REVSTAT – Statistical Journal
Volume 0, Number 0, Month 0000, 000–000
https://doi.org/00.00000/revstat.v00i0.000

Energy Distance and Kernel Mean Embeddings for Two-Sample Survival Testing with Applications in Immunotherapy Clinical Trials

Authors: Marcos Matabuena ✉
– CiTIUS, Centro Singular de Investigación en Tecnoloxías Intelixentes
Universidad de Santiago de Compostela,
(marcos.matabuena@usc.es)

Oscar Hernan Madrid Padilla
– Department of Statistics, University of California Los Angeles (oscar.madrid@stat.ucla.edu)

Received: Month 0000 Revised: Month 0000 Accepted: Month 0000

Abstract:

- We study the problem of comparing distribution equality between two random samples under a random censoring scheme. We design a series of tests based on energy distance and kernel mean embeddings to address this problem. We calibrate our tests using permutation methods and prove that they are consistent against all fixed continuous alternatives. To evaluate our proposed tests in real-world clinical scenarios, we simulate survival curves from immunotherapy clinical trials published in major medical journals. Additionally, we provide practitioners with recommendations on selecting parameters and distances for the crossing survival curves problem observed in the analyzed real data. Based on the parameter tuning method we propose, we demonstrate that our tests show a considerable gain in statistical power compared to classical survival tests. Furthermore, as our test depends on the chosen semi-metric or kernel, it can be adapted to other clinical settings or survival analysis problems.

Keywords:

- *survival analysis; clinical trials; two-sample tests; immunotherapy studies; kernel tests.*

AMS Subject Classification:

- 49A05, 78B26.

✉ Corresponding author

1. INTRODUCTION

One of the main objectives of survival analysis is to compare the distributions of the lifetimes of two populations. This is best illustrated by clinical trials when evaluating the efficacy of two treatments Su and Zhu (2018). In the context of right-censored data, the scientific community uses the log-rank test to test the equality between two distribution curves. Originally proposed by Mantel and Haenszel (1959), the log-rank test has been further studied by different authors, e.g., Schoenfeld (1981); Fleming and Harrington (2011). Importantly, the log-rank test is known to be the most powerful test when the hazard functions are proportional to each other Schoenfeld (1981). However, when this hypothesis is violated, the test suffers a significant loss of power Fleming et al. (1980); Lachin and Foulkes (1986); Lakatos (1988); Schoenfeld (1981).

An important area of statistical research is searching for new tests that guarantee high statistical power in real use cases where the log-rank test does not perform well. We refer the reader to Su and Zhu (2018), where the authors thoroughly discuss the lack of statistical power of the log-rank test found in numerous case studies. Recent cancer immunotherapy trials also provide a relevant example. These trials consist of situations where survival curves cross Melero et al. (2014); Xu et al. (2017, 2018); Su and Zhu (2018); Alexander et al. (2018).

We distinguish two types of tests in the right-censoring survival setting: directional and omnibus. Loosely speaking, the former seeks to obtain maximum power in specific scenarios, while the latter is consistent against all alternatives. Examples of directional tests are the log-rank test family, see e.g., Gehan (1965); Tarone and Ware (1977); Peto and Peto (1972); Fleming and Harrington (1981), where statistics are assigned a weight function that determines the optimality in specific directions. Other approaches include combinations of tests, such as those in Bathke et al. (2009) and Yang and Prentice (2010).

From a theoretical point of view, omnibus tests are often preferred over directional due to their ability to detect any alternative asymptotically. However, in practice, these tests have the disadvantage of having low local power versus a wide variety of alternatives. In addition, it is known that any test with finite samples can have high power only in a limited number of scenarios. In particular, Janssen (2000) proves that there exists no test with high power, except in a finite-dimensional space.

In the era of precision medicine, see Kosorok and Laber (2019) for a review, drugs are designed to be personalized. This makes the statistical analysis of treatment differences particularly challenging. For example, comparing two treatments in a group of individuals may present highly heterogeneous survival curves due to significant individual variability in response to the treatment. A particular instance of this can be seen in immunotherapy studies Ferris et al. (2016) (Figure 1, Image B), where the survival curves intersect several times. In this context, new statistical distances between random samples that support censoring can be significant to perform hypothesis testing in new clinical decisions or stratify patients' survival into different groups with cluster analysis.

In order to help in this challenge, this paper proposes a novel approach for the two-sample testing problem under right censoring. Our approach relies on energy distance Székely

(2003); Székely and Rizzo (2013) and maximum mean discrepancy estimation Gretton et al. (2012). We summarize our specific contributions next.

1.1. Summary of results

Formally, we consider the classical traditional framework of two-sample survival comparisons where we are given lifetimes $T_{j,i} \sim P_j$ $(j = 0, 1; i = 1, \dots, n_j)$ and censoring times $C_{j,i} \sim Q_j$ $(j = 0, 1; i = 1, \dots, n_j)$, with distributions P_j and Q_j $(j = 0, 1)$, defined in a subset of $\mathbb{R}^+$. Here, the index j represents a population, and the index i a particular sample within a population. Moreover, the random variables $T_{0,1}, \dots, T_{0,n_0}, \dots, T_{1,1}, \dots, T_{1,n_1}$; $C_{0,1}, \dots, C_{0,n_0}, \dots, C_{1,1}, \dots, C_{1,n_1}$ are assumed to be independent of each other. In practice, only the random variables $X_{j,i} = \min(T_{j,i}, C_{j,i})$ and $\delta_{j,i} = 1\{X_{j,i} = T_{j,i}\}$ $(j = 0, 1; i = 1, \dots, n_j)$ are observed.

On the basis of the observed data $\{(X_{j,i}, \delta_{j,i})\}_{j=0,1;i=1,\dots,n_j}$, the two-sample testing problem that we study can be formulated as

$$H_0 \,:\, P_0(t) = P_1(t),\ \forall t > 0, \quad \text{versus} \quad H_A \,:\, P_0(t) \neq P_1(t),\ \text{for some } t > 0. \tag{1.1}$$

Our main contributions are the following:

- We propose novel tests based on energy distance and maximum mean discrepancy. The resulting tests require minimum assumptions, involving only conditions on the moments of random variables. Specifically, we assume $E(T_{j,i}^2) < \infty$ and $E(C_{j,i}^2) < \infty$, and, for simplicity, that the variables $X_{j,i}, T_{j,i}, C_{j,i}$ $(j = 0, 1; i = 1, \dots, n_j)$ are continuous.
- Importantly, we show that the proposed tests are consistent against all alternatives. In addition, we present a permutation-based procedure to approximate the distribution of our test statistics under the null hypothesis.
- We provide guidance on how to tune parameters of our proposed tests in clinical situations of interest. Furthermore, we show that Gaussian and Laplacian kernels outperform energy distance with Euclidean distance and other tests of the logrank family in settings where there is a delay effect, a commonly found situation in contemporary clinical trials.

Finally, we extend the proposed method to the multivariate case (Appendix C) that can be very interesting for clustering analysis and demonstrate the theoretical properties of the proposed statistics (Appendix B). In particular, we show that these statistics behave as true distances between random samples.

1.2. Outline

The structure of the paper is as follows. Section 2 provides an introduction to energy distance-based methods. Next, in Section 3 the statistics for our tests are derived, establishing their connections with previous work on two-sample testing based on kernel methods. Subsequently, we propose a permutation method and provide recommendations on how to choose the test parameters. In Section 4, we show that our tests are consistent against all alternatives. Section 5 then provides a simulation study to compare the behavior of the proposed tests against state-of-the-art methods. To this end, we compare the type I error using known distributions. In addition, we consider real scenarios from clinical practice and evaluate performance based on the power of the tests. Finally, the validity of our methods is verified in practice using the previously collected data Stablein et al. (1981).

In order to increase readability of the present document, we place the proofs of the main theoretical contributions and complementary results in the appendices.

2. Background on energy distance

The energy distance is a statistical distance between two distribution functions proposed in 1984 by Gábor J. Székely Székely and Rizzo (2017); Székely and Rizzo (2013). This distance is inspired by the concept of gravitational energy between two bodies and has experienced a rise in appeal for modern statistical applications due to its applicability to data of a complex nature, such as functions, graphs, or objects that live in negative type space. In parallel, A. A. Zinger, A. V. Kakosyan, and L. B. Klebanov developed a similar notion of distance called N-distances Klebanov et al. (2005); Rachev et al. (2013) and applied it to some biological problems Klebanov et al. (2007).

Here, we extended the notion of energy distance for right-censored data for the first time. To arrive at our family of tests, we first recall some background on energy distance. To that end, let $X, X' \sim^{\text{i.i.d.}} P$ and $Y, Y' \sim^{\text{i.i.d.}} Q$ where P and Q are probability distribution functions in $\mathbb{R}^d$. Denoting by $\|\cdot\|$ the Euclidean distance in $\mathbb{R}^d$ and assuming that $\max\{E(||X||), E(||Y||)\} < \infty$, the energy distance between the distributions P and Q is defined, as in Székely (2003) and Székely and Rizzo (2013), by:

$$\epsilon(P, Q) = 2E||X - Y|| - E||X - X^{'}|| - E||Y - Y^{'}||. \tag{2.1}$$

It is fairly easy to see that $\epsilon(\cdot, \cdot)$ is invariant to rotations, non-negative, and $\epsilon(P, Q) = 0$ if and only if $P = Q$. In addition, (2.1) can be extended for a family of parameters $\alpha \in (0, 2]$ assuming in each case the existence of the moment of order α (see Székely and Rizzo (2013)). The corresponding α-energy distance is then given as

$$\epsilon_\alpha(P, Q) = 2E||X - Y||^\alpha - E||X - X^{'}||^\alpha - E||Y - Y^{'}||^\alpha. \tag{2.2}$$

It can be proved that $\epsilon_\alpha(P, Q) \geq 0$. Furthermore, $\epsilon_\alpha(P, Q) = 0$ if and only if $P = Q$. In the case of $\alpha = 2$, $\epsilon_2(P, Q) = 2||E(X) - E(Y)||^2$. Therefore, non-negativity is verified trivially, although $\epsilon_2(P, Q) = 0$ implies equality in means and not that $P = Q$.

For a characteristic kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ using properties of kernel mean embeddings Muandet et al. (2017), as in Gretton et al. (2012), we define the measure of maximum mean discrepancy (MMD) as

$$\gamma_K^2(P, Q) = E(K(X, X^{'}) + E(K(Y, Y^{'})) - 2E(K(X, Y)), \tag{2.3}$$

Table 1: **Characteristic kernels**. $\Gamma(\cdot)$ denotes the Gamma function and K_v is the modified Bessel function of the second order v (see explicit definitions in Appendix F)

Kernel Function	$K(x,y)$
Gaussian	$\exp\left(-\frac{\lvert\lvert x-y\rvert\rvert^2}{\sigma^2}\right), \sigma > 0$
Laplacian	$\exp\left(-\frac{\lvert\lvert x-y\rvert\rvert}{\sigma}\right), \sigma > 0$
Rational quadratic	$(\lvert\lvert x-y\rvert\rvert^2 + c)^{-\beta}, \beta, c > 0$
Matérn	$\frac{2^{1-v}}{\Gamma(v)}\left(\frac{\sqrt{2v}\lvert\lvert x-y\rvert\rvert}{\sigma}\right)^v K_v\left(\frac{\sqrt{2v}\lvert\lvert x-y\rvert\rvert}{\sigma}\right)$

where $X, X' \sim^{\text{i.i.d.}} P$ and $Y, Y' \sim^{\text{i.i.d.}} Q$. Intuitively, (2.3) can be thought of as a non-linear generalization of the energy distance (2.1) in an appropriate reproducing kernel Hilbert space (RKHS). The latter depends on the selected parameters/distances.

Following this line, if we consider the energy distance in metric spaces Lyons (2013) (with an arbitrary semi-metric of negative type instead of the Euclidean distance), we find it equivalent to the kernel methods just defined. This equivalence was established in Sejdinovic et al. (2013) and Shen and Vogelstein (2018), at both the population and sample level.

Finally, some typical examples of characteristic kernels Sriperumbudur et al. (2010) are provided in Table 1.

3. Methodology

In this section, we present a new family of tests which are the focus of this paper. We begin by providing constructions of the statistics that are the pillars of our tests. Then, we present a procedure for determining the distribution of the statistics under the null hypothesis.

3.1. Construction of statistics

In the context of right censoring with independent data, the maximum non-parametric likelihood approach is the Kaplan-Meier estimator originally introduced in Kaplan and Meier (1958). Notably, the Kaplan-Meier estimator is consistent Wang et al. (1987) and its asymptotic properties were studied in Cai (1998). However, Stute (1994a) showed that the Kaplan-Meier estimator suffered from negative bias, which can be large under high censoring.

To proceed with our construction, we exploit the Kaplan-Meier estimator, combining it with a kernel type of estimator based on energy distance. To this end, for each group $j \in \{0,1\}$, we consider its ordered sample

$$X_{j,(1:n_j)} < X_{j,(2:n_j)} < \cdots < X_{j,(n_j:n_j)},$$

and the corresponding censored indicators $\delta_{j,(1:n_j)}, \delta_{j,(2:n_j)}, \ldots, \delta_{j,(n_j:n_j)}$. In addition, we refer to the maximum possible lifetimes for each group as τ_0 and τ_1, respectively.

With the above notation in hand, we motivate the definition of our statistics. First, if we knew the distributions P_0 and P_1, then we could calculate the metrics defined in (2.2) or (2.3) to measure the distance between the two populations. Since these distributions are not available, it is then natural to estimate them with the Kaplan-Meier estimator and use a sample version of the distances (2.2) or (2.3). This leads to an energy distance statistic under right censoring:

$$\tilde{\epsilon}_\alpha(P_0, P_1) = 2\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} W^0_{i:n_0} W^1_{j:n_1} ||X_{0,(i:n_0)} - X_{1,(j:n_1)}||^\alpha \tag{3.1}$$

$$- \sum_{i=1}^{n_0}\sum_{j=1}^{n_0} W^0_{i:n_0} W^0_{j:n_0} ||X_{0,(i:n_0)} - X_{0,(j:n_0)}||^\alpha \tag{3.2}$$

$$- \sum_{i=1}^{n_1}\sum_{j=1}^{n_1} W^1_{i:n_1} W^1_{j:n_1} ||X_{1,(i:n_1)} - X_{1,(j:n_1)}||^\alpha,$$

and a kernel statistic under right censoring:

$$\tilde{\gamma}^2_K(P_0, P_1) = \sum_{i=1}^{n_0}\sum_{j=1}^{n_0} W^0_{i:n_0} W^0_{j:n_0} K(X_{0,(i:n_0)}, X_{0,(j:n_0)}) \tag{3.3}$$

$$+ \sum_{i=1}^{n_1}\sum_{j=1}^{n_1} W^1_{i:n_1} W^1_{j:n_1} K(X_{1,(i:n_1)}, X_{1,(j:n_j)}) \tag{3.4}$$

$$- 2\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} W^0_{i:n_0} W^1_{i:n_1} K(X_{0,(i:n_0)}, X_{1,(j:n_1)}),$$

where

$$W^0_{i:n_0} = \frac{\delta_{0,(i:n_0)}}{n_0 - i + 1} \prod_{j=1}^{i-1} \left[\frac{n_0 - j}{n_0 - j + 1}\right]^{\delta_{0,(j:n_0)}} \quad (i = 1, \ldots, n_0), \tag{3.5}$$

and

$$W^1_{i:n_1} = \frac{\delta_{1,(i:n_1)}}{n_1 - i + 1} \prod_{j=1}^{i-1} \left[\frac{n_1 - j}{n_1 - j + 1}\right]^{\delta_{1,(j:n_1)}} \quad (i = 1, \ldots, n_1) \tag{3.6}$$

are the Kaplan-Meier weights from Stute (2003). While the statistics $\tilde{\epsilon}_\alpha(P_0, P_1)$ and $\tilde{\gamma}^2_K(P_0, P_1)$ seem to capture the differences between two populations, it is possible to prove that, almost surely, $\tilde{\epsilon}_\alpha(P_0, P_1)$ and $\tilde{\gamma}^2_K(P_0, P_1)$ converge to quantities $\gamma_{c(K)}(P_0, P_1)$ and $\epsilon_{c(\alpha)}(P_0, P_1)$, respectively. However, they do not behave like distances between probability distributions. Specifically, there exist two different probability distributions P_0 and P_1 in $\mathbb{R}$ satisfying $\epsilon_{c(\alpha)}(P_0, P_1) < 0$. We can also find two different probability distributions P_0 and P_1 in $\mathbb{R}$ with $\epsilon_{c(\alpha)}(P_0, P_1) = 0$. We refer the reader to Appendix C for specific constructions of these examples.

The reason behind the odd behavior of the statistic $\tilde{\gamma}^2_K(P_0, P_1)$ ($\tilde{\epsilon}_\alpha(P_0, P_1)$) has to do with the fact that P_l is not completely supported in $[0, \tau_l]$, for $l \in \{0, 1\}$. We alleviate this problem by defining the conditional distributions $P'_0(x) = P_0(x)/\int_0^{\tau_0} dP_0(x)dx \ \forall x \in [0, \tau_0]$, and $P_1'(x) = P_1(x)/\int_0^{\tau_1} dP_1(x)dx \; \forall x \in [0,\tau_1]$. With P_0' and P_1' at hand, we construct conditional versions of the weights $W_{i:n_l}^{l}$ $(l = 0,1; i = 1,\ldots,n_j)$. Specifically, we consider the U-statistics under right censoring suggested in Bose and Sen (1999) and apply the aforementioned standardization, following Stute and Wang (1993). The resulting statistics are:

$$\begin{aligned}
\tilde{\epsilon}_\alpha(P_0,P_1) =& 2\frac{\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} W_{i:n_0}^{0} W_{j:n_1}^{1} ||X_{0,(i:n_0)} - X_{1,(j:n_1)}||^\alpha}{\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} W_{i:n_0}^{0} W_{j:n_1}^{1}} \tag{3.7}\\
&- \frac{\sum_{i=1}^{n_0}\sum_{j\neq i}^{n_0} W_{i:n_0}^{0} W_{j:n_0}^{0} ||X_{0,(i:n_0)} - X_{0,(j:n_0)}||^\alpha}{\sum_{i=1}^{n_0}\sum_{j\neq i}^{n_0} W_{i:n_0}^{0} W_{j:n_0}^{0}} \tag{3.8}\\
&- \frac{\sum_{i=1}^{n_1}\sum_{j\neq i}^{n_1} W_{i:n_1}^{1} W_{j:n_1}^{1} ||X_{1,(i:n_1)} - X_{1,(j:n_1)}||^\alpha}{\sum_{i=1}^{n_1}\sum_{j\neq i}^{n_1} W_{i:n_1}^{1} W_{j:n_1}^{1}}
\end{aligned}$$

(U-statistic α-energy distance under right censoring),

$$\begin{aligned}
\tilde{\gamma}_K^2(P_0,P_1) =& \frac{\sum_{i=1}^{n_0}\sum_{j\neq i}^{n_0} W_{i:n_0}^{0} W_{j:n_0}^{0} K(X_{0,(i:n_0)}, X_{0,(j:n_0)})}{\sum_{j=1}^{n_0}\sum_{j\neq i}^{n_0} W_{i:n_0}^{0} W_{j:n_0}^{0}} \tag{3.9}\\
&+ \frac{\sum_{i=1}^{n_1}\sum_{j\neq i}^{n_1} W_{i:n_1}^{1} W_{j:n_1}^{1} K(X_{1,(i:n_1)}, X_{1,(j:n_1)})}{\sum_{i=1}^{n_1}\sum_{j\neq i}^{n_1} W_{i:n_1}^{1} W_{j:n_1}^{1}} \tag{3.10}\\
&- 2\frac{\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} W_{i:n_0}^{0} W_{j:n_1}^{1} K(X_{0,(i:n_0)}, X_{1,(j:n_1)})}{\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} W_{i:n_0}^{0} W_{j:n_1}^{1}}
\end{aligned}$$

(U-statistic kernel method under right censoring).

Analogously, we can define V-statistics in the following manner:

$$\begin{aligned}
\tilde{\epsilon}_\alpha(P_0,P_1) =& 2\frac{\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} W_{i:n_0}^{0} W_{j:n_1}^{1} ||X_{0,(i:n_0)} - X_{1,(j:n_1)}||^\alpha}{\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} W_{i:n_0}^{0} W_{j:n_1}^{1}} \tag{3.11}\\
&- \frac{\sum_{i=1}^{n_0}\sum_{j=1}^{n_0} W_{i:n_0}^{0} W_{j:n_0}^{0} ||X_{0,(i:n_0)} - X_{0,(j:n_0)}||^\alpha}{\sum_{i=1}^{n_0}\sum_{j=1}^{n_0} W_{i:n_0}^{0} W_{j:n_0}^{0}} \tag{3.12}\\
&- \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} W_{i:n_1}^{1} W_{j:n_1}^{1} ||X_{1,(i:n_1)} - X_{1,(j:n_1)}||^\alpha}{\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} W_{i:n_1}^{1} W_{j:n_1}^{1}}
\end{aligned}$$

(V-statistic α-energy distance under right censoring),

$$\begin{aligned}
\tilde{\gamma}_K^2(P_0,P_1) =& \frac{\sum_{i=1}^{n_0}\sum_{j=1}^{n_0} W_{i:n_0}^{0} W_{j:n_0}^{0} K(X_{0,(i:n_0)}, X_{0,(j:n_0)})}{\sum_{j=1}^{n_0}\sum_{j=1}^{n_0} W_{i:n_0}^{0} W_{j:n_0}^{0}} \tag{3.13}\\
&+ \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} W_{i:n_1}^{1} W_{j:n_1}^{1} K(X_{1,(i:n_1)}, X_{1,(j:n_1)})}{\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} W_{i:n_1}^{1} W_{j:n_1}^{1}} \tag{3.14}\\
&- 2\frac{\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} W_{i:n_0}^{0} W_{j:n_1}^{1} K(X_{0,(i:n_0)}, X_{1,(j:n_1)})}{\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} W_{i:n_0}^{0} W_{j:n_1}^{1}}
\end{aligned}$$

(V-statistic kernel method under right censoring).

Finally, to establish consistency more easily, our final statistics are given as:

$$T_{\tilde{\epsilon}_\alpha} = \frac{n_0 n_1}{n_0 + n_1} \tilde{\epsilon}_\alpha(P_0, P_1) \text{ and } T_{\tilde{\gamma}_K^2} = \frac{n_0 n_1}{n_0 + n_1} \tilde{\gamma}_K^2(P_0, P_1). \tag{3.15}$$

In Appendix C, we can find, in some instances, an interpretation of the limits of these statistics. In particular, we show that the statistics behave as distances between distribution functions and the characteristic functions in a weighted Hilbert space $L^2(I)$.

3.2. Permutation tests

As in the case of the usual energy two-sample test from Székely (2003) and Székely and Rizzo (2013), the null distribution of our proposed statistics is approximated with a permutation method. If the censorship mechanism of the two groups is the same, the standard permutation method from Neuhaus et al. (1993) and Wang et al. (2010) is valid. However, when the censoring distributions differ, the standard permutation method does not perform well in small-sample settings or when the amount of censoring is large, see Heimann and Neuhaus (1998). In this case, one alternative is to use the re-sampling strategy proposed in Wang et al. (2010). Below we describe the steps of the classical permutation procedure.

We denote by $Z = (\overbrace{0, \cdots, 0}^{n_0}, \overbrace{1, \cdots, 1}^{n_1})$ a vector of size $n = n_0 + n_1$ that indicates the observed group membership. Thus, $z_i = 1$ ($z_i = 0$) indicates that the i-th subject belongs to group 1 (0). We then order the observed times and censorship indicators. Thus we construct vectors $U = (X_{0,1}, \cdots, X_{0,n_0}, X_{1,1}, \cdots, X_{1,n_1})$ and $\delta = (\delta_{0,1}, \cdots, \delta_{0,n_0}, \delta_{1,1}, \cdots, \delta_{1,n_1})$. Next, if we are interested in calculating the distribution of the statistic $\theta(Z, U, \delta)$ under the null distribution ($P_0 = P_1$), then we can proceed to construct permutations of the data. Specifically, let $\mathcal{S}$ be a collection of sets of size n_0 whose elements belong to $\{1, \ldots, n_0 + n_1\}$. For every $I \in \mathcal{S}$, we construct a vector $Z^I \in \mathbb{R}^n$ satisfying $Z_i^I = 0$ if $i \in I$ and $Z_i^I = 1$ if $i \notin I$. Next, we compare $\theta(Z, U, \delta)$ against $\theta(Z^I, U, \delta)$ for all $I \in \mathcal{S}$. The p-value is calculated as

$$\text{p-value} = \frac{\sum_{I \in \mathcal{S}} 1\{\theta(Z^I, U, \delta) \geq \theta(Z, U, \delta)\}}{\binom{n}{n_0}}. \tag{3.16}$$

In practice, we can reduce the number of operations in (3.16) by using a random subset $\mathcal{S}'$ of $\mathcal{S}$ to obtain

$$\text{p-value} \approx \frac{\sum_{I \in \mathcal{S}'} 1\{\theta(Z^I, U, \delta) \geq \theta(Z, U, \delta)\}}{|\mathcal{S}'|}.$$

3.3. Selection of tuning parameters/distances

Although the proposed methods are consistent against all alternatives from an asymptotic point of view (see Theorem 1), one of the main practical difficulties with finite samples is the selection of parameters/distances so that high statistical power is guaranteed. In fact, this problem is very common in kernel methods both in prediction models and hypothesis testing. Filippi et al. (2016) state that there exist few theoretical approaches to tackle this problem.

In this work, we only use the energy distance with the Euclidean distance and the Gaussian and Laplacian kernels (see Table 1). The main reason for this is that there is a corpus of previous work on how the selection of parameters influences the performance of different methods. There are also some heuristics that include theoretical results, see Ramdas et al. (2015) and Garreau et al. (2017).

Despite the fact that energy distance is more sensitive to the choice of the α parameter than to the choice of the kernel (see, for example, Sejdinovic et al. (2013)), there is no known formal criterion for selecting an optimal value of α.

In regard to the Gaussian and Laplacian kernels, there is a known ad-hoc rule called the Median heuristic that consists in selecting the median between the distance pairs of the aggregate sample. This procedure is explained in detail below.

Let $X = (X_1, \dots, X_{n_0}, X_{n_0+1}, \cdots, X_{n_0+n_1}) = (X_{0,1}, \cdots, X_{0,n_0}, X_{1,1}, \cdots, X_{1,n_1})$ be the aggregate sample vector. Consider $D \in \mathbb{R}^{(n_0+n_1)\times(n_0+n_1)}$ defined as $D_{ij} = |X_i - X_j|(i = 1, \dots, (n_0+n_1), j = 1, \dots, (n_0+n_1))$.

As in Garreau et al. (2017), we define

$$\sigma = \sqrt{H_n/2} \ , \quad \text{where} \quad H_n = \text{median}\{D_{ij}^2 : 1 \leq i < j \leq (n_0+n_1)\}.$$

In the literature, the resulting σ is known as kernel bandwidth. An intuitive explanation of how this works is given below:

- Given X_i, X_j $(i = 1, \dots, (n_0+n_1), j = 1, \dots, (n_0+n_1))$, if $\sigma \to 0$ or $\sigma \to \infty$, then $K(X_i, X_j) \to 1$ or $K(X_i, X_j) \to 0$ (see Table 1). Therefore, $\tilde{\gamma}_K^2(P_0, P_1)$ is almost always constant, and the statistical power of the test is low.

- It is reasonable to impose that the median of D_{ij} $(i = 1, \dots, (n_0+n_1), j = 1, \dots, (n_0+n_1))$ and σ are of the same order so that $K(X_i, X_j)$ $(i = 1, \dots, (n_0+n_1), j = 1, \dots, (n_0+n_1))$ does not take unnecessarily small or large values, so as not to suffer from the limitations mentioned above.

- Hence, a reasonable choice for σ is in the "middle range" of D_{ij} $(i = 1, \dots, (n_0+n_1), j = 1, \dots, (n_0+n_1))$. In this way, σ is of the same order as the median of D_{ij} $(i = 1, \dots, (n_0+n_1), j = 1, \dots, (n_0+n_1))$. The global dispersion between terms $K(X_i, X_j)$ $(i = 1, \dots, (n_0+n_1), j = 1, \dots, (n_0+n_1))$ is maximized, and therefore, the test has greater discrimination capacity.

Alternatively, σ is sometimes set to $\sqrt{H_n}$.

The influence of the suboptimal specification of the kernel bandwidth has mainly been studied in situations of high dimensionality. In this context, it has been shown to lead to important differences in the power of tests. For instance, Ramdas et al. (2015) noticed, using a simulation study and theoretical analysis, that the median heuristic σ maximized power with Gaussian kernel in several cases. However, power can be suboptimal with the Laplacian kernel, showing better results with some values of $\sigma = H_n^{\alpha}$ for $\alpha \in (0, 2]$ with $\alpha \neq 1/2$. In any case, we should be cautious in interpreting these results. As we do not consider the multidimensional case, the effects of a suboptimal kernel bandwidth specification may not be so dramatic in our setting.

In the case of censorship, in addition to the vector X, we also have to consider the vector $\delta = (\delta_{0,1}, \cdots, \delta_{0,n_0}, \delta_{1,1}, \cdots, \delta_{1,n_1})$ with censorship indicators. Now, we define the set of indices $I = \{i \in \{1, 2, \ldots, (n_0 + n_1)\} : \delta_i = 1\}$. A reasonable estimator for σ is given by $\sigma = \sqrt{H_n^*}$ or $\sigma = \sqrt{H_n^*/2}$ where

$$H_n^* = \text{median}\{D_{ij}^2 : 1 \leq i < j \leq (n_0 + n_1) \text{ with } i, j \in I\}.$$

The previous definition is justified because in equations (3.1)–(3.13), only the elements whose indices belong to I influence the corresponding expressions.

4. Theory

Next, we show that, under very mild conditions, our proposed tests are consistent against all alternatives. This is formally stated below and the proof can be found in Appendix A.

Theorem 1. Let $X_{j,i} = \min(T_{j,i}, C_{j,i}) \sim^{\text{i.i.d.}} P_{c(j)}$ and $\delta_{j,i} = 1\{X_{j,i} = T_{j,i}\}$ $(j = 0, 1; i = 1, \ldots, n_j)$ with $P_{c(j)}$ $(j = 0, 1)$. Suppose also that the conditions stated in Section 1.1 hold for the random variables $T_{j,i} \sim^{\text{i.i.d.}} P_j, \quad C_{j,i} \sim^{\text{i.i.d.}} Q_j$ $(j = 0, 1; i = 1, \ldots, n_j)$. Further assume that $\tau_0 = \tau_1$ or the support of the distribution functions P_0 and P_1 is contained in the intervals $[0, \tau_0]$ and $[0, \tau_1]$, respectively. Then, for testing the null $H_0 : P_0(t) = P_1(t) \quad \forall t \in [0, \tau_1]$, the statistics $T_{\tilde{\epsilon}_\alpha}$ and $T_{\tilde{\gamma}_K^2}$ determine tests that are consistent against all fixed alternatives with continuous random variables.

The Kolmogorov-Smirnov and Cramer Von-Mises tests under censorship have been proposed in the context of absolutely continuous random variables Schumacher (1984). However, unlike those tests, our results are also valid for discrete distributions provided that the second-order moments of the random variables exist. This can be very important in practice since many of the lifetimes collected in databases for simplification are truncated and discrete (see for example Cai et al. (2019) and `http://lce.biohpc.swmed.edu/lungcancer/dataset.php`).

5. Simulation study

The simulation study is divided into two phases. In the first part, we consider scenarios where the null hypothesis is true. Then, the performance of the proposed tests is compared with the log-rank family tests with different censorship rates and different sample sizes. In particular, the tests used are the energy distance ($\alpha = 1$), Gaussian kernel ($\sigma = 1$), Laplacian kernel ($\sigma = 1$), log-rank (Mantel and Haenszel (1959)), Gehan generalized Wilcoxon test (Gehan (1965)), Tarone-Ware (Tarone and Ware (1977)), Peto-Peto (Peto and Peto (1972)), Fleming & Harrington (Fleming and Harrington (1981)) (with $\rho = \gamma = 1$). For this purpose, parametric distributions such as normal exponential or lognormal are used.

In the second phase, the same tests are compared in scenarios where the null hypothesis is false. As in Guyot et al. (2012), we use the Digitizeit software (`https://www.digitizeit.de/`) to extract several survival curves from different clinical trials in which there was a delay effect, or there was no clear violation of the hypothesis that the hazard functions were not maintained. Survival curves were extracted from the studies analyzed in the following two papers: Su and Zhu (2018) and Alexander et al. (2018). We also consider simulations under the hypothesis that the hazard functions are proportional. This is to assess the power loss of our tests compared to the log-rank tests. In all comparisons, the σ parameter of the Gaussian and Laplacian kernels (see Table 1) is selected with the methodology defined in Section 3.3.

When the null hypothesis is true, the sample size $n \in \{20, 50\}$. Otherwise, $n \in \{20, 50, 100, 200\}$. The censorship mechanism was the same within each simulation performed.

All the tests are executed with the statistical software **R**. For the family of the log-rank test, the coin package Hothorn et al. (2008) is used while the new tests were implemented in **C++** and integrated in **R** with the "Rcpp" Eddelbuettel et al. (2011) and "Rcpp Armadillo" libraries. In all cases, the tests were calibrated by the permutation method, with 1000 permutations executed.

5.1. Null hypothesis

We perform 500 Monte Carlo simulations in which the null hypothesis is correct. The censoring rates are 10 and 30 percent and the sample size of 20 and 50 individuals. Since p-values are distributed uniformly (Uniform$(0,1)$) under the null hypothesis, the mean of the observed p-values obtained should be close to 0.5, and the standard deviation close to $\sqrt{1/12} = 0.2886751$. Similarly, approximately 5 percent of the observations should have a value less than 0.05. In the Appendix E Tables (1-3), we can see the results of calculations of the mean and standard deviation for each test. In Tables (4-6), the proportion of p-values is shown to be approximately less than or equal to 0.05 for the same cases.

The results of the proposed tests under the null hypothesis are consistent and similar to those of the log-rank test family. Certain discrepancies with the theoretical values are acceptable when doing the comparison with 500 Monte Carlo simulations in 8 different tests. In turn, the Kaplan-Meier estimator used in our models as well as in some of the log-rank family models presents a certain bias that is dependent on the censoring ratio, which produces small deviations under what is expected in a theoretical framework under the null hypothesis.

5.2. Alternative hypothesis

We perform 500 Monte Carlo simulations in different situations where the null hypothesis does not hold. In all cases, we simulated data from survival curves extracted from clinical trials by means of Digitizeit.

5.2.1. Survival curves from clinical trials

The curves extracted in this article for comparison are as follows: Figure 1-A from Borghaei et al. (2015), Figure 2-A from Rodriguez et al. (2016), Figure 2-B from Motzer et al. (2015), Figure 1-B from Ferris et al. (2016), Figure 1-B from Bellmunt et al. (2017), and Figure 1-C in Borghaei et al. (2015).

These articles were compiled from Su and Zhu (2018) and Alexander et al. (2018) who assessed the limitations of log-rank in many clinical situations or the problem of using summary measures to describe a survival curve. In addition, Alexander et al. (2018) focused on the field of immunotherapy where there was often a long-term delay effect on survival or the survival curves are crossed, which motivated the recent development of new tests for this situation, e.g., Xu et al. (2017) and Xu et al. (2018).

We use the presence of the survival curves are crossed with respect to the other as criteria for selecting the survival curves. Additionally, a curve was selected in which hypothesis that the function is hazard are proportional is not violated with experimental data, Figure 1-A in Rodriguez et al. (2016). In most of the selected curves, the tests used in the original papers did not show statistically significant differences.

The process of reconstructing each pair of curves is as follows:

1. Extraction of the numerical values of the curves through the software Digitizeit.
2. Reconstruction of the curves from the numerical values in the statistical software **R**.
3. Truncation of the support of the curves to the minimum right end of both curves, that is $\tau = \min\{\tau_0, \tau_1\}$, where τ_0 is the right end of the first curve, and, analogously, τ_1 for the second curve.
4. Smoothing curves with cubic smoothing spline, as in Hastie and Tibshirani (1990).
5. Applying piecewise anti-isotonic linear regression so that the generated curves decrease, see Robertson et al. (1988). Subsequently, data from the estimated curves are simulated. The censorship variable is $C \sim \text{Uniform}(0, \tau)$ where τ is the maximal value of support common to both curves by 3.

In Figures 1–3, we can see the Kaplan-Meier curves after simulating data from the generated curves (with sample sizes of 10000 individuals per population) along with an evaluation of power.

(5.1)

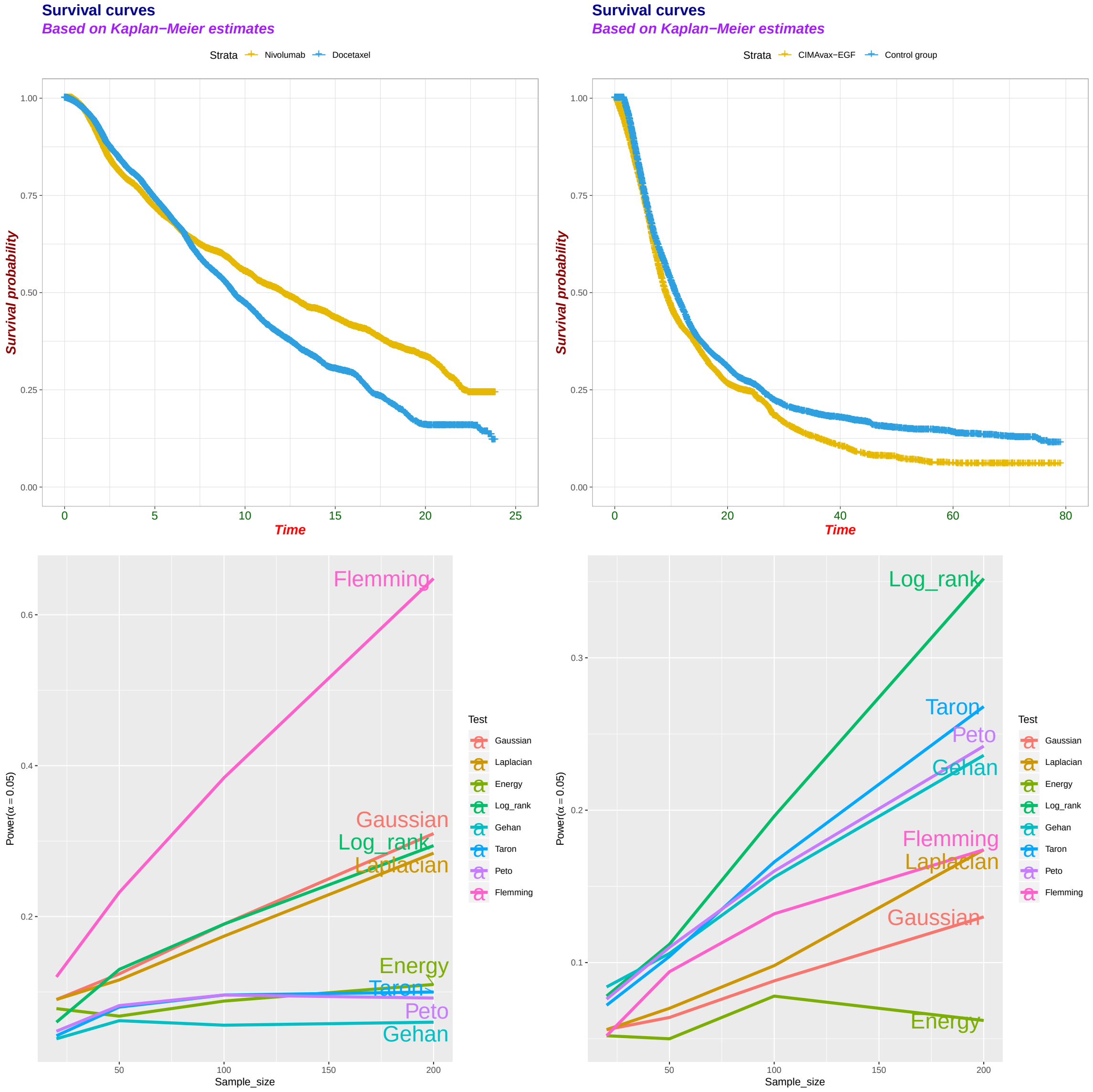

Figure 1: Statistical power of survival curves extracted Borghaei et al. (2015) Figure 1-*A* (left) and Rodriguez et al. (2016) Figure 2-*A* (right).

(5.2)

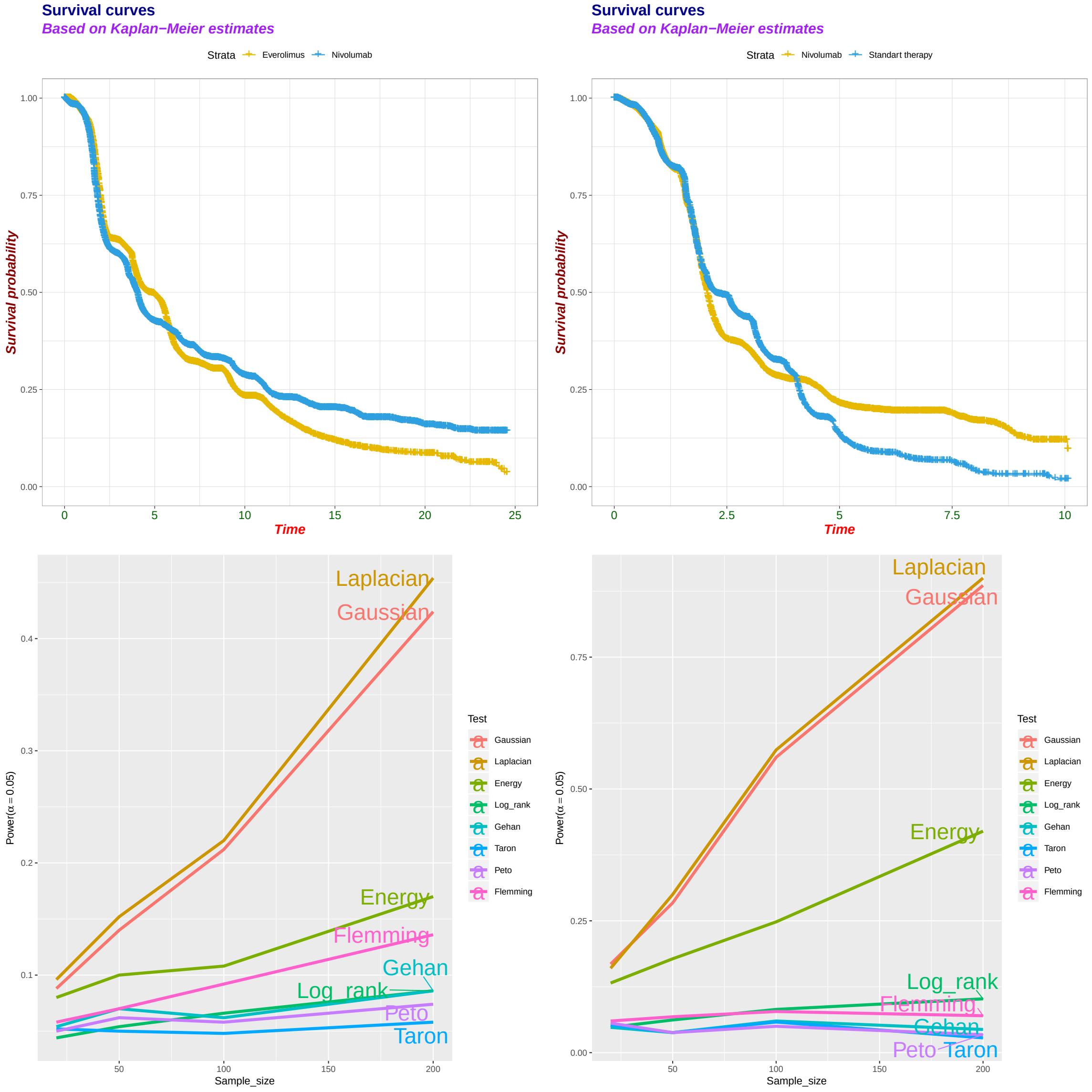

Figure 2: Statistical power of survival curves extracted Motzer et al. (2015) Figure 2-B (left) and Ferris et al. (2016) Figure 1-B (right).

(5.3)

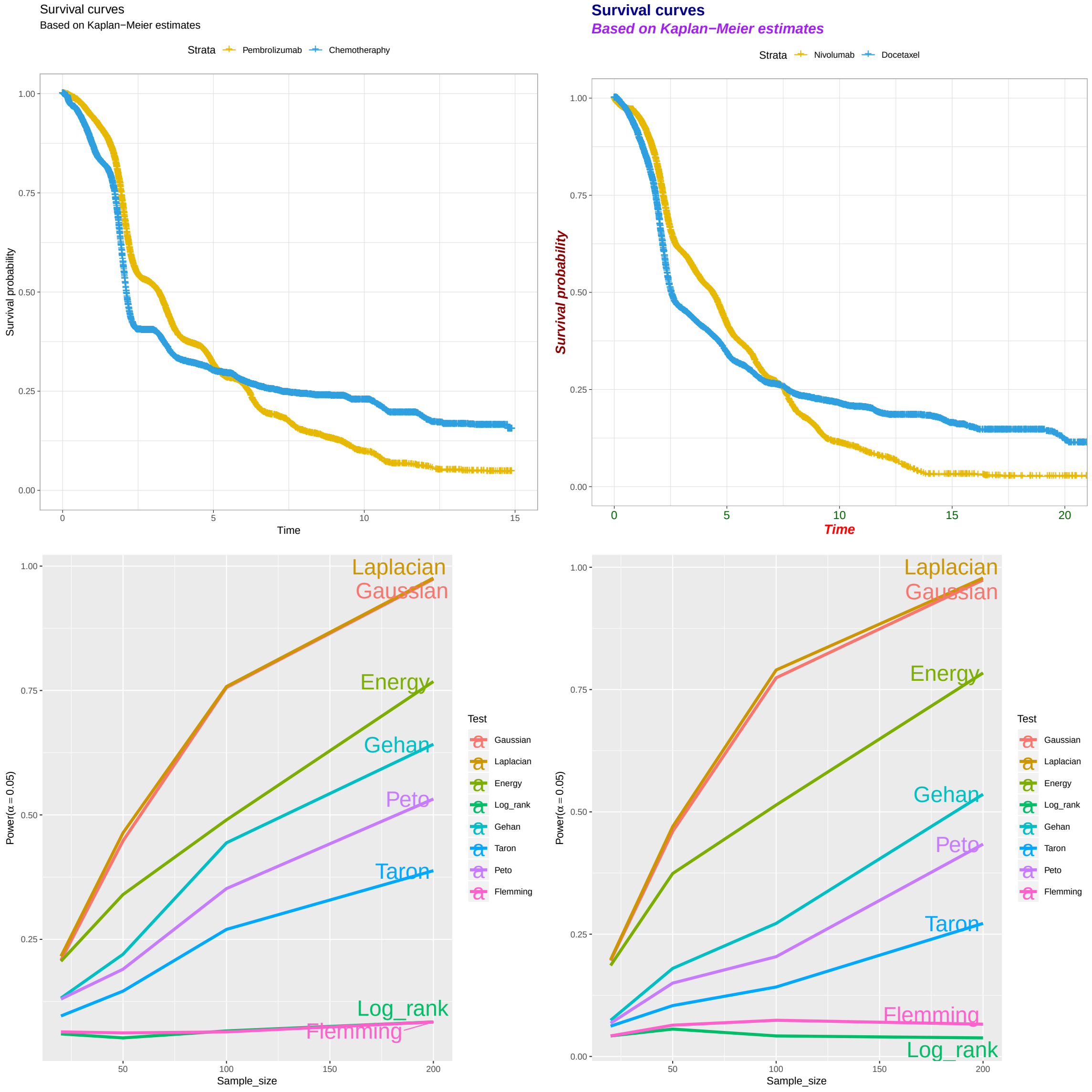

Figure 3: Statistical power of survival curves extracted Bellmunt et al. (2017) Figure 1-*B* (left) and Borghaei et al. (2015) Figure 1-*C* (right).

The results are discussed below:

- In Figures 2 and 3, all the images reflect a delay between the two treatments. In addition, almost all the patients die in the interval of time studied. In this situation, all our methods outperform the log-rank family test studied, especially those based on the Laplacian and Gaussian kernels.
- In Figure 1 (left), there is a small delay much smoother than those discussed above. In addition, there is a significant fraction of patients who survive. In this situation, all the tests have low power, even when the sample size is equal to 200. The Fleming &

Harrington test works better than our proposals.

- In Figure 1 (right), the situation where the hypothesis that hazard functions do not seem violated, our tests have low power. As expected, the best method, in this case, is the log-rank, although it does not present high power either. Graphically, it can be seen that the degree of discrepancy between both curves is low.

6. Example

To illustrate the potential of the newly proposed tests in real clinical cases, we use the database from a gastrointestinal tumor study by Stablein et al. (1981). This can be found in the **R** package "coin". The aim of this study is to test whether there are statistically significant differences in the survival curves of two treatments. In Figure 4, we present the survival curves between the two treatments, observing clear differences between the curves. At first glance, there appears to be a tendency that the first treatment increases the long-term survival compared to the second. The hypothesis of proportional hazards is strongly violated.

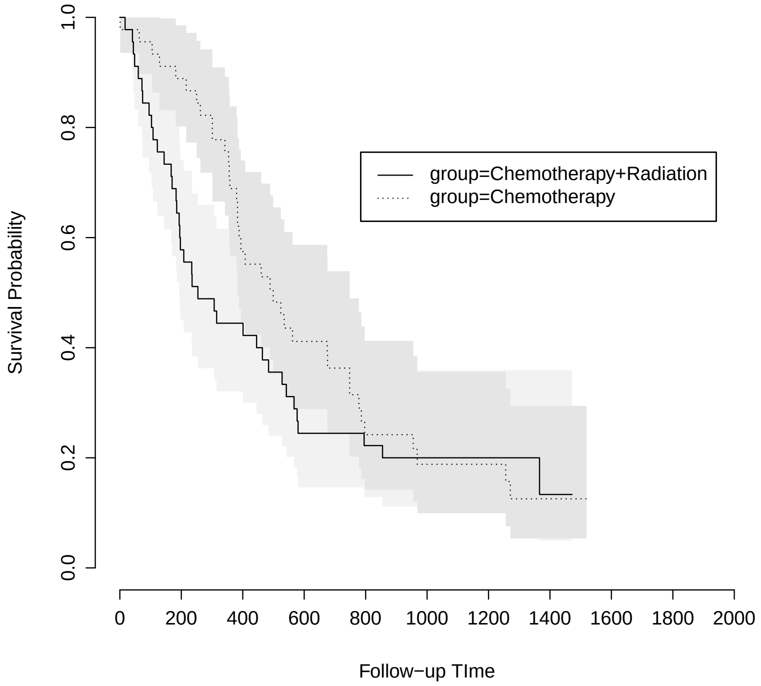

Figure 4: Survival curves real case

Table 2: p-values of the different methods used in the real case

	p-value		p-value		p-value
Energy distance $\alpha = 1$	0.018	Kernel Gaussian	0.004	Kernel Laplacian	**0.002**
Logrank	0.262	Gehan	0.024	Tarone	0.075
Peto	0.030	Flemming	0.753		

7. Discussion

In this article, we have proposed a family of consistent tests against all alternatives to compare the distribution equality between two samples based on energy distance and kernel mean embeddings. Additionally, several theoretical properties of the statistics have been established, along with a set of recommendations on how to select parameters and when to use our tests in clinical situations of interest.

Much work has been done in survival analysis in the context of hazard functions proportionality and for situations where alternatives do not differ much from this scenario. In such cases, the log-rank tests are known to be optimal Schoenfeld (1981), and tests such as Fleming & Harrington Fleming and Harrington (1981) offer a good alternative by choosing a suitable weight function in case of deviations.

If there is evidence that the above situation holds, we do not suggest implementing our tests with the distances/kernels used in this work because the performance difference with competitors is considerable.

In scenarios where there is a delay effect on survival in one treatment over another or where survival curves cross, our tests with the recommended parameters outperform classical tests. In cases where survival curves cross in a weak manner (as displayed in Figure 1 (left)), the performance of our tests is suboptimal. However, the situation where our tests perform excellently is quite common in clinical trials of immunotherapy Alexander et al. (2018), and therefore our tests can be considered an excellent alternative. Overall, we can say that in situations with crossed survival curves, the results show that the statistical power of using the Laplacian or Gaussian kernel hardly varies; however, it considerably improves the performance compared to using classical tests of the log-rank family.

The proposed estimators are based on the Kaplan-Meier estimator weights Stute (2003). If there is a high percentage of censored observations along with a small sample size, these methods may not work well (which is very common in all survival analysis methods). In this case, smoothing the weights may help increase the power. Alternatively, if there are apparent differences at the end of the survival curves, we recommend considering the last observation uncensored Efron (1967). In either case, this may increase the power of the tests and reduce the bias.

The overall cost of one evaluation of the test statistics is of the order $O(n^2)$ because it involves a V-statistic that depends on the weights of the Kaplan-Meier estimator (which has a linear estimation cost $O(n)$). Although this may be a concern, in epidemiological studies and clinical trials, the sample size is usually less than 500 subjects in each group.

A step forward might be to use incomplete U/V statistics to increase the computational efficiency of the method. However, this estimator for right-censored data is not available in the literature. Importantly, a recent calibration strategy that avoids using permutation methods has appeared in the context of distance correlation and complete information Shen et al. (2022). Unfortunately, in our setting, there is a bias in the estimators due to the censoring mechanism Fernández and Rivera (2020); Stute (1994b), which limits the direct application of the proposed bound in that paper.

The extension of the proposed tests to k-samples is analogous to non-censored methods. A large body of research exists in this field, such as Disco analysis Rizzo et al. (2010) or more recently proposed kernel methods Balogoun et al. (2018).

Our modeling strategy that modifies the sampling weights with Kaplan-survey weights can be adapted to handle other situations in survival analysis, such as truncated interval-censored data or even other types of incomplete information, such as missing data (see for example Matabuena et al. (2023)) . Therefore, our modeling strategy is general and can be used in other settings.

REFERENCES

Alexander, B. M., Schoenfeld, J. D., and Trippa, L. (2018). Hazards of hazard ratios—deviations from model assumptions in immunotherapy. *New England Journal of Medicine*, 378(12):1158–1159.

Balogoun, A. S. K., Nkiet, G. M., and Ogouyandjou, C. (2018). Kernel based method for the k-sample problem.

Bathke, A., Kim, M.-O., and Zhou, M. (2009). Combined multiple testing by censored empirical likelihood. *Journal of Statistical Planning and Inference*, 139(3):814–827.

Bellmunt, J., De Wit, R., Vaughn, D. J., Fradet, Y., Lee, J.-L., Fong, L., Vogelzang, N. J., Climent, M. A., Petrylak, D. P., Choueiri, T. K., et al. (2017). Pembrolizumab as second-line therapy for advanced urothelial carcinoma. *New England Journal of Medicine*, 376(11):1015–1026.

Borghaei, H., Paz-Ares, L., Horn, L., Spigel, D. R., Steins, M., Ready, N. E., Chow, L. Q., Vokes, E. E., Felip, E., Holgado, E., et al. (2015). Nivolumab versus docetaxel in advanced nonsquamous non–small-cell lung cancer. *New England Journal of Medicine*, 373(17):1627–1639.

Bose, A. and Sen, A. (1999). The strong law of large numbers for kaplan–meier u-statistics. *Journal of Theoretical Probability*, 12(1):181–200.

Cai, L., Lin, S., Girard, L., Zhou, Y., Yang, L., Ci, B., Zhou, Q., Luo, D., Yao, B., Tang, H., et al. (2019). Lce: an open web portal to explore gene expression and clinical associations in lung cancer. *Oncogene*, 38(14):2551.

Cai, Z. (1998). Asymptotic properties of kaplan-meier estimator for censored dependent data. *Statistics & probability letters*, 37(4):381–389.

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J., and Bates, D. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Efron, B. (1967). The two sample problem with censored data. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 4, pages 831–853.

Fernández, T. and Rivera, N. (2020). Kaplan-Meier V- and U-statistics. *Electronic Journal of Statistics*, 14(1):1872 – 1916.

Ferris, R. L., Blumenschein Jr, G., Fayette, J., Guigay, J., Colevas, A. D., Licitra, L., Harrington, K., Kasper, S., Vokes, E. E., Even, C., et al. (2016). Nivolumab for recurrent squamous-cell carcinoma of the head and neck. *New England Journal of Medicine*, 375(19):1856–1867.

Filippi, S., Flaxman, S., Sejdinovic, D., and Cunningham, J. (2016). Bayesian learning of kernel embeddings.

Fleming, T. R. and Harrington, D. P. (1981). A class of hypothesis tests for one and two sample censored survival data. *Communications in Statistics-Theory and Methods*, 10(8):763–794.

Fleming, T. R. and Harrington, D. P. (2011). *Counting processes and survival analysis*, volume 169. John Wiley & Sons.

Fleming, T. R., O'Fallon, J. R., O'Brien, P. C., and Harrington, D. P. (1980). Modified kolmogorov-smirnov test procedures with application to arbitrarily right-censored data. *Biometrics*, pages 607–625.

Garreau, D., Jitkrittum, W., and Kanagawa, M. (2017). Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*.

Gehan, E. A. (1965). A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52(1-2):203–224.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

Guyot, P., Ades, A., Ouwens, M. J., and Welton, N. J. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published kaplan-meier survival curves. *BMC medical research methodology*, 12(1):9.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*. London: Chapman & Hall.

Heimann, G. and Neuhaus, G. (1998). Permutational distribution of the log-rank statistic under random censorship with applications to carcinogenicity assays. *Biometrics*, pages 168–184.

Hothorn, T., Hornik, K., Van De Wiel, M. A., Zeileis, A., et al. (2008). Implementing a class of permutation pests: the coin package.

Janssen, A. (2000). Global power functions of goodness of fit tests. *Annals of Statistics*, pages 239–253.

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

Klebanov, L., Glazko, G., Salzman, P., Yakovlev, A., and Xiao, Y. (2007). A multivariate extension of the gene set enrichment analysis. *Journal of Bioinformatics and Computational Biology*, 5(5):1139–1153.

Klebanov, L. B., Beneš, V., and Saxl, I. (2005). *N-distances and their applications*. Charles University in Prague, the Karolinum Press Prague, Czech Republic.

Kosorok, M. R. and Laber, E. B. (2019). Precision medicine. *Annual review of statistics and its application*, 6:263–286.

Lachin, J. M. and Foulkes, M. A. (1986). Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics*, pages 507–519.

Lakatos, E. (1988). Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics*, pages 229–241.

Lyons, R. (2013). Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284–3305.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4):719–748.

Matabuena, M., Félix, P., Ditzhaus, M., Vidal, J., and Gude, F. (2023). Hypothesis testing for matched pairs with missing data by maximum mean discrepancy: An application to continuous glucose monitoring. *The American Statistician*, 77(4):357–369.

Melero, I., Gaudernack, G., Gerritsen, W., Huber, C., Parmiani, G., Scholl, S., Thatcher, N., Wagstaff, J., Zielinski, C., Faulkner, I., et al. (2014). Therapeutic vaccines for cancer: an overview of clinical trials. *Nature Reviews Clinical Oncology*, 11:509.

Motzer, R. J., Escudier, B., McDermott, D. F., George, S., Hammers, H. J., Srinivas, S., Tykodi, S. S., Sosman, J. A., Procopio, G., Plimack, E. R., et al. (2015). Nivolumab versus everolimus in advanced renal-cell carcinoma. *New England Journal of Medicine*, 373(19):1803–1813.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141.

Neuhaus, G. et al. (1993). Conditional rank tests for the two-sample problem under random censorship. *The Annals of Statistics*, 21(4):1760–1779.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, pages 185–207.

Rachev, S. T., Klebanov, L. B., Stoyanov, S. V., and Fabozzi, F. (2013). *The methods of distances in the theory of probability and statistics*, volume 10. Springer.

Ramdas, A., Reddi, S. J., Póczos, B., Singh, A., and Wasserman, L. (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Rizzo, M. L., Székely, G. J., et al. (2010). Disco analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4(2):1034–1055.

Robertson, T., Dykstra, R., and Wright, F. T. (1988). *Order restricted statistical inference*. Number . Probability and mathematical statistics in Wiley series in probability and mathematical statistics. Wiley.

Rodriguez, P. C., Popa, X., Martínez, O., Mendoza, S., Santiesteban, E., Crespo, T., Amador, R. M., Fleytas, R., Acosta, S. C., Otero, Y., et al. (2016). A phase iii clinical trial of the epidermal growth factor vaccine cimavax-egf as switch maintenance therapy in advanced non–small cell lung cancer patients. *Clinical Cancer Research*, 22(15):3782–3790.

Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, 68(1):316–319.

Schumacher, M. (1984). Two-sample tests of cramér–von mises-and kolmogorov–smirnov-type for randomly censored data. *International Statistical Review/Revue Internationale de Statistique*, pages 263–281.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291.

Shen, C., Panda, S., and Vogelstein, J. T. (2022). The chi-square test of distance correlation. *Journal of Computational and Graphical Statistics*, 31(1):254–262. PMID: 35707063.

Shen, C. and Vogelstein, J. T. (2018). The exact equivalence of distance and kernel methods for hypothesis testing. *arXiv preprint arXiv:1806.05514*.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561.

Stablein, D. M., Carter Jr, W. H., and Novak, J. W. (1981). Analysis of survival data with nonproportional hazard functions. *Controlled clinical trials*, 2(2):149–159.

Stute, W. (1994a). The bias of kaplan-meier integrals. *Scandinavian Journal of Statistics*.

Stute, W. (1994b). The bias of kaplan-meier integrals. *Scandinavian Journal of Statistics*, 21(4):475–484.

Stute, W. (2003). Kaplan–meier integrals. In *Advances in Survival Analysis*, volume 23 of *Handbook of Statistics*, pages 87 – 104. Elsevier.

Stute, W. and Wang, J.-L. (1993). Multi-sample u-statistics for censored data. *Scandinavian journal of statistics*, pages 369–374.

Su, Z. and Zhu, M. (2018). Is it time for the weighted log-rank test to play a more important role in confirmatory trials? *Contemporary Clinical Trials Communications*, 10:A1.

Székely, G. (2003). E-statistics: The energy of statistical samples. *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, 3(05):1–18.

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272.

Székely, G. J. and Rizzo, M. L. (2017). The energy of data. *Annual Review of Statistics and Its Application*, 4(1):447–479.

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272.

Tarone, R. E. and Ware, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika*, 64(1):156–160.

Wang, J.-G. et al. (1987). A note on the uniform consistency of the kaplan-meier estimator. *The Annals of Statistics*, 15(3):1313–1316.

Wang, R., Lagakos, S. W., and Gray, R. J. (2010). Testing and interval estimation for two-sample survival comparisons with small sample sizes and unequal censoring. *Biostatistics*, 11(4):676–692.

Xu, Z., Park, Y., Zhen, B., and Zhu, B. (2018). Designing cancer immunotherapy trials with random treatment time-lag effect. *Statistics in Medicine*.

Xu, Z., Zhen, B., Park, Y., and Zhu, B. (2017). Designing therapeutic cancer vaccine trials with delayed treatment effect. *Statistics in Medicine*, 36(4):592–605.

Yang, S. and Prentice, R. (2010). Improved logrank-type tests for survival data using adaptive weights. *Biometrics*, 66(1):30–38.