XVII Congreso Galego de Estatística e Investigación de Operacións
Ourense, 23 - 25 de outubro de 2025

Single-index mixture cure models: an application to cardiotoxicity in breast cancer patients

Beatriz Piñeiro-Lamas[1], Ana López-Cheda[1] and Ricardo Cao[1]

[1]Universidade da Coruña, Grupo de Modelización, Optimización e Inferencia Estatística, Departamento de Matemáticas, Facultade de Informática, Campus de Elviña, 15071 A Coruña, Spain

ABSTRACT

Cardiotoxicity is a critical side effect of breast cancer treatments that significantly impacts patient care and treatment outcomes. Understanding both the probability of developing cardiotoxicity and the timing of its onset is essential for optimizing therapeutic strategies and patient monitoring. However, analyzing cardiotoxicity data presents challenges: not all patients will develop this side effect and multiple patient characteristics of different nature are available. To address this real clinical problem, single-index mixture cure models are proposed for both vector and functional covariates. These models estimate the cure rate and the latency nonparametrically, assuming that they depend on patient characteristics via a single linear combination, avoiding the curse of dimensionality. The parameter vector that determines the single-index is estimated by maximizing the semiparametric likelihood, with bandwidth selection for the nonparametric components incorporated directly into the optimization process. The asymptotic normality of the parameter vector estimator is established. The methodology, implemented in the sicure `R` package, is applied to analyze cardiotoxicity in 531 women with breast cancer from the University Hospital of A Coruña.

Keywords: cardio-oncology, censored data, dimension reduction, semiparametric models, survival analysis.

1. INTRODUCTION

Cardiotoxicity represents one of the most significant complications of modern breast cancer treatment. The condition has important prognostic implications and may progress to heart failure over time, making its early detection a critical but ongoing clinical challenge. Current practice relies on standardized monitoring protocols that lead to inefficiencies: low-risk patients receive intensive monitoring that may be unnecessary, while high-risk patients may not be identified early enough for optimal intervention. Improving the baseline risk stratification of cardiotoxicity would allow for more personalized monitoring, with a focus on patients with higher risk, and for the optimization of resources.

Conventional survival models assume that all subjects would eventually experience the event of interest under infinite follow-up. This is reflected in survival functions that tend to zero as time increases. However, in many medical scenarios, including cardiotoxicity, this assumption is not realistic. Some patients, considered as *cured*, may never experience the event. In these cases, the survival function tends to a positive value as time increases, reflecting the cure rate. Mixture cure models address this by modeling the population as a mixture of two groups: susceptible individuals (who may experience the event) and cured individuals (who will never experience it). They allow to estimate the probability of cure and the survival function for the uncured or susceptible subjects (known as latency). In the literature, several parametric, semiparametric and nonparametric methods have been proposed to estimate both functions. This work focuses on the nonparametric estimators of the probability of cure and latency proposed and deeply studied in Xu and Peng (2014), López-Cheda et al. (2017a) and López-Cheda et al. (2017b). They are based on the Beran estimator of the conditional survival function (Beran, 1981) and are suitable when the covariate is one-dimensional.

In higher dimensions, the curse of dimensionality becomes a significant obstacle. We address this challenge using the single-index dimension reduction technique. A likelihood approach for estimating the conditional distribution or density under a single-index assumption and random right censoring was introduced in Strzalkowska-Kominiak and Cao (2013). While some recent work has considered hybrid approaches combining single-index models with Cox models in the presence of cure, we propose a full

single-index mixture cure model, applying this dimension reduction technique to both the cure rate and the latency.

The rest of the document is organized as follows. Section 2 describes a cardiotoxicity dataset related to breast cancer patients, providing the clinical context that motivates the methodological development. In Section 3, the single-index mixture cure model in the presence of a vector covariate is presented. This model extends the methodology developed in Xu and Peng (2014), López-Cheda et al. (2017a) and López-Cheda et al. (2017b) to covariates with dimension higher than one, and the single-index model in Strzalkowska-Kominiak and Cao (2013) to the presence of cure. A maximum likelihood estimator of the vector of parameters that determines the single-index is proposed, and its asymptotic normality is proved. Since the smoothing parameters involved in the model are incorporated in the optimization procedure, the problem of bandwidth selection, which is always a challenge in nonparametric or semiparametric estimation, is solved. In Section 4, single-index mixture cure models are extended to functional covariates, such as those arising from medical imaging tests. To make the optimization procedure feasible in this context, a basis representation of the functional covariate is considered. Finally, Section 5 presents the real data application, demonstrating the practical utility of the methodology and its clinical implications for personalized cardiotoxicity monitoring in breast cancer patients.

2. CARDIOTOXICITY DATASET

The dataset available in Piñeiro-Lamas et al. (2023a) and detailed in Piñeiro-Lamas et al. (2023b) is considered. It consists of information about 531 women diagnosed with breast cancer between 2007 and 2021 and treated with potentially cardiotoxic therapies at the University Hospital of A Coruña. A time variable, an uncensoring indicator and several baseline covariates are available for each woman. The time variable contains the time (in days) until cardiotoxicity onset or the lost of follow-up (the event which happens first). The uncensoring indicator denotes whether the patient did (54) or did not (477) experience cardiotoxicity during the follow-up period. Figure 1 shows the Kaplan-Meier estimation of the survival function. Given the presence of a plateau in the right tail and the medical evidence of a proportion of women not susceptible to cardiotoxicity, cure models seem plausible. The height of the plateau (around 0.7) estimates the probability of being cured (insusceptible to cardiotoxicity).

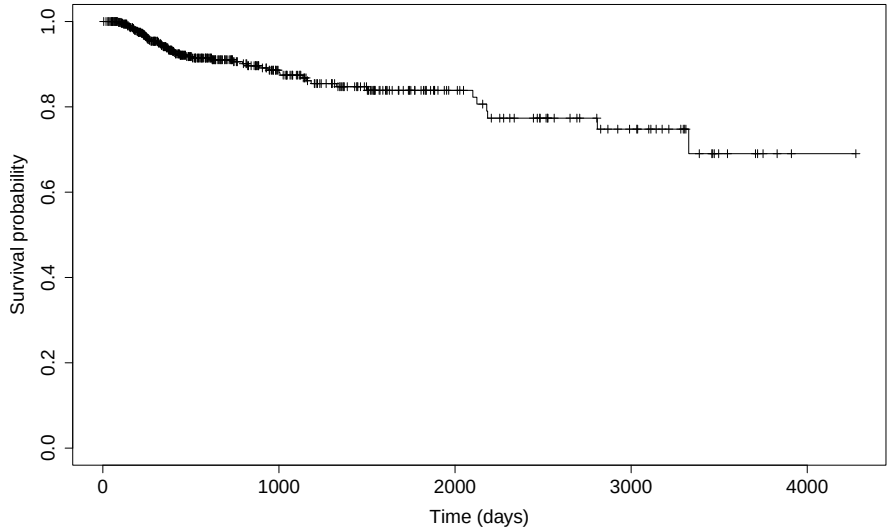

Figure 1: Kaplan-Meier estimation for the cardiotoxicity dataset.

For each woman, baseline covariates are available, including patient characteristics (age, height, weight), cardiac parameters, treatment information and comorbidity indicators. The dataset also includes Tissue Doppler Imaging (TDI) data, an echocardiographic technique that captures cardiac muscle velocity during contraction and relaxation, providing valuable information about the function and mechanics of the heart

(Figure 2). A preprocessing algorithm was developed to extract velocity functions from TDI images focusing on a single cardiac cycle (Piñeiro-Lamas et al., 2023b), obtaining the result in Figure 3. Each cycle contains three characteristic waves: S (systolic, above zero velocity) and E' and A' (diastolic, negative velocity). The difference between the cycles of both groups of patients (with and without cardiotoxicity) is more remarkable in the central zone. In particular, for the patients with cardiotoxicity detected during the follow-up period, the E' wave seems smoother and shifted to the right. This suggests the potential predictive role of the TDI in this context.

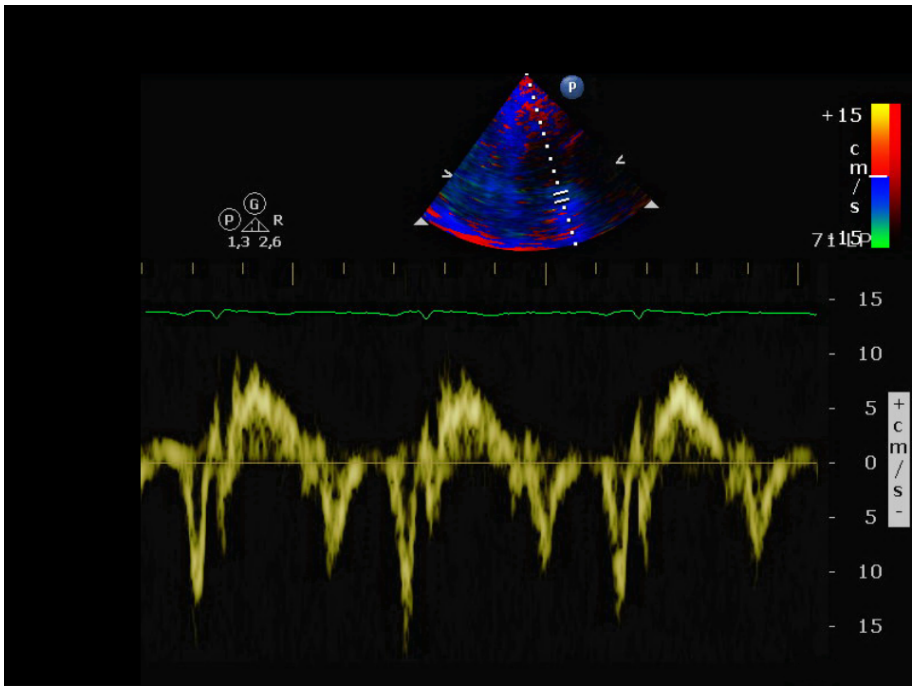

Figure 2: Example of a TDI.

Given the clinical importance of cardiotoxicity and the data available, several key questions arise. What is the probability of developing cardiotoxicity based on the patient characteristics? Which factors might increase or decrease this risk? For patients who do develop this complication, what is the distribution of the time until its onset? Does it appear early in treatment or as a long-term effect? Are there factors that might accelerate or delay its appearance? Answering these questions could enable more personalized treatment approaches, allowing clinicians to identify high-risk patients who may benefit from alternative therapies or more intensive monitoring. The statistical methodology presented in the following sections addresses these clinical questions by modeling both the probability of cardiotoxicity and the timing of its onset as functions of the available patient covariates.

3. SINGLE-INDEX MIXTURE CURE MODELS FOR VECTOR COVARIATES

To fix our notation, let Y denote the time variable of interest. In this work, we will assume that each individual is subject to random right censoring. For random right censored observations we know that the survival time, Y, is greater than a censoring time, C, but its exact value is unknown. Therefore, due to censoring, we do not always observe Y, but we observe instead (T, δ), being $T = \min(Y, C)$ the observed lifetime and $\delta = I(Y \leq C)$ the uncensoring indicator ($\delta = 1$ if the event occurs during the follow-up period, $\delta = 0$ otherwise). Moreover, let $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ be a vector covariate. In practice, the observations are $\{(\boldsymbol{X}_i, T_i, \delta_i), i = 1, \ldots, n\}$ (iid copies of the random vector $(\boldsymbol{X}, T, \delta)$). We assume that Y and C are conditionally independent given $\boldsymbol{X}$ ($Y \perp C|\boldsymbol{X}$). An important challenge in mixture cure models is the partial absence of the cure status. Let ν be a binary variable such that $\nu = 0$ if the subject is susceptible to the event and $\nu = 1$ if the subject is cured. While the uncensored observations are known to be uncured ($\nu = 0$), it is unknown if a censored individual will be eventually cured or not (ν is missing).

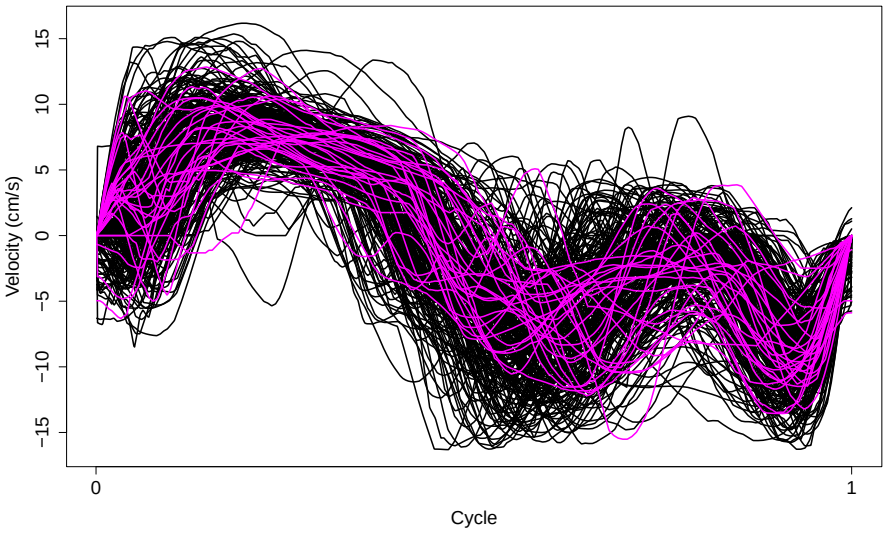

Figure 3: In magenta, the cardiac muscle velocity functions of the patients who experienced cardiotoxicity during the follow-up period. In black, the cardiac muscle velocity functions of the patients who did not experience the side effect.

Mixture cure models were proposed by Boag (1949) and, for a vector covariate $\boldsymbol{X}$, they can be expressed as follows:

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

where $\boldsymbol{x}$ denotes any possible value for the covariate $\boldsymbol{X}$, $S(t|\boldsymbol{x}) = P(Y > t|\boldsymbol{X} = \boldsymbol{x})$ is the conditional survival function of all the individuals (cured and uncured), $1 - p(\boldsymbol{x}) = 1 - P(\nu = 0|\boldsymbol{X} = \boldsymbol{x})$ is the probability of cure and $S_0(t|\boldsymbol{x}) = P(Y > t|\boldsymbol{X} = \boldsymbol{x}, \nu = 0)$ is the latency.

The probability of cure and the latency can be estimated parametrically, semiparametrically or nonparametrically. Xu and Peng (2014) first introduced a nonparametric cure probability estimator, which was extensively studied by López-Cheda et al. (2017a). López-Cheda et al. (2017b) proposed a nonparametric estimator for the latency. Both estimators work in the presence of a continuous univariate covariate. When dealing with a covariate vector of dimension greater than one, the curse of dimensionality can be problematic. To address this, we propose single-index mixture cure models that can handle vector covariates. Specifically, a single-index approach is considered for both the probability of cure and the latency. Under the single-index assumption, we assume that the functions $p(\boldsymbol{x})$ and $S_0(t|\boldsymbol{x})$ depend on the covariate vector only through a linear combination. This allows us to rewrite these functions as $p_{SI}(\boldsymbol{\theta_0}'\boldsymbol{x})$ and $S_{0,SI}(t|\boldsymbol{\theta_0}'\boldsymbol{x})$, where the subscript SI denotes the single-index versions that depend on the index $\boldsymbol{\theta_0}'\boldsymbol{x}$ rather than the full vector $\boldsymbol{x}$, leading to a single-index model for the survival function:

$$S(t|\boldsymbol{X} = \boldsymbol{x}) = 1 - p(\boldsymbol{x}) + p(\boldsymbol{x})S_0(t|\boldsymbol{x}) = 1 - p_{SI}(\boldsymbol{\theta_0}'\boldsymbol{x}) + p_{SI}(\boldsymbol{\theta_0}'\boldsymbol{x})S_{0,SI}(t|\boldsymbol{\theta_0}'\boldsymbol{x}) = S_{SI}(t|\boldsymbol{\theta_0}'\boldsymbol{X} = \boldsymbol{\theta_0}'\boldsymbol{x}),$$

where $\boldsymbol{v}'\boldsymbol{w}$ denotes the inner product for $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^d$, $S : \mathbb{R}^{d+1} \to [0,1]$, $p : \mathbb{R}^d \to [0,1]$, $S_0 : \mathbb{R}^{d+1} \to [0,1]$, $p_{SI} : \mathbb{R} \to [0,1]$, $S_{0,SI} : \mathbb{R}^2 \to [0,1]$ and $S_{SI} : \mathbb{R}^2 \to [0,1]$ is a function such that $S_{SI}(t|u) = 1 - p_{SI}(u) + p_{SI}(u)S_{0,SI}(t|u)$. For every u, $S_{0,SI}(t|u)$ is a proper survival function in t. Note that the single-index approach summarizes the vector covariate, $\boldsymbol{X}$, into one single score, often referred as the index. The cost of reducing the dimension is the estimation of $\boldsymbol{\theta_0} \in \mathbb{R}^d$, which is an unknown d-dimensional vector of parameters. To ensure identifiability, its first component is set to 1. The vector of parameters $\boldsymbol{\theta_0}$ can be estimated by maximum likelihood, where the logarithm of the theoretical conditional likelihood function can be expressed in terms of the sample $\{(\boldsymbol{X}_i, T_i, \delta_i), i = 1, \ldots, n\}$:

$$\tilde{\ell}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\delta_i \log\Big(p_{SI}(\boldsymbol{\theta}'\boldsymbol{X}_i)\Big) + \delta_i \log\Big(f_{0,SI}(T_i|\boldsymbol{\theta}'\boldsymbol{X}_i)\Big) + (1-\delta_i)\log\Big(1 - p_{SI}(\boldsymbol{\theta}'\boldsymbol{X}_i) + p_{SI}(\boldsymbol{\theta}'\boldsymbol{X}_i)S_{0,SI}(T_i|\boldsymbol{\theta}'\boldsymbol{X}_i)\Big)\Bigg], \quad (1)$$

where $f_{0,SI}$ is the density function of $Y|\boldsymbol{\theta_0}'\boldsymbol{X} = \boldsymbol{\theta_0}'\boldsymbol{x}, \nu = 0$. If p_{SI}, $S_{0,SI}$ and $f_{0,SI}$ were known, then $\boldsymbol{\theta_0}$ could be estimated by maximizing Equation (1) in $\boldsymbol{\theta}$. However, since these functions are unknown in practice, they need to be estimated. The logarithm of the estimated conditional likelihood function is obtained by replacing the unknown functions with their nonparametric leave-one-out cross-validation estimators. This function depends on $\boldsymbol{\theta}$ and four bandwidths:

$$\hat{\ell}_n(\boldsymbol{\theta}, h_1, h_2, h_3, h_4) = \sum_{i=1}^{n} \Bigg[\delta_i \log\left(\hat{p}^{-i}_{SI,h_1}(\boldsymbol{\theta}'\boldsymbol{X}_i)\right) + \delta_i \log\left(\hat{f}^{-i}_{0,SI,h_3,h_4}(T_i|\boldsymbol{\theta}'\boldsymbol{X}_i)\right) + (1-\delta_i)\log\left(1 - \hat{p}^{-i}_{SI,h_1}(\boldsymbol{\theta}'\boldsymbol{X}_i) + \hat{p}^{-i}_{SI,h_1}(\boldsymbol{\theta}'\boldsymbol{X}_i)\hat{S}^{-i}_{0,SI,h_2}(T_i|\boldsymbol{\theta}'\boldsymbol{X}_i)\right)\Bigg]. \quad (2)$$

To estimate the cure rate, since the covariate has dimension one, the estimator proposed by Xu and Peng (2014) can be used:

$$1 - \hat{p}_{h_1}(u) = \hat{S}_{h_1}(T^1_{\max}|u),$$

where $\hat{S}_{h_1}(t|u)$ is the Beran estimator of $S(t|u)$, $T^1_{\max} = \max_{i/\delta_i=1}(T_i)$ is the largest uncensored survival time and h_1 is a smoothing parameter. Regarding the latency, the nonparametric estimator proposed by López-Cheda et al. (2017b) can be used:

$$\hat{S}_{0,h_2}(t|u) = \frac{\hat{S}_{h_2}(t|u) - (1 - \hat{p}_{h_2}(u))}{\hat{p}_{h_2}(u)},$$

where $1 - \hat{p}_{h_2}(u)$ is the estimator of the cure probability by Xu and Peng (2014) and h_2 is a smoothing parameter. To estimate $f_{0,SI}$, we propose a kernel type estimator by generalizing the Parzen-Rosenblatt one to the presence of censoring, cure and covariates:

$$\hat{f}_{0,h_3,h_4}(t|u) = \sum_{i=1}^{n} K_{h_3}(t - T_i)\Big(\hat{F}_{0,h_4}(T_i|u) - \hat{F}^{-}_{0,h_4}(T_i|u)\Big),$$

where $K_{h_3}(\cdot) = \frac{1}{h_3}K\left(\frac{\cdot}{h_3}\right)$ is a rescaled kernel function, $1 - \hat{F}_{0,h_4}(\cdot|\cdot) = \hat{S}_{0,h_4}(\cdot|\cdot)$ is the nonparametric latency estimator in López-Cheda et al. (2017b), and h_3, h_4 are bandwidths. An almost sure representation for this estimator was derived (Piñeiro-Lamas, 2024). The estimator is implemented in the `cd.uncured` function of the `sicure` package, available on CRAN (Piñeiro-Lamas, López-Cheda and Cao, 2025). Note that, in Equation (2), we have considered the leave-one-out cross-validation estimators of the three involved functions, where the $-i$ in the superscript denotes that the ith observation is not considered to compute the estimators. This way, the sample used to construct the estimator and the value where it is evaluated are independent.

Four bandwidths are required for the estimation procedure: h_1 to estimate p_{SI}, h_2 to estimate $S_{0,SI}$, and h_3 and h_4 to estimate $f_{0,SI}$. The bandwidths h_1, h_2 and h_4 are used to smooth the covariate, while h_3 smooths the time variable. It is noteworthy that global bandwidths are taken into account, because choosing $4n$ local bandwidths would be very costly from a computational point of view.

The maximum likelihood estimator of $\boldsymbol{\theta_0}$ and the selected bandwidths can be obtained by maximizing the estimated log-likelihood function:

$$(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}, \hat{h}_1, \hat{h}_2, \hat{h}_3, \hat{h}_4) = \arg\max_{\boldsymbol{\theta}, h_1, h_2, h_3, h_4} \hat{\ell}_n(\boldsymbol{\theta}, h_1, h_2, h_3, h_4).$$

The optimization procedure is implemented in the `sicure.v` function, which is part of the `sicure` package. Since $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ and the selected smoothing parameters cannot be obtained explicitly, numerical optimization methods with a multi-start approach are used to address potential local maxima in the objective function. While incorporating bandwidth selection into the optimization procedure addresses the challenge of choosing appropriate smoothing parameters, it also introduces additional computational complexity.

The asymptotic normality of $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ was proved (Piñeiro-Lamas, 2024). Additionally, in order to empirically assess its performance, a simulation study was carried out.

4. SINGLE-INDEX MIXTURE CURE MODELS FOR FUNCTIONAL COVARIATES

Consider a functional covariate $\mathfrak{X} \in \mathcal{F}$, where $\mathcal{F}$ is a subset of the space of continuous functions on a compact interval τ. We consider the inner product $\langle f, g\rangle = \int_\tau f(u)g(u)du$, with associated norm $\|f\| = \sqrt{\langle f, f\rangle} = \sqrt{\int_\tau f^2(u)du}$. Furthermore, we assume that functions in $\mathcal{F}$ are uniformly bounded with respect to this norm: there exists $M > 0$ such that $\|f\| \leq M$ for all $f \in \mathcal{F}$. These conditions ensure desirable analytical properties. Then, mixture cure models can be written as follows:

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

Using a single-index approach, the functional covariate effect can be summarized into the inner product $\langle \theta_0, \mathfrak{X}\rangle = \int_\tau \theta_0(u)\mathfrak{X}(u)du$, for some function $\theta_0 \in \Theta$, where Θ is a bounded set of parameter functions defined on τ. Let us consider a single-index model for the cure rate and the latency. That is, let us assume that $\exists\theta_0 \in \Theta$ such that $1 - p(\boldsymbol{x}) = 1 - p_{SI}(\langle\theta_0, \boldsymbol{x}\rangle)$ and $S_0(t|\boldsymbol{x}) = S_{0,SI}(t|\langle\theta_0, \boldsymbol{x}\rangle)$, where $p : \mathcal{F} \to [0, 1]$, $p_{SI} : \mathbb{R} \to [0, 1]$, $S_0 : \mathbb{R} \times \mathcal{F} \to [0, 1]$ and $S_{0,SI} : \mathbb{R}^2 \to [0, 1]$. Thus, the single-index mixture cure model with a functional covariate is

$$S(t|\mathfrak{X} = x) = S_{SI}(t|\langle\theta_0, \mathfrak{X}\rangle = \langle\theta_0, x\rangle),$$

where $S : \mathbb{R} \times \mathcal{F} \to [0, 1]$ and $S_{SI} : \mathbb{R}^2 \to [0, 1]$.

The logarithm of the estimated conditional likelihood function is analogous to the one in the vector case (Equation (2)), replacing $\boldsymbol{\theta}'\boldsymbol{X}$ with $\langle\theta, \mathfrak{X}\rangle$. To make optimization feasible, the infinite dimension of the covariate is reduced via a basis representation. Using an orthonormal basis $\{\psi_k; k = 1, 2, \dots\}$, $\mathfrak{X}$ can be decomposed as follows:

$$\mathfrak{X}(t) = \sum_{k=1}^{\infty} \xi_k\psi_k(t) \approx \sum_{k=1}^{K} \xi_k\psi_k(t),$$

being K a sufficiently large integer that should be data-driven rather than arbitrary. This reduces each subject's information to a vector of scores $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_K)$, and the model then simplifies to

$$S_{SI}(t|\boldsymbol{\beta_0}'\boldsymbol{\xi}) = 1 - p_{SI}(\boldsymbol{\beta_0}'\boldsymbol{\xi}) + p_{SI}(\boldsymbol{\beta_0}'\boldsymbol{\xi})S_{0,SI}(t|\boldsymbol{\beta_0}'\boldsymbol{\xi}),$$

where $\boldsymbol{\beta_0} \in \mathbb{R}^K$ is a vector of parameters (analogous to $\boldsymbol{\theta_0}$ in the vector case).

By reducing the functional case to the vector one through a basis representation, we unlock the possibility of combining both types of variables in the same model. Specifically, we can create a combined vector $\boldsymbol{X_c} = (X_1, \dots, X_d, \xi_1, \dots, \xi_K)$ of dimension $d + K$, and apply the following single-index mixture cure model:

$$S_{SI}(t|\boldsymbol{\gamma_0}'\boldsymbol{X_c}) = 1 - p_{SI}(\boldsymbol{\gamma_0}'\boldsymbol{X_c}) + p_{SI}(\boldsymbol{\gamma_0}'\boldsymbol{X_c})S_{0,SI}(t|\boldsymbol{\gamma_0}'\boldsymbol{X_c}),$$

where $\boldsymbol{\gamma_0} \in \mathbb{R}^{d+K}$ is a vector of parameters (analogous to $\boldsymbol{\theta_0}$ in the vector case and $\boldsymbol{\beta_0}$ in the functional case).

The `sicure.f` function estimates the parameter vector $\boldsymbol{\beta_0}$ in a single-index mixture cure model for functional covariates, using Functional Principal Components Analysis (FPCA) for dimension reduction. The `sicure.vf` function estimates the parameter vector $\boldsymbol{\gamma_0}$ when both vector and functional covariates are present in the model.

5. DATA APPLICATION

The methodology presented in Sections 3 and 4 is applied to the cardiotoxicity dataset described in Section 2, demonstrating its practical utility for personalized risk assessment in clinical settings.

We first consider the vector covariate $\boldsymbol{X}$ = (heart rate, age, height, weight, LVEF, PWT, LAd, LVDd, LVSd) $\in \mathbb{R}^9$, where LVEF is the Left Ventricular Ejection Fraction, PWT the Posterior Wall Thickness, LAd the Left Atrial diameter and LVDd (LVSd) the Left Ventricular Diastolic (Systolic) diameter. Since $d = 9$, the optimization problem consists on the estimation of 12 parameters ($\theta_{02}, \dots, \theta_{09}$ and four bandwidths). A single-index mixture cure model is fitted to estimate $(1, \theta_{02}, \theta_{03}, \theta_{04}, \theta_{05}, \theta_{06}, \theta_{07}, \theta_{08}, \theta_{09})$ using the `sicure.v`, resulting in the following estimated single-index variable:

$$\begin{aligned}\text{heart rate} + 1.3742 \times \text{age} - 0.5597 \times \text{weight} + 0.0390 \times \text{height} - 0.3815 \times \text{LVEF}\\ - 0.0200 \times \text{PWT} - 0.1728 \times \text{LAd} - 0.0954 \times \text{LVDd} + 0.4931 \times \text{LVSd}.\end{aligned}$$

The largest estimated value (in absolute value) is the one associated to age, what indicates that it is the most important covariate. Also note that heart rate, age, height and LVSd have positive coefficients, while the remaining ones are negative. Besides, the selected bandwidths are $\hat{h}_1 = 0.3899$, $\hat{h}_2 = 0.5427$, $\hat{h}_3 = 144.3880$ and $\hat{h}_4 = 0.4345$. Although $\hat{h}_3$ seems considerably large compared with the others, it is important to highlight that it is used for smoothing the time variable, which is measured in days.

Figure 4 shows the estimated probability of cardiotoxicity (left) and latency curves (right) as functions of the single-index variable. Higher index values correspond to increased cardiotoxicity probability and earlier onset times, suggesting the variable's role as a risk factor. While the effect on the cure rate is not statistically significant ($p-$ value $= 0.26$), the clinical patterns suggest potential utility for risk stratification. Based on risk factors, doctors might develop personalized monitoring plans and recommend preventive measures to improve the patient's overall health outcomes.

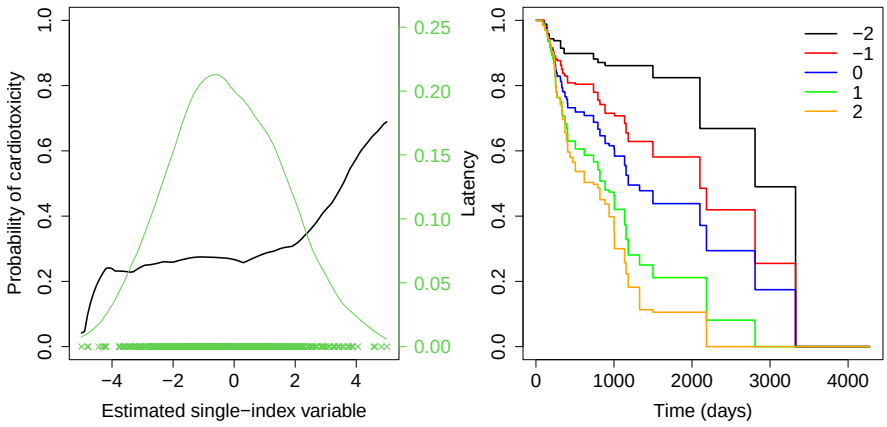

Figure 4: Left: Estimated probability of cardiotoxicity (black line) depending on the estimated single-index variable, considering the vector variable related to the cardiotoxicity dataset. The grey line represents the Parzen-Rosenblatt kernel density estimation of the estimated single-index variable, using Sheather and Jones' plug-in bandwidth. Right: Estimated latency depending on five values of the estimated single-index variable, considering the vector variable related to the cardiotoxicity dataset.

We next analyze the TDI data as functional covariates (Figure 3). To fit a single-index mixture cure model the `sicure.f` function is used, which internally reduces the dimensionality using FPCA with $K = 8$ components, explaining 90% of the data variability. This process transforms each patient's function into an 8-dimensional vector of scores. As in the vector case, higher values of the estimated single-index variable are associated with increased cardiotoxicity probabilities and earlier onset times (Figure 5), while the effect on the cure rate is not significant ($p-$ value $= 0.76$).

The most clinically relevant results emerge when combining both vector and functional covariates. This creates a 17-dimensional combined covariate vector (9 clinical variables + 8 functional scores). Using the `sicure.vf` function, we found that higher values of the estimated single-index variable are associated with higher probabilities of cardiotoxicity. In this case, the effect of the covariate on the cure rate was found to be significant ($p-$ value $= 0.047$). Based on these results, the estimated single-index variable may be used as a cardiotoxicity risk factor. This means that all components with positive estimated coefficients (as the age) may raise the risk of cardiotoxicity, while those with negative estimated coefficients (as the LVEF) may have a protective impact. These results are consistent with the cardio-oncology literature. However, the single-index variable seems to have no effect on the latency.

Figure 7 shows the estimated probability of cardiotoxicity for each patient, depending on her estimated single-index variable. Patients who developed cardiotoxicity have a higher mean estimated probability (0.38) compared to those who did not experience this cardiac problem (0.20). Note that all the patients with an estimated value of 0 did not experience cardiotoxicity during the follow-up period. Furthermore, among the patients with an estimated probability higher than 0.75, almost half presented cardiotoxicity during follow-up. Recall that censored observations (the ones in black) can correspond to uncured pa-

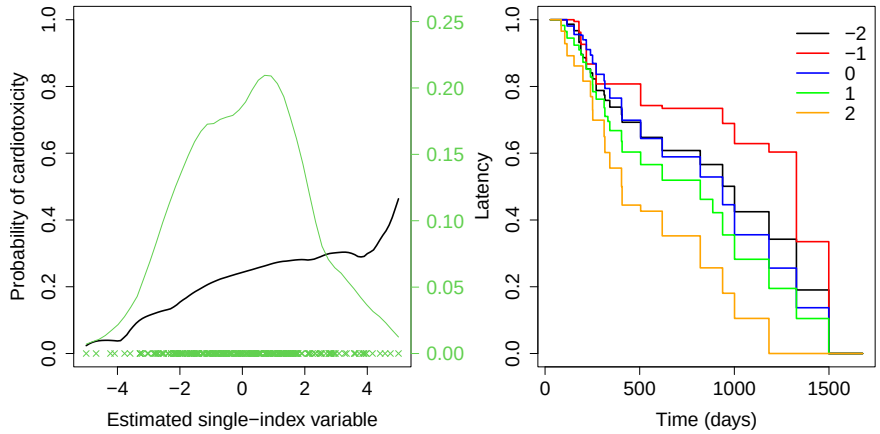

Figure 5: Left: Estimated probability of cardiotoxicity (black line) depending on the estimated single-index variable, considering the functional variable related to the cardiotoxicity dataset. The green line represents the Parzen-Rosenblatt kernel density estimation of the estimated single-index variable, using Sheather and Jones' plug-in bandwidth. Right: Estimated latency depending on five values of the estimated single-index variable, considering the functional variable related to the cardiotoxicity dataset.

tients. That is, if the follow-up time were longer, it is possible that the black points with high estimated probabilities would be actually magenta.

The computational complexity of fitting single-index mixture cure models deserves consideration. Using a computer with AMD Ryzen 9 5950X 16-Core Processor, 3.40 GHz, 128 GB RAM, the estimation procedure required reasonable computational times across all scenarios: the vector covariate case was the fastest, while the functional covariate alone and the combined vector-functional approach both required approximately 10 minutes, using multi-start Nelder-Mead optimization with 10 different initial values. This reasonable computational time makes the method feasible for clinical applications.

6. DISCUSSION

Single-index mixture cure models provide a powerful framework for analyzing survival data with a cure fraction, capable of handling both vector and functional covariates. The semiparametric nature of these models, combining nonparametric estimation of the cure probability and latency functions with parametric dimension reduction via the single-index approach, enables flexible modeling of complex relationships between covariates and survival outcomes while avoiding the curse of dimensionality. The sicure `R` package makes these models accessible to researchers.

From a clinical perspective, the application to cardiotoxicity demonstrates the practical potential of these methods. By simultaneously modeling both the probability of developing cardiotoxicity and the timing of its onset, the approach addresses a gap in current clinical practice where the temporal aspect is often overlooked. This enables more informed clinical decision-making, personalized monitoring approaches based on individual patient characteristics, and more efficient management of medical appointments. The clinical utility of this methodology has been further confirmed by Piñeiro-Lamas et al. (2025), who used insights from the single-index modeling of TDI functional data to develop simplified clinical predictors based on key E' wave characteristics.

There are a few limitations in our work. Since the same vector of parameters, $\boldsymbol{\theta_0}$, is considered in the single-index model for the cure rate and the latency, the vector covariate is assumed to have the same effect on both functions. In future work, we plan to make our models more flexible by considering that the survival function depends on $\boldsymbol{X}$ through two indices. Additionally, developing an appropriate goodness-of-fit test for the single-index model is crucial. This test would serve as a valuable tool to assess the adequacy of these models in capturing the underlying data structure.

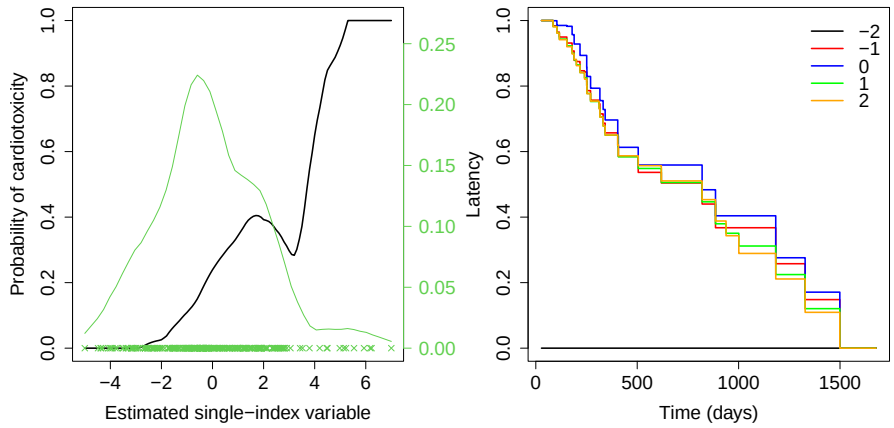

Figure 6: Left: Estimated probability of cardiotoxicity (black line) depending on the estimated single-index variable, considering both the vector and the functional variables related to the cardiotoxicity dataset. The green line represents the Parzen-Rosenblatt kernel density estimation of the estimated single-index variable, using Sheather and Jones' plug-in bandwidth. Right: Estimated latency depending on five values of the estimated single-index variable, considering both the vector and the functional variables related to the cardiotoxicity dataset.

Despite these limitations, the methodology presented here constitutes a novel contribution to evidence-based medicine and personalized treatment in cardio-oncology, providing clinicians with enhanced tools for risk stratification and patient monitoring optimization.

REFERENCES

Beran, R. (1981). Nonparametric regression with randomly censored survival data (Tech. Rep.). Berkeley: University of California, Berkeley.

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

López-Cheda, A., Cao, R., Jácome, M. A. and Van Keilegom, I. (2017a). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. Computational Statistics & Data Analysis, 105 , 144–165.

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

Piñeiro-Lamas, B., López-Cheda, A., Cao, R., Ramos-Alonso, L., González-Barbeito, G., Barbeito-Caamaño, C. and Bouzas-Mosquera, A. (2023a). BC_cardiotox: A cardiotoxicity dataset for breast cancer patients.

Piñeiro-Lamas, B., López-Cheda, A., Cao, R., Ramos-Alonso, L., González-Barbeito, G., Barbeito-Caamaño, C. and Bouzas-Mosquera, A. (2023b). A cardiotoxicity dataset for breast cancer patients. Scientific Data, 10, 527.

Piñeiro-Lamas, B. (2024). High dimensional single-index mixture cure models, PhD thesis, Universidade da Coruña. Available at https://ruc.udc.es/dspace/handle/2183/37035.

Piñeiro-Lamas, B., López-Cheda, A. and Cao, R. (2025). sicure: Single-Index Mixture Cure Models. http://CRAN.R-project.org/package=sicure. R package version 0.1.1.

Piñeiro-Lamas, B., López-Cheda, A., Cao, R., Lesta-Mellid, R., Bouzas-Mosquera, A. and Barbeito-Caamaño, C. (2025). Tissue Doppler Imaging as a predictor of therapy-related cardiac dysfunction in breast cancer patients. The International Journal of Cardiovascular Imaging, 41(8), 1505–1512.

Strzalkowska-Kominiak, E. and Cao, R. (2013). Maximum likelihood estimation for conditional distribution single-index models under censoring. Journal of Multivariate Analysis, 114, 74–98.

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

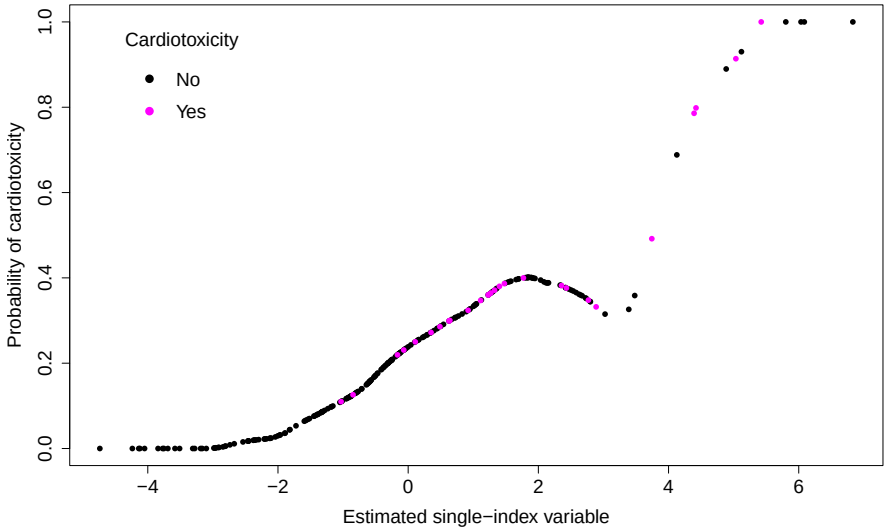

Figure 7: Estimated probability of cardiotoxicity for each patient depending on its estimated single-index variable, considering both the vector and the functional variables related to the cardiotoxicity dataset.