KU LEUVEN

FACULTEIT ECONOMIE EN BEDRIJFSWETENSCHAPPEN

Nonparametric covariate significance tests for the incidence in cure models

López-Cheda A, Jácome M, Van Keilegom I, Cao R.

KBI_1821

Biometrics 000, 000–000 DOI: 000
000 0000

Nonparametric covariate significance tests for the incidence in cure models

Ana López-Cheda[1,*], M. Amalia Jácome[1,], Ingrid Van Keilegom[2,***] and Ricardo Cao[1,****]**

[1]Department of Mathematics, University of A Coruña, Campus de Elviña s/n, 15071 A Coruña, Spain

[2]ORSTAT, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium

**email:* ana.lopez.cheda@udc.es

***email:* majacome@udc.es

****email:* ingrid.vankeilegom@kuleuven.be

*****email:* rcao@udc.es

Summary: In cancer studies there may be long term survivors, which lead to heavy censoring at the end of the follow-up period. Since a standard survival model can not handle this data, a cure model is needed. In the literature, covariate significance tests for cure models are limited to parametric and semiparametric methods. We fill this important gap by proposing a nonparametric covariate significance test for the probability of cure in mixture cure models. The procedure is based on the significance test by (Delgado and González-Manteiga, 2001), and it is extended to non continuous covariates: binary, discrete and qualitative. Its efficiency is evaluated in a Monte Carlo simulation study, in which the distribution of the test is approximated by bootstrap. The method is applied to a colorectal cancer dataset.

Key words: Bootstrap; Censored data; Survival analysis.

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

1. Cure models

Classical methods to analyze lifetime data assume that all subjects would experience the failure if there is no censoring and they are followed for long enough. They do not consider the possibility of a group of nonsusceptible individuals that will not develop such event and can be considered as cured. However, there is a increasingly large number of situations where there are long-term survivors who can be deemed to be immune to the event of interest. One well-known example of long-term survivors is cancer studies.

Cure models usually require long-term follow-up and large sample sizes, together with empirical and biological evidence of a nonsusceptible subpopulation (Farewell, 1986). In the literature, the most popular cure model is the mixture cure model (a recent review of the cure models can be found in (Peng and Taylor, 2014) and in (Amico and Van Keilegom, 2018), between others). Mixture cure models, proposed by (Boag, 1949), allow to estimate the probability of cure, $1 - p(\mathbf{x})$, also known as *incidence*, and the survival function of the uncured population or latency, denoted by $S_0(t|\mathbf{x})$. The model can be formulated as follows:

$$S(t|\mathbf{x}) = 1 - p(\mathbf{x}) + p(\mathbf{x})S_0(t|\mathbf{x}),$$

where $\mathbf{x}$ is a set of covariates and $S(t|\mathbf{x})$ is the survival function of all the (cured and uncured) patients. The main advantage of this model is that it allows covariates to have different influence on cured and uncured patients. (Maller and Zhou, 1996) provided a detailed review of this model. The estimation of cure models has been extensively studied for parametric and semiparametric models (Farewell, 1986; Peng and Dear, 2000; Peng and Dear, 2000; Sy and Taylor, 2000; Peng et al., 2007; Zhang and Peng, 2009; Peng and Taylor, 2011; Wang et al., 2012).

A nonparametric incidence estimator for the incidence in the mixture cure model was firstly

introduced by (Xu and Peng, 2014). (López-Cheda et al., 2017a) proposed a completely nonparametric mixture cure model, with nonparametric approaches for both the incidence and the latency functions. Even though it is considered only one covariate, the method can be directly extended to a case with multiple covariates. This enables the mixture cure model with covariates to be addressed in a completely nonparametric way. Nonparametric estimation of the latency has been also considered in (López-Cheda et al., 2017b).

It is interesting to test if a covariate has some influence on the cure rate or on the survival time of the susceptible patients. (Müller and Van Keilegom, 2018) propose a test statistic to assess whether the cure rate, $1 - p$ (as a function of the covariates) satisfies a certain parametric model. However, to the best of our knowledge, no significance testing has been proposed yet for nonparametric cure models. To fill this important gap, a covariate significance test for the incidence is presented in this paper. The method is based on the significance test by (Delgado and González-Manteiga, 2001). Its efficiency is evaluated in a Monte Carlo simulation study, in which the distribution of the test is approximated by bootstrap. Furthermore, the methodology is applied to a real dataset.

The rest of the article is organized as follows. In Section 2 we introduce the notation and we give a detailed description of the nonparametric mixture cure model by (López-Cheda et al., 2017a). In Section 3 we focus on the significance tests for the incidence. In Section 3.1 we study if the cure rate, as a function of the covariate Z, can be considered as a constant value versus if it depends on Z (Case 1). Moreover, in Section 3.2 we study if the cure probability, as a function of $(X, \mathbf{Z})$, only depends on the covariate X (Case 2). The good performance of the test is assessed in several simulation studies in Section 4. In Section 5 we apply the

proposed methodology to a real dataset related to colorectal cancer patients from CHUAC (Complexo Hospitalario Universitario de A Coruña, Spain).

2. Nonparametric mixture cure models

Throughout this paper we assume that individuals are subject to random right censoring and that the censoring time, C, is independent of the time to occurrence of the event, Y, given the set of covariates, $\mathbf{X}$. The conditional distribution function of Y is $F(t|\mathbf{x}) = P(Y \leqslant t|\mathbf{X} = \mathbf{x})$, and the corresponding survival function is $S(t|\mathbf{x})$. We define $T = \min(Y, C)$ as the observed time and $\delta = I(Y \leqslant C)$ the uncensoring indicator. Moreover, the distribution functions of C and T are G and H, respectively. Let us denote by ν the cure indicator, with $\nu = 0$ if the individual is susceptible and $\nu = 1$ otherwise. Therefore, the conditional probability of not being cured is $p(\mathbf{x}) = P(\nu = 0|\mathbf{X} = \mathbf{x})$. Note that if $\nu = 1$, it is assumed that $Y = \infty$. Then, the mixture cure model can be written as:

$$S(t|\mathbf{x}) = 1 - p(\mathbf{x}) + p(\mathbf{x})S_0(t|\mathbf{x}),$$

where $1 - p(\mathbf{x})$ is the incidence and $S_0(t|\mathbf{x})$ is the latency. Let X be a univariate continuous covariate with density function $f(x)$. The observations will be $\{(X_i, T_i, \delta_i), i = 1, \ldots, n\}$, i.i.d. copies of the random vector (X, T, δ).

(Xu and Peng, 2014) introduced the following kernel type incidence estimator:

$$1 - \hat{p}_h(x) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i]} B_{h(i)}(x)}{\sum_{r=i}^{n} B_{h(r)}(x)}\right) = \hat{S}_h(T^1_{\max}|x), \tag{1}$$

where $\hat{S}_h(t|x)$ is the conditional product-limit estimator by (Beran, 1981), and $T^1_{\max} = \max_{i:\delta_i=1}(T_i)$ is the largest uncensored failure time. Here $T_{(1)} \leqslant T_{(2)} \leqslant \ldots \leqslant T_{(n)}$ are the ordered T_i's, and $\delta_{[i]}$ and $X_{[i]}$ are the corresponding uncensoring indicator and covariate concomitants. These authors also proved the consistency and asymptotic normality of the estimator in (1). Furthermore, (López-Cheda et al., 2017a) obtained an i.i.d. representation, found an

asymptotic expression of the mean squared error, and proposed a bootstrap bandwidth selection method for (1).

3. Significance tests for the incidence

Significance testing is of primary importance in regression analysis, because the number of potential covariates to be included in the model can be large. In particular, in mixture cure models, variable selection is of outstanding interest, since the covariates having an effect on the survival of the uncured patients are not necessarily the same as those impacting the probability of cure. We propose a covariate significance test for the incidence based on the method by (Delgado and González-Manteiga, 2001), who introduced a test for selecting explanatory variables in nonparametric regression without censoring. The main advantage over other smoothed tests is that it only requires a smooth nonparametric estimator of the regression function depending on the explanatory variables which are significant under the null hypothesis. This feature is computationally convenient and solves, in part, the problem of the "curse of dimensionality" when selecting regressors in a nonparametric context.

Let us denote by $\mathbf{W} = (\mathbf{X}, \mathbf{Z}) = (X_1, \ldots, X_q, Z_1, \ldots, Z_m)$ the explanatory covariates. We would like to test if the cure probability, as a function of the covariate vector $\mathbf{W}$, only depends on $\mathbf{X}$, but not on $\mathbf{Z}$:

$$H_0 : E\left(\nu|\mathbf{X}, \mathbf{Z}\right) \equiv 1 - p\left(\mathbf{X}\right) \text{ vs. } H_1 : E\left(\nu|\mathbf{X}, \mathbf{Z}\right) \equiv 1 - p\left(\mathbf{X}, \mathbf{Z}\right),$$

where the function $p(\mathbf{X}, \mathbf{Z})$ depends not only on $\mathbf{X}$ but also on $\mathbf{Z}$.

Note that ν is not observed due to the censoring, since it is unknown if a censored individual will be eventually cured ($\nu = 1$) or not ($\nu = 0$). The idea is to express the regression function of an unobservable (and unestimable) response, ν, as a regression function with response η,

which is not observable but estimable. Let us define the variable η as follows:

$$\eta = \frac{\nu(1 - I(\delta = 0, T \leqslant \tau(\mathbf{X})))}{1 - G\left(\tau(\mathbf{X})|\mathbf{X}\right)},$$

where $\tau(\mathbf{X})$ is a time beyond which a subject is considered cured and $\mathbf{X}$ is the covariate vector that influences the cure rate under H_0. It is easy to check that $E(\eta|\mathbf{X}) = E(\nu|\mathbf{X})$ if the distribution of $(C|\mathbf{X}, \nu = 0)$ equals that of $(C|\mathbf{X}, \nu = 1)$.

Note that if there is no covariate $\mathbf{X}$, in practice, we consider $\tau(\mathbf{x}_i) = T^1_{\max}$. In any other case, without loss of generality, we estimate $\tau(\mathbf{x}_i)$ in the following way for a continuous univariate covariate X: using a bandwidth h_τ, we consider a subset of individuals j with $|x_j - x_i| < h_\tau$, and $\tau(x_i)$ will be estimated as the largest T_j with $\delta_j = 1$ in the subset. If there is no $\delta_j = 1$ in the subset, then $\tau(x_i)$ is equal to the available $\tau(x_l)$ for the nearest x_l to x_i. If there are several nearest values to determine x_l, then we estimate $\tau(x_i)$ as the mean of those. Preliminary studies suggested that a good bandwidth choice is $h_\tau = \left(X_{(n)} - X_{(1)}\right)\ 0.25\, n^{-1/9}$. Moreover, it is assumed that C does not depend on the covariates, $\mathbf{X}$, and then $G\left(\tau(\mathbf{X}_i)|\mathbf{X}\right)$ is estimated by the product limit estimator, $\hat{G}(\tau(\mathbf{X}_i))$. This gives the following estimations for the η_i. If $\delta_i = 1$, then we know that $\nu_i = 0$, so we define $\hat{\eta}_i = 0$. Furthermore, if $\delta_i = 0$ and $T_i \leqslant \tau(\mathbf{X}_i)$, then $\hat{\eta}_i = \frac{\nu_i(1-1)}{1-\hat{G}(\tau(\mathbf{X}_i))} = 0$; whereas if $\delta_i = 0$ and $T_i > \tau(\mathbf{X}_i)$, then we define $\hat{\eta}_i = \frac{1}{1-\hat{G}(\tau(\mathbf{X}_i))}$.

For $\mathbf{W} = (\mathbf{X}, \mathbf{Z})$ two cases are considered in this paper, depending on the dimension of the covariates: Case 1, where $\mathbf{W} = Z$ is one-dimensional, and Case 2, where $\mathbf{W} = (X, \mathbf{Z})$, with a one-dimensional covariate X and an m-dimensional covariate $\mathbf{Z}$. The third general case, with $\mathbf{W} = (\mathbf{X}, \mathbf{Z})$ where $\mathbf{X}$ is $\mathbb{R}^q$-valued and $\mathbf{Z}$ is $\mathbb{R}^m$-valued, can be easily generalized from Case 2.

3.1 *Case 1*

First, we study if the cure rate, as a function of Z, is a constant value versus if it depends on the covariate:

$$H_0 : E\left(\nu|Z\right) = 1 - p \;\text{ constant }\; \text{ vs } H_1 : E\left(\nu|Z\right) = 1 - p(Z), \tag{2}$$

where $p(Z)$ is not a constant function of Z. Our test will be based on the observations $\{(Z_i, \hat{\eta}_i), i = 1, \ldots, n\}$. Following (Delgado and González-Manteiga, 2001), the statistics we propose is based on the following process:

$$T_n(z) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\eta}_i - \left(\frac{1}{n}\sum_{j=1}^{n}\hat{\eta}_j\right)\right) I\left(Z_i \leqslant z\right), \tag{3}$$

which is a weighted mean of the difference between the observations of η and the conditional mean of η under the null hypothesis. Possible test statistics are the Cramér-von Mises (CvM) test, $C_n = \sum_{i=1}^{n} T_n^2(Z_i)$, or the Kolmogorov-Smirnov (KS) test, $K_n = \max_{i=1,\ldots,n} |n^{1/2}T_n(Z_i)|$. The null distribution of the test statistic is approximated by bootstrap, using an independent naive resampling. Specifically, the bootstrap procedure is the following:

1. For $i = 1, 2, \ldots, n$, obtain Z_i^* and $\hat{\eta}_i^*$ from $(Z_1, \ldots, Z_n)$ and $(\hat{\eta}_1, \ldots, \hat{\eta}_n)$ independently, by random resampling with replacement.
2. With the bootstrap resample, $\{(Z_i^*, \hat{\eta}_i^*), i = 1, \ldots, n\}$, obtain the bootstrap version of T_n:
$$T_n^*(z) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\eta}_i^* - \left(\frac{1}{n}\sum_{j=1}^{n}\hat{\eta}_j^*\right)\right) I(Z_i^* \leqslant z)$$
and the corresponding bootstrap version of the Cramér-von Mises and Kolmogorov-Smirnov statistics, C_n^* and K_n^*.
3. Repeat B times Steps 1-2 in order to generate B values of C_n^* and K_n^*. Define the critical values d_C^* and d_K^* as the values which are in position $\lceil(1-\alpha)B\rceil$ in the corresponding sorted vector.
4. Compare the value of the statistic, C_n (respectively, K_n), obtained with the original sample with d_C^* (respectively, d_K^*), and reject the null hypothesis if $C_n > d_C^*$ (respectively, $K_n > d_K^*$).

In addition, the p-value can be calculated as the proportion of resamples for which the bootstrap statistic, C_n^* (K_n^*) is larger than the value of the statistic with the original sample, C_n (K_n).

5. Repeat Steps 1-4 κ times. The power of the test is approximated as the proportion of rejections out of κ.

Note that since Z_i^* and $\hat{\eta}_i^*$ are resampled independently in Step 1, the bootstrap resampling plan mimics H_0.

3.1.1 *Z nominal.* In the case with only one non-ordinal qualitative covariate, $\mathbf{W} = Z$, there is no natural way to order the values of Z from lowest to highest. This makes impossible to compute the indicator function in the test statistic (3). We propose to consider all the possible $k!$ combinations of the values of Z and compute $T_n(z)$ (and also C_n and K_n) for each "ordered" combination. Finally, we compute the maximum of C_n and K_n along all these possible permutations and compare it with the critical point obtained by bootstrap likewise.

A different approach consists of working with $k-1$ dummy variables. The main advantage of this method is that we only need to compute $k-1$ times the value of the statistic, whereas with the previous method, we have to compute the statistic $k!$ times. Therefore, this approach is considerably less computationally expensive. On the other hand, by adressing the covariance testing using dummy variables, every new dummy variable has to be tested individually and it could be the case that the test leads to different conclusions for the dummy variables.

3.2 *Case 2*

In this case, $\mathbf{W} = (X, \mathbf{Z})$ has $m+1$ dimension, with a one-dimensional covariate X and an m-dimensional covariate $\mathbf{Z}$. We study if the cure probability, as a function of $(X, \mathbf{Z})$, only

depends on the covariate X, that is:

$$H_0 : E\left(\nu|X, \mathbf{Z}\right) = 1 - p(X), \quad \text{vs } H_1 : E\left(\nu|X, \mathbf{Z}\right) = 1 - p(X, \mathbf{Z}), \tag{4}$$

where $p(X, \mathbf{Z})$ depends on $\mathbf{Z}$. To do this, we use the observations $\{(X_i, \mathbf{Z}_i, \hat{\eta}_i), i = 1, \ldots, n\}$. Note that in Case 2, we estimate $\tau(X_i)$ as mentioned in Section 3.

The test statistic is different depending on whether X is continuous or not. For the sake of brevity, in the simulation study of Section 4.2 we only considered X continuous or qualitative, and a one-dimensional continuous covariate Z. The results for other types of covariate X can be found in the Supplementary Material.

3.3 *X continuous*

Following (Delgado and González-Manteiga, 2001), the statistic is defined as:

$$T_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_h(X_i) \left(\hat{\eta}_i - \hat{m}_h(X_i)\right) I\left(\mathbf{W}_i \leqslant \mathbf{w}\right), \tag{5}$$

where $\hat{f}_h$ is the Parzen-Rosenblatt estimator of the density function of the covariate X, which depends on the bandwidth h, $\hat{m}_h$ is the nonparametric estimator of the regression function $m(x) = E\left(\hat{\eta}|X = x\right)$, obtained by the Nadaraya-Watson kernel method with the same bandwidth h, and $\leqslant$ stands for component-wise inequality. Note that the process in (5) is a weighted mean of the difference between the $\hat{\eta}_i$ and their conditional mean under the null hypothesis. Similarly to Case 1, we consider the Cramér-von Mises, $C_n = \sum_{i=1}^{n} T_n^2(\mathbf{W}_i)$ and the Kolmogorov-Smirnov, $K_n = \max_{i=1,\ldots,n} |n^{1/2} T_n(\mathbf{W}_i)|$ statistics. The test distribution under H_0 is approximated by bootstrap, considering the following procedure:

1. We fix the covariate $X_i^* = X_i$ and we obtain $\mathbf{Z}_i^*$ from $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ by random resampling with replacement, for $i = 1, 2, \ldots, n$. Furthermore, we compute $\hat{\eta}_{gi}^* = \hat{\hat{\eta}}_{gi} + v_i \hat{\varepsilon}_i$, where $\hat{\hat{\eta}}_{gi} = \hat{m}_g(X_i)$ is the Nadaraya-Watson kernel regression estimate computed with the original sample and pilot bandwidth g, v_i is obtained from a $N(0, 1)$ and $\hat{\varepsilon}_i = \hat{\eta}_i - \hat{\hat{\eta}}_{gi}$ is the i-th

residual. Note that the Nadaraya-Watson estimation of $m(X_i)$ is bounded between 0 and 1, i.e. $0 \leqslant \hat{\hat{\eta}}_{gi} \leqslant 1$, $i = 1, 2, \ldots, n$.

2. With the bootstrap resample, $\{(X_i, \mathbf{Z}_i^*, \hat{\eta}_i^*), i = 1, \ldots, n\}$, obtain:

$$T_n^*(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_h(X_i) \left(\hat{\eta}_{gi}^* - \hat{m}_h(X_i)\right) I\left(\mathbf{W}_i^* \leqslant \mathbf{w}\right),$$

and the bootstrap version of the Cramér-von Mises and Kolmogorov-Smirnov statistics, C_n^* and K_n^*.

3. Repeat Steps 1-2 B times in order to generate B values of C_n^* and K_n^*. Define the critical values d_C^* and d_K^* as the values which are in position $\lceil (1-\alpha)B \rceil$ in the sorted vectors.

4. Compare the value of the statistic, C_n (respectively, K_n) obtained with the original sample with d_C^* (respectively, d_K^*), and reject the null hypothesis if $C_n > d_C^*$ (respectively, $K_n > d_K^*$).

5. Repeat Steps 1-4 κ times. The power of the test is approximated as the percentage of rejections out of κ.

To mimic H_0 in (4), the values $\hat{\eta}_{gi}^*$ defined in Step 1 do not depend on $\mathbf{Z}_i^*$ (just on X_i).

Note that in this procedure we need to select a bandwidth, h, and a pilot bandwidth, g. Although the results are quite insensitive to the value of the pilot bandwidth, preliminary studies showed that a good choice is $g = 2h$. Moreover, since we do not have a bandwidth selection method, in the simulation study in Section 4.2 we consider bandwidths of the order suggested by (Delgado and González-Manteiga, 2001), $h = Cn^{-1/3}$. In practice, we suggest to use any of the bandwidth selection methods for nonparametric tests proposed in the literature. There are two main approaches: (Kulasekera and Wang, 1997), among others, focuses on power maximization under the alternative hypothesis, whereas (Martínez-Camblor, 2010; Martínez-Camblor and de Uña-Álvarez, 2013) consider the idea of minimizing p-values. The two approaches are strongly related (see (Martínez-Camblor and de Uña-Álvarez, 2013)).

3.4 *X categorical or discrete*

For a categorical or discrete variable X, the estimated density, $\hat{f}_h(X_i)$, and the estimated regression function, $\hat{m}_h(X_i)$, in the test statistic in (5) are replaced by

$$\hat{\Pi}(X_i) = \frac{1}{n}\sum_{j=1}^{n} I(X_j = X_i) \text{ and } \hat{m}(X_i) = \frac{\frac{1}{n}\sum_{j=1}^{n} I(X_j = X_i)\hat{\eta}_j}{\hat{\Pi}(X_i)},$$

respectively. Similarly as in Case 1, for a qualitative variable in $\mathbf{W} = (X, \mathbf{Z})$ with no intrinsic order in its values, the indicator function $I(\mathbf{W}_i \leqslant \mathbf{w})$ in the test statistic is computed for all the possible "ordered" permutations of the values of $\mathbf{W}$.

4. Simulation studies

The purpose of the simulation studies is to assess the practical behavior of the proposed significance tests. We work with three different sample sizes: $n = 50$, $n = 100$ and $n = 200$. A total of $\kappa = 5000$ trials and $B = 2000$ bootstrap resamples are drawn. All the simulation studies were coded in R language. The software is available at `http://dm.udc.es/modes/es/node/256`. An R package will be developed and uploaded in CRAN.

4.1 *Case 1*

For the sake of brevity, only the results for Z continuous and Z nominal are given here. The details of the behaviour of the test in Case 1 with other types of variable Z, such as discrete and binary, are given in the Supplementary Material.

Under the null hypothesis, $H_0 : \ E(\nu|Z) = 1 - p$, we consider four different constant values for the incidence: $1 - p = 0.2,\ 0.3,\ 0.5$ and 0.7. Under the alternative hypothesis, $H_1 : \ E(\nu|Z) = 1 - p(Z)$, we study the following two models if Z is continuous, only Model 1 if Z is qualitative. For both models, the censoring variable follows an exponential distribution with mean $1/0.3$.

Model 1. The incidence is $1 - p(z)$, where

$$p(z) = \frac{\exp(\beta_0 + \beta_1 z)}{1 + \exp(\beta_0 + \beta_1 z)}, \tag{6}$$

with $\beta_0 = 0.476$ and $\beta_1 = 0.358$, and the latency is

$$S_0(t|z) = \frac{\exp(-\lambda(z)t) - \exp(-\lambda(z)\tau_0)}{1 - \exp(-\lambda(z)\tau_0)} I(t \leqslant \tau_0)$$

where $\tau_0 = 4.605$ and $\lambda(z) = \exp((z + 20)/40)$. The percentage of censored data is 54% and of cured data is 47%.

Model 2. The probability of uncure is:

$$p(z) = \frac{\exp(\beta_0 + \beta_1 z + \beta_2 z^2 + \beta_3 z^3)}{1 + \exp(\beta_0 + \beta_1 z + \beta_2 z^2 + \beta_3 z^3)}, \tag{7}$$

with $\beta_0 = 0.0476$, $\beta_1 = -0.2558$, $\beta_2 = -0.0027$ and $\beta_3 = 0.0020$ and the survival function of the susceptible population is

$$S_0(t|z) = \frac{1}{2}\left(\exp(-\alpha(z)t^5) + \exp(-100t^5)\right), \text{ with } \alpha(z) = \frac{1}{5}\exp((z + 20)/40).$$

The percentages of censored and cured data are 62% and 53%, respectively.

4.1.1 *Z continuous.* We consider a continuous covariate $Z \sim U(-20, 20)$. The results are given in Table 1. It is noteworthy that, under H_0, the size of the test is very similar to the significance level, α=0.05, for the different constant values for p. Furthermore, under H_1, the power of the test is very close (or even equal) to 1.

[Table 1 about here.]

4.1.2 *Z qualitative.* A qualitative covariate Z with three possible values $\{b_1, b_2, b_3\}$ was considered with $p(b_1) = p(b_2) = p(b_3) = 0.2$ and 0.5 under H_0, and $p(b_1) = 0.5$, $p(b_2) = 0.2$ and $p(b_3) = 0.7$ under H_1. Two situations, according to the probability mass function of Z given by $(1/3, 1/3, 1/3)$ and $(3/5, 1/5, 1/5)$, were studied. The observations (Z_i, T_i, δ_i), $i = 1, \ldots, n$ were simulated from Model 1, with functions $p(z)$, $S_0(t|z)$ and $G(t)$ defined there.

Remark. The computation of the probability of cure with a qualitative covariate deserves special attention, since that probability can not be obtained directly evaluating $p()$ *in (6) in the values of* Z*, provided they are not numerical. Therefore, let* (b'_1, b'_2, b'_3) *be the numerical values associated to the values* (b_1, b_2, b_3) *of* Z*, in the sense that the distribution of* Y *conditioned on* b_j *is* $S(t|b'_j)$ *in Model 1. Therefore, under the alternative hypothesis, the probability of cure derives from evaluating the function* $p()$ *in (6) not in* b_j*, but in the corresponding numerical values* $b'_j,\ j = 1, 2, 3$*.*

The results are shown in Table 2. Under the null hypothesis, the size of the test is very similar to the significance level, $\alpha = 0.05$. Regarding the alternative hypothesis, the power of the test is higher for large sample sizes and when the probability mass function of Z is equal in probability, as expected.

[Table 2 about here.]

4.2 *Case 2*

In this case, $\mathbf{W} = (X, \mathbf{Z})$ has dimension $m + 1$, with a one-dimensional covariate X and a m-dimensional covariate $\mathbf{Z}$. For the sake of simplicity, in this simulation study we suppose that Z is also one-dimensional. If there are only continuous covariates involved, we consider two different scenarios, Model 1 and Model 2. In any other case, just Model 1 is considered. The results for Case 2 when X and Z are continuous, X is continuous and Z is qualitative, X is qualitative and Z is continuous, and X and Z are qualitative are shown in Sections 4.2.1, 4.2.2, 4.2.3 and 4.2.4, respectively. The results for other type of covariates X and/or Z can be found in the Supplementary Material.

Model 1. Under the null hypothesis, $H_0 : E(\nu|X, Z) = 1 - p(X)$, the probability of uncure $p(x)$ is that in (6), corresponding to Model 1 in Section 4.1. Under the alternative,

the incidence is

$$1 - p(x, z) = 1 - \frac{\exp(\beta_0 + \beta_1 x(1 + \beta_2 z))}{1 + \exp(\beta_0 + \beta_1 x(1 + \beta_2 z))}, \tag{8}$$

with $\beta_0 = 0.476$, $\beta_1 = 0.358$ and $\beta_2 = 0.225$, and the latency is:

$$S_0(t|x, z) = \frac{\exp(-\lambda(x, z)t) - \exp(-\lambda(x, z)\tau_0)}{1 - \exp(-\lambda(x, z)\tau_0)} I(t \leqslant \tau_0),$$

where $\tau_0 = 4.605$ and $\lambda\,(x, z) = \exp\,((x + z + 20)/40)$.

Model 2. The probability of uncure, $p(x)$, under $H_0 : E(\nu|X, Z) = 1 - p(X)$, is that in (7), corresponding to Model 2 in Section 4.1. Under the alternative, that probability is:

$$p(x, z) = \frac{\exp\,(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3(1 + \beta_4 z))}{1 + \exp\,(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3(1 + \beta_4 z))},$$

with $\beta_0 = 0.0476$, $\beta_1 = -0.2558$, $\beta_2 = -0.0027$, $\beta_3 = 0.0020$ and $\beta_4 = 0.5$, and the survival function of the susceptible population is

$$S_0(t|x, z) = \frac{1}{2}\left(\exp(-\alpha(x, z)t^5) + \exp(-100t^5)\right), \text{ with } \alpha(x, z) = \frac{1}{5}\exp((x + z + 20)/40).$$

Furthermore, under H_0, if the covariates are not continuous, we define the distribution function of the variable Y:

$$F(y|x = 0) = F_1^0(y|x = a_1') \text{ and } F(y|x = 1) = F_1^0(y|x = a_2')$$

and the distribution function of the censoring variable,

$$G(y|x = 0) = G_1^0(y|x = a_1') \text{ and } G(y|x = 1) = G_1^0(y|x = a_2'),$$

where F_1^0 and G_1^0 are the conditional distribution functions for Model 1 under H_0. Analogously, under the alternative hypothesis:

$$F(y|x = 0, z = b_i) = F_1^1(y|x = a_1', z = b_i), \text{ with } i = 1, 2, 3$$

and

$$G(y|x = 1, z = b_i) = G_1^1(y|x = a_2', z = b_i), \text{ with } i = 1, 2, 3,$$

where F_1^1 and G_1^1 are the conditional distribution functions for Model 1 under H_1. Note that the distribution function of the variable Y, $F(y|x)$, and the distribution function of the

censoring variable, $G(y|x)$, are the corresponding distribution functions of Model 1 considered in the simulation studies for Case 2.

4.2.1 *X continuous, Z continuous.* We consider two continuous covariates X, Z with distribution $U(-20, 20)$. We simulated the data $(X_i, Z_i, T_i, \delta_i)$, $i = 1, ..., n$ from Models 1 and 2. Since we do not have a bandwidth selection method for h in (5), we follow the approach by (Delgado and González-Manteiga, 2001). They choose a bandwidth of the form $h = Cn^{-1/3m}$, for different values of C, where m is the dimension of the covariate vector Z that is being tested. Note that in our case, $m = 1$. (Delgado and González-Manteiga, 2001) explains that this bandwidth is compatible with assumptions $(A4)$ and $(A4')$ in their paper. Therefore, we decided to work with $h = Cn^{-1/3}$, where n is the sample size, and $C = 10, 20, 40, 60$. Our numerical experience shows that $g = 2h$ is a good choice for the pilot bandwidth. Under the null hypothesis, the results are very similar to the significance level, $\alpha = 0.05$, except for very large bandwidths. Furthermore, under the alternative hypothesis, the power of the test is considerably high. See the results in Table 3.

[Table 3 about here.]

4.2.2 *X continuous, Z qualitative.* As in the previous cases, X is $U(-20, 20)$, and Z is a categorical variable with values $\{b_1, b_2, b_3\}$, with probability mass functions $(1/3, 1/3, 1/3)$ and $(3/5, 1/5, 1/5)$. The probabilities of uncure, $p(x, b_1) = p(x, b_2) = p(x, b_3)$, under H_0, are the function $p(x)$ in (6). Under H_1, the incidence functions, $p(x, b_i)$ with $i = 1, 2, 3$, are given by the function $p(x, z)$ in (8), evaluated at (x, b_1'), (x, b_2') and (x, b_3'), where $(b_1', b_2', b_3') = (-5.2019, -1.3296, 1.0371)$.

The results under the null and the alternative hypothesis are shown in Table 4. Under the null hypothesis, the best choice of the bandwidth is $h = 5.43$ for $n = 50$, $h = 2.15$ for

$n = 100$ and $h = 1.71$ for $n = 200$. Moreover, under H_1 we obtain higher power when using $h = 16.28$ for $n = 50$, $h = 8.62$ or 12.93 for $n = 100$ and $h = 6.84$ or 10.26 for $n = 200$, that is, $h = Cn^{-1/3}$ with $C = 60$.

[Table 4 about here.]

4.2.3 *X qualitative, Z continuous.* The variable X is qualitative with values $\{a_1, a_2, a_3\}$, with probability mass functions for each scenario $(1/3, 1/3, 1/3)$ and $(3/5, 1/5, 1/5)$, and Z is $U(-20, 20)$. The observations were simulated from Model 1. Let a_i' be the numerical value x at which $p(x, z)$ in (8) is evaluated to get $p(a_i, z)$. We considered two scenarios depending on the values (a_1', a_2', a_3'): $(-3.6964, -1.3296, 1.0371)$ in the first scenario, and $(-7.4671, -1.3296, 4.8079)$ in the second one. The incidence reduces, under H_0, to $1 - p(a_i')$, $i = 1, 2, 3$, with $p(x)$ in (6).

Table 5 shows the results. Under H_0, the size of the test is close to the significance level, $\alpha = 0.05$, for both scenarios. Under H_1, the power is higher if we consider that the probability mass function of X is $(3/5, 1/5, 1/5)$.

[Table 5 about here.]

4.2.4 *X qualitative, Z qualitative.* Let both X and Z be qualitative variables with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$, respectively. We work with two different situations depending on the corresponding probability mass functions: in the first one, both for X and Z are $(1/3, 1/3, 1/3)$, whereas in the second one, both are $(3/5, 1/5, 1/5)$. The cure probabilities, $1 - p(a_i, b_j)$, $i, j = 1, 2, 3$, are computed from the function $p(x, z)$ in (8) evaluated at the numerical values $(a_i', \bar{b}_i)$, $i = 1, 2, 3$, given in Table 6 (top) under H_0, and (a_i', b_j'), $i, j = 1, 2, 3$, given in Table 6 (bottom) (under H_1).

[Table 6 about here.]

Table 7 shows the results under the null and the alternative hypothesis. In Scenario 1, the size of the test is close to the significance level ($\alpha = 0.05$), except for the CvM test, with $n = 50, 100$ and when the probability mass function of X is $(1/3, 1/3, 1/3)$. In the other 3 scenarios, the results are very competent regardless the probability mass function of X. Under the alternative hypothesis, in the 4 scenarios, the power is higher if the probability mass function of X is $(3/5, 1/5, 1/5)$.

[Table 7 about here.]

5. Application to a colorectal cancer example

We work with a dataset related to colorectal cancer patients from CHUAC, Spain. The variable of interest, Y, is the time, in months, since the diagnostic until death. The follow-up time is almost 19 years. An individual is considered cured if he or she will not die because of colorectal cancer. Censoring is caused by "cure", death due to any other cause different to colorectal cancer, dropout, or end of the study. The dataset contains 414 observations on 8 variables: the censoring indicator; the observed survival time; the location: colon (111 individuals) or rectum (303 individuals); the age: from 23 to 102 years; and the stage TNM, which is the main determinant in prognosis of these patients. The stage has 3 components: T, which describes the size of the tumor and whether it has invaded nearby tissue; N, which measures the lymph nodes that are involved; and M, which evaluates the presence (or not) of metastasis. The information of these 3 aspects can be combined and it lets us classificate each patient in a unique (and numeric) stage from 1 to 4. About 50% of the observations are censored, with the percentage of censoring varying from 30% to almost 71%, depending on the stage. The number of patients in Stage 1 is 62 (70.97% censored, aged $23 - 84$), in Stage 2 is 167 (55.09% censored, aged $36 - 102$), in Stage 3 is 133 (39.85% censored, aged

$30 - 88$) and 52 in Stage 4 (30.77% censored, aged $43 - 88$).

Age and tumour stage at diagnosis are known to strongly influence colorectal cancer treatment regimen and five-year survival ((Vercelli et al., 2006), (Guyot et al., 2005), among others). However, the effect of the age and stage on the probability of cure is rarely analyzed, since the few studies of cure for colorectal cancer patients have focused on the estimation of cure by age and stage at diagnosis (see, for example, (Shack et al., 2012)), but not on the statistical signficance of those covariates on the probability of cure. So the proposed nonparametric significance tests were applied to test the effect of the age and stage on the cure probability. We consider $B = 1000$ bootstrap resamples to approximate the distribution of the test statistic and a significance level $\alpha = 0.05$.

5.1 *Case 1*

We started studying the effect of the age on the probability of cure. The test did not find a significant effect of the age on the cure probability ($p_{CvM} = 0.142$, $p_{KS} = 0.146$). Hence, in order to avoid an interaction with the stage, the analysis was repeated for each stage separately. In Figure 1 we can appreciate how the nonparametric estimator of the incidence changes with the age for each stage. The semiparametric estimator of the incidence by (Peng and Dear, 2000) is also represented for comparison purposes. The cure probabilities in Stages 1 and 2 are higher than in Stages 3 and 4. The reason is that, in initial stages, most of the surgeries have healing purposes, whereas in advanced stages, surgeries are usually palliative treatments, and therefore the incidence for these patients is lower. Considering only this figure, it seems reasonable to suppose that in Stages 1 and 4 the cure rate could be a constant value, as a function of the age, whereas in Stages 2 and 3 the age may have some influence, since the cure probability decreases as the age increases.

[Figure 1 about here.]

The test is only significant in Stage 3 ($p_{CvM} = 0.002$, $p_{KS} = 0.000$); for patients above 60, in a 10 years gap the cure probability decreases considerably from 40% to almost 0%. In Stages 1 ($p_{CvM} = 0.396$, $p_{KS} = 0.257$) and 4 ($p_{CvM} = 0.587$, $p_{KS} = 0.551$) there is not enough evidence of an effect of the age on the cure probability. In Stage 1, the reason is that the estimated cure probability fluctuates around 25% for most patients regardless the age (see Figure 1). The results in Stage 4 deserve some comments. A total of 11 in the 12 greatest lifetimes, including the largest lifetime, are uncensored and, consequently, uncured. This causes that the nonparametric estimation of the probability of being cured, as a function of the age, is constant equal to 0 (see Figure 1). Although it should not be stated that it is impossible for a patient with Stage 4 colorectal cancer to survive, this estimation reinforces the assertion that long-term survival in patients with Stage 4 colorectal cancer is uncommon ((Miyamoto et al., 2015)). This fact, far from being a weakness of the nonparametric method, is an important advantage, since it allows to detect situations in which introducing the possibility of cure does not contribute to improve the model. Finally, in Stage 2, the probability of cure decreases with the age, as expected, from about 30% in patients with age at diagnosis 50-60, to 7% for patients above 80 years old (see Figure 1). This effect of the age in the cure probability is only borderline significant ($p_{CvM} = 0.082$, $p_{KS} = 0.067$).

Furthermore, note that the estimated cure probability for each stage is, regardless the age, 0.28 in Stage 1, 0.13 in Stage 2, 0.13 in Stage 3, and 0 in Stage 4. As expected, the probability of cure decreases as the stage of the cancer progresses. However, the differences of the cure probabilities among the four stages were not that large, and the test did not find them to be significant ($p_{CvM} = 0.581$ and $p_{KS} = 0.483$).

5.2 *Case 2*

For illustrative purposes only, we further apply the proposed test to assess the significance of one covariate, assuming that the other one has a clear effect on the cure probability. Although it was not the case, let us assume that the cure probability depends significantly on the stage. In such situation, it might be of interest to test if the cure probability is affected also by the age of the patient. The conclusion from the nonparametric test was that age was not significantlly associate with the cure probability ($p_{CvM} = 0.370$, $p_{KS} = 0.267$). Anagolously, suppose hypothetically that the probability of cure depends on the age of a patient. To study if it also depends on the cancer stage, the test statistic in (5) has to be computed, which requires a value for the bandwidth, h. Without a suitable bandwidth selector, we considered a bandwidth $h = Cn^{-1/3}$, with the following wide range of values $C = 10,\ 20,\ 40,\ 60,\ 120,\ 240,\ 300$ and 375. If the age is assumed to affect the cure probability, then the effect of the stage was not statistically significant with any of the values of h considered, from the smallest one $h = 1.342$ ($p_{CvM} = 0.721$, $p_{KS} = 0.815$) to the largest one $h = 50.315$ ($p_{CvM} = 0.137$, $p_{KS} = 0.153$).

6. Discussion

A nonparametric covariate significance test for the incidence in mixture cure models is introduced. The methodology can be applied to high dimensional datasets, including analysis of images, related to cancer for medical diagnosis. Although we do not have a bandwidth selection method for Case 2 when the covariate X is continuous, we do not have a bandwidth selection method. However, several bandwidth selectors for smoothed tests are proposed in the literature than can also be applied in this context. On the other hand, preliminary studies showed that the choice of the pilot bandwidth has a small effect on the results.

Acknowledgements

The first author's research was sponsored by the Spanish FPU (Formación de Profesorado Universitario) Grant from MECD (Ministerio de Educación, Cultura y Deporte) with reference FPU13/01371. All the authors acknowledge partial support by the MINECO (Ministerio de Economía y Competitividad) grant MTM2014-52876-R (EU ERDF support included) and the MICINN (Ministerio de Ciencia e Innovación) Grant MTM2017-82724-R (EU ERDF support included). The first, second and fourth authors acknowledge partial support of Xunta de Galicia (Centro Singular de Investigación de Galicia accreditation ED431G/01 2016-2019 and Grupos de Referencia Competitiva CN2012/130 and ED431C2016-015) and the European Union (European Regional Development Fund - ERDF). Financial support from the European Research Council (2016-2021, Horizon 2020 / ERC grant agreement No. 694409) is gratefully acknowledged. The authors are grateful to Dr. S. Pértega and Dr. S. Pita, at the University Hospital of A Coruña, for providing the colorectal cancer dataset.

References

Amico, M. and Van Keilegom, I. (2018). Cure models in survival analysis. Review paper, KU Leuven, Leuven, Belgium.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, University of California, Berkeley, Berkeley.

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.

Delgado, M. A. and González-Manteiga, W. (2001). Significance testing in nonparametric regression based on the bootstrap. *Ann. Stat.*, 29:1469–1507.

Farewell, V. T. (1986). Mixture models in survival analysis: are they worth the risk? *Can. J. Stat.*, 14:257–262.

Guyot, F., Faivre, J., Manfredi, S., Meny, B., Bonithon-Kopp, C., and Bouvier, A. M. (2005).

Time trends in the treatment and survival from recurrence of colorectal cancer. *Annals of Oncology*, 16:756–761.

Kulasekera, K. B. and Wang, J. (1997). Smoothing parameter selection for power optimality in testing of regression curves. *J. Am. Stat. Assoc.*, 92:500–511.

López-Cheda, A., Cao, R., Jácome, M. A., and Van Keilegom, I. (2017a). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Comput. Stat. Data Anal.*, 105:144–165.

López-Cheda, A., Jácome, M. A., and Cao, R. (2017b). Nonparametric latency estimation for mixture cure models. *Test*, 26:353–376.

Maller, R. A. and Zhou, S. (1996). *Survival Analysis with Long-Term Survivors*. Wiley, Chichester, U. K.

Martínez-Camblor, P. (2010). Nonparametric k-sample test based on kernel density estimator for paired design. *Comput. Stat. Data Anal.*, 54:2035–2045.

Martínez-Camblor, P. and de Uña-Álvarez, J. (2013). Studying the bandwidth in k-sample smooth tests. *Comput. Stat.*, 28:875–892.

Miyamoto, Y., Hayashi, N., Sakamoto, Y., Ohuchi, M., Tokunagam, R., Kurashige, Y., Hiyoshi, Y., Baba, S., Iwagami, N., Yoshida, M., Yoshida, J., and Baba, H. (2015). Predictors of long-term survival in patients with stage IV colorectal cancer with multi-organ metastases: a single-center retrospective analysis. *Int. J. Clin. Oncol.*, 20:1140–1146.

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

Peng, Y. and Dear, K. B. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, 56:237–243.

Peng, Y. and Taylor, J. M. G. (2011). Mixture cure model with random effects for the

analysis of a multi-centre tonsil cancer study. *Statistics in Medicine*, 30:211–223.

Peng, Y. and Taylor, J. M. G. (2014). Cure models. In Klein, J., van Houwelingen, H., Ibrahim, J. G., and Scheike, T. H., editors, *Handbook of Survival Analysis*, pages 113–134. Chapman & Hall, Boca Raton, FL, USA.

Peng, Y., Taylor, J. M. G., and Yu, B. (2007). A marginal regression model for multivariate failure time data with a surviving fraction. *Lifetime Data Analysis*, 13:351–369.

Shack, L. G., Shah, A., Lambert, P. C., and Rachet, B. (2012). Cure by age and stage at diagnosis for colorectal cancer patients in north west england, 19972004: A population-based study. *Cancer Epidemiology*, 36:548–553.

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

Vercelli, M., Lillini, R., Capocaccia, R., Micheli, A., Coebergh, J. W., Quinn, M., Martínez-García, C., and Quaglia, A. (2006). Cancer survival in the elderly: effects of socio-economic factors and health care systems features (eldcare project). *European Journal of Cancer*, 42:234–242.

Wang, L., Du, P., and Lian, H. (2012). Two-component mixture cure rate model with spline estimated nonparametric components. *Biometrics*, 68:726–735.

Xu, J. and Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Can. J. Stat.*, 42:1–17.

Zhang, J. and Peng, Y. (2009). Accelerated hazards mixture cure model. *Lifetime Data Analysis*, 15:455–467.

Figure 1. Semiparametric (black line) and nonparametric estimations of the incidence in Stages 1-4 depending on the age, computed with the bootstrap bandwidth (solid blue line) and with a smoothed bootstrap bandwidth (dashed blue line). The green line represents the Parzen-Rosenblatt kernel density estimations of the covariate age, using Sheather and Jones' plug-in bandwidth.

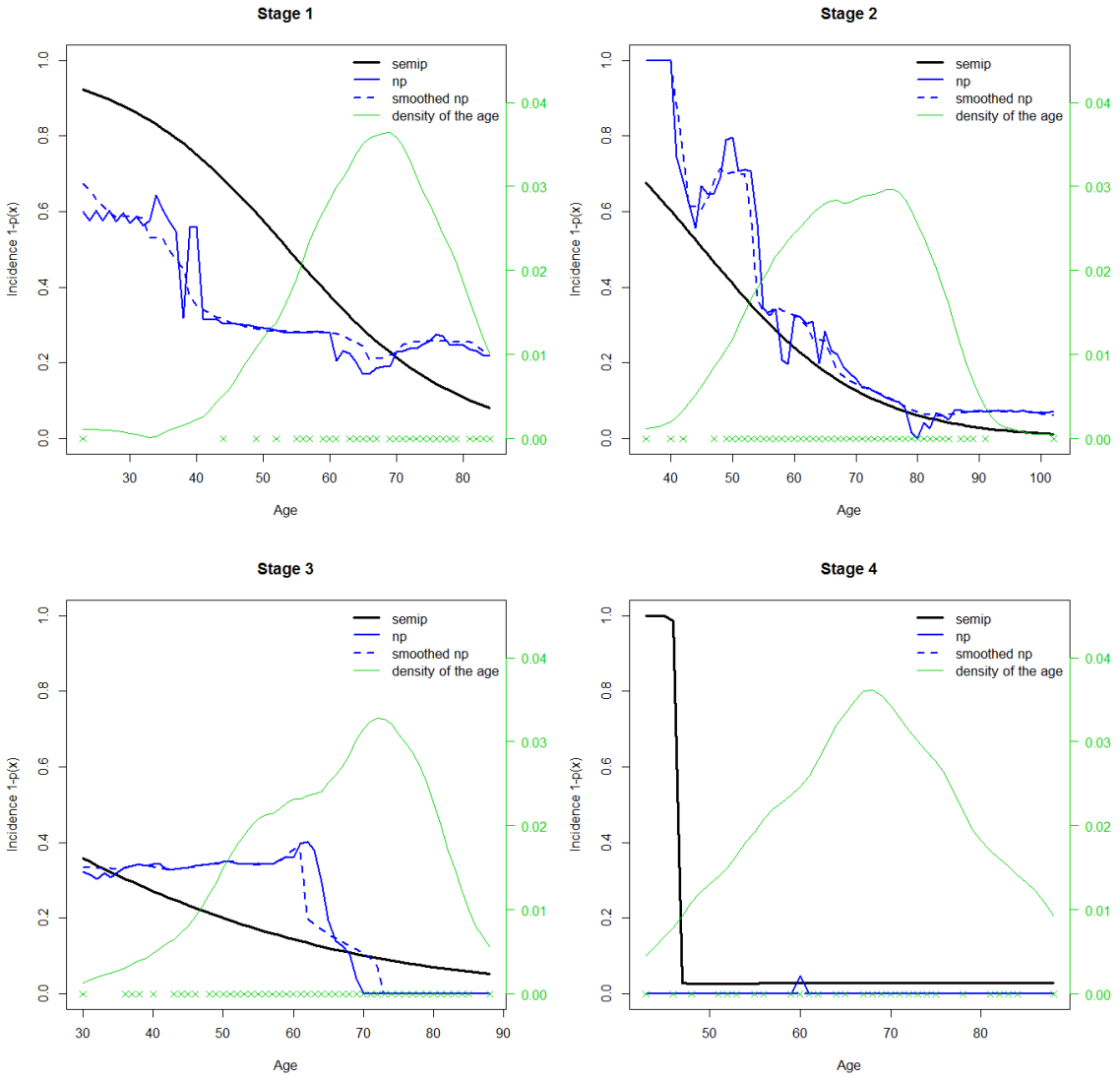

Table 1
Size (top) and power (bottom) of the test for Case 1 with Z continuous with distribution $U(-20, 20)$ under the null and the alternative hypothesis, respectively.

		Model 1		Model 2	
n	p	CvM	KS	CvM	KS
Under H_0					
50	0.3	0.0484	0.0600	0.0550	0.0636
	0.5	0.0500	0.0564	0.0490	0.0582
	0.7	0.0446	0.0472	0.0510	0.0504
	0.8	0.0418	0.0360	0.0400	0.0412
100	0.3	0.0608	0.0636	0.0544	0.0584
	0.5	0.0494	0.0562	0.0488	0.0586
	0.7	0.0444	0.0520	0.0474	0.0468
	0.8	0.0414	0.0396	0.0484	0.0470
200	0.3	0.0490	0.0530	0.0528	0.0534
	0.5	0.0552	0.0616	0.0540	0.0572
	0.7	0.0498	0.0508	0.0516	0.0514
	0.8	0.0446	0.0432	0.0466	0.0480
Under H_1					
50		0.9890	0.9862	0.4200	0.4148
100		0.9994	0.9992	0.7330	0.7402
200		1	1	0.9670	0.9746

Table 2
Size (top) and power (bottom) of the test for Case 1 with Z qualitative under the null and the alternative hypothesis, respectively.

n	$(\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3))$	p			CvM	KS
Under H_0						
50	$(1/3, 1/3, 1/3)$	0.2			0.0512	0.0526
	$(1/3, 1/3, 1/3)$	0.5			0.0494	0.0520
100	$(1/3, 1/3, 1/3)$	0.2			0.0544	0.0538
	$(1/3, 1/3, 1/3)$	0.5			0.0488	0.0532
200	$(1/3, 1/3, 1/3)$	0.2			0.0486	0.0500
	$(1/3, 1/3, 1/3)$	0.5			0.0456	0.0516
Under H_1						
n	$(\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3))$	$p(b_1)$	$p(b_2)$	$p(b_3)$	CvM	KS
50	$(1/3, 1/3, 1/3)$	0.5	0.2	0.7	0.3402	0.3408
	$(3/5, 1/5, 1/5)$	0.5	0.2	0.7	0.1680	0.1606
100	$(1/3, 1/3, 1/3)$	0.5	0.2	0.7	0.5588	0.5600
	$(3/5, 1/5, 1/5)$	0.5	0.2	0.7	0.2748	0.2690
200	$(1/3, 1/3, 1/3)$	0.5	0.2	0.7	0.8028	0.7994
	$(3/5, 1/5, 1/5)$	0.5	0.2	0.7	0.4552	0.4448

Table 3
Size (top) and power (bottom) of the test for Case 2 with X and Z continuous with distribution $U(-20, 20)$, under the null and the alternative hypothesis, respectively.

		Model 1		Model 2	
n	h	CvM	KS	CvM	KS
Under H_0					
50	2.71	0.0504	0.0524	0.0494	0.0500
	5.43	0.0428	0.0436	0.0464	0.0444
	10.86	0.0372	0.0486	0.0448	0.0438
	16.28	0.0442	0.0782	0.0692	0.0734
100	2.15	0.0494	0.0514	0.0526	0.0520
	4.31	0.0476	0.0556	0.0544	0.0542
	8.62	0.0580	0.0708	0.0514	0.0490
	12.93	0.0590	0.0988	0.0700	0.0782
200	1.71	0.0436	0.0412	0.0510	0.0496
	3.42	0.0420	0.0448	0.0558	0.0542
	6.84	0.0522	0.0606	0.0508	0.0494
	10.26	0.0590	0.0948	0.0614	0.0708
Under H_1					
50	2.71	0.2016	0.2182	0.2632	0.2466
	5.43	0.2596	0.2698	0.3244	0.3074
	10.86	0.2610	0.2696	0.3210	0.3150
	16.28	0.2278	0.2442	0.3084	0.3152
100	2.15	0.3906	0.4736	0.6228	0.5662
	4.31	0.5132	0.5556	0.6918	0.6128
	8.62	0.5160	0.5564	0.6864	0.6320
	12.93	0.4718	0.5206	0.6746	0.6438
200	1.71	0.7428	0.8554	0.9730	0.9364
	3.42	0.8492	0.9156	0.9832	0.9486
	6.84	0.8582	0.9114	0.9830	0.9568
	10.26	0.8358	0.8914	0.9852	0.9576

Table 4
Size (top) and power (bottom) of the test for Case 2 with X continuous with distribution $U(-20, 20)$, and Z qualitative with values $\{b_1, b_2, b_3\}$.

n	h	$(\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3))$	CvM	KS
Under H_0				
50	2.71	$(1/3, 1/3, 1/3)$	0.0520	0.0600
	5.43	$(1/3, 1/3, 1/3)$	0.0448	0.0554
	10.86	$(1/3, 1/3, 1/3)$	0.0450	0.0832
	16.28	$(1/3, 1/3, 1/3)$	0.0558	0.1546
100	2.15	$(1/3, 1/3, 1/3)$	0.0436	0.0530
	4.31	$(1/3, 1/3, 1/3)$	0.0414	0.0560
	8.62	$(1/3, 1/3, 1/3)$	0.0584	0.0886
	12.93	$(1/3, 1/3, 1/3)$	0.0656	0.1538
200	1.71	$(1/3, 1/3, 1/3)$	0.0406	0.0480
	3.42	$(1/3, 1/3, 1/3)$	0.0374	0.0480
	6.84	$(1/3, 1/3, 1/3)$	0.0584	0.0800
	10.26	$(1/3, 1/3, 1/3)$	0.0780	0.1462
Under H_1				
50	2.71	$(1/3, 1/3, 1/3)$	0.0976	0.1270
	5.43	$(1/3, 1/3, 1/3)$	0.1382	0.1598
	10.86	$(1/3, 1/3, 1/3)$	0.1562	0.1990
	16.28	$(1/3, 1/3, 1/3)$	0.1686	0.2448
	2.71	$(3/5, 1/5, 1/5)$	0.1592	0.1536
	5.43	$(3/5, 1/5, 1/5)$	0.2086	0.2018
	10.86	$(3/5, 1/5, 1/5)$	0.2090	0.2176
	16.28	$(3/5, 1/5, 1/5)$	0.1920	0.2200
100	2.15	$(1/3, 1/3, 1/3)$	0.1796	0.2524
	4.31	$(1/3, 1/3, 1/3)$	0.2452	0.3160
	8.62	$(1/3, 1/3, 1/3)$	0.2796	0.3730
	12.93	$(1/3, 1/3, 1/3)$	0.2858	0.4276
	2.15	$(3/5, 1/5, 1/5)$	0.3202	0.3336
	4.31	$(3/5, 1/5, 1/5)$	0.3848	0.3876
	8.62	$(3/5, 1/5, 1/5)$	0.3868	0.4038
	12.93	$(3/5, 1/5, 1/5)$	0.3550	0.4060
200	1.71	$(1/3, 1/3, 1/3)$	0.3666	0.5278
	3.42	$(1/3, 1/3, 1/3)$	0.4698	0.6044
	6.84	$(1/3, 1/3, 1/3)$	0.5112	0.6566
	10.26	$(1/3, 1/3, 1/3)$	0.5212	0.6988
	1.71	$(3/5, 1/5, 1/5)$	0.6048	0.6342
	3.42	$(3/5, 1/5, 1/5)$	0.6532	0.6804
	6.84	$(3/5, 1/5, 1/5)$	0.6410	0.6850
	10.26	$(3/5, 1/5, 1/5)$	0.6068	0.6772

Table 5
Size (top) and power (bottom) of the test for Case 2 with X qualitative with values $\{a_1, a_2, a_3\}$, and Z continuous with distribution $U(-20, 20)$.

		Scenario 1		Scenario 2	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	CvM	KS	CvM	KS
Under H_0					
50	$(1/3, 1/3, 1/3)$	0.0502	0.0658	0.0498	0.0624
	$(3/5, 1/5, 1/5)$	0.0524	0.0696	0.0458	0.0616
100	$(1/3, 1/3, 1/3)$	0.0494	0.0632	0.0446	0.0574
	$(3/5, 1/5, 1/5)$	0.0468	0.0558	0.0392	0.0560
200	$(1/3, 1/3, 1/3)$	0.0496	0.0572	0.0486	0.0614
	$(3/5, 1/5, 1/5)$	0.0512	0.0550	0.0538	0.0612
Under H_1					
50	$(1/3, 1/3, 1/3)$	0.3888	0.4148	0.4136	0.4866
	$(3/5, 1/5, 1/5)$	0.7196	0.7418	0.7436	0.7906
100	$(1/3, 1/3, 1/3)$	0.6552	0.6834	0.7058	0.7958
	$(3/5, 1/5, 1/5)$	0.9290	0.9348	0.9380	0.9554
200	$(1/3, 1/3, 1/3)$	0.9126	0.9280	0.9460	0.9742
	$(3/5, 1/5, 1/5)$	0.9938	0.9940	0.9984	0.9992

Table 6

Uncure probabilities, $p(a_i, b_j)$, considered under H_1, for Case 2 when X and Z are qualitative with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$, respectively. See Remark in Section 4.2 for details.

	Under H_0		Under H_1		
	$(1/3, 1/3, 1/3)$	$(3/5, 1/5, 1/5)$	$b'_1 = 0.6157$	$b'_2 = -3.5434$	$b'_3 = -7.7026$
Scenario 1					
$a'_1 = -6.5585$	0.5000	0.3400	0.1	0.5000	0.9000
$a'_2 = -1.1678$	0.5943	0.5566	0.5	0.5966	0.6862
$a'_3 = 0.9109$	0.6304	0.6581	0.7	0.6323	0.5590
Scenario 2					
$a'_1 = -6.5585$	0.5000	0.3400	0.1	0.5000	0.9000
$a'_2 = -4.5690$	0.5261	0.3957	0.2	0.5360	0.8423
$a'_3 = 0.9109$	0.6304	0.6583	0.7	0.6323	0.5590
Scenario 3					
$a'_1 = -6.5585$	0.5000	0.3400	0.1	0.5000	0.9000
$a'_2 = -4.5690$	0.5261	0.3957	0.2	0.5360	0.8423
$a'_3 = 4.2229$	0.6444	0.7466	0.9	0.6862	0.3470
Scenario 4					
$a'_1 = -6.5585$	0.5000	0.3400	0.1	0.5000	0.9000
$a'_2 = -4.5690$	0.5261	0.3957	0.2	0.5360	0.8423
$a'_3 = -3.2466$	0.5501	0.4501	0.3	0.5598	0.7905

Table 7
Size (top) and power (bottom) of the test for Case 2 with X and Z qualitative with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$, respectively. The probability mass function of Z equals that of X.

		Scenario 1		Scenario 2		Scenario 3		Scenario 4	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	CvM	KS	CvM	KS	CvM	KS	CvM	KS
Under H_0									
50	$(1/3, 1/3, 1/3)$	0.0392	0.0490	0.0452	0.0548	0.0480	0.0558	0.0432	0.0526
	$(3/5, 1/5, 1/5)$	0.0496	0.0528	0.0490	0.0510	0.0524	0.0524	0.0530	0.0512
100	$(1/3, 1/3, 1/3)$	0.0392	0.0512	0.0500	0.0552	0.0472	0.0562	0.0452	0.0546
	$(3/5, 1/5, 1/5)$	0.0524	0.0502	0.0454	0.0496	0.0520	0.0546	0.0514	0.0512
200	$(1/3, 1/3, 1/3)$	0.0478	0.0552	0.0496	0.0588	0.0470	0.0496	0.0486	0.0526
	$(3/5, 1/5, 1/5)$	0.0496	0.0520	0.0540	0.0544	0.0532	0.0514	0.0506	0.0490
Under H_1									
50	$(1/3, 1/3, 1/3)$	0.1658	0.1802	0.2968	0.3224	0.2410	0.2780	0.4586	0.5164
	$(3/5, 1/5, 1/5)$	0.4874	0.4818	0.5054	0.5104	0.5056	0.5126	0.5482	0.5740
100	$(1/3, 1/3, 1/3)$	0.2932	0.2888	0.5358	0.5478	0.4556	0.5206	0.7584	0.7896
	$(3/5, 1/5, 1/5)$	0.7324	0.7220	0.7536	0.7612	0.7500	0.7616	0.7774	0.8060
200	$(1/3, 1/3, 1/3)$	0.5128	0.4854	0.8232	0.8324	0.7350	0.8266	0.9578	0.9640
	$(3/5, 1/5, 1/5)$	0.9218	0.9142	0.9340	0.9382	0.9354	0.9408	0.9412	0.9516

FACULTY OF ECONOMICS AND BUSINESS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be