.e., for all $\theta, \theta' \in \Theta$, $|\dot{p}_\theta(x) - \dot{p}_{\theta'}(x)| \leq L(x)|\theta - \theta'|$ with $\sup_x L(x) < \infty$.

(A7) The estimator $\hat{\theta}$ is root-n consistent, i.e. $\hat{\theta} - \theta = O_p(n^{-1/2})$.

Assumptions (A3)(i) and (A3)(iii) are satisfied by a wide class of distributions. For instance, if $H(\cdot \mid x)$ is an exponential distribution with mean a_x^{-1}, then (A3)(i) is satisfied for $L_1(t) = 1 - \exp(-mt)$ with $m = \inf_x a_x$, provided $m > 0$ and $\sup_x a_x < \infty$.

Remark 1. In order to handle the finite-sample bias we choose, similarly to Härdle and Mammen (1993) and Cao and González-Manteiga (2008), the estimator $\hat{\theta}$, which is used to construct the cure rate model p_θ, as follows. First we create a new data set $\{X_1, \hat{p}(X_1)\}, \ldots, \{X_n, \hat{p}(X_n)\}$, where $\hat{p}$ is again Xu and Peng's estimator given in (4). Then, based on these new data, the estimator $\hat{\theta}$ is obtained by maximising a version of a Bernoulli log-likelihood,

$$\hat{\theta} = \text{argmax}_\theta \sum_{i=1}^{n} \big[\hat{p}(X_i) \log p_\theta(X_i) + \{1 - \hat{p}(X_i)\} \log\{1 - p_\theta(X_i)\}\big]. \tag{7}$$

The root-n consistency of $\hat{\theta}$ follows from Theorems 1 and 2 in Chen et al. (2003).

Theorem 1 only requires that there is some root-n consistent estimator for the parameter θ in the cure rate model p_θ (assumption A7). For the construction of an alternative root-n-consistent estimator that is not based on the new data $\{X_i, \hat{p}(X_i)\}$, we refer to the 2018 preprint by Patilea and Van Keilegom, who propose a profile likelihood estimator of θ, and show the root-n consistency and asymptotic normality of their estimator. In particular they provide a complete list of assumptions and a detailed proof.

We now state our main result, the limiting normality of the test statistic $\mathcal{T}_n = nh^{1/2} \int \{\hat{p}(x) - p_{\hat{\theta}}(x)\}^2 \pi(x)\, dx$; see equation (2). The proof is given in the Appendix. Recommendations on bandwidth selection are provided in Remark 3.

THEOREM 1. *Suppose assumptions (A1)-(A7) and (5) hold true and that the bandwidth $h = h_n$ satisfies $nh^3(\log n)^{-5} \to \infty$ and $nh^5/(\log n) = O(1)$ as $n \to \infty$. Then, under the null hypothesis, the statistic $\mathcal{T}_n$ from equation (2) is asymptotically normally distributed, $\mathcal{T}_n - b_h \to N(0, V)$ in distribution as $n \to \infty$. The asymptotic bias is $b_h = h^{-1/2} R(K) \int_0^1 p^2(x)\pi(x)\mu_{xx}(x)/f(x)\, dx = O(h^{-1/2})$, with $R(g) = \int g^2(t)\, dt$. The asymptotic variance is*

$$V = 2K^{(4)}(0) \int \Big\{ \frac{p^2(x)\pi(x)\mu_{xx}(x)}{f(x)} \Big\}^2 dx,$$

where $K^{(4)}$ denotes the fourth convolution product of K.

Under local alternatives of the form $p(x) = p_\theta(x) + n^{-1/2}h^{-1/4}\Delta_n(x)$, with $\Delta_n(x)$ bounded uniformly in x and n, $\mathcal{T}_n$ is asymptotically normally distributed,

$$\mathcal{T}_n - b_h - \int_0^1 \Delta_n(x)^2 \pi(x)\, dx \to N(0, V)$$

in distribution as $n \to \infty$. This reduces to the first statement under the null hypothesis, i.e., when $p(x) = p_\theta(x)$ and $\Delta_n \equiv 0$.

The asymptotic limit is degenerate under the hypothesis that there is no cure fraction ($p \equiv 0$). The test can only detect deviations of $p(x)$ from the null hypothesis larger than $n^{-1/2}$: the

assumptions on the bandwidths, $nh^3(\log n)^{-5} \to \infty$ and $nh^5/(\log n) = O(1)$, are, for example, satisfied if $h = n^{-1/5}$. For this choice the deviation from H_0 has order $n^{-1/2}h^{-1/4} = n^{-9/20}$, i.e., larger than $n^{-1/2}$.

Remark 2. The second test statistic $\tilde{\mathcal{T}}_n$ from equation (3) can be regarded an empirical version of $\mathcal{T}_n$ for the special case $\pi(\cdot) = f(\cdot)$, i.e., with $\pi(x)dx = dF(x)$ replaced by the $d\hat{F}(x)$, with $\hat{F}$ denoting the empirical distribution function. Using a smoothed version of $\hat{F}$ for an intermediate step, one can show that $\mathcal{T}_n$ and $\tilde{\mathcal{T}}_n$ are asymptotically equivalent,

$$\tilde{\mathcal{T}}_n - \mathcal{T}_n = nh^{1/2} \int \{\hat{p}(x) - p_{\hat{\theta}}(x)\}^2 \, d\{\hat{F}(x) - F(x)\} = o_p(1). \tag{8}$$

The proof of (8) is outlined in the Supplementary Material.

Applying Theorem 1 to the case $\pi(\cdot) = f(\cdot)$, it follows that $\tilde{\mathcal{T}}_n$ is also asymptotically normally distributed under H_0, $\tilde{\mathcal{T}}_n - \tilde{b}_h \to N(0, \tilde{V})$ in distribution as $n \to \infty$, with asymptotic bias $\tilde{b}_h = h^{-1/2}R(K)\int_0^1 p^2(x)\mu_{xx}(x)\,dx = O(h^{-1/2})$ and asymptotic variance

$$\tilde{V} = 2K^{(4)}(0) \int \{p^2(x)\mu_{xx}(x)\}^2 dx.$$

3. BOOTSTRAP PROCEDURE

Extensive simulations for a scenario similar to that in Table 1 revealed that the normal approximation given in Theorem 1 does not work well in practice: we obtained histograms of the test statistic for 2000 samples of size $n = 50$ up to size $n = 8000$, which were roughly bell-shaped for sample sizes $n = 2000$ and larger, but skewed throughout. This suggests that the normal approximation is only justified for very large samples. We therefore prefer to approximate the critical values of our test using a bootstrap procedure given below. This is in line with Härdle and Mammen (1993), who also noticed that a test based on the asymptotic result using normal quantiles cannot be recommended due to the slow convergence rate and since the approximation is only of first order. The bootstrap method has the additional advantage that mean and variance do not need to be estimated. Estimating mean and variance of the normal limit is quite difficult since both quantities involve the unknown function $\mu_{xx}(x)$. Since the bootstrap approximation gives an accurate estimation of the level of the test even for samples of size as small as $n = 50$, as can be seen from Table 1 in Section 4, we would prefer the bootstrap method in practice. The consistency of the bootstrap is stated after the description of the procedure in Theorem 2.

In what follows pr^* denotes the probability with respect to the bootstrap data and conditionally on the original data.

Step 1. Use the data and a pilot bandwidth $h = h_0$ to calculate, for $i = 1, \ldots, n$, the nonparametric estimator $\hat{p}_{h_0}(X_i)$.

Step 2. Fit the parametric model p_θ for p into the new data that were obtained in the previous step, i.e., use the $\hat{p}_{h_0}(X_i)$'s to calculate the parametric estimators $p_{h_0,\hat{\theta}}(X_i)$, $i = 1, \ldots, n$.

Step 3. Write $\hat{S}_{h_0}(\cdot \mid x)$ for $\hat{S}(\cdot \mid x)$ with bandwidth $h = h_0$. For $b = 1, \ldots, B$ we proceed as follows. First we generate binary bootstrap data $Z^*_{i,b}$ $(i = 1, \ldots, n)$ such that $\mathrm{pr}^*(Z^*_{i,b} = 0) = p_{h_0,\hat{\theta}}(X_i)$. For $i = 1, \ldots, n$ set $T^*_{i,b} = \infty$ or a large number if $Z^*_{i,b} = 0$; if $Z^*_{i,b} = 1$ generate

$$T^*_{i,b} \sim \frac{\hat{S}_{h_0}(\cdot \mid X_i) - \hat{p}_{h_0}(X_i)}{1 - \hat{p}_{h_0}(X_i)}.$$

For $i = 1, \ldots, n$ generate $C^*_{i,b} \sim \hat{G}_{h_0}(\cdot \mid X_i)$, where

$$\hat{G}_{h_0}(t \mid x) = 1(t \leq Y_{(n)}) \prod_{j=1}^{n} \Big\{ \frac{1 - \sum_{i=1}^{n} 1(Y_i \leq Y_j) w_{h_0 i}(x)}{1 - \sum_{i=1}^{n} 1(Y_i < Y_j) w_{h_0 i}(x)} \Big\}^{1(Y_j \leq t; \delta_j = 0)}.$$

The bootstrap data are $(Y^*_{i,b}, \delta^*_{i,b}, X_i)$ $(i = 1, \ldots, n)$ with

$$Y^*_{i,b} = \min(T^*_{i,b}, C^*_{i,b}), \quad \delta^*_{i,b} = 1(T^*_{i,b} \leq C^*_{i,b}).$$

Use the bootstrap data and a bandwidth h to calculate the nonparametric bootstrap estimator $\hat{p}^*_{hh_0,b}(X_i)$ for $p(X_i)$ $(i = 1, \ldots, n)$. Analogously to step (2), use these estimates to obtain the parametric estimates $p_{hh_0,\hat{\theta}^*,b}(X_i)$ $(i = 1, \ldots, n)$, where $\hat{\theta}^*$ is the bootstrap version of the maximum likelihood estimator from Remark 1, equation 7, with $\hat{p}^*_{hh_0,b}$ in place of $\hat{p}$. Alternatively, a bootstrap version of Patilea and Van Keilegom's profile likelihood estimator mentioned in the same remark can be used.

Step 4. Order $\mathcal{T}^*_{n,1}, \ldots, \mathcal{T}^*_{n,B}$, where

$$\mathcal{T}^*_{n,b} = nh^{1/2} \frac{1}{n} \sum_{i=1}^{n} \{\hat{p}^*_{hh_0,b}(X_i) - p_{hh_0,\hat{\theta}^*,b}(X_i)\}^2, \quad b = 1 \ldots, B, \tag{9}$$

and select the $\{(1-\alpha)B\}$-th order statistic as the critical value of the test.

For the following theorem, which shows the consistency of the above bootstrap procedure, we need to introduce additional assumptions.

(A8) The estimator $\hat{\theta}^*$ from Step 3 is root-n consistent, i.e., $\hat{\theta}^* - \hat{\theta} = O_{p^*}(n^{-1/2})$ almost surely.

(A9) The derivatives $\partial^2/\partial x^2 H(t \mid x)$, $\partial^2/\partial t^2 H(t \mid x)$, $\partial^2/(\partial x \partial t) H(t \mid x)$, $\partial^2/\partial x^2 H_1(t \mid x)$, $\partial^2/\partial t^2 H_1(t \mid x)$ and $\partial^2/(\partial x \partial t) H_1(t \mid x)$ exist and are continuous in (x, t).

To see that assumption (A8) is reasonable, we refer to Theorem B in Chen et al (2003), who show the root-n consistency and limiting distribution of the bootstrap version $\hat{\theta}^*$ of a general semiparametric estimator $\hat{\theta}$. In particular, we can take $\hat{\theta}$ equal to any of the two estimators of θ considered in Remark 1, provided we adapt the bootstrap procedure in Chen et al (2003) to the model based bootstrap procedure described above. A formal proof of assumption (A8) is straightforward but lengthy. We omit it for reasons of brevity.

THEOREM 2. *Consider the bootstrap statistic $\mathcal{T}^*_n$ introduced in equation (9). Suppose assumptions (A1)-(A9) and (5) hold true and that the bandwidths $h = h_n$ and $h_0 = h_{0n}$ satisfy $h = Cn^{-1/5}$ for some $C > 0$, $h_0 \to 0$, $nh_0^5 / \log n \to \infty$ and $(nh_0^5 / \log n)(h/h_0) = O(1)$ as $n \to \infty$. Then,*

$$\sup_t \Big| \mathrm{pr}^*(\mathcal{T}^*_n \leq t) - \Phi\Big(\frac{t - b_h}{V}\Big) \Big| = o(1)$$

almost surely, where Φ is the distribution function of a standard normal random variable. In particular we have under H_0,

$$\sup_t \Big| \mathrm{pr}^*(\mathcal{T}^*_n \leq t) - \mathrm{pr}(\mathcal{T}_n \leq t) \Big| = o(1)$$

almost surely.

The proof of this result is given in the Appendix.

Remark 3. The bandwidths h and h_0 can be chosen as $h = c_h n^{-0.2}$ and $h_0 = c_h n^{-0.11}$, i.e., $h_0 = h \times n^{0.09}$, with, for example, $c_h = 1$, 1.5 or 2. The choice for h is motivated by the conditions stated in Theorem 1, whereas the choice of h_0 comes from Li and Datta (2001, Remark 2.1, p. 714). For practical applications we recommend selecting the bandwidth h of the nonparametric estimator $\hat{p}(x)$ via cross-validation, which is data-driven and straightforward to apply. The cross-validation criterion can be found in, e.g., Geerdens et al. (2018, (2.6)). Their criterion has been proposed for the Beran estimator $\hat{S}(\cdot \mid x)$, and hence it can also be applied to $\hat{p}(x) = \hat{S}(Y^1_{(n)} \mid x)$. For the pilot bandwidth one can simply take $h_0 = h \times n^{.09}$, in accordance with Li and Datta (2001).

4. SIMULATIONS

4·1. *Scenario*

We now check the performance of our test procedure. We test whether the logistic model is appropriate for p, and we also consider if a cure fraction p exists. For the simulations we used R and the following model. Let X be a uniform random variable on the interval $[-1, 1]$ and, for a given value x of X, generate a survival time T from model (1) with conditional survival function

$$S_0(t \mid x) = \begin{cases} 1 - 0.98\frac{F(t|x)}{F(\tau|x)} - 0.02I(t = \tau), & 0 \le t \le \tau, \\ 0, & t > \tau, \end{cases}$$

where $F(t \mid x) = 1 - \exp\{-(t/b_x)^{a_x}\}$ $(t \ge 0)$ is a Weibull distribution function with shape parameter $a_x = (1 + x)^{c_T}$, $c_T \ge 0$, and scale parameter $b_x = \exp\{-(1 + x)/(2a_x)\}$. Further $\tau = F^{-1}(0.9)$, with $F(t) = 1 - \exp\{-(t/b)^a\}$, $a = E(a_X)$ and $b = E(b_X)$, which means that $\tau = b(\log 10)^{1/a}$. We truncate at τ to make sure that condition (5) is satisfied. In addition, we generate C independently of X from an exponential distribution with mean 1.5, i.e., $G(t \mid x) = G(t) = \exp(-2t/3)$ for $t \ge 0$.

4·2. *Standard logistic vs. extended logistic*

For our first illustration we assume a standard logistic model under the null hypothesis, i.e.,

$$H_0 : p(x) = p_\theta(x) = \frac{1}{1 + \exp(\theta_0 + \theta_1 x)}, \quad x \in \mathbb{R},$$

and study the performance of our test if in fact an extended logistic model is true, i.e., if

$$p(x) = \frac{1}{1 + \exp(\theta_0 + \theta_1 x - \theta_2 x^2)} \neq p_\theta(x),$$

for appropriate values of θ_0, θ_1 and $\theta_2 \neq 0$. Note that $\theta_2 = 0$ yields the standard logistic model. The proportion of censoring $p_{\text{cens}} = \text{pr}(C < T)$ and the average cure proportion $p_{\text{cure}} = E\{p(X)\}$ are given in Table 1 for several values of c_T, θ_0, θ_1 and θ_2. Table 1 also shows the rejection proportions of the test statistic $\tilde{\mathcal{T}}_n$ defined in (3) for each of these settings. We work with $n = 50, 100$ and 200, and with bandwidths h and h_0 as stated in Remark 3 above. We used the R package *np*, which provides the cross-validation bandwidth for estimating a conditional distribution. For the kernel function K we use the Epanechnikov kernel on $[-1, 1]$.

c_T	θ_0	θ_1	θ_2	p_{cens}	p_{cure}	$n=50$ c_h 1	1.5	2	CV	$n=100$ c_h 1	1.5	2	CV	$n=200$ c_h 1	1.5	2	CV
0	1	1	0	46	28	6	5	6	6	5	4	5	5	5	4	5	4
			1	51	35	8	8	8	8	10	10	11	10	16	22	21	18
			2	56	42	19	20	20	19	36	39	40	41	64	66	69	67
			4	66	54	59	65	65	64	90	92	93	92	100	100	100	100
	5	5	0	54	38	5	6	4	5	6	6	6	5	5	5	5	5
			1	60	46	8	8	8	7	15	17	17	17	26	28	29	29
			2	65	53	22	25	27	26	48	51	53	53	73	76	79	78
			4	74	64	52	63	63	61	89	92	94	92	100	100	100	100
1	1	1	0	45	28	5	5	6	4	5	4	5	4	4	5	5	4
			1	51	35	7	7	8	8	10	10	11	10	19	19	20	19
			2	56	42	19	21	19	21	37	39	41	39	64	71	72	67
			4	66	54	61	65	66	65	91	93	94	93	100	100	100	100
	5	5	0	52	38	5	5	4	4	4	4	5	5	5	5	6	5
			1	59	46	5	7	7	7	14	18	18	16	27	32	33	30
			2	65	53	22	24	24	24	50	53	55	52	74	79	82	78
			4	73	64	53	62	61	61	90	92	95	94	99	100	100	100

Table 1. *Proportion (%) of censoring* $p_{\text{cens}} = \text{pr}(C < T)$*, proportion of cure* $p_{\text{cure}} = E\{p(X)\}$ *and rejection proportion of* $\tilde{\mathcal{T}}_n$ *for several values of* n*,* c_h*,* c_T*,* θ_0*,* θ_1 *and* θ_2*. The null hypothesis corresponds to* $\theta_2 = 0$*; the significance level is* $\alpha = 0.05$*.*

Table 1 is based on 500 simulation runs, where for each sample the bootstrap critical value was obtained from 500 bootstrap samples. The nominal level is $\alpha = 0.05$. The table rows corresponding to $\theta_2 = 0$ show that under the null hypothesis that the logistic model is true the level is well respected, even for sample size $n = 50$. If a binomial distribution with success probability 0.05 is appropriate, the standard deviation of the simulated rejection probabilities is $\{(0.05)(0.95)/500\}^{1/2} = 0.01$. Table 1 also shows that the rejection proportion increases with the sample size and with the value of θ_2, as is to be expected. The constant c_h, which determines the value of the bandwidths h and h_0, does not seem to have an important impact on the rejection proportion. The data-driven cross-validation method yields very similar results, so both approaches to bandwidth selection appear to be acceptable. These findings are true for both values of c_T and for both choices of the vector (θ_0, θ_1) considered in Table 1. The simulated rejection probabilities have a standard deviation bounded above by 0.02.

A degenerate special case is given if $p \equiv 0$, which can be regarded as a standard logistic model with $\theta_1 = 0$ and a very large θ_0. We therefore expect that in this case the test will have difficulties rejecting the null hypothesis of a logistic cure rate, $p(x) = 1/\{1 + \exp(\theta_0 + \theta_1 x)\}$. This is indeed confirmed by a small simulation study: see Table 2, with $c_T = 1$ and cross-validation bandwidth as before, which shows power close to the nominal level $\alpha = 0.05$.

n	50	100	200
$p \equiv 0$	11	5	5
probit model	5	5	6

Table 2. *Power (%) of the test for a logistic model, when in reality there is no cure fraction (row 1) and when the true cure rate is a probit model (row 2).*

4·3. *Logistic vs. probit*

We also carried out a small simulation study to explore what happens if we test again for a logistic model, but the probit model is more appropriate for the cure rate. We therefore generated data from a probit model $1 - p(x) = \Phi(\theta_0 + \theta_1 x)$ with $\theta_0 = \theta_1 = 0.5$. The model for T, C and X is as in Table 1 with $c_T = 1$ and a bandwidth chosen by cross-validation. The results are in Table 2. The power figures are close to the nominal level $\alpha = 0.05$. This is, however, not surprising: the probit and the logistic model are very close to each other, except in the tails, so that large samples are necessary "for even modest sensitivity" (Chambers and Cox, 1967).

4·4. *Logistic vs. mixture model*

The previous scenario presents, of course, a special case. Further simulations, with models for the alternative that are clearly different from the logistic curve, yielded much better results. We generated, for example, data from a mixture of a logistic model and a polynomial of order three:

$$1 - p(x) = (1-\alpha)\frac{\exp(\theta_0 + \theta_1 x)}{1 + \exp(\theta_0 + \theta_1 x)} + \alpha\big\{0.9(\theta_0 + \theta_1 x) + (\theta_0 + \theta_1 x)^2 - (\theta_0 + \theta_1 x)^3\big\},$$

for $-1 \le x \le 1$, $0 \le \alpha \le 1$ and $\theta_0 = \theta_1 = 0.75$. Hence $0 \le \theta_0 + \theta_1 x \le 1.5$ for all $-1 \le x \le 1$. On the interval $[0, 1.5]$ the function $u \to 0.9u + u^2 - u^3$ takes values between 0 and 1 and is not monotone. We use the cross-validation bandwidth as before, and the same scenario as for Table 1, with $c_T = 1$. Table 3 shows that the power of this test increases with sample size and with the value of α, as can be expected.

α	$n = 50$	$n = 100$	$n = 200$
0	5	6	6
0.25	5	10	14
0.50	12	28	51
0.75	27	57	84
1	46	78	99

Table 3. *Power (%) of the test for a logistic model, when the true cure rate is a mixture of a logistic model and a polynomial model.*

4·5. *Testing for a cure rate*

Finally we used simulations to test whether $p \equiv 0$, i.e., whether a cure rate exists, which is an important question in applications. Although the asymptotic result shows that the limit is degenerate, so the rate of convergence is faster than the rate $nh^{1/2}$ in Theorem 1, the bootstrap is capable of detecting deviations from the null hypothesis.

We considered again the scenario from Table 1, in particular the same survival function S_0. Only the probability $p(\cdot)$ differs: we consider the null hypothesis, $p \equiv 0$, and alternatives $p \equiv 0.05, 0.15, 0.25$ in Table 4, as well as $p(x) = 1/[1 + \exp\{\theta(1 + 0.2x)\}]$ in Table 5, with θ chosen such that $p(0) = 0.05, 0.15, 0.25$. Apart from the case $n = 50$, the test again respects the level $\alpha = 5\%$. It also appears to be very powerful for all alternatives and sample sizes.

c_T	p	p_{cens}	$n=50$				$n=100$				$n=200$			
			c_h			CV	c_h			CV	c_h			CV
			1	1.5	2		1	1.5	2		1	1.5	2	
0	0	0.25	11	9	7	8	6	5	4	5	5	5	4	4
	0.05	0.28	40	46	52	50	57	65	73	69	84	90	92	91
	0.15	0.36	76	83	86	83	97	99	99	99	100	100	100	100
	0.25	0.44	94	96	97	95	100	100	100	100	100	100	100	100
1	0	0.22	11	10	10	10	7	5	4	5	5	4	3	4
	0.05	0.26	47	49	53	51	63	69	73	67	88	94	95	92
	0.15	0.34	79	81	85	80	98	98	99	98	100	100	100	100
	0.25	0.42	95	97	97	97	100	100	100	100	100	100	100	100

Table 4. *Proportion of censoring p_{cens} and rejection proportion (%) of $\tilde{\mathcal{T}}_n$ for several values of n, c_h and c_T. The null hypothesis corresponds to $p \equiv 0$, the alternatives are $p \equiv 0.05, 0.15, 0.25$; the significance level is $\alpha = 0.05$.*

c_T	$p(0)$	p_{cens}	$n=50$				$n=100$				$n=200$			
			c_h			CV	c_h			CV	c_h			CV
			1	1.5	2		1	1.5	2		1	1.5	2	
0	0.05	0.28	39	47	52	51	58	66	72	68	88	91	95	93
	0.15	0.36	80	84	88	85	97	98	99	99	100	100	100	100
	0.25	0.44	94	95	96	95	100	100	100	100	100	100	100	100
1	0.05	0.26	47	53	58	52	66	74	78	72	88	93	96	92
	0.15	0.34	80	83	86	83	97	98	99	98	100	100	100	100
	0.25	0.42	93	95	96	95	100	100	100	100	100	100	100	100

Table 5. *Proportion of censoring p_{cens} and rejection proportion of $\tilde{\mathcal{T}}_n$ as in Table 4, now with alternatives $p(x) = 1/[1 + \exp\{\theta(1 + 0.2x)\}]$, with θ chosen such that $p(0) = 0.05, 0.15, 0.25$, which is also equal to $E\{p(X)\}$.*

5. Application

As an illustration of our method we analysed data from by the Medical Birth Registry of Norway (see http://folk.uio.no/borgan/abg-2008/data/data.html). The data contain information on births in Norway since 1967, related to a total of 53,558 women. We are interested in the time between the births of the first two children, and how this is affected if the first child died within one year. The covariate of interest is *age*, X, which is the age of the mother at the birth of the first child. The censoring indicator equals 1 if the mother had a second birth, and 0 if the observation is censored. We want to test whether the logistic model is suitable for the cure rate p, i.e., the fraction of women who gave birth only once.

We consider two subpopulations: the first subpopulation, case 1, is concerned with the $n = 262$ observations for which the first child has died within one year; for case 2 we consider a random subset of size $n = 300$ of the entire data set. Our main reason for working with a sample and not with the complete data set is feasibility. The subset is obtained using the random number generator in R. We expect that the cure rate, i.e., the probability of no second birth, is different in the two groups. This is confirmed by our analysis; see Figures 1 and 2 for case 1 and Figures 3 and 4 for case 2. For case 1, i.e., the first child did not survive the first 12 months, the logistic

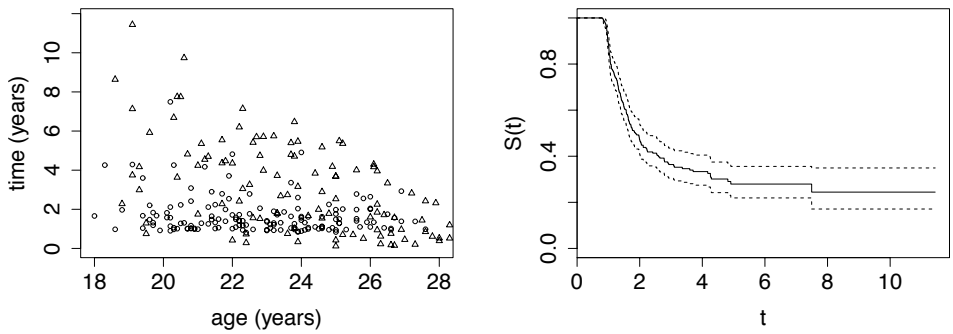

Fig. 1. *Left: Raw (univariate) data (case 1); circles are uncensored, triangles are censored data. Right: Kaplan–Meier estimator of all data; the plateau (containing 7 observations) suggests that a cure fraction is present (sufficiently long follow-up).*

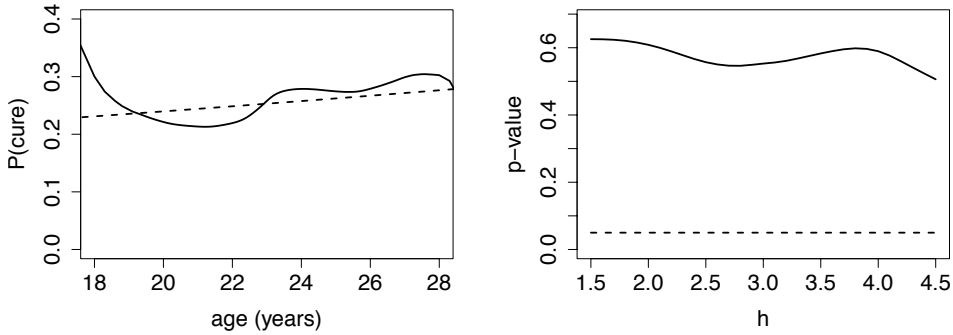

Fig. 2. *Left: Nonparametric estimator of $p(x)$ for case 1 with $h = 3$ (solid curve); logistic estimator of $p(x)$ (dashed curve). Right: Significance trace (plot of p-value as a function of h), clearly showing that the null hypothesis of logistic cure rate is not rejected for a wide range of possible bandwidths; the dashed line is the 5% level.*

model seems to be appropriate: the null hypothesis is not rejected for all values of the bandwidth h. For case 2, however, the logistic model is not appropriate: the p-value is close to zero for all values of h. A larger subset of the entire data set of size $n > 300$ would lead to even smaller p-values than the ones shown in Figure 4, and would hence lead to an even stronger rejection of the null hypothesis. For our analysis we chose $h_0 = h \times n^{0.09}$, which was motivated by the simulations. The number of bootstrap samples is 500, the time between the first and second births is expressed in years. Further comments are provided in the figure captions.

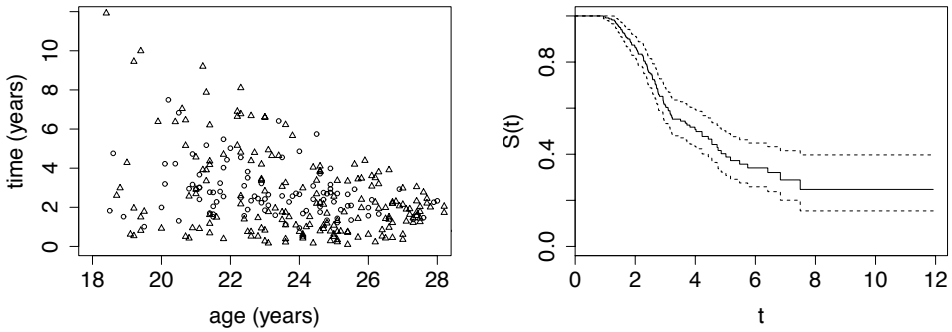

Fig. 3. *Raw data (case 2) and Kaplan–Meier estimator as in Figure 1; again there is a plateau indicating that a cure fraction is present.*

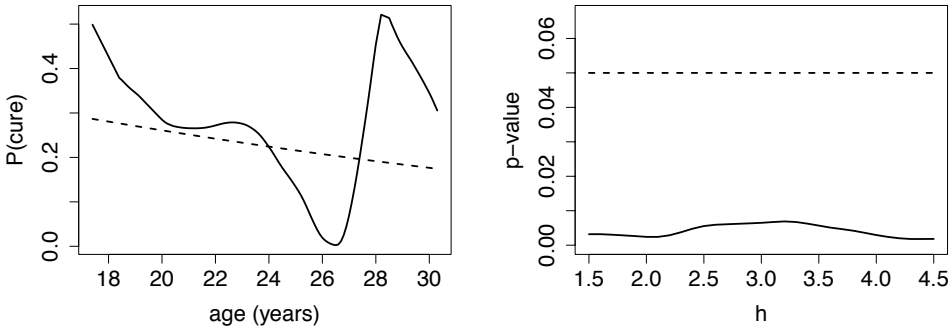

Fig. 4. *Left: Nonparametric estimator of $p(x)$ for case 2 with $h = 3$ (solid curve); logistic estimator of $p(x)$ (dashed curve). Right: Significance trace as in Figure 2, this time showing that the null hypothesis of logistic cure rate is rejected for all bandwidths in a wide range; the dashed line is again the 5% level.*

6. CONCLUDING REMARKS

A reasonable next step would be to target models with more than one covariate, since most data sets contain information in the form of many relevant covariates. In order to avoid the curse of dimensionality, it would make sense to work with dimension-reducing techniques. For example, one could replace Xu and Peng's estimator $\hat{p}(X)$ by a single index model estimator $\hat{p}(S)$ based on an estimator of the index $S = \beta^\top X$, now with X multi-dimensional containing various types of covariates, and a parameter vector β of matching dimension. The estimation of such a model, or any other semiparametric model for the cure rate $p(x)$ that allows for multi-dimensional covariates, has not been studied so far in the literature. So the next step will be to tackle estimation of $p(x)$ in such models. The implementation of the test is then straightforward using our approach with the new cure rate estimator $\hat{p}(X)$ in place of Xu and Peng's estimator.

ACKNOWLEDGMENT

The authors would like to thank the Referees and the Editors for helpful comments that greatly improved the paper. I. Van Keilegom acknowledges financial support from the European Research Council and from the Interuniversity Attraction Pole research network of the Belgian government (Belgian Science Policy).

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online contains the proofs of Theorem 1 and of equation (8).

APPENDIX

Proof of Theorem 1

We only give an outline. A detailed proof can be found in the Supplementary Material.

Consider the test statistic $\mathcal{T}_n$ from equation (2), which can be written in the form

$$\mathcal{T}_n = nh^{1/2} \int_0^1 [\{\hat{p}(x) - p(x)\}^2 + \{p(x) - p_{\hat{\theta}}(x)\}^2 + 2\{p(x) - p_{\hat{\theta}}(x)\}\{\hat{p}(x) - p(x)\}]\pi(x)\, dx.$$

We write $\mathcal{T}_{0n}$ for the first term of $\mathcal{T}_n$,

$$\mathcal{T}_{0n} = nh^{1/2} \int_0^1 \{\hat{p}(x) - p(x)\}^2 \pi(x)\, dx,$$

and show that $\mathcal{T}_{0n}$ determines the distribution of the test statistic under the null hypothesis, i.e. the second and third term of $\mathcal{T}_n$ above are asymptotically negligible.

To prove asymptotic normality, we expand $(nh)^{1/2}\{\hat{p}(x) - p(x)\}$ using results by Xu and Peng (2014) and Du and Akritas (2002), which allow us to write $\mathcal{T}_{0n} = \mathcal{T}_{1n} + \mathcal{T}_{2n} + o_p(1)$, where

$$\mathcal{T}_{1n} = \frac{h^{1/2}}{n} \int_0^1 p^2(x) \sum_{i=1}^n \frac{K_h^2(x - X_i)}{f^2(x)} \zeta_i^2\{\tau_0(x) \mid x\}\pi(x)\, dx,$$

$$\mathcal{T}_{2n} = \frac{h^{1/2}}{n} \int_0^1 p^2(x) \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \frac{K_h(x - X_i)K_h(x - X_j)}{f^2(x)} \zeta_i\{\tau_0(x) \mid x\}\zeta_j\{\tau_0(x) \mid x\}\pi(x)\, dx.$$

The first term, $\mathcal{T}_{1n}$, yields the bias, $T_{1n} = b_h + o_p(1)$, whereas T_{2n} has a limiting normal distribution, $\mathcal{T}_{2n} \to N(0, V)$ in distribution. The latter is derived analogously to the proof in Härdle and Mammen (1993), using a result by de Jong (1987). This yields the desired normal approximation $\mathcal{T}_{0n} - b_h \approx N(0, V)$, with mean $b_h = O(h^{-1/2})$ tending to infinity as n increases.

Under local alternatives of the form $p(x) = p_\theta(x) + n^{-1/2}h^{-1/4}\Delta_n(x)$, the second term of $\mathcal{T}_n$ produces an additional shift,

$$nh^{1/2} \int_0^1 \{p(x) - p_{\hat{\theta}}(x)\}^2 = \int_0^1 \Delta_n(x)^2 \pi(x)\, dx + o_p(1),$$

which is not present if $\Delta_n = 0$, i.e., under the null hypothesis. □

Proof of Theorem 2

We use a decomposition and notation as in the proof of Theorem 1:

$$\begin{aligned}\mathcal{T}_n^* &= nh^{1/2}\int_0^1 \{\hat{p}^*_{hh_0}(x) - p_{hh_0,\hat{\theta}^*}(x)\}^2\pi(x)\,dx \\ &= nh^{1/2}\int_0^1 \Big[\big\{\hat{p}^*_{hh_0}(x) - p_{h_0,\hat{\theta}}(x)\big\}^2 + \big\{p_{h_0,\hat{\theta}}(x) - p_{hh_0,\hat{\theta}^*}(x)\big\}^2 \\ &\quad +2\big\{\hat{p}^*_{hh_0}(x) - p_{h_0,\hat{\theta}}(x)\big\}\big\{p_{h_0,\hat{\theta}}(x) - p_{hh_0,\hat{\theta}^*}(x)\big\}\Big]\pi(x)\,dx. \qquad (10)\end{aligned}$$

We start with the first term of (10), and in particular we want to show first why we center $\hat{p}^*_{hh_0}(x)$ by means of $p_{h_0,\hat{\theta}}(x)$. We know that $\hat{p}^*_{hh_0}(x) = \hat{S}^*_{hh_0}(Y^{*1}_{(n)} \mid x)$ and that $\hat{S}^*_{hh_0}(Y^{*1}_{(n)} \mid x) - \hat{S}^*_{hh_0}(Y^1_{(n)} \mid x) = o_{p^*}(1)$ almost surely, where $\hat{S}^*_{hh_0}(\cdot \mid x)$ is the Beran estimator based on the bootstrap data $(Y_1^*, \delta_1^*), \ldots, (Y_n^*, \delta_n^*)$, and where $Y^{*1}_{(n)} = \max_{i:\delta_i^*=1} Y_i^*$. We also know that for given x, the variable T^* is drawn from the survival function

$$\mathrm{pr}^*(T^* > t \mid x) = \frac{\hat{S}_{h_0}(t \mid x) - \hat{p}_{h_0}(x)}{1 - \hat{p}_{h_0}(x)}\{1 - p_{h_0,\hat{\theta}}(x)\} + p_{h_0,\hat{\theta}}(x).$$

Hence, $\mathrm{pr}^*(T^* > Y^1_{(n)} \mid x) = p_{h_0,\hat{\theta}}(x)$, since $\hat{p}_{h_0}(x) = \hat{S}_{h_0}(Y^1_{(n)} \mid x)$. This shows that $p_{h_0,\hat{\theta}}(x)$ is the correct centering term for $\hat{p}^*_{hh_0}(x)$.

As our bootstrap procedure is similar to that studied by Van Keilegom and Veraverbeke (1997) in the context of nonparametric regression with right-censored data without cure fraction, it can be shown, similarly as in Theorem 4.1 in their paper, that

$$\begin{aligned}&\hat{p}^*_{hh_0}(x) - p_{h_0,\hat{\theta}}(x) \qquad (11)\\ &= p(x)\Big[\sum_{i=1}^n w_{hi}(x)\zeta_i^*\{\tau_0(x) \mid x\} - \sum_{i=1}^n w_{h_0 i}(x)\zeta_i\{\tau_0(x) \mid x\}\Big] + O_{p^*}\{(nh)^{-3/4}(\log n)^{3/4}\},\end{aligned}$$

almost surely and uniformly in x, where $\zeta_i^*(\cdot \mid x)$ is as $\zeta_i(\cdot \mid x)$ except that (Y_i, δ_i) is replaced by (Y_i^*, δ_i^*). Defining $\eta_i^*(\cdot \mid x) = \zeta_i^*(\cdot \mid x) - \sum_{j=1}^n w_{h_0 j}(x)\zeta_j\{\tau_0(x) \mid x\}$, the latter can be written as

$$p(x)\sum_{i=1}^n w_{hi}(x)\eta_i^*\{\tau_0(x) \mid x\} + O_{p^*}\{(nh)^{-3/4}(\log n)^{3/4}\}. \qquad (12)$$

Now, we follow the arguments in the proof of Theorem 1, and write

$$\mathcal{T}_n^* = nh^{1/2}\int_0^1 \big\{\hat{p}^*_{hh_0}(x) - p_{h_0,\hat{\theta}}(x)\big\}^2\pi(x)\,dx + o_{p^*}(1) = \mathcal{T}_{1n}^* + \mathcal{T}_{2n}^* + o_{p^*}(1)$$

almost surely, where

$$\begin{aligned}\mathcal{T}_{1n}^* &= \frac{h^{1/2}}{n}\int_0^1 p^2(x)\sum_{i=1}^n \frac{K_h^2(x - X_i)}{f^2(x)}\eta_i^{*2}\{\tau_0(x) \mid x\}\pi(x)\,dx, \\ \mathcal{T}_{2n}^* &= \frac{h^{1/2}}{n}\int_0^1 p^2(x)\sum_{i=1}^n \sum_{\substack{j=1\\ j\neq i}}^n \frac{K_h(x - X_i)K_h(x - X_j)}{f^2(x)}\eta_i^*\{\tau_0(x) \mid x\}\eta_j^*\{\tau_0(x) \mid x\}\pi(x)\,dx.\end{aligned}$$

For $\mathcal{T}_{1n}^*$ we find that

$$E^*(\mathcal{T}_{1n}^*) = h^{-1/2} R(K) \int_0^1 \frac{p^2(x)\pi(x)}{f(x)} \mu_{xx}^*(x)\, dx,$$

where

$$\begin{aligned}\mu_{xx}^*(x) &= E^*\{\eta^*(\tau_0(x) \mid x)^2 \mid X = x\} = \mathrm{var}^*\{\zeta^*(\tau_0(x) \mid x)^2 \mid X = x\} \\ &= \mathrm{var}[\zeta\{\tau_0(x) \mid x\}^2 \mid X = x] + o(1)\end{aligned}$$

almost surely, using arguments similar to those in the proof of Lemma 8(b) in Van Keilegom and Veraverbeke (1997), and the latter equals $\mu_{xx}(x) + o(1)$ almost surely Hence, $\mathcal{T}_{1n}^* = E^*(\mathcal{T}_{1n}^*) + o_{p^*}(1) = b_h + o_{p^*}(1)$ almost surely Using similar arguments we obtain that $\mathrm{pr}^*(\mathcal{T}_{2n}^* \leq t) - \Phi(t/V) \to 0$ almost surely, which shows the result.

It remains to show that the second and third term of (10) are negligible. For the second term note that by assumption (A8) this term is $O_{p^*}(nh^{1/2}n^{-1}) = o_{p^*}(1)$. The third term can be decomposed in a similar way as in the proof of Theorem 1, which together with assumption (A8) and equations (11) and (12) shows that this term also vanishes asymptotically. □

REFERENCES

BOAG, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 11, 15-53.

CAO, R. AND GONZÁLEZ-MANTEIGA, W. (2008). Goodness-of-fit tests for conditional models under censoring and truncation. *J. Econometrics*, 143, 166-190.

CHAMBERS E.A. AND COX D.R. (1967). Discrimination between alternative binary response models, *Biometrika* **54**, 573-578.

CHEN, X., LINTON, O. AND VAN KEILEGOM, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* **71**, 1591-1608.

DE JONG, P. (1987). A central limit theorem for generalized quadratic forms. *Probab. Theory Related Fields.*, 75, 261-277.

DU, Y. AND AKRITAS, M.G. (2002). Uniform strong representation of the conditional Kaplan–Meier process. *Math. Methods Statist.*, 11, 152-182.

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

GEERDENS, C., ACAR, E. AND JANSSEN, P. (2018). Conditional copula models for right-censored clustered event time data. *Biostatistics*, 19, 247-262.

HÄRDLE, W. AND MAMMEN, E. (1993) Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, 21, 1926-1947.

KUK, A.Y.C. AND CHEN, C. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79, 531-541.

LASKA, E.M. AND MEISNER, M.J. (1992). Nonparametric estimation and testing in a cure model. *Biometrics*, 48, 1223-1234.

LI, G. AND DATTA, S. (2001). A bootstrap approach to nonparametric regression for right censored data. *Ann. Inst. Statist. Math.*, 53, 708-729.

LI, Y., TIWARI, R.C. AND GUHA, S. (2007). Mixture cure survival models with dependent censoring. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69, 285-306.

LU, W. (2008). Maximum likelihood estimation in the proportional hazards cure model. *Ann. Inst. Statist. Math.*, 60, 545-574.

MALLER, R.A. AND ZHOU, S. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika*, 79, 731-739.

MALLER, R.A. AND ZHOU, S. (1994). Testing for sufficient follow-up and outliers in survival data. *J. Amer. Statist. Assoc.*, 89, 1499-1506.

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.

TSODIKOV, A. (1998). Asymptotic efficiency of a proportional hazards model with cure. *Statist. Probab. Lett.*, 39, 237-244.

TSODIKOV, A. (2003). Semiparametric models: A generalized self-consistency approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 65, 759-774.

VAN KEILEGOM, I. AND VERAVERBEKE, N. (1997). Estimation and bootstrap with censored data in fixed design nonparametric regression. *Ann. Inst. Statist. Math.*, 49, 467-491.

YAKOVLEV, A.Y., CANTOR, A.B. AND SHUSTER, J.J. (1994). Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Stat. Med.*, 13, 983-986.

XU, J. AND PENG, Y. (2014). Nonparametric cure rate estimation with covariates. *Canad. J. Statist.*, 42, 1-17.

[*Received 2 January* 2017. *Editorial decision on* 1 *April* 2017]