RESEARCH ARTICLE

Nonparametric covariate significance tests for the cure rate in mixture cure models

Ana López-Cheda*¹ | M. Amalia Jácome^{2,3} | Ingrid Van Keilegom⁴ | Ricardo Cao^{2,3,5}

¹EPIUnit-Instituto de Saúde Pública da Universidade do Porto, University of Porto, Rua das Taipas 135, 4050-600 Porto, Portugal

²Department of Mathematics, University of A Coruña, Campus de Elviña s/n, 15071 A Coruña, Spain

³CITIC, INIBIC, Spain

⁴ORSTAT, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium

⁵ITMATI, Spain

Correspondence

*Ana López-Cheda Email: ana.lopez.cheda@ispup.up.pt

Present Address

Rua das Taipas 135, 4050-600 Porto, Portugal

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 is not appropriate to handle these 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. A bootstrap method is proposed to approximate the null distribution of the test statistic. The procedure can be applied to any type of covariate, and could be extended to the multivariate setting. Its efficiency is evaluated in a Monte Carlo simulation study. Finally, the method is applied to a colorectal cancer dataset.

KEYWORDS:

bootstrap, censored data, cure models, significance tests, survival analysis

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 an increasingly large number of situations where there are individuals 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 subpopulation¹. In the literature, the most popular cure model is the mixture cure model (recent reviews of cure models can be found in^{2,3}, among others). Mixture cure models, proposed by⁴, split the population in two groups: the cured, who will never experience the event of interest, and the susceptible group. These models allow to estimate the probability of being cured and the survival function of the uncured population, or latency, depending on a set of covariates. The main advantage of this model is that it allows covariates to have different influence on cured and uncured patients. A detailed review of this model is provided by⁵. The estimation of cure models has been extensively studied using parametric and semiparametric methods^{1 6 7 8 9 10 11}.

A nonparametric estimator for the probability of cure in the mixture cure model was first introduced by¹², with some asymptotic properties further studied by¹³. The nonparametric estimator of the latency was introduced and analyzed by^{13 14}. This enables the mixture cure model with covariates to be addressed in a completely nonparametric way.

⁰**Abbreviations:** CHUAC, Complejo Hospitalario Universitario de A Coruña; CM, Cramér-von Mises; CRC, colorectal cancer; K, Kolmogorov-Smirnov; KM, Kaplan-Meier; NW, Nadaraya-Watson; PR, Parzen-Rosenblatt; pmf, probability mass function

It is always of interest to test if a covariate has some influence on the cure rate or on the survival time of the susceptible patients.¹⁵ propose a test statistic to assess whether the cure rate, as a function of the covariates, satisfies a certain parametric model. However, to the best of our knowledge, no significance testing for mixture cure models has been proposed yet in a completely nonparametric way. To fill this important gap, a covariate significance test for the probability of cure is presented in this paper. The method, evaluated in a Monte Carlo simulation study, is based on the significance test by¹⁶. Furthermore, the methodology is applied to a real dataset related to colorectal cancer patients.

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^{12 13 14}. In Section 3 we focus on the significance tests for the probability of cure. According to the number of covariates, we address three situations: in Section 3.1 we consider a one-dimensional covariate Z , and we study if the probability of cure depends on Z (Case 1); in Section 3.2, we assume that the probability of cure depends on the one-dimensional covariate X , and study if it also depends on the m -dimensional covariate Z (Case 2). Section 3.3 addresses the general case of (X, Z) , where X is $\mathbb{R}^q$ -valued and Z is $\mathbb{R}^m$ -valued. 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 (Complejo 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 , and the time to occurrence of the event, Y , are conditionally independent given a set of covariates, say X . The conditional distribution function of Y is $F(t|\mathbf{x}) = P(Y \leq t|X = \mathbf{x})$, and the corresponding survival function is $S(t|\mathbf{x}) = 1 - F(t|\mathbf{x})$. The observed time is $T = \min(Y, C)$ and $\delta = I(Y \leq C)$ is the uncensoring indicator. Moreover, the distribution functions of C and T are G and H , respectively. Let us denote by v the cure indicator, with $v = 0$ if the individual is susceptible and $v = 1$ otherwise. Note that if $v = 1$, it is assumed that $Y = \infty$. Therefore, the conditional probability of not being cured (incidence) is $p(\mathbf{x}) = P(v = 0|X = \mathbf{x})$. Then, the mixture cure model becomes:

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

where $1 - p(\mathbf{x})$ is the probability of cure and $S_0(t|\mathbf{x})$ is the conditional survival function of the uncured group, also called latency. Without loss of generality, let X be a univariate continuous covariate with density function f_X . The observations will be $\{(X_i, T_i, \delta_i), i = 1, \dots, n\}$, i.i.d. copies of the random vector (X, T, δ) .

¹² introduced the following kernel type cure rate estimator:

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

where $\hat{S}_h(t|x)$ is the conditional product-limit estimator by¹⁷, $B_{h[i]}$ are the Nadaraya-Watson (NW) weights depending on the bandwidth h , and $T_{\max}^1 = \max_{i: \delta_i=1} (T_i)$ is the largest uncensored failure time. Here $T_{(1)} \leq T_{(2)} \leq \dots \leq T_{(n)}$ are the ordered T_i 's, and $\delta_{[i]}$ and $X_{[i]}$ are the corresponding uncensoring indicator and covariate concomitants.¹² also proved the consistency and asymptotic normality of the estimator in (1). Furthermore,¹³ showed that it is the local maximum likelihood estimator of the cure rate, obtained an i.i.d. representation, found an asymptotic expression of the mean squared error, and proposed a bootstrap selector for the parameter h .

Mixture cure models might have identifiability issues due to problems associated with the tail of the distribution S . If the follow-up is not long enough, events could plausibly occur after the longest observed time, making very difficult to distinguish between cured subjects and long-term uncures. As pointed out by^{18 19}, some conditions on the latency component S_0 should be imposed to make the mixture cure model identifiable when the cure rate $1 - p(x)$ is completely unspecified. Some of these conditions cover the conditional latency function to be fully parametrically specified, or the dependence of S_0 on the covariate through a proportional hazards form¹⁰ or an accelerated failure time structure²⁰. When the latency function is completely unspecified, identifiability is guaranteed if S_0 is a proper survival function, that is, if $S_0(\tau^-|x) = 0$ for all x , where $\tau > 0$ is the length of the observation period, including the possibility of $\tau = \infty$ ¹⁹. According to¹⁹, the assumption of a proper latency function S_0 is similar to the zero-tail constraint in^{21 6} (see also¹³):

$$\tau_0 \leq \tau_G(x) \text{ for all } x, \quad (2)$$

where $\tau_0 = \sup_x \tau_0(x)$, and $\tau_0(x)$ and $\tau_G(x)$ are the right extremes of the supports of $S_0(t|x)$ and $G(t|x)$, respectively. Assumption (2) states that there is zero probability of a susceptible individual surviving longer than the largest possible censoring time. Identifiability of the model when applying the proposed test is entailed with assumption (2), that is, essentially assuming that all the observations after the largest failure time are cured.

To be confident that condition (2) is true, the length of follow-up should be chosen with considerable care. A nonparametric test for assumption (2) was proposed by²². The idea is based on the difference between the largest observed time, $T_{(n)}$, and the largest uncensored time, $T_{\max}^1$, that is, the interval at the right tail of the distribution where the Kaplan-Meier (KM) estimator of the survival function, S , has got a long stable plateau. A large interval with heavy censoring is considered an evidence that the follow-up period has been sufficiently long for the assumption (2) to be true.

3 | SIGNIFICANCE TESTS FOR THE CURE RATE

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 great 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 nonparametric covariate significance test for the cure rate based on the method by¹⁶, 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 present under the null hypothesis. This feature is computationally convenient and partially solves the problem of the “curse of dimensionality” when selecting regressors in a nonparametric context.

Let us denote by $\mathbf{W} = (\mathbf{X}, \mathbf{Z}) = (X_1, \dots, X_q, Z_1, \dots, Z_m)$ 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 $\mathbf{Z}$:

$$H_0 : E(v|\mathbf{X}, \mathbf{Z}) \equiv 1 - p(\mathbf{X}) \text{ vs. } H_1 : E(v|\mathbf{X}, \mathbf{Z}) \equiv 1 - p(\mathbf{X}, \mathbf{Z}), \quad (3)$$

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

Two cases are considered in this paper, depending on the dimension of the covariates: Case 1, where $\mathbf{W} = \mathbf{Z}$ is univariate, and Case 2, where $\mathbf{W} = (\mathbf{X}, \mathbf{Z})$, with a one-dimensional covariate $\mathbf{X}$ 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 generalized from Case 2 (see Section 3.3).

The main challenge of testing (3) is that the response variable (the cure indicator, v), is only partially observed due to the censoring. The complete observations are known to be uncured ($v = 0$), but it is unknown if a censored individual will be eventually cured or not (v is missing). The novelty of the proposed test is that this inconvenience is overcome expressing the regression function of the unobservable (and unestimable) response, v , as a regression function with response η , which is not observable but estimable. Let us define the variable η , which is a conditional proxy for v , given $\mathbf{W}$, as follows:

$$\eta = \frac{v(1 - I(\delta = 0, T \leq \tau))}{1 - G(\tau|\mathbf{W})}, \quad (4)$$

where τ is an unknown time beyond which a subject can be considered cured. It is easy to check that $E(\eta|\mathbf{W}) = E(v|\mathbf{W})$ if the distribution of $(C|\mathbf{W}, v = 0)$ equals that of $(C|\mathbf{W}, v = 1)$. Specifically,

$$E(\eta|\mathbf{W}) = E(\eta|\mathbf{W}, v = 0)P(v = 0|\mathbf{W}) + E(\eta|\mathbf{W}, v = 1)P(v = 1|\mathbf{W}).$$

Since $v = 0$ implies $\eta = 0$, then $E(\eta|\mathbf{W}, v = 0) = 0$ and $E(\eta|\mathbf{W})$ reduces to

$$E(\eta|\mathbf{W}) = \frac{E(v(1 - I(\delta = 0, T \leq \tau))|\mathbf{W}, v = 1)}{1 - G(\tau|\mathbf{W})} P(v = 1|\mathbf{W}). \quad (5)$$

Note that $v = 1$ implies $\delta = 0$, hence $T = C$ and the numerator in (5) is,

$$\begin{aligned} & E(v(1 - I(\delta = 0, T \leq \tau))|\mathbf{W}, v = 1) \\ &= E(1 - I(C \leq \tau)|\mathbf{W}, v = 1) = E(I(C > \tau)|\mathbf{W}, v = 1) \\ &= P(C > \tau|\mathbf{W}, v = 1) = 1 - G(\tau|\mathbf{W}, v = 1). \end{aligned}$$

Therefore, if C and ν are independent conditionally on $\mathbf{W}$, then

$$\begin{aligned} & 1 - G(\tau|\mathbf{W}, \nu = 1) \\ &= P(C > \tau|\mathbf{W}, \nu = 1)[P(\nu = 1|\mathbf{W}) + P(\nu = 0|\mathbf{W})] \\ &= P(C > \tau|\mathbf{W}, \nu = 1)P(\nu = 1|\mathbf{W}) + P(C > \tau|\mathbf{W}, \nu = 0)P(\nu = 0|\mathbf{W}) \\ &= P(C > \tau|\mathbf{W}) = 1 - G(\tau|\mathbf{W}). \end{aligned}$$

As a consequence, $E(\eta|\mathbf{W})$ in (5) is

$$E(\eta|\mathbf{W}) = \frac{1 - G(\tau|\mathbf{W}, \nu = 1)}{1 - G(\tau|\mathbf{W})} P(\nu = 1|\mathbf{W}) = P(\nu = 1|\mathbf{W}) = E(\nu|\mathbf{W}).$$

As a result, the values of η_i , $i = 1, \dots, n$ are the following: if $\delta_i = 1$ or if $\delta_i = 0$ and $T_i \leq \tau$, then $\eta_i = 0$; otherwise $\eta_i = 1/(1 - G(\tau|\mathbf{W}_i))$. Note that τ and the conditional distribution function $G(t|\mathbf{W})$ have to be estimated to obtain an estimation of η . Several estimates of the conditional distribution $G(t|\mathbf{W})$ can be considered according to the dimension and type of the covariate $\mathbf{W}$. As an alternative to, among others, the popular Cox proportional hazards model, we propose a completely nonparametric approach. Specifically, in Case 1, if Z is continuous, $G(t|Z)$ can be estimated using the conditional product-limit estimator by¹⁷. Otherwise, when Z is discrete or qualitative with values $\{z_1, \dots, z_k\}$, the stratified KM estimator can be used for every subsample $Z = z_j, j = 1, \dots, k$. In Case 2, the conditional distribution $G(t|X, Z)$ is estimated according to the type of the covariates (X, Z) following the general ideas in²³. For example, in the simplest case (X, Z) , if both covariates are continuous, $G(t|X, Z)$ is estimated with the generalization of the conditional product-limit estimator in²⁴, whereas if both (X, Z) are discrete or qualitative, with the stratified KM estimator using the corresponding subsamples. We suggest to select the bandwidth needed in the computation of the conditional product-limit estimator with the cross-validation (CV) bandwidth selector in²⁵.

The estimation of τ , a cure threshold beyond which a censored observation can be assumed as cured, would seem an ill-posed problem since whether a patient is cured or not is generally unobservable. However, note that under condition (2), then the largest observed survival time, $T_{\max}^1$, converges in probability to τ_0 (see¹²). Therefore, condition (2) guarantees asymptotically that all subjects censored after $T_{\max}^1$ can be assumed to be cured. As a consequence, we suggest to estimate in practice τ as the largest uncensored failure time, $T_{\max}^1$.

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:

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

where $p(Z)$ is not a constant function of Z . Our test will be based on the observations $\{(Z_i, \hat{\eta}_i), i = 1, \dots, n\}$. This implies that the test is carried out in a context without censoring, where the response is not observed but estimated. The test we propose is based on the following process:

$$U_n(z) = \frac{1}{n} \sum_{i=1}^n \left(\hat{\eta}_i - \left(\frac{1}{n} \sum_{j=1}^n \hat{\eta}_j \right) \right) I(Z_i \leq z), \quad (6)$$

which is a mean of the difference between the estimates of η and the conditional mean of η under the null hypothesis. Possible test statistics are the Cramér-von Mises (CM) test, $CM_n = \sum_{i=1}^n U_n^2(Z_i)$, or the Kolmogorov-Smirnov (K) test, $K_n = \max_{i=1, \dots, n} |n^{1/2} U_n(Z_i)|$. The null distribution of the test statistic is approximated by the bootstrap procedure. The bootstrap resampling plan used is similar to the one proposed by^{13 14} for bandwidth selection in nonparametric incidence and latency estimation but mimicking the null hypothesis. The method proceeds as follows:

1. For $i = 1, 2, \dots, n$, obtain Z_i^* from $(Z_1, \dots, Z_n)$ by random resampling with replacement.
2. Let $1 - \hat{p} = \hat{S}_n^{KM}(\infty)$, the value of the unconditional KM estimator of the survival function $S(t) = P(Y > t)$ evaluated at the largest uncensored observation. For $i = 1, 2, \dots, n$:
 - 2.1 Obtain a bootstrap cured observation $Y_i^* = \infty$ with probability $1 - \hat{p}$, and draw Y_i^* from a nonparametric estimator of the conditional distribution $F_0(t|Z_i^*) = 1 - S_0(t|Z_i^*)$ otherwise.
 - 2.2 Draw C_i^* from a nonparametric estimator of the conditional distribution $G(t|Z_i^*)$.

2.3 Compute $T_i^* = \min(Y_i^*, C_i^*)$ and $\delta_i^* = I(Y_i^* \leq C_i^*)$.

3. With the bootstrap resample, $\{(Z_i^*, T_i^*, \delta_i^*), i = 1, \dots, n\}$, compute $\hat{\eta}_i^*$ in (4) and obtain the bootstrap version of U_n in (6), and the corresponding bootstrap version of the Cramér-von Mises and Kolmogorov-Smirnov statistics, CM_n^* and K_n^* .
4. Repeat B times Steps 1-3 in order to generate B values of CM_n^* and K_n^* . Define the critical values d_{CM}^* and d_K^* as the values which are in position $\lceil (1 - \alpha)B \rceil$ in the corresponding sorted vector.
5. Compare the value of the statistic, CM_n (respectively, K_n), obtained with the original sample with d_{CM}^* (respectively, d_K^*), and reject the null hypothesis if $CM_n > d_{CM}^*$ (respectively, $K_n > d_K^*$). In addition, the p -value can be calculated as the proportion of resamples for which the bootstrap statistic, CM_n^* (respectively K_n^*) is larger than the value of the statistic with the original sample, CM_n (respectively K_n).
6. Repeat Steps 1-5 κ times. The power of the test is approximated as the proportion of rejections out of the κ trials.

Note that the nonparametric estimator of the conditional distributions $F_0(t|Z_i^*)$ and $G(t|Z_i^*)$ in Steps 2.1 and 2.2. can be the estimator of the latency function by¹³ and the conditional PL estimator by¹⁷, respectively, if Z is continuous, and the corresponding stratified estimators otherwise. Additionally, since we are assuming in Step 2 that the cure rate $1 - p$ is constant, the bootstrap resampling plan mimics H_0 .

3.1.1 | Z nominal

In the case with only one non-ordinal qualitative covariate $\mathbf{W} = Z$ with values $\{z_1, \dots, z_k\}$, 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 (6). We propose to consider all the possible $k!$ permutations of the values of Z and compute $U_n(z)$ (and the corresponding CM_n and K_n statistics) for each “ordered” permutation. Finally, the maximum of the $k!$ values CM_n and K_n is computed and compared with the critical point obtained by the bootstrap likewise.

A different approach consists of working with $k - 1$ dummy variables. The main benefit of this method would be that the value of the statistic is computed $k - 1$ times, whereas with the previous method, the statistic should be computed $k!$ times. Therefore, when the number k of levels is high, this approach is considerably less computationally expensive. However, the clear advantage of the first approach is that the categorical covariate is tested as a whole, regardless the number of levels.

3.2 | Case 2

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

$$H_0 : E(v|X, Z) = 1 - p(X) \text{ vs } H_1 : E(v|X, Z) = 1 - p(X, Z),$$

where $p(X, Z)$ depends on Z . To do this, we use the observations $\{(X_i, Z_i, \hat{\eta}_i), i = 1, \dots, n\}$.

3.2.1 | X continuous

Following¹⁶, the statistic is defined as:

$$U_{n,h}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \hat{f}_{Xh}(X_i) (\hat{\eta}_i - \hat{m}_h(X_i)) I(\mathbf{W}_i \leq \mathbf{w}), \quad (7)$$

where $\hat{f}_{Xh}$ is the Parzen-Rosenblatt (PR) 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 $m(x) = E(\eta|X = x)$, obtained by the NW kernel method with the same bandwidth h , and $\leq$ stands for component-wise inequality. Note that the process in (7) 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, $CM_{n,h} = \sum_{i=1}^n U_{n,h}^2(\mathbf{W}_i)$ and the Kolmogorov-Smirnov, $K_{n,h} = \max_{i=1, \dots, n} |n^{1/2} U_{n,h}(\mathbf{W}_i)|$ statistics. The test distribution under H_0 is approximated by the bootstrap, considering the following procedure, parallel to the algorithm presented in Section 3.1:

1. For $i = 1, 2, \dots, n$, obtain (X_i^*, Z_i^*) from $\{(X_1, Z_1), \dots, (X_n, Z_n)\}$ by random resampling with replacement.

2. For $i = 1, 2, \dots, n$, compute the value of the estimator of the cure probability, $1 - \hat{p}_{h_{CV}}(X_i^*)$, using the CV bandwidth, h_{CV} , proposed by²⁵. Then:
 - 2.1 With probability $1 - \hat{p}_{h_{CV}}(X_i^*)$ generate a bootstrap cured observation $Y_i^* = \infty$. Otherwise, Y_i^* is drawn from a nonparametric estimator of $F_0(t|Z_i^*, X_i^*) = 1 - S_0(t|Z_i^*, X_i^*)$.
 - 2.2 Draw C_i^* from a nonparametric estimator of the conditional distribution $G(t|X_i^*, Z_i^*)$.
 - 2.3 Compute $T_i^* = \min(Y_i^*, C_i^*)$ and $\delta_i^* = I(Y_i^* \leq C_i^*)$.
3. With the bootstrap resample, $\{(X_i^*, Z_i^*, T_i^*, \delta_i^*), i = 1, \dots, n\}$, compute $\hat{\eta}_i^*$ in (4) and obtain the bootstrap version of $U_{n,h}$ in (7), and the bootstrap version of the Cramér-von Mises and Kolmogorov-Smirnov statistics, say $CM_{n,h}^*$ and $K_{n,h}^*$.

Steps 4-6 are the same as those in the bootstrap procedure of Section 3.1. Note that in order to mimic H_0 , the bootstrap resamples are generated assuming in Step 2 that the cure rate does not depend on Z . To apply the test in practice, the estimated values $\hat{\eta}_i$ in (4) have to be computed. They depend on a suitable estimator of the conditional censoring distribution $G(t|\mathbf{w})$. Besides, in Steps 2.1 and 2.2, nonparametric estimators of $F_0(t|Z_i^*, X_i^*) = 1 - S_0(t|Z_i^*, X_i^*)$ and $G(t|X_i^*, Z_i^*)$ are required. In all cases, the estimators are the generalized conditional product-limit estimator in²⁴ when (X, Z) are continuous. When Z is not continuous, the functions are estimated with the conditional product-limit estimator in¹⁷ stratified on every value of Z . All the aforementioned estimators are computed with the corresponding CV bandwidth selector in²⁵.

In Case 2, a bandwidth h needs to be selected for the computation of the test statistic $U_{n,h}$ in (7). There are two main approaches for bandwidth selection in smoothing-based hypothesis tests: power maximization under the alternative hypothesis²⁶; and minimization of p -values^{27,28}. The two approaches are strongly related²⁹. In practice, we suggest to use any of the aforementioned bandwidth selection methods for nonparametric tests proposed in the literature.

3.2.2 | X categorical or discrete

For a categorical or discrete variable X , the estimated density, $\hat{f}_{Xh}(X_i)$, and the estimated regression function, $\hat{m}_h(X_i)$, in the test statistic in (7) 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. In this case, the estimation of the conditional distributions of Y and C is carried out with the Beran estimator using the CV bandwidth if Z is continuous, or with the stratified KM estimator otherwise.

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

3.3 | Case 3

The test can be generalized to a covariate vector $\mathbf{W} = (X, Z)$, where X is $\mathbb{R}^q$ -valued and Z is $\mathbb{R}^m$ -valued, as follows:

$$U_{n,h}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \hat{f}_{Xh}(X_i) (\hat{\eta}_i - \hat{m}_h(X_i)) I(\mathbf{W}_i \leq \mathbf{w}),$$

with

$$\hat{f}_{Xh}(X_i) = \frac{1}{nh^q} \sum_{j=1}^n \mathbb{K}\left(\frac{X_i - X_j}{h}\right) \text{ and } \hat{m}_h(X_i) = \frac{1}{nh^q} \frac{1}{\hat{f}_{Xh}(X_i)} \sum_{j=1}^n \mathbb{K}\left(\frac{X_i - X_j}{h}\right) \hat{\eta}_j,$$

where $\mathbb{K}$ denotes a kernel function on $\mathbb{R}^q$ such as $\mathbb{K}(\mathbf{x}) = \prod_{j=1}^q K(x_j)$, and $h \rightarrow 0$ is the bandwidth parameter. This case is not considered in the simulations.

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: $n = 50, 100$, and 200 . A total of $\kappa = 1000$ trials and $B = 2000$ bootstrap resamples are drawn. Since Case 2 with (X, Z) continuous involves a time consuming process, only $B = 1000$ bootstrap resamples are considered for it. The nominal significance level is $\alpha = 0.05$ along the simulations. All the simulation studies were coded in R language³⁰.

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 hypotheses, respectively.

H_0	% Cens	% Cure	n	CM	K
$p(z) = 0.5$	60.4	50.0	50	0.055	0.072
			100	0.056	0.052
			200	0.069	0.063
$p(z) = 0.6$	52.5	40.0	50	0.049	0.058
			100	0.069	0.082
			200	0.061	0.068
$p(z) = 0.7$	44.6	30.0	50	0.050	0.059
			100	0.051	0.053
			200	0.087	0.075
$p(z) = 0.8$	36.6	20.0	50	0.031	0.034
			100	0.026	0.033
			200	0.094	0.092
H_1	% Cens	% Cure		CM	K
$p(z)$	54.6	46.7	50	0.987	0.987
			100	0.998	0.997
			200	1	1

4.1 | Case 1

We investigate the finite sample behavior of the test for Z continuous, discrete and nominal.

4.1.1 | Z continuous

Let Z be a continuous random variable with distribution $U(-20, 20)$. The censoring variable C has conditional distribution $G(t|z) \sim \text{Exp}(\lambda(z))$, with $\lambda(z) = 0.6/(2 + (z - 20)/40)$, and the survival function of the uncured individuals is

$$S_0(t|z) = \frac{\exp(-\alpha(z)t) - \exp(-\alpha(z)\tau_0)}{1 - \exp(-\alpha(z)\tau_0)} I(t \leq \tau_0), \quad (8)$$

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

Under the null hypothesis, $H_0 : E(v|Z) = 1 - p$, we consider the four constant values $p = 0.5, 0.6, 0.7, 0.8$. Under the alternative hypothesis, the cure probability considered is

$$1 - p(z) = 1 - \frac{\exp(\beta_0 + \beta_1 z)}{1 + \exp(\beta_0 + \beta_1 z)}, \quad (9)$$

with $\beta_0 = 0.476$ and $\beta_1 = 0.358$. The average percentage of censored data is 54.65% and of cured data is 46.66%.

For the estimation of the conditional distribution functions $F_0(t|z)$ and $G(t|z)$ required in Steps 2.1 and 2.2 of the bootstrap procedure and in the estimation of the values of η in (4), we use the conditional product-limit estimator by¹³ and¹⁷, respectively, with the CV bandwidth selector by²⁵, selected from a grid of 10 equispaced bandwidths $h_j = D_j n^{-1/5}$, from $D_1 = 4$ to $D_{10} = 60$.

The results are given in Table 1. It is noteworthy that, under H_0 , the size of the test is not far from the nominal level, for different constant values of p . The effect of the value of p in the results of the test can be seen in Table 1. As the value of p decreases the higher percentage of cured observations, so the more information about the cure rate is expected to be available in the sample, and the better performance of the test (the size under H_0 gets closer to the nominal level). Furthermore, under H_1 , the power of the test is very close to 1.

4.1.2 | Z discrete

We work with a discrete covariate Z with 3 ordered values $b_1 < b_2 < b_3$. We consider two situations according to the probability mass function of Z given by $\Pi_z(b_i) = P(Z = b_i)$: $(\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3)) = (1/3, 1/3, 1/3)$ and $(1/5, 1/5, 3/5)$.

The observations (Z_i, T_i, δ_i) , $i = 1, \dots, n$ were simulated from the model for Z continuous in Section 4.1.1. Since Z is not a continuous covariate, the values of Z , $\{b_1, b_2, b_3\}$, will be those that, in that model, yield fixed probabilities $p(b_1) = p(b_2) = p(b_3) \in \{0.5, 0.6, 0.7, 0.8\}$, under H_0 , and $p(b_1) \in \{0.1, 0.3\}$, $p(b_2) = 0.5$ and $p(b_3) \in \{0.7, 0.9\}$ under H_1 .

The results in Table 2 were obtained under the null and under the alternative hypotheses. The results under the null hypothesis are very similar to the significance level, $\alpha = 0.05$, regardless the value of $p(b_i)$, $i = 1, 2, 3$. Moreover, under the alternative hypothesis, the power of the test is very close to 1 for large sample sizes and when the values of the incidence, as a function of z , are more distant, that is, $p(b_1) = 0.1$, $p(b_2) = 0.5$ and $p(b_3) = 0.9$.

4.1.3 | Z qualitative

A qualitative covariate Z with three possible values $\{b_1, b_2, b_3\}$ was considered, with cure probabilities $p(b_1) = p(b_2) = p(b_3) \in \{0.5, 0.6, 0.7, 0.8\}$ under H_0 . Under the alternative hypothesis, two scenarios were considered: $(p(b_1), p(b_2), p(b_3)) = (0.3, 0.5, 0.7)$, and $(0.1, 0.5, 0.9)$. Each scenario was simulated with two possible probability mass functions (pmf) for Z , $(1/3, 1/3, 1/3)$ and $(1/5, 1/5, 3/5)$. The distribution of the censoring variable is $C|Z = b_1 \sim \text{Exp}(0.6)$, $C|Z = b_2 \sim \text{Exp}(0.45)$ and $C|Z = b_3 \sim \text{Exp}(0.3)$. Finally, the conditional survival functions of the uncured individuals $S_0(t|b_1)$, $S_0(t|b_2)$ and $S_0(t|b_3)$ are given by $S_0(t|z)$ in (8) at 3 numerical values z_1 , z_2 and z_3 , selected in every scenario in such a way that the function p in (9) evaluated at z_1 , z_2 and z_3 equals $p(b_1)$, $p(b_2)$ and $p(b_3)$ respectively. In this case, the conditional distribution functions $F_0(t|b_j)$ and $G(t|b_j)$, $j = 1, 2, 3$ are estimated with the corresponding stratified latency and KM estimators computed with the corresponding subsample $Z = b_j$, $j = 1, 2, 3$.

The results are shown in Table 3. Under the null hypothesis, the size of the test is very similar to the significance level, $\alpha = 0.05$, mainly for larger sample sizes. Regarding the alternative hypothesis, as expected, the power of the test is higher for large sample sizes and when H_1 is easier to detect, that is, in the most extreme case $(p(b_1), p(b_2), p(b_3)) = (0.1, 0.5, 0.9)$.

4.2 | Case 2

In this case, $\mathbf{W} = (X, Z)$ has dimension $m + 1$, with a univariate X and a m -dimensional covariate Z . For the sake of simplicity, in this simulation study we suppose that Z is also one-dimensional. In Section 4.2.1 we can see the results for Case 2 with X and Z continuous, and in Section 4.2.2, the corresponding results for Case 2 with X and Z discrete.

4.2.1 | (X, Z) continuous

We consider two continuous covariates X and Z with distribution $U(-20, 20)$. The censoring variable C has conditional distribution $C|X = x, Z = z \sim \text{Exp}(\lambda(x, z))$, with $\lambda(x, z) = 0.6/(2 + (0.5(x + z) - 20)/40)$, and the latency is

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

where $\tau_0 = 4.605$ and $\alpha(x, z) = \exp((z + 20)/40)$ under H_0 and $\alpha(x, z) = \exp((x + z + 20)/40)$ under H_1 . The incidence is

$$p(x, z) = \frac{\exp(\beta_0 + \beta_1 x(1 + \beta_2 z))}{1 + \exp(\beta_0 + \beta_1 x(1 + \beta_2 z))}, \quad (10)$$

with $\beta_0 = 0.476$, $\beta_1 = 0.358$, and $\beta_2 = 0$ under H_0 and $\beta_2 = 0.225$ under H_1 .

The conditional distribution functions $F_0(t|x, z)$ and $G(t|x, z)$ are estimated with the generalized product-limit estimator by¹³ and²⁴, respectively, and the CV bandwidth selector in²⁵, using the search grid of equispaced bandwidths $h_j = D_j n^{-1/6}$, from $D_1 = 7$ to $D_{10} = 60$. The performance of the test is assessed in a grid of bandwidths $h = C n^{-1/3m}$, following¹⁶, where $C=10$, 20, 40, 60, 100 and m is the dimension of the tested covariate Z (note that in our case, $m = 1$).

TABLE 2 Size (top) and power (bottom) of the test for Case 1 with Z discrete, under the null and the alternative hypotheses, respectively.

H_0	n	% Cens	% Cure	CM	K
$p(z_1) = p(z_2) = p(z_3) = 0.5$		60.8	50.0		
	50			0.042	0.043
	100			0.039	0.040
	200			0.047	0.054
$p(z_1) = p(z_2) = p(z_3) = 0.6$		52.7	40.0		
	50			0.036	0.034
	100			0.048	0.042
	200			0.046	0.044
$p(z_1) = p(z_2) = p(z_3) = 0.7$		44.5	30.0		
	50			0.032	0.036
	100			0.044	0.035
	200			0.046	0.045
$p(z_1) = p(z_2) = p(z_3) = 0.8$		36.1	20.0		
	50			0.036	0.035
	100			0.026	0.023
	200			0.039	0.047
H_1	n	% Cens	% Cure	CM	K
$\Pi_z = (1/3, 1/3, 1/3)$					
$p(z_1) = 0.3, p(z_2) = 0.5, p(z_3) = 0.7$		60.0	50.0		
	50			0.380	0.334
	100			0.673	0.631
	200			0.918	0.874
$\Pi_z = (1/3, 1/3, 1/3)$					
$p(z_1) = 0.1, p(z_2) = 0.5, p(z_3) = 0.9$		58.8	50.0		
	50			0.928	0.888
	100			0.995	0.992
	200			0.999	0.999
$\Pi_z = (1/5, 1/5, 3/5)$					
$p(z_1) = 0.3, p(z_2) = 0.5, p(z_3) = 0.7$		52.2	42.0		
	50			0.282	0.283
	100			0.626	0.609
	200			0.840	0.821
$\Pi_z = (1/5, 1/5, 3/5)$					
$p(z_1) = 0.1, p(z_2) = 0.5, p(z_3) = 0.9$		44.3	34.0		
	50			0.892	0.885
	100			0.985	0.985
	200			0.998	0.998

Under the null hypothesis, the test is conservative using the Cramér-von Mises statistic. The performance for the Kolmogorov-Smirnov version is good, provided that the smoothing parameter is selected in an appropriate way. See the results in Table 4.

TABLE 3 Size (top) and power (bottom) of the test for Case 1 with Z qualitative with values $\{b_1, b_2, b_3\}$ and pmf $\Pi_z = (\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3))$ under the null and the alternative hypotheses, respectively.

H_0	% Cens	% Cure	n	CM	K
$p(z_1) = p(z_2) = p(z_3) = 0.5$	60.8	50.0	50	0.041	0.047
			100	0.045	0.050
			200	0.055	0.053
			50	0.026	0.028
			100	0.047	0.044
			200	0.040	0.043
			50	0.041	0.046
			100	0.038	0.040
			200	0.039	0.041
$p(z_1) = p(z_2) = p(z_3) = 0.6$	52.7	40.0	50	0.028	0.030
			100	0.026	0.029
			200	0.044	0.045
			50	0.041	0.046
			100	0.038	0.040
			200	0.039	0.041
			50	0.041	0.046
			100	0.038	0.040
			200	0.039	0.041
$p(z_1) = p(z_2) = p(z_3) = 0.7$	44.5	30.0	50	0.041	0.046
			100	0.038	0.040
			200	0.039	0.041
			50	0.041	0.046
			100	0.038	0.040
			200	0.039	0.041
			50	0.041	0.046
			100	0.038	0.040
			200	0.039	0.041
$p(z_1) = p(z_2) = p(z_3) = 0.8$	36.1	20.0	50	0.028	0.030
			100	0.026	0.029
			200	0.044	0.045
			50	0.041	0.046
			100	0.038	0.040
			200	0.039	0.041
			50	0.041	0.046
			100	0.038	0.040
			200	0.039	0.041
H_1	% Cens	% Cure		CM	K
$(p(b_1), p(b_2), p(b_3)) = (0.3, 0.5, 0.7)$ $\Pi_z = (1/3, 1/3, 1/3)$	60.1	50.0	50	0.277	0.287
			100	0.566	0.574
			200	0.859	0.851
			50	0.280	0.269
			100	0.598	0.588
			200	0.817	0.814
			50	0.280	0.269
			100	0.598	0.588
			200	0.817	0.814
$(p(b_1), p(b_2), p(b_3)) = (0.1, 0.5, 0.9)$ $\Pi_z = (1/3, 1/3, 1/3)$	58.8	50.0	50	0.865	0.855
			100	0.993	0.993
			200	1	1
			50	0.893	0.885
			100	0.987	0.989
			200	0.998	0.998
			50	0.893	0.885
			100	0.987	0.989
			200	0.998	0.998

4.2.2 | (X, Z) discrete

The covariates X and Z are discrete 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 for X and Z : in the first one, both are $(1/3, 1/3, 1/3)$, whereas in the second one, both are $(1/5, 1/5, 3/5)$.

Under H_0 , the values of X are $\{a_1, a_2, a_3\} = \{-2.4622, -0.19702, 1.0371\}$ and the values of Z are $b = b_1 = b_2 = b_3 = 0.6157$. The probabilities $p(a_i, b_1) = p(a_i, b_2) = p(a_i, b_3)$, with $i = 1, 2, 3$, are given by function (10) evaluated at (a_i, b) . Under H_1 , values $\{a_1, a_2, a_3\}$ are the same as the ones under H_0 . The values of Z are $\{b_1, b_2, b_3\} = \{-13.123, 0, 4.9454\}$. See Table 5 for details.

Table 6 (top) shows the results under the null hypothesis, which are very close to the nominal significance level. The results under the alternative hypothesis are given in Table 6 (bottom). It is important to highlight that the power of the test increases considerably with the sample size (i.e. for CM, from 0.083 with $n=50$, to 0.492 with $n=200$).

TABLE 4 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 hypotheses, respectively.

H_0	% Cens	% Cure	n	h	CM	K
	54.3	45.8	50	2.71	0.012	0.009
				5.43	0.015	0.022
				10.86	0.018	0.028
				16.28	0.017	0.035
				27.14	0.024	0.058
			100	2.15	0.011	0.009
				4.31	0.014	0.025
				8.62	0.018	0.030
				12.93	0.018	0.038
				21.54	0.019	0.063
			200	1.71	0.022	0.033
				3.42	0.029	0.035
				6.84	0.035	0.039
				10.26	0.028	0.052
				17.10	0.019	0.066
H_1	% Cens	% Cure	n	h	CM	K
	58.1	47.3	50	2.71	0.069	0.074
				5.43	0.130	0.155
				10.86	0.164	0.208
				16.28	0.167	0.210
				27.14	0.119	0.172
			100	2.15	0.227	0.238
				4.31	0.367	0.384
				8.62	0.430	0.432
				12.93	0.424	0.433
				21.54	0.320	0.357
			200	1.71	0.575	0.582
				3.42	0.767	0.756
				6.84	0.820	0.800
				10.26	0.830	0.793
				17.10	0.753	0.735

5 | APPLICATION TO A COLORECTAL DATASET

We work with a dataset related to 414 colorectal cancer patients from the University Hospital of A Coruña (CHUAC), Spain. The variable of interest, Y , is the time, in months, since diagnosis until death from cancer. An individual is considered long-term survivor (or 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 information provided is, along with the observed lifetime and the censoring indicator, the location (colon $n_1 = 111$, rectum $n_2 = 303$), the age (23-102 years) and the stage TNM, which is the main determinant in prognosis of these patients. The stage has 3 components which can be combined so that we can classify each patient in a unique (and numeric) stage from 1 to 4. About 50% of the observations are censored, with the percentage of censoring 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).

In Stage 4, the nonparametric estimation of the cure rate is constant equal to 0 (see Figure 3), sustaining that long-term survival for individuals at this stage is not usual³¹. Since the cure probability is almost 0, a cure model would not be useful to study

TABLE 5 Scenario for Case 2 when X and Z are discrete with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$, respectively. The probability mass function, for both X and Z , is $(1/3, 1/3, 1/3)$. The uncure probabilities, $p(a_i, b_j)$, $i, j = 1, 2, 3$, are obtained evaluating $p()$ in Equation (10) in (a_i, b_j) , $i, j = 1, 2, 3$.

H_0	$b_1 = 0.6157$	$b_2 = 0.6157$	$b_3 = 0.6157$
$a_1 = -2.4622$	0.40	0.40	0.40
$a_2 = -0.19702$	0.60	0.60	0.60
$a_3 = 1.0371$	0.70	0.70	0.70
H_1	$b_1 = -13.123$	$b_2 = 0$	$b_3 = 4.9454$
$a_1 = -2.4622$	0.10	0.60	0.80
$a_2 = -0.19702$	0.35	0.40	0.42
$a_3 = 1.0371$	0.56	0.30	0.22

TABLE 6 Size (top) and power (bottom) of the test for Case 2 with X and Z discrete, with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$, under the null and the alternative hypotheses, respectively. The probability mass function of Z equals that of X .

H_0	n	% Cens	% Cure	CM	K
$\Pi_x = (1/3, 1/3, 1/3)$	50	55.4	42.9	0.029	0.041
	100			0.054	0.064
	200			0.037	0.052
$\Pi_x = (1/5, 1/5, 3/5)$	50	52.6	37.6	0.044	0.081
	100			0.040	0.045
	200			0.043	0.049
H_1	n	% Cens	% Cure	CM	K
$\Pi_x = (1/3, 1/3, 1/3)$	50	54.4	41.1	0.083	0.096
	100			0.225	0.212
	200			0.492	0.370
$\Pi_x = (1/5, 1/5, 3/5)$	50	52.9	38.0	0.090	0.111
	100			0.167	0.158
	200			0.236	0.222

this subset and the results for Stage 4 are not included in the analysis. Hence, we work with a dataset of 362 observations. We considered $B = 1000$ bootstrap resamples to approximate the distribution of the test statistic and a significance level $\alpha = 0.05$.

5.1 | Suitability of the real-data to a cure model

Cure models should be applied when there is a strong rationale for the existence of cured subjects. Colorectal cancer (CRC) is one of the leading causes of cancer mortality and morbidity worldwide, accounting for 9.4% of all cancer cases and 1 million new cases annually³². Death rates from CRC have declined progressively since the last decades, due to improvements such as earlier diagnosis and better treatments, most notably surgical techniques, such as laparoscopic surgery and total mesorectal excision³³. This increased the rates of long-term survivors, conventionally been defined by those with at least 5-year survival times after cancer diagnosis³⁴. The long-term survival rate for colorectal cancer patients is, for example, 90.3% when colorectal cancers are detected at a localized stage³⁵. This shows feasible the possibility of cure in the CRC patients of the analyzed dataset.

Cure models usually require, not only biological evidence for the possibility of cure, but also large sample sizes and a reasonably long follow-up time. This is particularly important if censoring is heavy, since too much censoring or insufficient follow-up time can lead to overestimated cure rates³⁶. The present dataset contains a sample size of $n = 362$ patients.

There are several ways to guess if the follow-up period (almost 19 years in the dataset) is long enough. A characteristic of a cure model is that the limit of $S(t) = P(Y > t)$ is non-zero as t tends to infinity. Then if the Kaplan-Meier (KM) plot suggests

FIGURE 1 Survival function estimation for the colorectal cancer dataset computed with the Kaplan-Meier estimator. The black crosses correspond to censored observations.

a non-zero asymptote, then a cure model may be appropriate and useful. Figure 1 shows the KM estimate, $\hat{S}$, for the survival function for the colorectal cancer dataset. We can appreciate that the survival curve $\hat{S}$ has a plateau at the end of the study. This non-zero asymptote could be taken as an estimator of the cure rate, that is, the proportion of patients who will not die from colorectal cancer, so they can be considered as “cured”.

Besides,²² proposed a simple nonparametric test to devise whether the follow-up time is enough. The procedure is based on the length of the range $(T_{\max}^1, T_{(n)})$, that is, the interval at the right tail where the KM estimator of the survival function $S(t)$ is constant. A long and stable plateau with heavy censoring at the tail of the KM curve is taken as an evidence that the follow-up period has been sufficient. The results of the test by²² suggest that there is a significant evidence for the existence of a cure rate ($p = 0.0008589$).

5.2 | Case 1

We started studying the effect of the age and the stage, separately, on the probability of cure. Age at diagnosis and tumor stage are known to strongly influence colorectal cancer treatment regimen and five-year survival^{37,38}. However, the effect of the age and stage on the probability of cure is rarely analyzed, since a few studies of cure for colorectal cancer patients have focused on the estimation of cure by age and stage at diagnosis³⁹, but not on the statistical significance of those covariates on the probability of cure.⁴⁰ propose a multi-state Markov model with an incorporated cured fraction to assess how specific covariates influence the cure rate. They state that the covariate age does not have any influence, unlike the covariate stage. Furthermore, the studies of cure for colorectal cancer patients usually categorize the age into intervals, not treating it as a continuous covariate. So the purpose is to test whether the age or the stage have a significant effect on the cure rate, using the nonparametric significance test.

As for the age, in Figure 2 we can appreciate how the nonparametric estimator of the cure rate changes with the age. The cure probability decreases, in general, with increasing age, suggesting that the age may have some influence on the cure rate for patients of Stages 1-3. Younger patients are likely to tolerate the intensive cancer treatments better than older patients and are therefore to achieve cure, while elderly patients are substantially less likely to receive surgery and chemotherapy³⁹. The test found this effect of the age on the cure probability as significant ($p_{CM} = 0.028$, $p_K = 0.030$).

We also tested the effect of the age on the cure probability within each stage (see Figure 3). In Stage 1 the cure rate remains almost constant regardless the age, whereas for Stages 2 and 3 the age may have some influence since the cure probability decreases as the age increases. Specifically, for Stage 1 the estimated cure probability fluctuates around 25% for most patients, whereas in Stage 2, this probability decreases with the age, from about 30% in patients with age at diagnosis 50-60 to 7% for

FIGURE 2 Left panel: Nonparametric estimation of the cure probability depending on the age for patients in Stages 1 – 3 computed with the bootstrap bandwidth (solid line) and with a smoothed bootstrap bandwidth (dashed line). The thin solid line represents the Parzen-Rosenblatt kernel density estimation of the covariate age, using Sheather and Jones’ plug-in bandwidth. Right panel: Estimated KM survival curves for age groups.

patients above 80. Regarding Stage 3, the cure probability is around 35% for individuals younger than 60, whereas for patients above this age that probability decreases dramatically. When the test was applied for each stage separately, the effect of the age on the cure probability was not found significant in any stage ($p_{CM} = 0.410$, $p_K = 0.276$ for Stage 1; $p_{CM} = 0.418$, $p_K = 0.498$ for Stage 2; and $p_{CM} = 0.186$, $p_K = 0.166$ for Stage 3).

Regarding the effect of the stage on the probability of cure, the more tumors are detected at earlier stage, the more curative resections are possible⁴¹. The estimated cure probabilities for each stage support that statement, as expected they decrease as the stage of the cancer progresses: 0.28 for Stage 1; 0.13 for Stage 2, and 0.13 for Stage 3, see Figure 4. However, the differences of the cure probabilities among the three stages were not significant ($p_{CM} = 0.733$ and $p_K = 0.690$).

5.3 | Case 2

Since the probability of cure was found to depend on the age of the patient (see Section 5.2), the test was performed to study if it also depends on the cancer stage. Note that in this case a value for the bandwidth h is required to compute the test statistic in (7). The test, using a set of bandwidths $h = Cn^{-1/3}$ with $C = 10, 20, 40, 60, 100, 125, 150, 200, 250, 300, 350, 400, 450, 500, 550$ and 600 , indicates that 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 $h_1 = 1.40$ ($p_{CM} = 0.272$, $p_K = 0.271$) to the largest one $h_{16} = 84.19$ ($p_{CM} = 0.489$, $p_K = 0.256$).

6 | CONCLUSIONS

A nonparametric covariate significance test for the probability of cure in mixture cure models is introduced. The methodology, that can be applied to any type of covariates, enjoys the flexibility of nonparametric hypothesis tests with the advantage of getting rid of the need for a bandwidth parameter in the simplest case where the effect of one univariate Z is being tested. For more complex hypotheses, where a continuous covariate X is assumed to affect the cure rate under the null hypothesis, and the influence of other covariate Z is studied, the choice of a smoothing parameter h is required. Several bandwidth selectors for smoothed tests are proposed in the literature that can also be applied in this context.

FIGURE 3 Nonparametric estimation of the cure probability depending on the age for the patients in every stage separately, computed with the bootstrap bandwidth (solid line) and with a smoothed bootstrap bandwidth (dashed line). The thin solid line represents the Parzen-Rosenblatt kernel density estimation of the covariate age, using Sheather and Jones' plug-in bandwidth.

The test is based on the estimation of an unobservable variable η (a proxy for the cure indicator, v) through suitable estimation of the conditional distribution of the censoring variable. The estimation of the conditional distribution of the lifetime is also needed to approximate the distribution of the test using the bootstrap. The results of the simulation study support the use of nonparametric estimators for both conditional distributions, such as the conditional product-limit estimator with the CV bandwidth selector in²⁵. The method is applied to a colorectal dataset, and the results show that the covariate age has a significant influence on the cure probability.

FIGURE 4 Estimated survival functions according to the stage.

ACKNOWLEDGMENTS

The first author's research was sponsored by the Spanish FPU (Formación de Profesorado Universitario) Grant FPU13/01371 from MECD (Ministerio de Educación, Cultura y Deporte). 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, Innovación y Universidades) 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) for the third author is gratefully acknowledged. The authors are grateful to Dr. S. Pértiga and Dr. S. Pita, at the University Hospital of A Coruña, for providing the colorectal cancer dataset.

Conflict of interest

The authors declare no potential conflict of interests.

References

1. Farewell VT. Mixture models in survival analysis: are they worth the risk?. *Can J Stat* 1986; 14: 257–262. doi: 10.2307/3314804
2. Peng Y, Taylor JMG. Cure models. In: Klein J, Van Houwelingen HC, Ibrahim JG, Scheike TH., eds. *Handbook of Survival Analysis* Boca Raton, FL, USA: Chapman & Hall. 2014 (pp. 113–134).
3. Amico M, Van Keilegom I. Cure models in survival analysis. Review paper, KU Leuven; Leuven, Belgium: 2018.
4. Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J R Stat Soc Ser B - Stat Methodol* 1949; 11: 15–53. doi: 10.2307/2983694
5. Maller RA, Zhou S. *Survival Analysis with Long-Term Survivors*. Chichester, U. K.: Wiley . 1996

6. Peng Y, Dear KB. A nonparametric mixture model for cure rate estimation. *Biometrics* 2000; 56: 237–243. doi: 10.1111/j.0006-341X.2000.00237.x
7. Sy JP, Taylor JMG. Estimation in a Cox proportional hazards cure model. *Biometrics* 2000; 56: 227–236. doi: 10.1111/j.0006-341X.2000.00227.x
8. Zhang J, Peng Y. Accelerated hazards mixture cure model. *Lifetime Data Anal* 2009; 15: 455–467. doi: 10.1007/s10985-009-9126-4
9. Peng Y, Taylor JMG. Mixture cure model with random effects for the analysis of a multi-centre tonsil cancer study. *Stat Med* 2011; 30: 211–223. doi: 10.1002/sim.4098
10. Wang L, Du P, Lian H. Two-component mixture cure rate model with spline estimated nonparametric components. *Biometrics* 2012; 68: 726–735. doi: 10.1111/j.1541-0420.2011.01715.x
11. Patilea V, Van Keilegom I. A general approach for cure models in survival analysis. *arXiv preprint arXiv:1701.03769* 2017.
12. Xu J, Peng Y. Nonparametric cure rate estimation with covariates. *Can J Stat* 2014; 42: 1–17. doi: 10.1002/cjs.11197
13. López-Cheda A, Cao R, Jácome MA, Van Keilegom I. Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Comput Stat Data Anal* 2017; 105: 144–165. doi: 10.1016/j.csda.2016.08.002
14. López-Cheda A, Jácome MA, Cao R. Nonparametric latency estimation for mixture cure models. *TEST* 2017; 26: 353–376. doi: 10.1007/s11749-016-0515-1
15. Müller UU, Van Keilegom I. Goodness-of-fit tests for the cure rate in a mixture cure model. *Biometrika* 2018; 106: 211–227. doi: 10.1093/biomet/asy058
16. Delgado MA, González-Manteiga W. Significance testing in nonparametric regression based on the bootstrap. *Ann Stat* 2001; 29: 1469–1507. doi: 10.1214/aos/1013203462
17. Beran R. Nonparametric regression with randomly censored survival data. tech. rep., University of California, Berkeley; Berkeley: 1981.
18. Li CS, Taylor JMG, Sy JP. Identifiability of cure models. *Stat Probab Lett* 2001; 54: 389–395. doi: 10.1016/S0167-7152(01)00105-5
19. Hanin L, Huang LS. Identifiability of cure models revisited. *J Multivar Anal* 2014; 130: 261–274. doi: 10.1016/j.jmva.2014.06.002
20. Chen T, Du P. Mixture cure rate models with accelerated failures and nonparametric form of covariate effects. *J Nonparametr Stat* 2018; 30: 216–237. doi: 10.1080/10485252.2017.1404599
21. Taylor JMG. Semi-parametric estimation in failure time mixture models. *Biometrics* 1995; 51: 899–907. doi: 10.2307/2532991
22. Maller RA, Zhou S. Estimating the proportion of immunes in a censored sample. *Biometrika* 1992; 79: 731–739. doi: 10.1093/biomet/79.4.731
23. Racine J, Li Q. Nonparametric estimation of regression functions with both categorical and continuous data. *J Econom* 2004; 119: 99–130. doi: 10.1016/S0304-4076(03)00157-X
24. Liang HY, de Uña-Álvarez J, Iglesias-Pérez MC. Asymptotic properties of conditional distribution estimator with truncated, censored and dependent data. *TEST* 2012; 21: 790–810. doi: 10.1007/s11749-012-0281-7
25. Geerdens C, Acar EF, Janssen P. Conditional copula models for right censored clustered event time data. *Biostatistics* 2018; 19: 247–262. doi: 10.1093/biostatistics/kxx034
26. Kulasekera KB, Wang J. Smoothing parameter selection for power optimality in testing of regression curves. *J Am Stat Assoc* 1997; 92: 500–511. doi: 10.2307/2965699

27. Martínez-Camblor P. Nonparametric k -sample test based on kernel density estimator for paired design. *Comput Stat Data Anal* 2010; 54: 2035–2045. doi: 10.1016/j.csda.2010.03.009
28. Martínez-Camblor P, de Uña-Álvarez J. Studying the bandwidth in k -sample smooth tests. *Comput Stat* 2013; 28: 875–892. doi: 10.1007/s00180-012-0333-1
29. de Uña-Álvarez J. Comments on “An updated review of Goodness-of-Fit tests for regression models. *Test* 2013; 22: 414–418. doi: 10.1007/s11749-013-0327-5
30. R Core Team . *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing . 2018.
31. Miyamoto Y, Hayashi N, Sakamoto Y, et al. 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* 2015; 20: 1140–1146. doi: 10.1007/s10147-015-0835-2.
32. Boyle P, Levin B. *World Cancer Report*. Lyon, France: IARC . 2018.
33. Mitry E, Bouvier AM, Esteve J, Faivre J. Improvement in colorectal cancer survival: A population-based study. *Eur J Cancer* 2005; 41: 2297–2303. doi: 10.1016/j.ejca.2005.01.028
34. American Cancer Society (ACS) . *Cancer Facts and Figures*. Atlanta: American Cancer Society . 2017.
35. DeSantis CE, Lin CC, Mariotto AB, et al. Cancer treatment and survivorship statistics. *CA Cancer J Clin* 2014; 64: 252–271. doi: 10.3322/caac.21235
36. Laska EM, Meisner MJ. Nonparametric estimation and testing in a cure model. *Biometrics* 1992; 48: 1223–1234. doi: 10.2307/2532714
37. Vercelli M, Lillini R, Capocaccia R, et al. Cancer survival in the elderly: effects of socio-economic factors and health care system features (ELDCARE project). *Eur J Cancer* 2006; 42: 234–242. doi: 10.1016/j.ejca.2005.07.032
38. Guyot F, Faivre J, Manfredi S, Meny B, Bonithon-Kopp C, Bouvier AM. Time trends in the treatment and survival from recurrence of colorectal cancer. *Ann Oncol* 2005; 16: 756–761. doi: 10.1093/annonc/mdi151
39. Shack LG, Shah A, Lambert PC, Rachet B. Cure by age and stage at diagnosis for colorectal cancer patients in North West England, 1997-2004: A population-based study. *Cancer Epidemiol* 2012; 36: 548–553. doi: 10.1016/j.canep.2012.06.011
40. Conlon AS, Taylor JM, Sargent DJ. Multi-state models for colon cancer recurrence and death with a cured fraction. *Stat Med* 2014; 10: 1750–1766. doi: 10.1002/sim.6056
41. Ito Y, Nakayama T, Miyashiro I, et al. Trends in ‘cure’ fraction from colorectal cancer by age and tumour stage between 1975 and 2000, using population-based data, Osaka, Japan. *Jpn J Clin Oncol* 2012; 42: 974–983. doi: 10.1093/jjco/hys132

How to cite this article: A. López-Cheda, M. A. Jácome, I. Van Keilegom and R. Cao (—), Nonparametric covariate significance tests for the cure rate in mixture cure models, —, —.