Presmoothed Latency Estimation of Cardiotoxicity in Breast Cancer Patients

Samuel Saavedra, Ana López-Cheda, and María Amalia Jácome

Grupo de Modelización, Optimización e Inferencia Estatística (MODES), Departamento de Matemáticas, CITIC, Facultade de Informática, Universidade da Coruña, A Coruña, Spain
Grupo de Modelización, Optimización e Inferencia Estatística (MODES), Departamento de Matemáticas, CITIC, Facultade de Ciencias, Universidade da Coruña, A Coruña, Spain
Correspondence: samuel.saavedra@udc.gal

DOI: *https://doi.org/10.17979/spu.23.c12*

Abstract: Breast cancer remains the leading cancer diagnosis among women. While treatment advances have significantly improved patient outcomes, some of the adverse effects associated with these therapies have been linked to cardiotoxicity, potentially compromising heart function. Understanding the time when such toxicity emerges is crucial to optimizing patient care and monitoring strategies. This study aims to estimate the time until cardiotoxicity appears, known as latency. The idea is to apply an innovative adaptation of traditional cure models considering a presmoothing method.

1 Cardiotoxicity in breast cancer patients

Cancer is a group of diseases characterized by the uncontrolled growth and spread of abnormal cells in the body. Among women, breast cancer is the most common cancer worldwide. In 2020, it is estimated that there were 2.3 million women were diagnosed and nearly 685,000 died from the disease. In 2040, the estimation is to increase this number up to over 3 million new cases and 1 million deaths per year Arnold et al. (2022).

A significant subset of these diagnoses, approximately 20%, are classified as HER2-positive (HER2+), characterized by the overexpression of the human epidermal growth factor receptor 2 protein. These cancers are often characterized by faster growth and spread than their HER2-negative counterparts (Owens et al., 2004). However, the development of monoclonal antibodies like trastuzumab, which specifically targets the HER2 protein, has dramatically improved both overall survival and progression-free survival for these patients (Mazzotta et al., 2019). Current clinical guidelines strongly advocate for anti-HER2 therapies, often administered for a duration of one year in the adjuvant setting, frequently combined with other antineoplastic drugs, including taxanes and sometimes anthracyclines (Bešlija et al., 2021).

The success of HER2-targeted therapy is shadowed by the risk of cardiovascular adverse events, collectively termed cardiotoxicity. In tumour cells, trastuzumab binds to the subdomain IV of the extracellular domain of HER2, blocking HER2 cleavage, stimulating antibody-dependent cellular cytotoxicity, and inhibiting HER2-mediated mitogenic signaling (Florido et al., 2017). Although trastuzumab-mediated cardiotoxicity is generally considered Type II (often reversible), its co-administration with traditional chemotherapeutics, particularly anthracyclines, significantly magnifies the risk. The concurrent or sequential use of trastuzumab and doxorubicin, an anthracycline, has been shown to enhance the downregulation of key proteins,

increase apoptosis, and promote reactive oxidative species production in human cardiomyocytes (Mohan et al., 2018).

The manifestation of this toxicity is generally defined as Cancer Therapy-Related Cardiac Dysfunction (CTRCD). Clinical consensus relies primarily on echocardiographic assessment of the Left Ventricular Ejection Fraction (LVEF), which measures the percentage of blood leaving the heart during contraction. CTRCD is diagnosed when two criteria are met: the LVEF falls below 50% and there is a decline of at least 10 percentage points from the patient's baseline LVEF measured before treatment initiation (Piñeiro-Lamas et al., 2023).

Identifying patients at risk remains a major challenge. Known clinical factors contributing to CTRCD risk include demographic features such as advanced age, pre-existing cardiovascular conditions like hypertension (HTA) and diabetes mellitus (DM), and treatment history, specifically previous exposure to anthracyclines (ACprev) (Piñeiro-Lamas et al., 2023). However, these established risk factors are insufficient for accurate prediction of which patients will develop the dysfunction. This gap necessitates the exploration of more granular, pre-treatment biomarkers. The potential utility of functional variables, such as those derived from baseline Tissue Doppler Imaging (TDI), which captures the velocity of myocardial contraction and relaxation, is substantial but remains largely unexamined in the existing literature (Kaboré et al., 2023).

2 Presmoothed Mixture Cure Models

Survival analysis is a branch of statistics that focuses on studying the time until an event of interest occurs. The event could be death, relapse of an illness, equipment failure, loss of customers, or any other well-defined outcome, such as cardiotoxicity. Classical survival analysis assumes that all the individuals will suffer the event of interest, but in some cases this will not happen. Taking this consideration into account, cure models arise. One of the most common cure models is Mixture Cure Models (MCM) (Boag, 1949). This framework proposes that the population consists of two distinct sub-groups: the susceptible group, who will eventually experience the event, and the cured, who will not.

We start by defining the survival function. Let Y be the lifetime of interest, and denote by $S(t) = P(Y > t)$ the survival function.

In the MCM framework, the survival function can be written as:

$$S(t) = (1 - p) + pS_0(t),$$

where $p = P(Y < \infty)$ is the probability of suffering the event of interest, $1 - p$ is the cure rate and $S_0(t) = P(Y > t|Y < \infty)$ is the survival function of the susceptible (non-cured) individuals, also known as latency.

In practice, we do not observe Y directly, but instead we observe (T, δ), where $T = \min(Y, C)$ is the observed time and C is the censoring variable. Finally, $\delta = I(Y \leqslant C)$ is the uncensoring indicator.

In this work we propose to improve the nonparametric MCM submitted by López-Cheda et al. (2017) with the application of a presmoothing method. In order to achieve this goal, we propose to replace the uncensoring indicator by an estimation of the $q(t)$ function. The estimation of the latency can be obtained as:

$$\hat{S}_0 = \frac{\hat{S}(t) - (1 - \hat{p})}{\hat{p}},$$

where $\hat{S}(t)$ is Kaplan and Meier (1958) (KM) estimator, also known as product-limit estimator, and $1 - \hat{p} = \hat{S}(T^1_{\max})$ is the estimation of the probability of suffering the event of interest.

Presmoothing is a preliminary step that requires to estimate the conditional probability of non-censoring beforehand, denoted by $q(t) = E(\delta|T = t)$. That is, instead of considering the

uncensoring indicator (which can take values 0 or 1), we consider a continuous function that can take values between 0 and 1. This leads to a smoother and more efficient estimation of the survival function.

Analogously to the classical approach, one can define the presmoothed latency estimator as:

$$\hat{S}^P_{0,b}(t) = \frac{\hat{S}^P_b(t) - (1 - \hat{p}^P_b)}{\hat{p}^P_b},$$

where $1 - \hat{p}^P_b = \hat{S}^P_b(T^1_{\max})$.

The KM estimator, in the context of presmoothing (Cao et al., 2005) is defined as:

$$\hat{S}^P_b(t) = \prod_{i:T_{(i)} \leqslant t}^{n} \left(1 - \frac{\hat{q}_b(T_{(i)})}{n - i + 1}\right).$$

The function $\hat{q}_b(t)$ (Nadaraya, 1964; Watson, 1964) is estimated as:

$$\hat{q}_b(t) = \frac{\frac{1}{n}\sum_{i=1}^{n} K_b(t - T_i)\delta_i}{\frac{1}{n}\sum_{i=1}^{n} K_b(t - T_i)},$$

where $K_b(\cdot)$ is the kernel function and b is the presmoothing bandwidth. When $b = 0$, $\hat{q}_b(T_i) = \delta_i$ thus behaving as the classical estimator, meanwhile when $b > 0$ the $\hat{q}_b(t)$ adopts any value between 0 and 1.

The bootstrap bandwidth selector for the presmoothing b bandwidth is obtained by minimizing a bootstrap estimate of the mean integrated squared error (MISE).

3 Real data application

3.1 Dataset

The dataset analyzed in this study was compiled by Piñeiro-Lamas et al. (2023) and comprises information from 531 patients diagnosed with HER2-positive breast cancer between 2007 and 2021 at the University Hospital Complex of A Coruña (CHUAC). The dataset integrates heterogeneous data types, including numerical, binary, and imaging variables.

The event of interest is the onset of cardiotoxicity. Therefore, the time variable contains the time (in days) until the appearance of this side effect for patients who suffered it, and the length of the follow-up period for the individuals who did not experience the event.

The analysis was restricted to binary variables with at least 30 patients in each category. Each variable was coded using two values: 0 indicating absence of the characteristic and 1 indicating presence. The variables, listed alphabetically, were: *AC* (use of anthracyclines), *ACprev* (prior use of anthracyclines), *antiHER*2 (use of anti-HER2 therapies), *antiHER*2*prev* (prior use of anti-HER2 therapies), *DL* (dyslipidemia), *exsmoker* (former smoker), *HTA* (hypertension), *RTprev* (prior chest radiotherapy), and *smoker* (current smoker), for a total of 9 variables.

The Table 1 shows the different samples in each class:

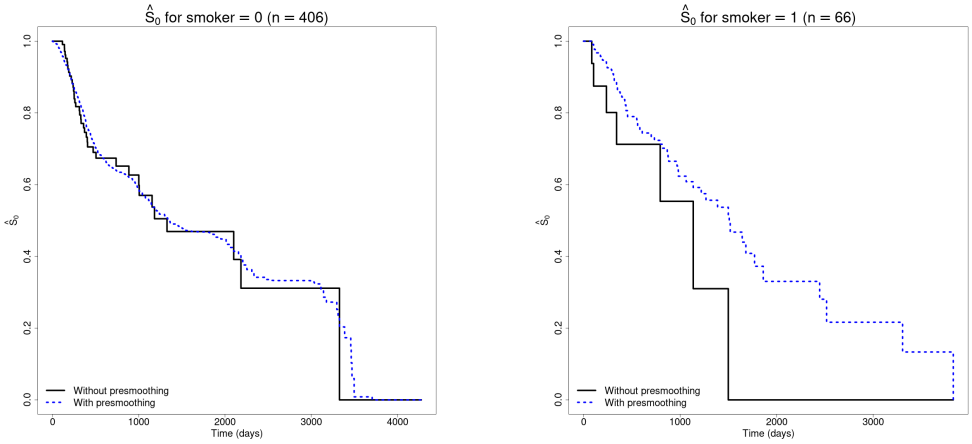
The objective is to compare the estimation of the latency with two different methods: without presmoothing and presmoothing with a bootstrap bandwidth selector for each subset of the presence or absence of the variable.

3.2 Results

The results have been processed using the R programming language (R Core Team, 2025) in its 4.5.1 version for the Debian GNU/Linux 12 (bookworm) operative system. The presmoothed estimator of the survival function, $S(t)$, was obtained obtained using the *presmooth* function of the `survPresmooth` package (López-de Ullibarri and Jácome, 2013), where a bootstrap bandwidth selector is available. The resample size to obtain the b bandwidth is $B = 1000$.

Table 1: Sample size of each variable

Variable	Absence	Presence
AC	149	323
ACprev	417	55
antiHER2	136	336
antiHER2prev	440	32
DL	382	90
exsmoker	406	66
HTA	373	99
RTprev	409	63
smoker	406	66

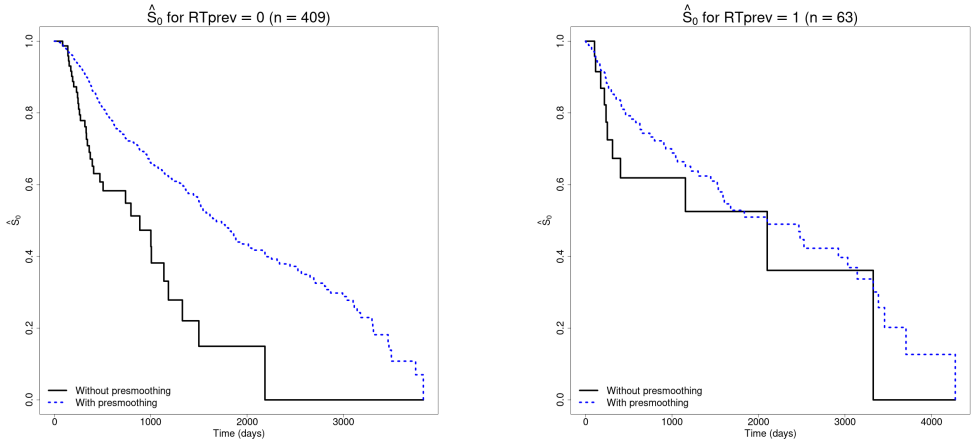


(a) Non-smokers (b) Smokers

Figure 1: Comparison of the classical latency estimator (**black** solid line) and the presmoothed latency estimator with bootstrap bandwidth selector (**blue** dashed line) separated by non-smoker (1a) and smoker (1b) patients.

Figure 1 displays the estimated latency survival functions $\hat{S}_0(t)$ for the subgroup of non-smoker patients, comparing the classical KM latency estimator. The presmoothed estimator exhibits smoother transitions, particularly in regions with sparse events, while maintaining close agreement with the non-presmoothed curve. This reduction in abrupt fluctuations highlights the advantage of presmoothing in stabilizing the estimation of latency, leading to more reliable inference without altering the overall survival trend. These results illustrate that presmoothing provides a practical refinement to traditional methods, especially when dealing with censored data in small to moderate sample sizes as it shows in the Figure 1b where $n = 66$. The presmoothed estimator reveals that smokers experience cardiotoxicity later than the classic estimator.

Figure 2, in contrast to Figure 1, shows a greater difference between the latency estimations of the greater sample size (2a), in counterpoint to the previous figure. Therefore, the sample size seems to not be determinant on how close or far is the difference between the two latency estimations, meanwhile it is determinant on how the latency is smoothed. The smaller sample sizes exhibit greater steps in the latency function than the bigger sample sizes, even in the presmoothed estimator. The presmoothed estimator indicates that the patients that did not received



(a) Non prior chest radiotherapy

(b) Prior chest radiotherapy

Figure 2: Comparison of the classical Kaplan-Meier latency estimator (**black** solid line) and the presmoothed Kaplan-Meier latency estimator with bootstrap bandwidth selector (**blue** dashed line) separated by non prior chest radiotherapy (2a) and prior chest radiotherapy (2b) classes.

prior chest radiotherapy experience cardiotoxicity later than indicated by the classic estimator.

4 Conclusions

In this study, we introduced a presmoothed estimator for latency, which assigns weight to all observations, including censored data, without relying on parametric assumptions.

In the real data application, we compared the behaviour of the non presmoothed classical Kaplan-Meier estimator versus the presmoothed version with a bootstrap bandwidth selector. We observed that the proposed estimator and the classical one differ. In all cases the presmoothed estimator shows a smoother transition than the classical one due presmoothing asigns weights to each observation in contrast to classical estimator where only the uncensored observations has. The proximity between estimations does not seem to be affected by the sample size as we could see in the Figure 1 and 2, where a larger sample size did not lead a better result. Although further research is needed, these results show that differences between estimators exists.

Acknowledgements

This work, integrated into the framework of PERTE for Vanguard Health, has been co-financed by the Spanish Ministry of Science, Innovation and Universities with funds from the European Union Next Generation EU, from the Recovery, Transformation and Resilience Plan (PRTR-C17.I1) and from the Autonomous Community of Galicia within the framework of the Biotechnology Plan Applied to Health. Besides, this work is part of the grant PID2023-147127OB-I00 "ERDF/EU",
funded by MCIN/AEI/10.13039/501100011033/. It has also been supported by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2024/14) and by CITIC as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01).

Bibliography

M. Arnold, E. Morgan, H. Rumgay, A. Mafra, D. Singh, M. Laversanne, J. Vignat, J. R. Gralow, F. Cardoso, S. Siesling, et al. Current and future burden of breast cancer: Global statistics for 2020 and 2040. *The Breast*, 66:15–23, 2022.

S. Bešlija, Z. Gojković, T. Cerić, A. M. Abazović, I. Marijanović, S. Vranić, J. Mustedanagić-Mujanović, F. Skenderi, I. Rakita, A. Guzijan, et al. 2020 consensus guideline for optimal approach to the diagnosis and treatment of HER2-positive breast cancer in Bosnia and Herzegovina. *Bosnian Journal of Basic Medical Sciences*, 21(2):120, 2021.

J. W. Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):15–53, 1949.

R. Cao, I. López-de Ullibarri, P. Janssen, and N. Veraverbeke. Presmoothed Kaplan–Meier and Nelson–Aalen estimators. *Journal of Nonparametric Statistics*, 17(1):31–56, 2005.

R. Florido, K. L. Smith, K. K. Cuomo, and S. D. Russell. Cardiotoxicity from human epidermal growth factor receptor-2 (HER2) targeted therapies. *Journal of the American Heart Association*, 6(9):e006915, 2017.

E. G. Kaboré, C. Macdonald, A. Kaboré, R. Didier, P. Arveux, N. Meda, M.-C. Boutron-Ruault, and C. Guenancia. Risk prediction models for cardiotoxicity of chemotherapy among patients with breast cancer: a systematic review. *JAMA Network Open*, 6(2):e230569, 2023.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

A. López-Cheda, M. A. Jácome, and R. Cao. Nonparametric latency estimation for mixture cure models. *TEST*, 26(2):353–376, 2017.

I. López-de Ullibarri and M. A. Jácome. survPresmooth: an R package for presmoothed estimation in survival analysis. *Journal of Statistical Software*, 54:1–26, 2013.

M. Mazzotta, E. Krasniqi, G. Barchiesi, L. Pizzuti, F. Tomao, M. Barba, and P. Vici. Long-term safety and real-world effectiveness of trastuzumab in breast cancer. *Journal of Clinical Medicine*, 8(2):254, 2019.

N. Mohan, J. Jiang, M. Dokmanovic, and W. J. Wu. Trastuzumab-mediated cardiotoxicity: current understanding, challenges, and frontiers. *Antibody Therapeutics*, 1(1):13–17, 2018.

E. Nadaraya. On estimating regression. *Theory of Probability & its Applications*, 9(1):141–142, 1964.

M. A. Owens, B. C. Horten, and M. M. Da Silva. HER2 amplification ratios by fluorescence in situ hybridization and correlation with immunohistochemistry in a cohort of 6556 breast cancer tissues. *Clinical Breast Cancer*, 5(1):63–69, 2004.

B. Piñeiro-Lamas, A. López-Cheda, R. Cao, L. Ramos-Alonso, G. González-Barbeito, C. Barbeito-Caamaño, and A. Bouzas-Mosquera. A cardiotoxicity dataset for breast cancer patients. *Scientific Data*, 10(1):527, 2023.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2025. URL *https://www.R-project.org/*.

G. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26: 359–372, 1964.