A class of two-sample nonparametric statistics for binary and time-to-event outcomes

Marta Bofill Roig[1,2] [iD] and **Guadalupe Gómez Melis[1]** [iD]

Abstract

We propose a class of two-sample statistics for testing the equality of proportions and the equality of survival functions. We build our proposal on a weighted combination of a score test for the difference in proportions and a weighted Kaplan–Meier statistic-based test for the difference of survival functions. The proposed statistics are fully non-parametric and do not rely on the proportional hazards assumption for the survival outcome. We present the asymptotic distribution of these statistics, propose a variance estimator, and show their asymptotic properties under fixed and local alternatives. We discuss different choices of weights including those that control the relative relevance of each outcome and emphasize the type of difference to be detected in the survival outcome. We evaluate the performance of these statistics with small sample sizes through a simulation study and illustrate their use with a randomized phase III cancer vaccine trial. We have implemented the proposed statistics in the R package `SurvBin`, available on GitHub (https://github.com/MartaBofillRoig/SurvBin).

Keywords

Clinical trials, mixed outcomes, multiple endpoints, non-proportional hazards, survival analysis, weighted Mean survival test

Introduction

In many clinical studies, two or more endpoints are investigated aiming to provide a comprehensive picture of the treatment's benefits and harms. Survival analysis has often been the sharp focus of clinical trial research. However, when there is more than one event of interest, the time until the appearance of the event is not always the unique center of attention; often the occurrence of an event over a fixed time period is as well an outcome of interest.

In the context of cancer immunotherapies trials, short-term binary endpoints based on the tumor size, such as objective response, are common in early-phase trials, whereas overall survival remains the gold standard in late-phase trials.[1,2] Since traditional oncology endpoints may not capture the clinical benefit of cancer immunotherapies, the idea of looking at both tumor response and survival has grown from the belief that together may achieve a better characterization of the clinical response.[3]

Several authors have considered both objective response and overall survival as primary endpoints in cancer trials. Lai and Zee[4] proposed a single-arm phase II trial design with a tumor response rate and a time-to-event outcome, such as overall survival or progression progression-free survival. In their design, the dependence between the binary response and the time-to-event outcome is modeled through a Gaussian copula. Lai et al.[5] proposed a two-step sequential design in which the response rate and the time to the event are jointly modeled. Their approach relates the response rate and the time to the event by means of a mixture model build on the basis of the Cox proportional hazards model assumption.

[1]Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain
[2]Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria

Corresponding author:
Marta Bofill Roig, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria.
Email: marta.bofillroig@meduniwien.ac.at

Chen and Wang[6] presented a joint model for binary marker responses and survival outcomes for clustered data. They based the statistical inference on a multivariate penalized likelihood method and estimated the standard errors using a jackknife resampling method.

An additional challenge in immunotherapy trials lies in the fact that delayed effects are likely to be found, bringing the need of alternative methods accounting for the non-proportionality of the hazards.[7] Statistics that look at differences between integrated weighted survival curves, such as those defined by Pepe and Fleming[8,9] and extended by Gu et al.[10], are better suited to detect early or late survival differences and do not depend on the proportional hazards assumption. In this paper, we aim to propose a class of two-sample statistics that could be used in phase II/III design to jointly evaluate the efficacy on binary and survival endpoints, even in the presence of delayed treatment effects.

The problem of how to analyze multiple outcomes has been widely discussed in the literature.[11,12] Depending on the inferential goals, this problem can be addressed in different ways. When the trial's objective is to test the null hypothesis of no effect on any of the endpoints, one can use the univariate tests together with a multiplicity adjustment to control the inflation of the type I error. The most popular multiple testing procedures (e.g., Bonferroni procedure[13]) do not make any assumption regarding the joint distribution of the univariate tests, and therefore may lead to conservative designs when the endpoints are correlated. When aiming at demonstrating that the treatment has an effect across the endpoints without necessarily specifying the minimum effect on any of the endpoints, one can consider methods based on a combination of the univariate statistics, which generally require assumptions about the joint test distribution.

Several approaches have been developed allowing for the joint distribution of test statistics. O'Brien[14] and Pocock et al.[15] proposed global test statistics through the sum of individual statistics. O'Brien[14] developed a generalized least squares method by combining multiple statistics into a single hypothesis test when variables are normally distributed; whereas Pocock et al.[15] extended O'Brien's work to asymptotically normal test statistics. Hothorn et al.[16] and Pipper et al.[17] approached the problem of testing multiple hypotheses using parametric and semi-parametric models. Hothorn et al.[16] used the limiting distribution of the parameter estimators to build upon the corresponding test statistics and their joint distribution. Based on that, their approach corrects the significance level by means of the simultaneous asymptotic normality of the test statistics. Pipper et al.[17] proposed a procedure for evaluating the efficacy in trials with multiple endpoints of different types. Their procedure is based on simultaneous asymptotic normality of the effect estimators from the single models for each endpoint together with multiple testing adjustments.

Extensive research has been done on joint modeling of longitudinal measurements and survival data (comprehensive overviews can be found in Tsiatis and Davidian,[18] Rizopoulos[19] and Papageorgiou et al.[20]). In most cases, the primary focus is on characterizing the association between the longitudinal and event time processes. The common framework is to relate the time-to-event and longitudinal outcomes through the proportional hazard model. Nevertheless, the relationship between binary response at a specific time point and survival outcome has received less attention.[6]

In this paper, we have followed the idea launched by Pocock et al.[15] of combining multiple test statistics into a single hypothesis test. Specifically, we propose a class of statistics based on a weighted sum of a difference in proportions test and a weighted Kaplan–Meier test-based on the difference of survival functions to evaluate an overall effect on the binary and survival endpoints. Our proposal adds versatility to the study design by enabling different follow-up periods for each endpoint, and flexibility by incorporating weights. We define these weights to specify unequal priorities to the different endpoints and to anticipate the type of time-to-event difference to be detected.

This article is organized as follows. In Section 'A general class of binary and survival test statistics', we present the class of statistics for binary and time-to-event outcomes. In Section 'Large sample results', we set out the assumptions and present the large sample distribution theory for the proposed statistics. In Section 'On the choice of weights', we introduce different weights and discuss their choice. We give an overview of our R package `SurvBin` in Section 'Implementation' and illustrate our proposal with a recent immunotherapy trial in Section 'Example'. In Section 'Simulation study', we evaluate the performance of these statistics in terms of the significance level and the statistical power with a simulation study. We conclude with a discussion.

All the required functions to use these statistics have been implemented in R and have been made available at: https://github.com/MartaBofillRoig/SurvBin.

A general class of binary and survival test statistics

Consider a randomized controlled trial comparing two treatment groups, control group ($i = 0$) and intervention group ($i = 1$), each composed of $n^{(i)}$ individuals, and denote by $n = n^{(0)} + n^{(1)}$ the total sample size. Suppose that both groups are followed over the time interval $[0, \tau]$ and are compared on the basis of the following two endpoints: the occurrence of an event ε_b before τ_b ($0 < \tau_b \leq \tau$), and the time to a different event ε_s within the interval $[\tau_0, \tau]$ ($0 \leq \tau_0 < \tau$). For the

i-th group ($i = 0, 1$), let $p^{(i)}(\tau_b)$ be the probability of having the event ε_b before τ_b, and $S^{(i)}(\cdot)$ be the survival function of the time to the event ε_s.

We consider the problem of testing simultaneously $H_{b,0}$: $p^{(0)}(\tau_b) = p^{(1)}(\tau_b)$ and $H_{s,0}$: $S^{(0)}(t) = S^{(1)}(t)$, $\forall t \in [\tau_0, \tau]$, aiming to demonstrate an overall effect across the binary and survival endpoints, either by a higher probability of the occurrence of ε_b, $H_{b,1}$: $p^{(0)}(\tau_b) < p^{(1)}(\tau_b)$, or an improved survival with respect to ε_s in the intervention group, $H_{s,1}$: $S^{(0)}(t) \leq S^{(1)}(t)$, $\forall t \in [\tau_0, \tau]$, $S^{(0)}(\cdot) \neq S^{(1)}(\cdot)$. The hypothesis problem can then be formalized as follows:

$$\begin{cases} H_0: & p^{(0)}(\tau_b) = p^{(1)}(\tau_b) \text{ and } S^{(0)}(t) = S^{(1)}(t), \ \forall t \in [\tau_0, \tau] \\ H_1: & p^{(0)}(\tau_b) < p^{(1)}(\tau_b) \text{ or } S^{(0)}(t) \leq S^{(1)}(t), \ \forall t \in [\tau_0, \tau], \\ & \exists t^* \in [\tau_0, \tau], \ S^{(0)}(t^*) < S^{(1)}(t^*) \end{cases} \tag{1}$$

We propose a class of statistics —hereafter called $\mathcal{L}$-class— as a weighted linear combination of the difference of proportions statistic for the binary outcome and the integrated weighted difference of two survival functions for the time-to-event outcome, as follows:

$$\mathbf{U}_n^\omega(\tau_0, \tau_b, \tau; \hat{Q}) = \omega_b \cdot \frac{U_{b,n}(\tau_b)}{\hat{\sigma}_b} + \omega_s \cdot \frac{U_{s,n}(\tau_0, \tau; \hat{Q})}{\hat{\sigma}_s} \tag{2}$$

for some real numbers $\omega_b, \omega_s \in (0, 1)$, such that $\omega_b + \omega_s = 1$, and where:

$$U_{b,n}(\tau_b) = \sqrt{\frac{n^{(0)}n^{(1)}}{n}} \left(\hat{p}^{(1)}(\tau_b) - \hat{p}^{(0)}(\tau_b) \right) \tag{3}$$

$$U_{s,n}(\tau_0, \tau; \hat{Q}) = \sqrt{\frac{n^{(0)}n^{(1)}}{n}} \left(\int_{\tau_0}^\tau \hat{Q}(t) \cdot \left(\hat{S}^{(1)}(t) - \hat{S}^{(0)}(t) \right) dt \right) \tag{4}$$

denoting by $\hat{p}^{(i)}(\tau_b)$ the estimated proportion of events ε_b before τ_b, and by $\hat{S}^{(i)}(\cdot)$ the Kaplan–Meier estimator of $S^{(i)}(\cdot)$ for group i. The estimates $\hat{\sigma}_b^2$ and $\hat{\sigma}_s^2$ are such that converge in probability to σ_b^2 and σ_s^2, respectively, as $n \to +\infty$, where σ_b^2 and σ_s^2 represent the asymptotic variances of $U_{b,n}(\tau_b)$ and $U_{s,n}(\tau_0, \tau; \hat{Q})$, respectively. Both theoretical and estimated expressions for the variances of $U_{b,n}(\tau_b)$ and $U_{s,n}(\tau_0, \tau; \hat{Q})$ will be given in Section 'Large sample results' (see equations (5), (6) for the theoretical expressions and (10), (11) for the estimates). The term $\hat{Q}(\cdot)$ is a possibly random function that converges pointwise in probability to a deterministic function $Q(\cdot)$. For ease of notation, and letting $\omega = (\omega_b, \omega_s)$, we will suppress the dependence on τ_0, τ_b, τ and use instead $\mathbf{U}_n^\omega(\hat{Q})$, $U_{b,n}$, $U_{s,n}(\hat{Q})$. Note that $\hat{p}^{(i)}(\tau_b)$, $\hat{S}^{(i)}(\cdot)$, $\hat{\sigma}_b$ and $\hat{\sigma}_s$ depend on the sample size $n^{(i)}$, but it has been omitted in notation for short.

The weights ω control the relative relevance of each outcome—if any—and the random weight function $\hat{Q}(\cdot)$ serves two purposes: to specify the type of survival difference that may exist between groups and to stabilize the variance of the difference of the two Kaplan–Meier functions. Some well-known special cases of $\hat{Q}(\cdot)$ are follows:

(i) $\hat{Q}(t) = \hat{G}(t-)$, where $\hat{G}(t-)$ is the pooled Kaplan–Meier estimator for the censoring distribution. This choice of $\hat{Q}(t)$ down-weights the contributions on those times where the censoring is heavy.

(ii) $\hat{Q}(t) = \hat{S}(t-)^\rho \cdot (1 - \hat{S}(t-))^\gamma$, where $\rho, \gamma \geq 0$ and $\hat{S}(t-)$ is the pooled Kaplan–Meier estimator for the survival function. This $\hat{Q}(t)$ corresponds to the weights of the Fleming–Harrington $G^{\rho,q}$ family.[21] Then, for instance, if $\rho = 1$ and $\gamma = 0$, $\hat{Q}(t)$ emphasizes early differences between survival functions; whereas late differences could be highlighted with $\rho = 0$ and $\gamma = 1$.

(iii) $\hat{Q}(t) = \bar{Y}(t-)$, where $\bar{Y}(t-)$ denotes the number of individuals at risk of ε_s at time t. In this case, $\hat{Q}(t)$ accentuates the information at the beginning of the survival curve allowing early failures to receive more weight than later failures.

We state the precise conditions for the weight function $\hat{Q}(\cdot)$ in Section 'Large sample results' and postpone the discussion about the choice of $\hat{Q}(\cdot)$ and $\omega = (\omega_b, \omega_s)$ to Section 'On the choice of weights'.

The statistics in the $\mathcal{L}$-class are defined for possible different follow-up configurations based on different choices of: the overall follow-up period, τ; the time where the binary event is evaluated, τ_b; and the origin time for the survival outcome, τ_0; taking into account that $0 < \max\{\tau_0, \tau_b\} < \tau$. There are however no restrictions on whether or not these periods overlap and, if they do, how much and when. We illustrate two different situations with different configurations for τ_0, τ_b, τ in Figure 1. The first case is exemplified by an HIV therapeutic vaccination study where safety-tolerability response (binary outcome) and time-to-viral rebound (survival outcome) are outcomes of interest. Whereas the safety–tolerability
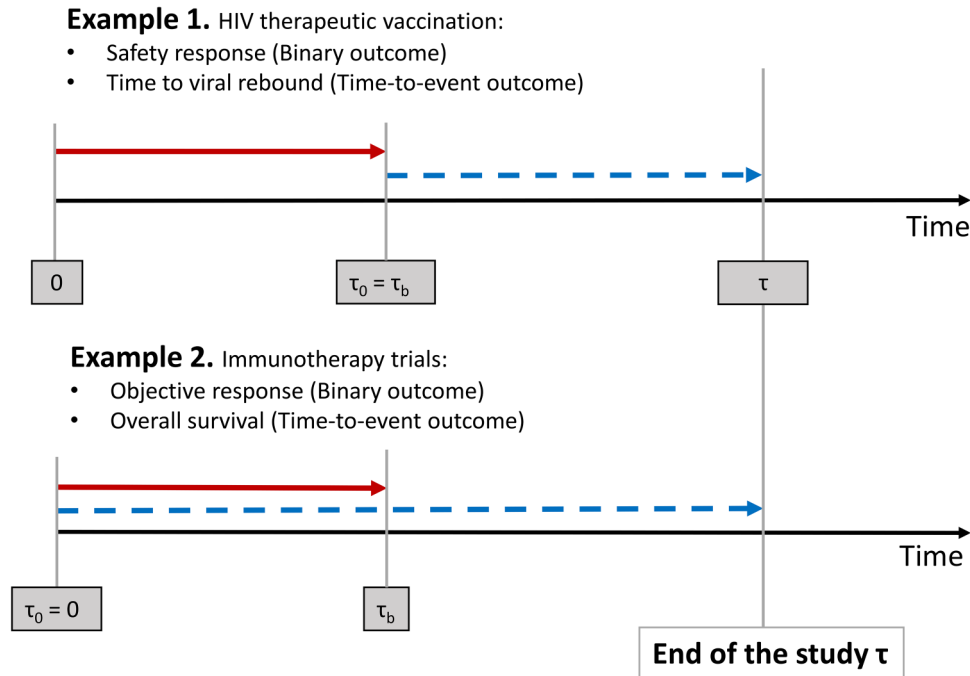
Example 1. HIV therapeutic vaccination:
- Safety response (Binary outcome)
- Time to viral rebound (Time-to-event outcome)

Example 2. Immunotherapy trials:
- Objective response (Binary outcome)
- Overall survival (Time-to-event outcome)

Figure 1. Illustration of two different follow-up configurations, the red and blue arrows represent the time-frame for binary and time-to-event outcomes, respectively. The red line goes from the start of the study (at time-point 0) until the binary outcome is evaluated at time τ_b. The blue (dashed) line goes from when the time-to-event information begins to be collected (τ_0) to the end of the study (τ).

is evaluated at week 6 ($\tau_b = 6$), the time-to-viral rebound is evaluated from week 6 to week 18 ($\tau_0 = 6$ and $\tau = 18$).[22] The second example in the area of immunotherapy trials includes a binary outcome (objective response), evaluated at month 6, and overall survival, evaluated from randomization until year 4 ($\tau_0 = 0$, $\tau_b = 0.5$ and $\tau = 4$).[23]

The $\mathcal{L}$-class statistics includes several statistical tests. If $\tau_0 = 0$, $\tau_b = \tau$ and $\omega_b = \omega_s$, then, $\mathbf{U}_n^\omega(\hat{Q})$ corresponds to the global test statistic proposed by Pocock et al.[15]. If $\varepsilon_b = \varepsilon_s$, $\tau_0 = \tau_b$, and $\omega_b = \omega_s$, the statistic $\mathbf{U}_n^\omega(\hat{Q})$ is the equivalent of the linear combination test of Logan et al.[24] when there is no censorship until τ_b for testing for differences in survival curves after a pre-specified time-point.

Large sample results

In this section, we derive the asymptotic distribution of the $\mathcal{L}$-class of statistics given in (2) under the null hypothesis and under contiguous alternatives, present an estimator of their asymptotic variance, and discuss the consistency of the $\mathcal{L}$-statistics against any alternative hypothesis of the form of H_1 in (1). We start the section with the conditions we require for the $\mathcal{L}$-class of statistics. To make the paper more concise and more readable, proofs and technical details are in the Supplemental material. Moreover, stratified $\mathcal{L}$-tests are also presented and discussed in the Supplemental material.

Further notation and assumptions

We consider two independent random samples of $n^{(i)}$ ($i = 0, 1$) individuals and for each we denote the binary response by $X_{ij} = I\{\varepsilon_b \text{ has occurred}\}$, the time to ε_s by T_{ij} and the censoring time by C_{ij} for $j = 1, \ldots, n^{(i)}$ and where $I\{\cdot\}$ is the usual 0/1 indicator function. Assume that T_{ij} is non-informatively right-censored by C_{ij}, that X_{ij} is independent of C_{ij}, and that the occurrence of the survival and censoring times, T_{ij} and C_{ij}, does not prevent to assess the binary response, X_{ij}. The observable data are summarized by $\{X_{ij}, T_{ij} \wedge C_{ij}, \delta_{ij}\}$, where $\delta_{ij} = I\{T_{ij} \wedge C_{ij} = T_{ij}\}$ and $a \wedge b = \min(a, b)$.

Denote by $G^{(i)}(\cdot)$ and $\hat{G}^{(i)}(\cdot)$ the censoring survival function and the Kaplan–Meier estimator for the censoring times, respectively. As we will see in the next section, the distribution of the $\mathcal{L}$-statistics relies, among others, on the survival function for those individuals who respond ($X_{ij} = 1$) to the binary endpoint. We then introduce here the survival function for responders as $S_X^{(i)}(t) = P(T_{ij} > t | X_{ij} = 1)$ ($t > \tau_b$).

Furthermore we assume that: (i) at the end of follow-up, $S^{(i)}(\tau) > 0$, $S_X^{(i)}(\tau) > 0$, and $G^{(i)}(\tau) > 0$; (ii) the limiting fraction of the total sample size is non-negligible, that is, $\lim_{n \to +\infty} n^{(i)}/n = \pi^{(i)} \in (0, 1)$; and (iii) $Q(\cdot)$ is a nonnegative piecewise continuous with finitely discontinuity points. For all the continuity points in $[0, \tau]$, $\hat{Q}(t)$ converges in probability to $Q(t)$ as $n \to +\infty$. Moreover, $\hat{Q}(\cdot)$ and $Q(\cdot)$ are functions of total variation bounded in probability.

Finally, we introduce the counting process $\bar{N}^{(i)}(t) = \sum_{j=1}^{n^{(i)}} N_{ij}(t) = \sum_{j=1}^{n^{(i)}} I\{T_{ij} \wedge C_{ij} \leq t, \delta_{ij} = 1\}$ as the number of observed events that have occurred by time t for the i-th group ($i = 0, 1$) and $\bar{Y}^{(i)}(t) = \sum_{j=1}^{n^{(i)}} Y_{ij}(t) = \sum_{j=1}^{n^{(i)}} I\{T_{ij} \wedge C_{ij} \geq t\}$ as the number of subjects at risk at time t for the i-th group. We define $y^{(i)}(s) = E(Y_{ij}(s))$ and suppose that $y^{(i)}(\tau) > 0$.

Remark: Throughout the paper and to refer to the group i ($i = 0, 1$), we will use subindexes for the individual observations and stochastic processes, as in X_{ij}, while we will use superindexes in parentheses for the functions and parameters, as in $S^{(i)}(\cdot)$.

Asymptotic distribution

To derive the asymptotic distribution of the statistic $\mathbf{U}_n^\omega(\hat{Q})$, we use that $\mathbf{U}_n^\omega(\hat{Q})$ can be approximated by $\mathbf{U}_n^\omega(Q)$, the same statistic with the weights replaced by its deterministic function (see Lemma 1 in the Supplemental material). Roughly speaking, thanks to this approximation, we can ignore the randomness of $\hat{Q}(\cdot)$ and use $\mathbf{U}_n^\omega(Q)$ to obtain the limiting distribution of $\mathbf{U}_n^\omega(\hat{Q})$. In what follows, we state the asymptotic distributions under the null hypothesis in Theorem 1 and under a sequence of contiguous alternatives in Theorem 2.

Theorem 1 Let $\mathbf{U}_n^\omega(\hat{Q})$ be the statistic defined in (2). Under the conditions outlined in 'Further Notation and Assumptions', if the null hypothesis $H_0 : H_{s,0} \cap H_{b,0}$ holds, $\mathbf{U}_n^\omega(\hat{Q})$ converges in distribution, as $n \to +\infty$, to a normal distribution as follows:

$$\mathbf{U}_n^\omega(\hat{Q}) \to N\left(0, \omega_b^2 + \omega_s^2 + 2\omega_b\omega_s \cdot \frac{\sigma_{bs}}{\sigma_b \cdot \sigma_s}\right)$$

where σ_b^2, σ_s^2 stand for the asymptotic variances of $U_{b,n}$ and $U_{s,n}(Q)$, respectively, and σ_{bs} is the covariance between $U_{b,n}$ and $U_{s,n}(Q)$. Their corresponding expressions are given by:

$$\sigma_b^2 = \sum_{i=0,1} \left(1 - \pi^{(i)}\right) p^{(i)}(\tau_b)\left(1 - p^{(i)}(\tau_b)\right) \tag{5}$$

$$\sigma_s^2 = -\sum_{i=0,1} \left(1 - \pi^{(i)}\right) \int_{\tau_0}^{\tau} \frac{(K_\tau^{(i)}(t))^2}{(S^{(i)}(t))^2 G^{(i)}(t)} dS^{(i)}(t) \tag{6}$$

$$\sigma_{bs} = \sum_{i=0,1} \left(1 - \pi^{(i)}\right) \cdot \left(I\{\tau_{\max} = \tau_b\} \cdot \int_{\tau_0}^{\tau_b} \frac{K_{\tau_b}^{(i)}(t)}{S^{(i)}(t)} \cdot \left(p_N^{(i)}(t) - p^{(i)}(\tau_b)\right) dS^{(i)}(t) \right.$$

$$\left. + \int_{\tau_{\max}}^{\tau} \frac{K_\tau^{(i)}(t)}{S^{(i)}(t)} \cdot p^{(i)}(\tau_b)\left(dS_X^{(i)}(t) - -dS^{(i)}(t)\right)\right) \tag{7}$$

where $\tau_{\max} = \max(\tau_0, \tau_b)$, $K_{\tau_*}^{(i)}(t) = \int_t^{\tau_*} Q(u)S^{(i)}(u)du$ ($\tau_* = \tau$ or τ_b), $p_N^{(i)}(t) = P(X_{ij} = 1|dN_{ij}(t) = 1)$, and $S_X^{(i)}(t) = P(T_{ij} > t|X_{ij} = 1)$ for $i = 0, 1$.

Recall that σ_b^2, σ_s^2, and σ_{bs} depend on τ_0, τ_b, τ, but we omit them for notational simplicity.

Theorem 2 Let $\mathbf{U}_n^\omega(\hat{Q})$ be the statistic defined in (2). Under the conditions outlined in the "Further notation and assumptions" section, consider the following sequences of contiguous alternatives for both binary and time-to-event hypotheses satisfying, as $n \to +\infty$:

$$\sqrt{n}(p_n^{(1)} - p^{(0)}) \to g$$

and

$$\sqrt{n}(S_n^{(1)}(t) - S^{(0)}(t)) \to \mathcal{G}(t)$$

for some constant $g \in \mathbb{R}^+$ and bounded function $\mathcal{G}(\cdot) \in \mathbb{R}^+$, and $\forall t \in [\tau_0, \tau]$. Then, under contiguous alternatives of the

form:

$$\mathrm{H}_{1,n}: \sqrt{n}(p_n^{(1)} - p^{(0)}) = g \text{ and } \sqrt{n}(S_n^{(1)}(t) - S^{(0)}(t)) = \mathcal{G}(t), \quad \forall t \in [\tau_0, \tau] \tag{8}$$

we have that:

$$\mathbf{U}_n^\omega(\hat{Q}) \to N\left(\omega_b g + \omega_s \int_{\tau_0}^\tau Q(t)\mathcal{G}(t)dt, \omega_b^2 + \omega_s^2 + 2\omega_b\omega_s \frac{\sigma_{bs}}{\sigma_b \cdot \sigma_s}\right)$$

in distribution as $n \to +\infty$, where σ_b^2, σ_s^2 and σ_{bs} are given in (5)–(7), respectively.

The covariance in (7) involves the conditional probabilities $S_X^{(i)}(t)$ and $p_N^{(i)}(t)$, while $S_X^{(i)}(t)$ represents the survival function for responders –individuals that have had the binary event ε_b–, $p_N^{(i)}(t)$ stands for the probability of being a responder among individuals experiencing ε_s at t. Also note that, if $\tau_b < \tau_0$, the survival experience starts after the binary event has been evaluated and only involves the second integral in (7).

We notice that the efficiency of the $\mathcal{L}$-statistics, $\mathbf{U}_n^\omega(\hat{Q})$, under contiguous alternatives is driven by the non-centrality parameter $\mu_c = \omega_b g + \omega_s \int_{\tau_0}^\tau Q(t)\mathcal{G}(t)\mathrm{d}t$, that is, by the sum of the weighted non-centrality parameters of $U_{b,n}$ and $U_{s,n}(\hat{Q})$.

Variance estimation and consistency

We now describe how to use the $\mathcal{L}$-statistics to test H_0 versus H_1 given in (1). In particular, we propose a consistent estimator of the asymptotic variance of $\mathbf{U}_n^\omega(\hat{Q})$, and present the standardized $\mathcal{L}$-statistics to test $\mathrm{H}_0 : \mathrm{H}_{s,0} \cap \mathrm{H}_{b,0}$.

The asymptotic variance of $\mathbf{U}_n^\omega(\hat{Q})$, given in Theorem 1, can be consistently estimated by:

$$\widehat{\mathrm{Var}}(\mathbf{U}_n^\omega(\hat{Q})) = \omega_b^2 + \omega_s^2 + 2\omega_b\omega_s \frac{\hat{\sigma}_{bs}}{\hat{\sigma}_b \cdot \hat{\sigma}_s} \tag{9}$$

where $\hat{\sigma}_b$, $\hat{\sigma}_s$, and $\hat{\sigma}_{bs}$ denote the estimates of σ_b, σ_s and σ_{bs}, and are given by:

$$\hat{\sigma}_b^2 = \hat{p}(\tau_b)\left(1 - \hat{p}(\tau_b)\right) \tag{10}$$

$$\hat{\sigma}_s^2 = -\int_{\tau_0}^\tau \frac{(\hat{K}_\tau(t))^2}{\hat{S}(t)\hat{S}(t-)} \cdot \frac{n^{(0)}\hat{G}^{(0)}(t-) + n^{(1)}\hat{G}^{(1)}(t-)}{\hat{G}^{(0)}(t-)\hat{G}^{(1)}(t-)} d\hat{S}(t) \tag{11}$$

$$\hat{\sigma}_{bs} = -\int_{\tau_0}^{\tau_b} \hat{K}_{\tau_b}(t)\left(\sum_{i=0,1} \frac{n - n^{(i)}}{n} \cdot \hat{\lambda}_{X,T}^{(i)}(t)dt + \frac{\hat{p}(\tau_b) \cdot d\hat{S}(t)}{\hat{S}(t)}\right)$$
$$+ \int_{\tau_b}^\tau \frac{\hat{K}_\tau(t) \cdot \hat{p}(\tau_b)}{\hat{S}(t-)}\left(-\frac{\hat{S}(t-) \cdot d\hat{S}(t)}{\hat{S}(t)} + \sum_{i=0,1} \frac{n - n^{(i)}}{n} \cdot \frac{\hat{S}_X^{(i)}(t-) \cdot d\hat{S}_X^{(i)}(t)}{\hat{S}_X^{(i)}(t)}\right) \tag{12}$$

where $\hat{K}_{\tau_*}(t) = \int_t^{\tau_*} \hat{Q}(u)\hat{S}(u)\mathrm{d}u$ $(\tau_* = \tau \text{ or } \tau_b)$, $\hat{S}(t)$ is the pooled Kaplan–Meier estimator of the survival functions, $\hat{p}(\tau_b)$ is the pooled estimator of the probabilities $p^{(i)}(\tau_b)$, $\hat{S}_X^{(i)}(t)$ is the Kaplan–Meier estimator of $S_X^{(i)}(t)$, and $\hat{\lambda}_{X,T}^{(i)}(t)$ is the estimator of $\lambda_{X,T}^{(i)}(t) = \lim_{\mathrm{d}t\to 0} P(X_{ij} = 1, t \le T_{ij} < t + \mathrm{d}t|T_{ij} > t)/\mathrm{d}t$.

The variance estimator presented in (9) is obtained assuming that the variances of the two groups are equal (pooled estimator). An unpooled variance estimator is proposed in the Supplemental material. For both pooled and unpooled estimators, smoothing techniques are used to estimate $\lambda_{X,T}^{(i)}(t)$ over the time period $[\tau_0, \tau_b]$. In this work, we have chosen kernel smoothing methods. Note that resampling methods can also be used to get an estimator of the variance of $\mathbf{U}_n^\omega(\hat{Q})$. In the simulation section, we will discuss the results using the pooled, unpooled, and bootstrap variance estimators.

To test the global null hypothesis $\mathrm{H}_0 : \mathrm{H}_{s,0} \cap \mathrm{H}_{b,0}$ in (1), we consider the normalized statistic of $\mathbf{U}_n^\omega(\hat{Q})$:

$$\mathbf{U}_n^\omega(\hat{Q})\Big/\sqrt{\widehat{\mathrm{Var}}(\mathbf{U}_n^\omega(\hat{Q}))} \tag{13}$$

Because this statistic (13) converges in distribution to a standard normal distribution, it can be used to test $\mathrm{H}_0 : \mathrm{H}_{s,0} \cap \mathrm{H}_{b,0}$ by comparing $\mathbf{U}_n^\omega(\hat{Q})\big/\sqrt{\widehat{\mathrm{Var}}(\mathbf{U}_n^\omega(\hat{Q}))}$ to a standard normal distribution. Moreover, for positive $Q(\cdot)$, the statistic is consistent against any alternative hypothesis of the form of H_1 in (1).

On the choice of weights

An important consideration when applying the statistics proposed in this paper is the choice of the weight functions. The $\mathcal{L}$-class of statistics involves the already mentioned random weight function $\hat{Q}(t)$ and deterministic weights $\omega = (\omega_b, \omega_s)$. These weights are defined according to different purposes and have different roles in the statistic $\mathbf{U}_n^\omega(Q)$. In this section, we include different weights and discuss some of their strengths as well as shortcomings. The list provided is not exhaustive, other weights are possible and might be useful in specific circumstances.

Choice of $\omega = (\omega_b, \omega_s)$

The purpose of the weights ω is to prioritize the binary and the time-to-event outcomes. They have to be specified in advance according to the research questions. Whenever the two outcomes are equally relevant, we should choose $\omega_b = \omega_s = 0.5$. In this case, the statistics will be optimal whenever the standardized effects on both outcomes coincide.

Choice of $\hat{Q}(\cdot)$

The choice of $\hat{Q}(\cdot)$ might be very general as long as $\hat{Q}(\cdot)$ converges in probability to a function $Q(\cdot)$, and both $\hat{Q}(\cdot)$ and $Q(\cdot)$ satisfy the conditions outlined in the 'Further notation and assumptions' section. In this section, we center our attention on a family of $\hat{Q}(\cdot)$ weights of the form:

$$\hat{Q}(t) = \hat{f}(t) \cdot \hat{v}(t),$$

where: (i) $\hat{f}(\cdot)$ is a data-dependent function that converges, in probability to $f(\cdot)$, a nonnegative piecewise continuous function with bounded variation on [0, 1]. The term $\hat{f}(t)$ takes care of the expected differences between survival functions and can be used as well to emphasize some parts of the follow-up according to the time-points (τ_0, τ_b, τ_s); (ii) the weights $\hat{v}(\cdot)$ converge in probability to a deterministic positive bounded weight function $v(\cdot)$. The main purpose of the weight $\hat{v}(t)$ is to ensure the stability of the variance of the difference of the two Kaplan–Meier functions. To do so, we make the additional assumption that:

$$|v(t)| \leq \Gamma \cdot G^{(i)}(t)^{1/2+\delta} \quad \text{and} \quad |\hat{v}(t)| \leq \Gamma \cdot \hat{G}^{(i)}(t)^{1/2+\delta}$$

for all $t \in [\tau_0, \tau]$, $i = 0, 1$ and for some constants $\Gamma, \delta > 0$.

Different choices of $\hat{f}(t)$ yield other known statistics. For instance, if $f(\cdot) = 1$, $U_{s,n}(\hat{Q})$ corresponds to the Weighted Kaplan–Meier statistics.[8,9] Whenever $\hat{f}$ and $\hat{v}$ correspond to the weights (15) and (14), respectively, introduced below, we have the statistic proposed by Shen and Cai.[25] Furthermore, note that the weight functions of the form $\hat{Q}(t) = \hat{f}(t) \cdot \hat{v}(t)$ are similar to those proposed by Shen and Cai[25]; while they assume that $\hat{f}$ is a bounded continuous function, we assume that $\hat{f}(\cdot)$ is a nonnegative piecewise continuous function with bounded variation on [0, 1], and instead of only considering the Pepe and Fleming weight function corresponding to (15), we also allow for different weight functions $\hat{v}(t)$. Finally, if we do not consider any weight, that is, if $\hat{Q}(t) = 1, \forall t, U_{s,n}(\hat{Q})$ corresponds to the difference of restricted mean survival times from τ_0 to τ.

In what follows, we outline different choices of $\hat{v}(t)$ and $\hat{f}(t)$, together with a brief discussion for each one:

- We require $\hat{v}(t)$ to be small towards the end of the observation period if censoring is heavy. The usual weight functions $\hat{v}(t)$ involve Kaplan–Meier estimators of the censoring survival functions. The most common weight functions are as follows:

$$\hat{v}_c(t) = \frac{n\hat{G}^{(0)}(t-)\hat{G}^{(1)}(t-)}{n^{(0)}\hat{G}^{(0)}(t-) + n^{(1)}\hat{G}^{(1)}(t-)} \tag{14}$$

and $\hat{v}_{\sqrt{}}(t) = \sqrt{\hat{v}_c(t)}$, both proposed by Pepe and Fleming.[8] Among other properties, $\hat{v}_c(\cdot)$ has been proved to be a competitor to the log-rank test for the proportional hazards alternative.[8] Note that if the censoring survival functions are equal for both groups and the sampling design is balanced ($n^{(0)} = n^{(1)}$), then, the differences in Kaplan–Meier estimators are weighted by the censoring survival function, that is, $w(t) = C(t) = C^{(i)}(t)$ for $i = 0, 1$. Also note that $w(t) = 1$ for uncensored data.

- Analogously to Fleming and Harrington[21] statistics, $\hat{f}(t)$ could be used to specify the type of expected differences between survival functions. That is, if we set:

$$f(\hat{S}(t-)) = \hat{S}(t-)^{\rho}(1 - \hat{S}(t-))^{\gamma}, \quad \rho, \gamma \geq 0 \tag{15}$$

the choice $\rho > 0$, $\gamma = 0$ leads to a test to detect early differences, while $\rho = 0$, $\gamma > 0$ leads to a test to detect late differences; and $\rho = \gamma = 0$ leads to a test evenly distributed over time and corresponds to the weight function of the log-rank.

- To put more emphasis on those times after the binary follow-up period we might consider:

$$f(t) = \begin{cases} a, & t < \tau_b \\ 1 - a, & t \geq \tau_b \end{cases}$$

for $a < 0.5$.

Implementation

We have developed the `SurvBin` package to facilitate the use of the $\mathcal{L}$-statistics and is now available on GitHub (https://github.com/MartaBofillRoig/SurvBin). The `SurvBin` package contains two key functions: `lstats` to compute the standardized $\mathcal{L}$-statistic, $\mathbf{U}_n^{\omega}(\hat{Q})\big/\sqrt{\widehat{\mathrm{Var}}(\mathbf{U}_n^{\omega}(\hat{Q}))}$, using the variance estimator given in Section 'Large sample results'; and `lstats_boots` to compute the standardized $\mathcal{L}$-statistic by using a bootstrap procedure to estimate the variance.

The `SurvBin` package also provides the functions `survbinCov` to calculate $\hat{\sigma}_{bs}$; and `bintest` and `survtest` to compute the univariate binary and survival statistics (3) and (4), $U_{b,n}(\tau_b)\big/\hat{\sigma}_b$ and $U_{s,n}(\tau_0, \tau; \hat{Q})\big/\hat{\sigma}_s$, respectively. In addition, the `SurvBin` package includes the function `simsurvbin` that can be used to simulate bivariate binary and survival data in a variety of situations.

The main function `lstats` can be called by:

```
lstats(time, status, binary, treat,
       tau0, tau, taub, rho, gam, eta, wb, ws, var_est)
```

(time, status, binary, treat,tau0, tau, taub, rho, gam, eta, wb, ws, var_est)

where `time`, `status`, `binary` and `treat` are vectors of the right-censored data, the status indicator, the binary data and the treatment group indicator, respectively; `tau0`, `tau`, `taub` denote the follow-up configuration; `wb`, `ws` are the weights ω; `rho`, `gam`, `eta` are scalar parameters that controls the weight $\hat{Q}(t)$, which is given by $\hat{Q}(t) = \hat{G}(t-)^{\eta} \cdot \hat{S}(t-)^{\rho} \cdot (1 - \hat{S}(t-))^{\gamma}$; and `var_est` indicates the variance estimate to use (`pooled` or `unpooled`).

In this work, we estimate $\lambda_{X,T}^{(i)}(t)$ by means of the Epanechnikov kernel function, and the local bandwidth selection and the boundary correction described by Muller and Wang[26] by using the `muhaz` package.[27]

Example

Melanoma has been considered a good target for immunotherapy and its treatment has been a key goal in recent years. Here we consider a randomized, double-blind, phase III trial whose primary objective was to determine the safety and efficacy of the combination of a melanoma immunotherapy (gp100) together with an antibody vaccine (ipilimumab) in patients with previously treated metastatic melanoma.[23] Despite the original endpoint was objective response rate at week 12, it was amended to overall survival and then considered a secondary endpoint. A total of 676 patients were randomly assigned to receive ipilimumab plus gp100, ipilimumab alone, or gp100 alone. The study was designed to have at least 90% power to detect a difference in overall survival between the ipilimumab-plus-gp100 and gp100-alone groups at a two-sided α level of 0.05, using a log-rank test. Cox proportional-hazards models were used to estimate hazard ratios and to test their significance. The results showed that ipilimumab with gp100 improved overall survival as compared with gp100 alone in patients with metastatic melanoma. However, the treatment had a delayed effect and an overlap between the Kaplan–Meier curves was observed during the first six months. Hence, the proportional hazards assumption appeared to be no longer valid, and a different approach would have been advisable.

To illustrate our proposal, we consider the comparison between the ipilimumab-plus-gp100 and gp100-alone groups based on the overall survival and objective response as multiple primary endpoints of the study. For this purpose, we have reconstructed individual observed times by scanning the overall survival Kaplan–Meier curves reported in
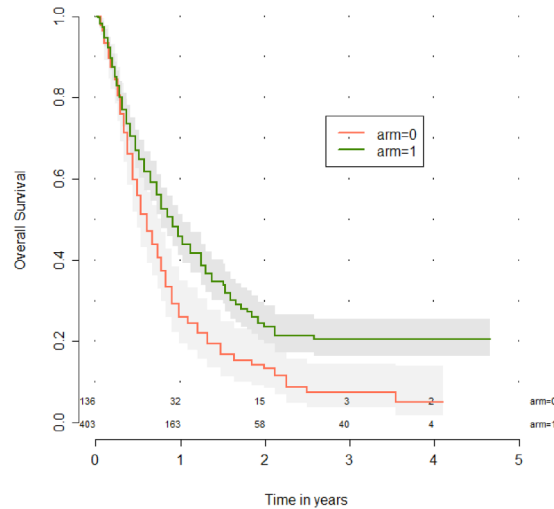
Figure 2. Kaplan–Meier curves for overall survival for ipilimumab-plus-gp100 and gp100-alone groups (arms 1 and 0, respectively).

Figure 1A of Hodi et al.[23] using the `reconstructKM` package[28] (see Figure 2), and, afterwards, we have simulated the binary response to mimic the percentage of responses obtained in the study.

Using the data obtained, we employ the $\mathcal{L}$-statistic by means of the function `lstats` in the `SurvBin` package. To do so, we need to specify the weights ($\hat{Q}$, ω) to be used, and the time-points (τ_0, τ_b, τ). In our particular case, we take $\tau_0 = 0$, $\tau_b = 0.5$, $\tau = 4$ according to the trial design, choose $\hat{Q}(t) = \hat{G}(t-) \cdot (1 - \hat{S}(t-))$ to account for censoring and delayed effects in late times, and (ω_b, ω_s) = (0.25, 0.75) to emphasize the importance of overall survival over objective response.

As shown below, the function `lstats` returns the standardized $\mathcal{L}$-statistic, together with the $\mathcal{L}$-statistic and its standard deviation, and the individual statistics.

```
## Parameter                     Value
## 1 (Standardized) L-Test  4.0950273
## 2               L-Test  3.2362929
## 3    Standard deviation  0.7902982
##
## $Binary_Tests
## Parameter                     Value
## Test Standardized L-Test  1.8678088
## Ub          Binary Test  0.4540763
## sd    Standard deviation  0.2431064
##
## $Survival_Tests
## Parameter                     Value
## Test  Standardized Test  3.6924543
## Us        Survival Test  2.4398019
## sd    Standard deviation  0.6607534
##
## $Covariance
## Parameter     Value
## 1 Covariance  −0.0001836297
```

The value of the $\mathcal{L}$-statistic, $\mathbf{U}_n^\omega(\hat{Q})$ in (2), is 3.24 and is obtained by using the values of $U_{b,n}(\tau_b)$ and $U_{s,n}(\tau_0, \tau; \hat{Q})$ (0.45 and 2.44, respectively), and $\hat{\sigma}_b$ and $\hat{\sigma}_s$ (0.24 and 0.66). The statistic $\mathbf{U}_n^\omega(\hat{Q})/\sqrt{\widehat{\mathrm{Var}}(\mathbf{U}_n^\omega(\hat{Q}))}$ equals 4.10 and is computed by using the variance estimator in (9) and then by means of $\hat{\sigma}_b$ and $\hat{\sigma}_{bs}$ together with the estimated covariance $\hat{\sigma}_{bs}$ (−0.0002).

Since we obtained $\mathbf{U}_n^\omega(\hat{Q})/\sqrt{\widehat{\mathrm{Var}}(\mathbf{U}_n^\omega(\hat{Q}))} = 4.10 > z_{\alpha=0.05}$, we have a basis to reject H_0 and conclude that there is an overall effect of ipilimumab across the binary endpoint of tumor reduction and overall survival in patients with metastatic melanoma. Note that, in this example, we have been using the pooled variance estimator for the $\mathcal{L}$-statistic. We have also

Table 1. Comparison of statistics (at 4 years) using different weights: $\mathcal{L}$-statistics with different combinations of (ω_b, ω_s); and weighted Kaplan–Meier (WKM) statistics and weighted log-rank (WLR) statistics with different (ρ, γ).

Statistics	Weights	Standardized test
$\mathcal{L}$-statistics	$(\omega_b, \omega_s) = (0.25, 0.75)$	4.11
$\mathcal{L}$-statistics	$(\omega_b, \omega_s) = (0.5, 0.5)$	3.93
$\mathcal{L}$-statistics	$(\omega_b, \omega_s) = (0.75, 0.25)$	2.94
WKM statistics	$(\rho, \gamma, \eta) = (0, 0, 0)$	3.70
WKM statistics	$(\rho, \gamma, \eta) = (0, 1, 0)$	3.69
WLR statistics	$(\rho, \gamma) = (0, 0)$	3.13
WLR statistics	$(\rho, \gamma) = (0, 1)$	2.75

calculated the statistic using the unpooled and bootstrap variance estimators (see Supplemental material) and notice that the results were not substantially different.

We also explored how much the results might vary if we change the weights assigned. Table 1 shows the values of the $\mathcal{L}$-statistic when we modify the weights ω. Moreover, we compared the values of the statistics for the survival endpoint when using the weighted Kaplan–Meier statistics and weighted log-rank statistics. We note no major differences when using weighted Kaplan–Meier tests with different weights, nor in the weighted log-rank tests. This is because the survivals are the same at the beginning of the curves, and therefore putting a higher weight at the end of the curve does not have as much influence as if there was a small effect at the beginning of the trial. On the other hand, we observe that the weighted Kaplan–Meier statistics take higher values than the weighted log-rank tests for all combinations of weights.

Simulation study

Design

We conducted a simulation study to evaluate our proposal in terms of the statistical power and the type-I error with small sample sizes. We generated bivariate binary and time-to-event data through a copula-based framework and used conditional sampling as described in Trivedi and Zimmer.[29]

The parameters used for the simulation (summarized in Table 2) have been the following:

- Clayton's copula with association parameter between the marginal distributions of the binary and time-to-event outcomes equal to $\theta = 0.001, 0.51, 0.91$. These values correspond, respectively, to Spearman's rank correlation values equal to 0.0002, 0.32, 0.45, which represent increasing associations between the binary and time-to-event outcomes. We have not considered higher values of θ as they do not fulfill the condition that $S_X^{(i)}(\tau) > 0$ $(i = 0, 1)$.
- Weibull survival functions, $S_{b,a}^{(0)}(t) = e^{-(t/b)^a}$, with $a = 0.5, 1, 2$ and $b = 1$.
- Probability of having the binary endpoint $p^{(0)} = 0.1, 0.3$.; and
- Sample size per arm $n^{(i)} = 250$ for $i = 0, 1$ and total sample size $n = 500$.
- The censoring distributions between groups were assumed equal and uniform $U(0, c)$ with $c = 3$.
- Two different follow-up configurations were considered for $\tau_0 < \tau_b \leq \tau$: (i) $\tau_0 = 0, \tau_b = 0.5, \tau = 1$; and (ii) $\tau_0 = 0, \tau_b = \tau = 1$.
- We have considered the weights: $\hat{Q}(t) = \hat{G}(t-)^\eta \cdot \hat{S}(t-)^\rho \cdot (1 - \hat{S}(t-))^\gamma$ with $\eta = 1$ and $\rho, \gamma = 0, 1$. When simulating under the null hypothesis, we considered (ω_b, ω_s) equal to $(0.5, 0.5)$; whereas when simulating under the alternative hypotheses, we considered (ω_b, ω_s) equal to $(0.25, 0.75)$, $(0.5, 0.5)$, and $(0.75, 0.25)$.

The simulations under the alternative hypothesis considered four different situations depending on whether there is a treatment effect on both endpoints and the type of difference between the survival curves. Specifically, the following cases were considered:

1. Effect on both binary and survival endpoints. The effect on the survival endpoint satisfies the proportional hazards assumption, that is, the hazard ratio (HR) between treatment groups is constant over the study duration;
2. Effect on the binary endpoint and non-effect on the survival endpoint ($H_{s,0} : S^{(0)}(t) = S^{(1)}(t), \ \forall t$);

Table 2. Scenarios used in the simulation study.

Parameter	Value	Parameter	Value
$p^{(0)}$	0.1, 0.3	a	0.5, 1, 2
b	1	c	3
θ	0.001, 0.51, 0.91	$n^{(i)}$ $(i = 0, 1)$	250
τ_b	0.5, 1	τ	1
ρ, γ	0, 1	η	1
d	0, 0.075	HR	0.75, 1
(ω_b, ω_s)	(0.25, 0.75),(0.5, 0.5),(0.75, 0.25)	t_*	0, 0.5

3. Non-effect on the binary endpoint ($H_{b,0} : p^{(0)}(\tau_b) = p^{(1)}(\tau_b)$) and effect on the survival endpoint with HR constant over the study duration;
4. Effect on both binary and survival endpoints. The treatment differences on the survival endpoint have a delayed effect, that is, the survival functions are assumed to be equal until time t_*, and there is a constant hazard ratio (HR) between treatment groups from t_* to τ.

We used $d = p^{(1)}(\tau_b) - p^{(0)}(\tau_b) = 0.075$ to simulate the effects on the binary endpoint. For the survival endpoint, we considered HR = 0.75 under proportional hazards, and HR = 0.70 and $t_* = 0.5$ under delayed effects.

We evaluated the empirical significance level and the statistical power using the $\mathcal{L}$-statistics with pooled, unpooled, and bootstrap variance estimators, and for the sake of the comparison, using the Bonferroni procedure. In addition, we presented the empirical results for testing the individual hypothesis $H_{b,0}$ and $H_{s,0}$ by using the statistics (3) and (4).

The total number of scenarios was 1504 (144 under the null hypothesis and 1360 under the alternative hypothesis). We ran 100, 000 replicates and estimated the significance level ($\alpha = 0.05$) for each scenario under the null hypothesis. We ran 1000 replicates and estimated the statistical power for each scenario under alternative hypotheses. We performed all computations using the R software (version 4.0.2).

Power properties

When there is a treatment effect on both endpoints and the proportional hazards assumption is fulfilled (case 1), we obtained empirical powers with medians 0.85, 0.85, 0.83 using the $\mathcal{L}$-statistics with pooled, unpooled, and bootstrap variance estimators, respectively; whereas the median of the empirical powers using Bonferroni was 0.80. When there is a treatment effect on both endpoints and there are delayed effects (case 4), the empirical powers for the $\mathcal{L}$-statistics have medians 0.73, 0.75, 0.70 using the pooled, unpooled, and bootstrap variance estimators, respectively; whereas the median of the empirical powers using Bonferroni was 0.63.

Table 3 summarizes the simulation results on the power across different parameters in case 1. We compared the performance of the different variance estimators and noticed that the empirical powers do not substantially differ between them. We also observed that the power is not affected by the different weight functions in the case of proportional hazards. We obtained higher powers when emphasizing late-differences between the survival curves ($\gamma = 1$) in the case of delayed effects (median powers of 0.85, 0.85, 0.84 for unpooled, pooled, and bootstrap variance estimators, respectively, with $\gamma = 1$ against 0.83, 0.83, 0.82 for unpooled, pooled, and bootstrap variance estimators with $\gamma = 0$).

Figure 3 shows boxplots for the empirical powers using the pooled, unpooled, and bootstrap variance estimators and the Bonferroni procedure. These simulations show the superiority of the $\mathcal{L}$-statistics over the Bonferroni procedure, in terms of power, both under proportional hazards and under delayed effects and regardless of the choice of the weights (ω_b, ω_s).

When there is a treatment effect on only one of the endpoints (cases 2 and 3), the behavior of the power mainly relies on the pre-specified weights (ω_b, ω_s) (see Figure 3). If the survival endpoint is considered clinically more important than the binary endpoint and we use the weights ($\omega_b = 0.25, \omega_s = 0.75$), then the median of the empirical powers is around 0.42 in case 2 (i.e., when there is treatment effect on the survival endpoint) and around 0.40 in case 3 (i.e., when there is no effect on the survival endpoint). We found a similar behavior when the binary endpoint is considered more important and ($\omega_b = 0.75, \omega_s = 0.25$).

If both endpoints are equally important and there is treatment effect in only one of them, the empirical powers using the $\mathcal{L}$-statistics take values between the power would have had used the two individual statistics. Given that the Bonferroni procedure assigns more importance to the more highly significant of the endpoints,[15] the powers are in this case higher using Bonferroni than using $\mathcal{L}$-statistics.

Table 3. Median empirical size and median empirical power from 100, 000 and 1000 replications, respectively. The empirical size and powers are calculated using: the $\mathcal{L}$-statistics (in (2)) according to the pooled, unpooled, bootstrap variance estimators (labeled as Pooled, Unpooled, and Boots.); and the Bonferroni procedure (Bonf.). Under the null hypothesis there is no effect on any of the endpoints ($d = 0$, HR$= 1$). Under the alternative hypothesis there is effect on both endpoints (Case 1: $d = 0.075$, HR$= 0.75$) and the effect on the survival endpoint satisfies the proportional hazards assumptions ($t_* = 0$).

		Empirical size				Empirical powers (Case 1)			
		Pooled	Unpooled	Boots.	Bonf.	Pooled	Unpooled	Boots.	Bonf.
(τ_b, τ)	(0.5, 1)	0.053	0.053	0.050	0.050	0.84	0.85	0.83	0.80
	(1, 1)	0.056	0.055	0.050	0.050	0.85	0.85	0.83	0.79
θ	0.001	0.051	0.051	0.052	0.051	0.87	0.88	0.87	0.82
	0.510	0.054	0.055	0.049	0.050	0.84	0.83	0.81	0.79
	0.910	0.056	0.056	0.048	0.049	0.82	0.82	0.79	0.78
$p^{(0)}$	0.1	0.053	0.054	0.050	0.051	0.89	0.89	0.88	0.84
	0.3	0.053	0.055	0.050	0.049	0.78	0.78	0.76	0.73
a	0.5	0.053	0.054	0.051	0.050	0.87	0.87	0.86	0.82
	1	0.053	0.054	0.050	0.050	0.84	0.85	0.83	0.80
	2	0.054	0.055	0.048	0.050	0.82	0.81	0.79	0.78
(ρ, γ, η)	(0,0,1)	0.054	0.053	0.051	0.050	0.84	0.85	0.83	0.80
	(0,1,1)	0.054	0.055	0.051	0.050	0.84	0.85	0.83	0.80
	(1,0,1)	0.053	0.055	0.050	0.050	0.83	0.82	0.80	0.79
	(1,1,1)	0.054	0.053	0.050	0.050	0.86	0.86	0.84	0.81

We have also evaluated the empirical powers if we have had an endpoint with a small effect instead of no effect. We observe that the difference in powers between the $\mathcal{L}$-statistics and Bonferroni procedure is smaller, and that the powers using $\mathcal{L}$-statistics could be even higher than the ones using Bonferroni (see Supplemental material). Furthermore, we have assessed how the association between the binary and survival endpoints might affect the results by considering Frank's copula instead of Clayton's copula. We notice that there is no substantial difference between the results using Frank's and Clayton's copulas. Results using Frank's copula can be found in the Supplemental material.

Size properties

The empirical results show that the type I error is very close to the nominal level $\alpha = 0.050$ across a broad range of situations. The empirical size resulted in type I errors with a median of 0.054, 0.053, and 0.050 using the unpooled, pooled, and bootstrap variance estimators, respectively. Table 3 summarizes the results according to the parameters of the simulation study. The results show that the $\mathcal{L}$-statistics have the appropriate size and are not specially influenced by the selection of weights (η, ρ, γ).

We observed that when using the unpooled and pooled estimators, the empirical size is slightly larger than 0.050, especially when $\tau_b = 1$. This can be explained mainly by the number of individuals at risk at the end of the follow-up. Having a small number of individuals make it difficult for the smooth estimation of the probability $p_N^{(i)}(t)$ in (7). Therefore, we recommend the use of the bootstrap variance estimator for studies with small sample sizes with long follow-ups for both the binary and survival endpoints and where the probability of observing the binary endpoint is low.

Discussion

We have proposed a class of statistics for a two-sample comparison based on two different outcomes: one dichotomous taking care, in most occasions, of short short-term effects, and a second one addressed to detect long long-term differences in a survival endpoint. Such statistics test the equality of proportions and the equality of survival functions. The approach combines a score test for the difference in proportions and a weighted Kaplan-Meier test-based for the difference of survival functions. The statistics are fully non-parametric and α level for testing the null hypothesis of no effect on any of these two outcomes. The statistics in the $\mathcal{L}$-class are appealing in situations when both outcomes are relevant, regardless of how the follow-up periods of each outcome are, and even when the hazards are not proportional with respect to the time-to-event outcome or in the presence of delayed treatment effects, albeit the survival curves are supposed not to cross.
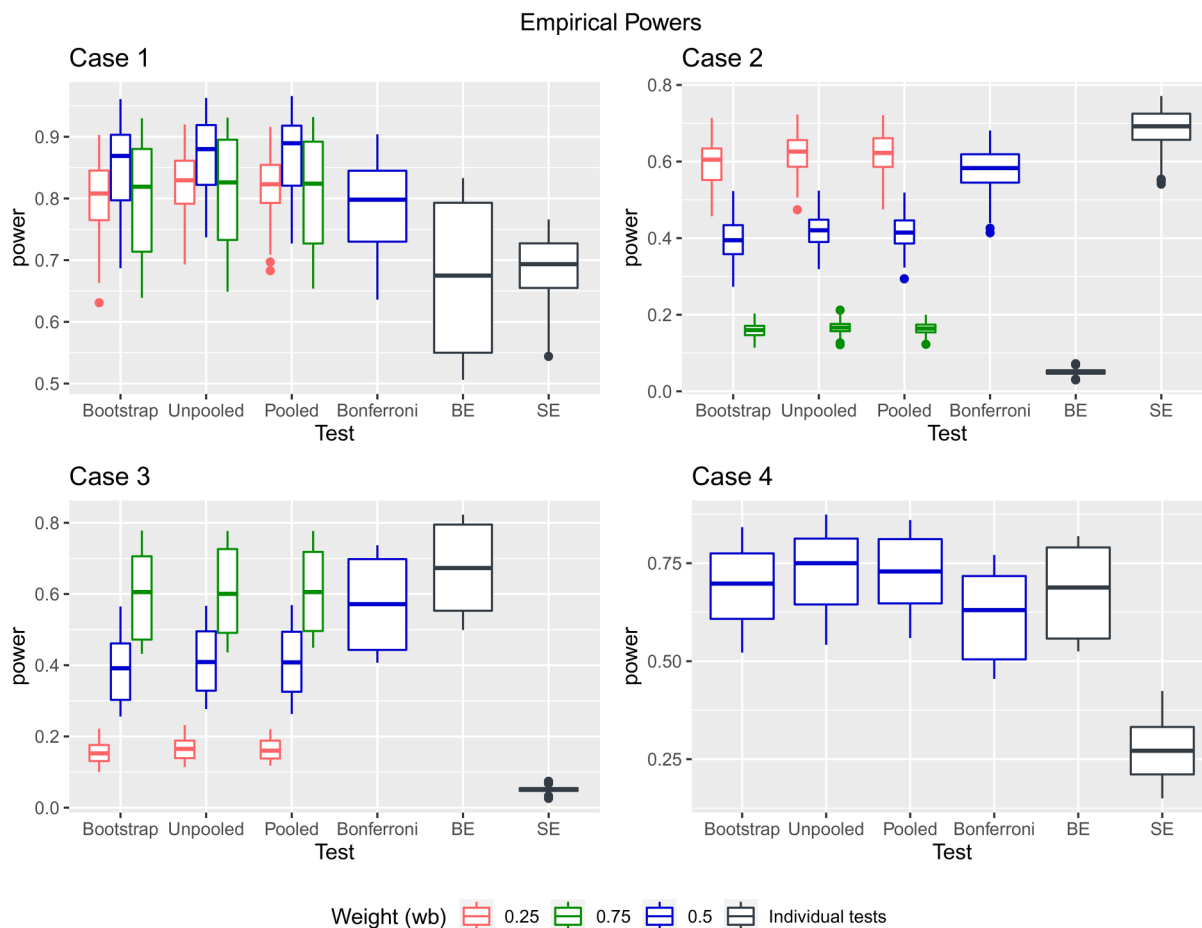
Figure 3. Boxplot of empirical powers based on scenarios in Table 2. The empirical powers are calculated using: the $\mathcal{L}$-statistics (in (2)) according to the pooled, unpooled, bootstrap variance estimators; the Bonferroni procedure; and the individual statistics (3) and (4). The individual statistics for the binary and survival endpoints are labeled, respectively, as BE and SE. The color indicates which combination of weights (ω_b, ω_s) were used: red for $(\omega_b = 0.25, \omega_s = 0.75)$; blue for $(\omega_b = 0.5, \omega_s = 0.5)$; and green for $(\omega_b = 0.75, \omega_s = 0.25)$.

In this work, we focus mainly on clinical trials where the groups to be compared are supposed to only differ with respect to the treatment they receive. However, in observational studies and also sometimes in clinical trials, there is a need to adjust for other covariates than treatment. Whenever the number of covariates is not too large, stratified tests can be useful. The stratified version of the $\mathcal{L}$-test is deferred to the Supplemental material.

We have incorporated weighted functions in the $\mathcal{L}$-statistics to control the relative relevance of each outcome and to specify the type of survival differences that may exist between groups. In our proposed statistics, the weights (ω_b, ω_s) have been defined with the goal of incorporating the potential difference in clinical importance between the binary and survival endpoint, and therefore they must be fixed in the planning stage. As shown in the simulation study, the power of the trial will depend on the trial objectives and then on the relevance of each of the endpoints by means of (ω_b, ω_s). A sensitivity analysis could be carried out calculating the $\mathcal{L}$-statistics based on a pre-specified set of weights, as it is often done with the weighted log-rank statistics. [30] The extension of these statistics incorporating data-driven weights to maximize the power will be considered in future works.

The testing procedure using the $\mathcal{L}$-class of statistics satisfies a property called coherence that says that the nonrejection of an intersection hypothesis implies the nonrejection of any sub-hypothesis it implies, that is, $H_{s,0}$ and $H_{b,0}$.[31] However, the testing procedure based on the $\mathcal{L}$-class of statistics does not fulfil the consonant property that states that the rejection of the global null hypothesis implies the rejection of at least one of its sub-hypothesis. Bittman et al.[32] faced the problem of how to combine tests into a multiple testing procedure for obtaining a procedure that satisfies the coherence and consonance principles. An extension of this work to obtain a testing procedure that satisfies both properties could be an important research line to consider. If, however, it is desired to conclude an efficacy claim for the individual endpoints, alternative

methods could also be considered, such as an adjusted Bonferroni (taking into account the joint distribution of the univariate tests stated in Section 'Large sample results'), or a two-stage method in which efficacy is assessed in the univariate tests after detecting an overall effect.[33]

In this paper, we have considered weighted Kaplan–Meier tests-based for comparing the survival curves motivated by recent immunotherapy trials in which the proportional hazards assumption is in doubt. However, if the survival endpoint satisfies the proportional hazards assumption, the combination of a weighted log-rank test and difference in proportion test could also be considered. Previous works by Wei and Lachin[34,35] could then be used for this problem. The comparison of the relative efficiency between tests combining binary and survival data using weighted log-rank tests or weighted Kaplan–Meier tests and differences in proportions under a range of alternative hypotheses is open for future research.

This work has been restricted to those cases in which censoring and survival do not prevent assessing the binary endpoint response. Two questions remain open for future research concerning this point. The first is the possibility of censorship occurring earlier and making it impossible to assess the binary endpoint. The second is the fact that the survival event may preclude observing the binary endpoint. While the second case refers to a competing risk problem, the first refers to a censoring regarding binary endpoint. Regarding the competing-risk problem, it should also be noted that the time-to-event is usually more relevant in practice. And, in this situation, it is generally considered that if the survival endpoint happens before the binary endpoint, the binary endpoint would have taken the value representing the worst outcome. Further research is needed to study the implications of competing risks or censorship on this type of design. We plan to extend the $\mathcal{L}$-class of statistics to more general censoring schemes where the binary endpoint could be censored or where there may be a case of competing risk. Last but not least, extensions to sequential and adaptive procedures in which the binary outcome could be tested at more than one time-point remain open for future research.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ORCID iD

Marta Bofill Roig https://orcid.org/0000-0002-4400-7541
Guadalupe Gómez Melis https://orcid.org/0000-0003-4252-4884

Supplemental Material

Supplemental material for this article is available online.

References

1. Ananthakrishnan R and Menon S. Design of oncology clinical trials: A review. *Crit Rev Oncol Hematol* 2013; **88**: 144–153.
2. Wilson MK, Collyar D, Chingos DT, et al. Outcomes and endpoints in cancer trials: Bridging the divide. *Lancet Oncol* 2015; **16**: e43–e52.
3. Thall PF. A review of phase 2-3 clinical trial designs. *Lifetime Data Anal* 2008; **14**: 37–53.
4. Lai X and Zee BCY. Mixed response and time-to-event endpoints for multistage single-arm phase II design. *Trials* 2015; **16**: 1–10.
5. Lai TL, Lavori PW and Shih MC. Sequential design of phase II-III cancer trials. *Stat Med* 2012; **31**: 1944–1960.(18)
6. Chen BE and Wang J. Joint modeling of binary response and survival for clustered data in clinical trials. *Stat Med* 2020; **39**: 326–339.
7. Mick R and Chen TT. Statistical challenges in the design of late-stage cancer immunotherapy studies. *Cancer Immunol Res* 2015; **3**: 1292–1298.

8. Pepe MS and Fleming TR. Weighted Kaplan-Meier statistics: A class of distance tests for censored survival data. *Biometrics* 1989; **45**: 497–507.

9. Pepe MS and Fleming TR. Weighted Kaplan-Meier statistics: Large sample and optimality considerations. *Journal of the Royal Statistical Society Series B (Methodological)* 1991; **53**: 341–352.

10. Gu M, Follmann D and Geller NL. Monitoring a general class of two-sample survival statistics with applications. *Biometrika* 1999; **86**: 45–57.

11. Dmitrienko A and Agostino RD. Traditional multiplicity adjustment methods in clinical trials. *Stat Med* 2013; **32**: 5172–5218.

12. Alosh M, Bretz F and Huque M. Advanced multiplicity adjustment methods in clinical trials. *Stat Med* 2014; **33**: 693–713.

13. Bland JM and Altman DG. Multiple significance tests: the Bonferroni method. *BMJ* 1995; **310**: 170.

14. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics* 1984; **04**: 1079–1087.

15. Pocock SJ, Geller NL and Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987; **43**: 487–498.

16. Hothorn T, Bretz F and Westfall P. Simultaneous inference in general parametric models. *Biom J* 2008; **50**: 346–363.

17. Pipper CB, Ritz C and Bisgaard H. A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. *Journal of the Royal Statistical Society Series C: Applied Statistics* 2012; **61**: 315–326.

18. Tsiatis AA and Davidian M. Joint modeling of longitudinal and time-to-event data: An overview. *Stat Sin* 2004; **14**: 809–834.(3)

19. Rizopoulos D. *Joint models for longitudinal and time-to-event data, with applications*. R. Boca Raton: Chapman & Hall/CRC, 2012.

20. Papageorgiou G, Mauff K, Tomer A, et al. An overview of joint modeling of time-to-event and longitudinal outcomes. *Annu Rev Stat Appl* 2019; **6**: 1–18.

21. Fleming TR and Harrington DP. *Counting processes and survival analysis*. New York: Wiley, 1991.

22. de Jong W, Aerts J, Allard S, et al. . iHIVARNA phase IIa, a randomized, placebo-controlled, double-blinded trial to evaluate the safety and immunogenicity of iHIVARNA-01 in chronically HIV-infected patients under stable combined antiretroviral therapy. *Trials* 2019; **20**: 361.

23. Hodi FS, O'Day SJ and McDermott DF. Improved survival with ipilimumab in patients with metastatic melanoma. *N Engl J Med* 2010; **363**: 711–723.

24. Logan BR, Klein JP and Zhang MJ. Comparing treatments in the presence of crossing survival curves: An application to bone marrow transplantation. *Biometrics* 2008; **64**: 733–740.

25. Shen Y and Cai J. Maximum of the weighted Kaplan-Meier tests with application to cancer prevention and screening trials. *Biometrics* 2001; **57**: 837–843.

26. Muller HG and Wang JL. Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics* 1994; **50**: 61–76.

27. Hess K and Gentleman R. R Package 'muhaz': Hazard Function Estimation in Survival Analysis. Version 1.2.6.1, 2019.

28. Sun R. R package. GitHub Repository: https://github.com/ryanrsun/reconstructkm, 2020.

29. Trivedi PK and Zimmer DM. Copula modeling: an introduction for practitioners. *Found Trends Econ* 2007; **1**: 1–111.

30. Prior TJ. Group sequential monitoring based on the maximum of weighted log-rank statistics with the Fleming-Harrington class of weights in oncology clinical trials. *Stat Methods Med Res* 2020; **29**: 3525–3532.

31. Romano JP and Wolf M. Exact and approximate stepdown methods for multiple hypothesis testing. *J Am Stat Assoc* 2005; **100**: 94–108.

32. Bittman RM, Romano JP, Vallarino C, et al. Optimal testing of multiple hypotheses with common effect direction. *Biometrika* 2009; **96**: 399–410.

33. Dmitrienko A, Tamhane A and Bretz F (eds). *Multiple testing problems in pharmaceutical statistics*. Taylor & Francis: New York, 2009.

34. Wei LJ and Lachin JM. Two-sample asymptotically observations for incomplete multivariate. *J Am Stat Assoc* 1984; **79**: 653–661.

35. Lachin JM and Bebu I. Application of the Wei-Lachin multivariate one-directional test to multiple event-time outcomes. *Clinical Trials* 2015; **12**: 627–633.