SORT 38 (1) January-June 2014, 73-88

The asymptotic relative efficiency and the ratio of sample sizes when testing two different null hypotheses

Guadalupe Gómez*, [1] and Moisés Gómez-Mateu[1]

Abstract

Composite endpoints, consisting of the union of two or more outcomes, are often used as the primary endpoint in time-to-event randomized clinical trials. Previously, Gómez and Lagakos provided a method to guide the decision between using a composite endpoint instead of one of its components when testing the effect of a treatment in a randomized clinical trial. Consider the problem of testing the null hypotheses of no treatment effect by means of either the single component or the composite endpoint. In this paper we prove that the usual interpretation of the asymptotic relative efficiency as the reciprocal ratio of the sample sizes required for two test procedures, for the same null and alternative hypothesis, and attaining the same power at the same significance level, can be extended to the test procedures considered here for two different null and alternative hypotheses. A simulation to study the relationship between asymptotic relative efficiency and finite sample sizes is carried out.

MSC: 62N03, 62P10

Keywords: Asymptotic relative efficiency, composite endpoint, logrank test, sample size, simulation, survival analysis.

1. Introduction

In clinical trials research, one of the most important issues that investigators have to solve at the design stage of the study is the appropriate choice of the primary endpoint. Composite endpoints (CE) consisting of the union of two or more outcomes are commonly used as primary endpoints. For example, in the cardiovascular area the

* Corresponding author e-mail: lupe.gomez@upc.edu

[1] Departament d'Estadística i I.O., Universitat Politècnica de Catalunya, Jordi Girona 1–3, 08034. Barcelona, Spain.

Received: May 2013

Accepted: January 2014

relevant endpoint of death is often combined with other additional endpoints such as myocardial infarction, stroke or hospitalization. Pros and cons on the use of CE have been extensively discussed (Freemantle et al., 2003; Ferreira-González et al., 2007, among many others). One of the main advantages of using a CE relies in the fact that by means of a CE the problem of multiplicity is adequately addressed and the bias associated with competing risks (Wittkop et al., 2010) is avoided. Also, with a CE the number of observed events will be higher and, hopefully, the power of the test will increase. However, as it has been discussed (Montori et al., 2005) and shown in Gómez and Lagakos (2013), adding inappropriate components to the relevant endpoint might actually lead to a decrease in the power of the test statistic, consequently having a larger chance to fail in detecting a real effect of the treatment under study.

Gómez and Lagakos (2013) developed a methodology to help to decide when it is worthwhile to base the analysis on the composite endpoint $\mathscr{E}_* = \mathscr{E}_1 \cup \mathscr{E}_2$ where $\mathscr{E}_1$ and $\mathscr{E}_2$ are two candidate relevant endpoints to evaluate the effect of a treatment instead of sticking to one of them, $\mathscr{E}_1$, say. In order to do so, they compared how more efficient than $\mathscr{E}_1$ would $\mathscr{E}_*$ be to justify its use. Let H_0 be the null hypothesis of no treatment effect evaluated on $\mathscr{E}_1$ and denote by H_a an alternative hypothesis, for instance, claiming to delay the event $\mathscr{E}_1$. Analogously, define H_0^* and H_a^* the null and alternative hypotheses if the treatment effect is to be evaluated on $\mathscr{E}_*$. Since when comparing two treatment groups based on time-to-event endpoints, the primary analysis would be based, very commonly, on a logrank test, their method considers the logrank test Z to test H_0 versus H_a and the logrank test Z_* to test H_0^* versus H_a^*. The asymptotic relative efficiency (ARE) of Z_* versus Z is the measure proposed to choose between $\mathscr{E}_1$ and $\mathscr{E}_*$, with values larger than 1 in favour of $\mathscr{E}_*$. This relative measure can be computed as $(\mu_*/\mu)^2$ where μ and μ_* are, respectively, the asymptotic means of Z and Z_*, under alternative contiguous hypotheses to H_0 and H_0^*. The purpose of this paper is to prove that the usual interpretation of the ARE, as the ratio of sample sizes, n and n_*, needed to attain the same power for a given significance level, still holds even though two different sets of hypothesis (H_0 versus H_a and H_0^* versus H_a^*) are compared.

To clarify the purpose of our investigation consider the following. If we were to test H_0 versus H_a with two different test statistics S_n and T_m, Pitman's relative efficiency would be defined as the ratio m/n, where n and m are the required sample sizes for S_n and T_m, respectively, to attain the same power for a given significance level. Furthermore, if both S_n and T_m are asymptotically normal with unit variance and means μ_S and μ_T, it can be proved that Pitman's ARE corresponds to the square of the ratio of the noncentrality parameters, that is $(\mu_S/\mu_T)^2$. Gómez and Lagakos' method compares the logrank statistics: Z and Z_* derived for two different set of hypotheses H_0 versus H_a and H_0^* versus H_a^* and do so using, as definition of the ARE, the ratio $(\mu_*/\mu)^2$ where μ and μ_* are, respectively, the asymptotic means of Z and Z_*, under alternative contiguous hypotheses to H_0 and H_0^*.

This paper is organized as follows. In Section 2 the notation, assumptions and main results from Gómez and Lagakos' paper are introduced. Section 3 establishes

the limiting relationship between ARE and sample sizes and proves that the usual interpretation of the ARE as the ratio of sample sizes holds. Section 4 presents a simulation to study under which conditions and for finite sample sizes, the relationship $\mathrm{ARE}(Z_*, Z) = (\mu_*/\mu)^2 = n/n_*$ holds where n and n_* are the needed sample sizes for Z and Z_*, respectively, to attain the same power for a given significance level. Section 5 concludes the paper with a discussion.

2. Notation, the logrank test and the asymptotic relative efficiency

2.1. The logrank tests for the relevant and for the composite endpoints

Assume that we have a two-arm study involving random assignment to an active ($X = 1$) or control treatment ($X = 0$) aiming to prove the efficacy of the new active treatment. The effect of treatment is to be evaluated on the time $T_1^{(j)}$ to a relevant event $\mathscr{E}_1$, where the superscript j indicates the treatment group ($j = 0$ for the control group and $j = 1$ for the treatment group). Let $\lambda_1^{(j)}(t)$ denote the hazard function of $T_1^{(j)}$ ($j = 0, 1$). The null hypothesis of no effect is given by $H_0 : \mathrm{HR}_1(t) = \lambda_1^{(1)}(t)/\lambda_1^{(0)}(t) = 1$ and the alternative that the new treatment improves survival by $H_a : \mathrm{HR}_1(t) < 1$. The logrank test Z is used to test that the new treatment improves survival.

Assume now that an additional endpoint $\mathscr{E}_2$ is considered as component of the primary endpoint and the composite endpoint $\mathscr{E}_* = \mathscr{E}_1 \cup \mathscr{E}_2$ is to be used, instead, to prove the efficacy of the new treatment. The effect of treatment would then be evaluated on the time $T_*^{(j)}$ to $\mathscr{E}_*$ where $T_*^{(j)} = \min\{T_1^{(j)}, T_2^{(j)}\}$ and $T_2^{(j)}$ stands for the time to $\mathscr{E}_2$ ($j = 0, 1$). Let $\lambda_2^{(j)}(t)$ and $\lambda_*^{(j)}(t)$ denote, respectively, the hazard functions of $T_2^{(j)}$ and $T_*^{(j)}$ ($j = 0, 1$). The treatment effect on $\mathscr{E}_*$ would then be tested with the logrank test Z_* to compare $H_0^* : \mathrm{HR}_*(t) = \lambda_*^{(1)}(t)/\lambda_*^{(0)}(t) = 1$ versus $H_a^* : \mathrm{HR}_*(t) < 1$.

Observation of endpoints $\mathscr{E}_1$ and $\mathscr{E}_2$ depends on whether or not they include a terminating event and yield four different situations referred, in Gómez and Lagakos (2013), as Cases 1, 2, 3 and 4. In this paper we assume that the additional endpoint does not include a terminating event, which corresponds to Case 1 when neither the relevant nor the additional endpoint includes a terminating event, and Case 3, when the relevant endpoint includes a terminating event.

Schoenfeld (1981) studies the asymptotic behaviour of the logrank statistic and proves that under the null hypothesis of no treatment difference, the logrank is asymptotically $N(0, 1)$ and, under a sequence of alternatives contiguous to the null, the logrank is asymptotically normal with unit variance and finite mean. Gómez and Lagakos apply Schoenfeld's results and proceed as follows. They consider $\lambda_1^{(0)}(t)$ as fixed and define a sequence of alternatives $H_{a,n}$ consisting of instantaneous hazard functions close enough to $\lambda_1^{(0)}(t)$, for instance taking $\lambda_{1,n}^{(1)}(t) = \lambda_1^{(0)}(t) e^{g(t)/\sqrt{n}}$ for some $g(t)$ function. These sequence of alternatives, formulated equivalently as $\mathrm{HR}_{1,n}(t) = e^{g(t)/\sqrt{n}}$, include pro-

portional hazard alternatives, i.e, taking $g(t) = \beta$ for a fixed real value β. Logrank Z is asymptotically $N(0,1)$ under the null hypothesis of no treatment difference (H_0 : $\mathrm{HR}_1(t) = 1$) and asymptotically normal with unit variance and mean μ given in equation (1) under the sequence of alternatives $H_{a,n} : \mathrm{HR}_{1,n}(t) = e^{g(t)/\sqrt{n}} < 1$. Analogously, fix $\lambda_*^{(0)}(t)$ and define $H_0^* : \mathrm{HR}_*(t) = 1$ and the sequence of alternatives $H_{a,n}^* : \mathrm{HR}_{*,n}(t) = e^{g_*(t)/\sqrt{n}} < 1$ for a given function $g_*(t)$. It follows that Z^* is asymptotically $N(0,1)$ under H_0^* and asymptotically normal with unit variance and mean μ_* given in equation (2) under the sequence $H_{a,n}^*$. The asymptotic means of Z and Z^* are given by

$$\mu = \frac{\int_0^\infty g(t)p(t)[1-p(t)]\mathrm{Pr}_{H_0}\{U \geq t\}\lambda_1^{(0)}(t)dt}{\sqrt{\int_0^\infty p(t)[1-p(t)]\mathrm{Pr}_{H_0}\{U \geq t\}\lambda_1^{(0)}(t)dt}}, \tag{1}$$

$$\mu_* = \frac{\int_0^\infty g_*(t))p_*(t)[1-p_*(t)]\mathrm{Pr}_{H_0^*}\{U_* \geq t\}\lambda_*^{(0)}(t)dt}{\sqrt{\int_0^\infty p_*(t)[1-p_*(t)]\mathrm{Pr}_{H_0^*}\{U_* \geq t\}\lambda_*^{(0)}(t)dt}}, \tag{2}$$

where $U = \min\{T_1, C\}$ (in Cases 1 and 3) and $U_* = \min\{T_*, C\}$ denote the observed outcome; C denotes the censoring time; $p(t) = \mathrm{Pr}_{H_0}\{X = 1 | U \geq t\}$ and $p_*(t) = \mathrm{Pr}_{H_0^*}\{X = 1 | U_* \geq t\}$ are the null probabilities that someone at risk at time t is in treatment group 1; $\mathrm{Pr}_{H_0}\{U \geq t\}$ and $\mathrm{Pr}_{H_0^*}\{U_* \geq t\}$ are the null probabilities that someone is still at risk at time t and $\mathrm{Pr}_{H_0}\{U \geq t\}\lambda_1^{(0)}(t)$ and $\mathrm{Pr}_{H_0^*}\{U_* \geq t\}\lambda_*^{(0)}(t)$ correspond to the probabilities, under the null hypothesis, of observing events $\mathscr{E}_1$ and $\mathscr{E}_*$, respectively, by time t.

2.2. Asymptotic relative efficiency

Efficiency calculations throughout the paper will assume that end-of-study censoring at time τ ($\tau = 1$ without loss of generality) is the only non-informative censoring cause for both groups; this assumption indirectly implies that the censoring mechanism is the same for both groups. It is as well assumed that the hazard functions $\lambda_1^{(j)}(t)$ and $\lambda_2^{(j)}(t)$ ($j = 0, 1$) are proportional, that is, $\mathrm{HR}_1(t) = \mathrm{HR}_1$ and $\mathrm{HR}_2(t) = \mathrm{HR}_2$, for all t, where $\mathrm{HR}_1(t) = \lambda_1^{(1)}(t)/\lambda_1^{(0)}(t)$ and $\mathrm{HR}_2(t) = \lambda_2^{(1)}(t)/\lambda_2^{(0)}(t)$ are the hazard ratios between $T_1^{(0)}$ and $T_1^{(1)}$ and between $T_2^{(0)}$ and $T_2^{(1)}$, respectively. Note that although we are assuming that the hazard functions $\lambda_1^{(j)}(t)$ and $\lambda_2^{(j)}(t)$ ($j = 0, 1$) are proportional, this does not imply the proportionality of hazards $\lambda_*^{(0)}(t)$ and $\lambda_*^{(1)}(t)$ for the composite endpoint T_* (see Figure 1).

To assess the difference in efficiency between using logrank test Z, based on the relevant endpoint $\mathscr{E}_1$, and logrank test Z_*, based on the composite endpoint $\mathscr{E}_*$, Gómez and Lagakos base their strategy on the behaviour of the asymptotic relative efficiency (ARE) of Z_* versus Z. The ARE is a measure of the relative power of two tests that can

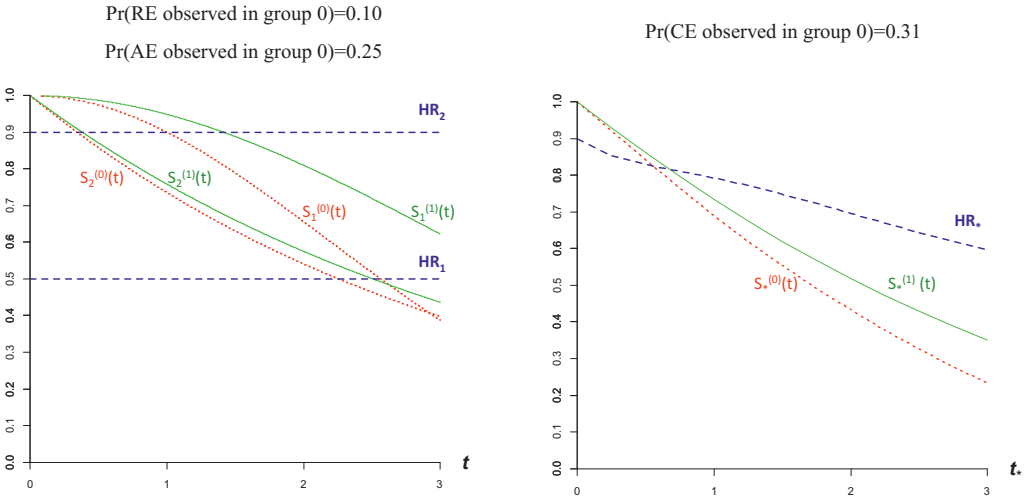

Figure 1: *Survival and hazard ratio for the relevant endpoint (RE), T_1, for the additional endpoint (AE), T_2 and for the composite endpoint (CE), $T_* = \min\{T_1, T_2\}$. $T_1 \sim$ Weibull with shape parameter $\beta_1 = 2$ (increasing hazard) for treatment groups 0 and 1 and $T_2 \sim$ Weibull with shape parameter $\beta_2 = 1$ (constant hazard) for treatment groups 0 and 1. Scale parameters for T_1 and T_2 have been calculated such that* $\Pr\{T_1$ *observed in group 0*$\}$*=0.1,* $\Pr\{T_2$ *observed in group 0*$\}$*=0.25,* $\mathrm{HR}_1 = 0.5$*,* $\mathrm{HR}_2 = 0.9$ *and Spearman's* $\rho(T_1, T_2) = 0.45$ *assuming Frank's copula between T_1 and T_2. Considering the RE as a terminating event (case 3), in this setting* $\mathrm{ARE}(Z_*, Z) = 0.21$.

be interpreted, when the two tests are for the same null and alternative hypothesis, as the ratio of the required sample sizes to detect a specific treatment effect to attain the same power for a given significance level (Lehmann and Romano, 2005). In this case, a value of ARE$=0.6$ would mean that we only need 60% as many cases to reach a given power if we use $\mathscr{E}_1$ as we would need if we used $\mathscr{E}_*$. Whenever the tests under consideration, Z and Z_*, are asymptotically N(0,1) under H_0 and H_0^*, respectively, and asymptotically normal with variance 1 under a sequence of contiguous alternatives to the null hypothesis, a different definition for Pitman's relative efficiency as the square of the ratio of the non-centrality parameters μ and μ_* is appropriate

$$\mathrm{ARE}(Z_*, Z) = \left(\frac{\mu_*}{\mu}\right)^2, \tag{3}$$

where μ and μ_* are to be replaced by expressions (1) and (2).

Before providing the expression that is being used to evaluate the ARE, and for the sake of clarity, we enumerate the assumptions that have been taken into account:

- End-of-study censoring at time τ is the only non-informative censoring cause for both groups.
- The additional endpoint does not include a terminating event.

- The hazard ratios between $T_1^{(0)}$ and $T_1^{(1)}$ and between $T_2^{(0)}$ and $T_2^{(1)}$ are proportional, that is, $\mathrm{HR}_1(t) = \lambda_1^{(1)}(t)/\lambda_1^{(0)}(t) = \mathrm{HR}_1$ and $\mathrm{HR}_2(t) = \lambda_2^{(1)}(t)/\lambda_2^{(0)}(t) = \mathrm{HR}_2$ for all t.
- Effect of treatment on $\mathscr{E}_1$ is tested establishing $H_0 : \mathrm{HR}_1 = 1$ versus a sequence of alternatives $H_{a,n} : \lambda_{1,n}^{(1)}(t) = \lambda_1^{(0)}(t)e^{g(t)/\sqrt{n}}$ for some $g(t)$ function. Note that $g(t)/\sqrt{n} = \log\{\lambda_{1,n}^{(1)}(t)/\lambda_1^{(0)}(t)\}$.
- Effect of treatment on $\mathscr{E}_*$ is tested establishing $H_0^* : \mathrm{HR}_*(t) = 1$ versus a sequence of alternatives $H_{a,n}^* : \mathrm{HR}_{*,n}(t) = e^{g_*(t)/\sqrt{n}} < 1$ for a given function $g_*(t)$. Note that $g_*(t)/\sqrt{n} = \log\{\mathrm{HR}_{*,n}(t)\}$.

Under the above assumptions expression (3) becomes

$$\mathrm{ARE}(Z_*, Z) = \frac{\left(\int_0^1 \log\left\{\lambda_*^{(1)}(t)/\lambda_*^{(0)}(t)\right\} f_*^{(0)}(t)dt\right)^2}{\left(\log\left\{\mathrm{HR}_1\right\}\right)^2 \left(\int_0^1 f_*^{(0)}(t)dt\right)\left(\int_0^1 f_1^{(0)}(t)dt\right)}, \tag{4}$$

where $f_1^{(0)}(t)$ and $f_*^{(0)}(t)$ are the density functions of $T_1^{(0)}$ and $T_*^{(0)}$, respectively.

Remark The density function $f_*^{(0)}(t)$ is the density of the $T_*^{(0)} = \min\{T_1^{(0)}, T_2^{(0)}\}$, computed from the joint density between $T_1^{(0)}$ and $T_2^{(0)}$, which itself is built from the marginals of $T_1^{(0)}$ and $T_2^{(0)}$ by means of a bivariate copula.

3. Relationship between ARE and sample sizes

We start establishing that if the hazard ratios for $T_1^{(j)}$ $(j = 0, 1)$ and for $T_2^{(j)}$ $(j = 0, 1)$ approach the unity as n gets large, so does the hazard ratio of the minimum $T_*^{(j)}$ between $T_1^{(j)}$ and $T_2^{(j)}$ $(j = 0, 1)$.

Lemma 1 *Given two sequences of hazard ratios* $\{\mathrm{HR}_{1,n}(t) = \lambda_{1,n}^{(1)}(t)/\lambda_1^{(0)}(t)\}$ *and* $\{\mathrm{HR}_{2,n}(t) = \lambda_{2,n}^{(1)}(t)/\lambda_2^{(0)}(t)\}$, *both converging uniformly to 1 as* $n \to \infty$, *the sequence corresponding to the hazard ratio of* $T_*^{(j)} = \min\{T_1^{(j)}, T_2^{(j)}\}$, *namely* $\{\mathrm{HR}_{*,n}(t) = = \lambda_{*,n}^{(1)}(t)/\lambda_*^{(0)}(t)\}$, *tends to* 1 *as* $n \to \infty$. *In particular, this lemma holds whenever* $\log(\lambda_{k,n}^{(1)}(t)/\lambda_k^{(0)}(t)\}) = O(n^{-1/2})$, *which in turn, is true if* $\log(\lambda_{k,n}^{(1)}(t)/\lambda_k^{(0)}(t)\}) = = g_k(t)/\sqrt{n}$, *for any bounded real function* $g_k(t)$ $(k = 1, 2)$.

Proof 1 It follows immediately that for fixed t, $\lim_{n\to\infty} \lambda_{1,n}^{(1)}(t) = \lambda_1^{(0)}(t)$ and $\lim_{n\to\infty} \lambda_{2,n}^{(1)}(t) = \lambda_2^{(0)}(t)$. Furthermore, it follows that the corresponding densities and

survival functions $f_{1,n}^{(1)}(t)$, $f_{2,n}^{(1)}(t)$, $S_{1,n}^{(1)}(t)$ and $S_{2,n}^{(1)}(t)$, converge to $f_1^{(0)}(t)$, $f_2^{(0)}(t)$, $S_1^{(0)}(t)$ and $S_2^{(0)}(t)$, respectively. Taking into account that the survival function of the minimum, $S_{*,n}^{(1)}(t)$ is expressed in terms of the marginal survival functions $S_{1,n}^{(1)}(t)$ and $S_{2,n}^{(1)}(t)$ of $T_1^{(1)}$ and $T_2^{(1)}$ via a copula C, that is,
$S_{*,n}^{(1)}(t) = C(S_{1,n}^{(1)}(t), S_{2,n}^{(1)}(t))$, it remains to prove that $\lim_{n\to\infty} S_{*,n}^{(1)}(t) = S_*^{(0)}(t)$. This result will imply that

$\lim_{n\to\infty} f_{*,n}^{(1)}(t) = f_*^{(0)}(t)$, $\lim_{n\to\infty} \lambda_{*,n}^{(1)}(t) = \lambda_*^{(0)}(t)$ and hence the sequence $\mathrm{HR}_{*,n}(t) \to 1$ as $n \to \infty$, as we wanted to prove.
The convergence of $S_{*,n}^{(1)}(t)$ to $S_*^{(0)}(t)$ is guaranteed by the convergence of $S_{1,n}^{(1)}(t)$ and $S_{2,n}^{(1)}(t)$ to $S_1^{(0)}(t)$ and $S_2^{(0)}(t)$, respectively, together with the fact that bivariate copulas C are bivariate distribution functions with uniform marginals. The reader is referred to Lindner and Szimayer (2005) for the corresponding technical proofs. □

Proposition 1 *Consider two test procedures ϕ_n and ϕ_n^* to test $H_0 : \mathrm{HR}_1(t) = 1$ against $H_{a,n} : \mathrm{HR}_{1,n}(t) < 1$ and $H_0^* : \mathrm{HR}_*(t) = 1$ against $H_{a,n}^* : \mathrm{HR}_{*,n}(t) < 1$, respectively. Let n and n_* be the sample sizes required for ϕ_n and ϕ_n^*, respectively, to have power at least Π at level α. Assume the sequences $\phi = \{\phi_n\}$ and $\phi^* = \{\phi_n^*\}$ are based on the logrank statistics Z and Z^*, respectively, converging, to Normal $(\mu, 1)$ and Normal $(\mu_*, 1)$ with μ and μ_* given in (1) and (2), under sequences of local alternatives $\mathrm{HR}_{k,n}(t)$ $(k = 1, 2)$ converging uniformly to 1 as $n \to \infty$. Given $0 < \alpha < \Pi < 1$,*

$$\lim_{\substack{\mathrm{HR}_{1,n}(t)\to 1 \\ \mathrm{HR}_{2,n}(t)\to 1}} \frac{n}{n_*} = \mathrm{ARE}(Z_*, Z).$$

The usual interpretation of the ARE as the reciprocal ratio of the sample sizes holds even when two different sets of hypotheses (H_0 versus $H_{a,n}$ and H_0^* versus $H_{a,n}^*$) are tested. As a consequence of this proposition, the interpretation of the ARE is the following. If $\mathrm{ARE}(Z_*, Z) = 0.7$, then, asymptotically, we only need 70% as many cases to attain a given power if we use Z as we would need if we used Z_*.

Proof 2 By Lemma 1, uniform convergence to 1 of $\{\mathrm{HR}_{1,n}(t)\}$ and $\{\mathrm{HR}_{2,n}(t)\}$ imply that $\lim \mathrm{HR}_{*,n}(t) \to 1$. Under the sequence of contiguous alternatives to the null $H_{a,n}$: $\{\mathrm{HR}_{1,n}(t) = \lambda_{1,n}^{(1)}(t)/\lambda_1^{(0)}(t)\} \to 1$ and $H_{a,n}^* : \{\mathrm{HR}_{*,n}(t) = \lambda_{*,n}^{(1)}(t)/\lambda_*^{(0)}(t)\} \to 1$, both Z and Z^* are asymptotically $N(\mu, 1)$ and $N(\mu_*, 1)$, respectively. The power function for a one-sided test with size α is therefore given, respectively, by

$$\Pi_1 = \lim_{n\to\infty} \mathrm{Prob}\{Z < z_{1-\alpha} | H_{a,n}\} = 1 - \Phi(-z_{1-\alpha} + \mu)$$

$$\Pi_* = \lim_{n\to\infty} \mathrm{Prob}\{Z_* < z_{1-\alpha} | H_{a,n}^*\} = 1 - \Phi(-z_{1-\alpha} + \mu_*) \tag{5}$$

where Φ is the distribution function of the standard normal and $z_{1-\alpha}$ is the standard normal quantile corresponding to the left tail probability α. It immediately follows that $\Pi_1 = \Pi_*$ is equivalent to $\mu = \mu_*$.

The equivalence of powers ($\Pi_1 = \Pi_*$) implies that $\mu = \mu_*$, given by (1) and (2). Equivalently

$$\left(\frac{\mu_*}{\mu}\right)^2 = 1 \quad \Longleftrightarrow \quad \left(\frac{\frac{\int_0^\infty g(t)p(t)[1-p(t)]\Pr_{H_0}\{U\geq t\}\lambda_1^{(0)}(t)dt}{\sqrt{\int_0^\infty p(t)[1-p(t)]\Pr_{H_0}\{U\geq t\}\lambda_1^{(0)}(t)dt}}}{\frac{\int_0^\infty g_*(t)p_*(t)[1-p_*(t)]\Pr_{H_0^*}\{U_*\geq t\}\lambda_*^{(0)}(t)dt}{\sqrt{\int_0^\infty p_*(t)[1-p_*(t)]\Pr_{H_0^*}\{U_*\geq t\}\lambda_*^{(0)}(t)dt}}} \right)^2 = 1. \tag{6}$$

Since

$$p(t) = \frac{\Pr_{H_0}\{U \geq t | X = 1\}\pi}{\Pr_{H_0}\{U \geq t\}} = \frac{\Pr_{H_0}\{U^{(j)} \geq t\}\pi}{\Pr_{H_0}\{U \geq t\}}$$

where $\pi = \Pr_{H_0}\{X = 1\}$, we have

$$p(t)(1-p(t))\Pr_{H_0}\{U \geq t\} = \frac{\Pr_{H_0}\{U^{(1)} \geq t\}\pi \Pr_{H_0}\{U^{(0)} \geq t\}(1-\pi)}{\Pr_{H_0}\{U^{(0)} \geq t\}(1-\pi) + \Pr_{H_0}\{U^{(1)} \geq t\}\pi}.$$

Based on the stated assumptions, because $T_1^{(j)}$ is right-censored by the end-of-study at time τ, and under the null hypothesis of no effect ($S_1^{(0)}(t) = S_1^{(1)}(t)$), we have $\Pr_{H_0}\{U^{(j)} \geq t\} = S_1^{(0)}(t)1\{[0,1]\}(t)$, for $j = 0, 1$. Replacing in (1), the noncentrality parameter μ becomes

$$\mu = \frac{\sqrt{\pi(1-\pi)}\int_0^1 g(t)S_1^{(0)}(t)\lambda_1^{(0)}(t)dt}{\sqrt{\int_0^1 S_1^{(0)}(t)\lambda_1^{(0)}(t)dt}} = \frac{\sqrt{\pi(1-\pi)}\int_0^1 g(t)f_1^{(0)}(t)dt}{\sqrt{\int_0^1 f_1^{(0)}(t)dt}}$$

where $f_1^{(0)}(t)$ is the marginal density function for $T_1^{(0)}$. Analogously, it can be seen that

$$\mu_* = \frac{\sqrt{\pi(1-\pi)}\int_0^1 g_*(t)f_*^{(0)}(t)dt}{\sqrt{\int_0^1 f_*^{(0)}(t)dt}}$$

where $f_*^{(0)}(t)$ is the density function for $T_*^{(0)}$. The reader is addressed to the online supporting material of Gómez and Lagakos paper for other technical details.

If we would replace $g(t)$ and $g_*(t)$ by $\sqrt{n}\log\left(\frac{\lambda_{1,n}^{(1)}(t)}{\lambda_1^{(0)}(t)}\right) = \sqrt{n}\log(\mathrm{HR}_1)$ and $\sqrt{n_*}\log\left(\frac{\lambda_{*,n}^{(1)}(t)}{\lambda_*^{(0)}(t)}\right)$, respectively, equality (6), after cancelling $\pi(1-\pi)$, becomes equal to

$$\lim_{\substack{\mathrm{HR}_{1,n}(t)\to 1\\ \mathrm{HR}_{2,n}(t)\to 1}} \frac{\sqrt{n_*}}{\sqrt{n}} \frac{\frac{\int_0^1 \log\left\{\lambda_*^{(1)}(t)/\lambda_*^{(0)}(t)\right\} f_*^{(0)}(t)dt}{\sqrt{\int_0^1 f_*^{(0)}(t)dt}}}{\log(\mathrm{HR}_1)\sqrt{\int_0^1 f_1^{(0)}(t)dt}} = 1$$

which in turn is equivalent to

$$\lim_{\substack{\mathrm{HR}_{1,n}(t)\to 1\\ \mathrm{HR}_{2,n}(t)\to 1}} \frac{n}{n_*} = \frac{\left(\int_0^1 \log\left\{\lambda_*^{(1)}(t)/\lambda_*^{(0)}(t)\right\} f_*^{(0)}(t)dt\right)^2}{(\log(\mathrm{HR}_1))^2 \left(\int_0^1 f_*^{(0)}(t)dt\right)\left(\int_0^1 f_1^{(0)}(t)dt\right)} \tag{7}$$

and it follows that $\mathrm{ARE}(Z_*, Z) = \lim_{\substack{\mathrm{HR}_{1,n}(t)\to 1\\ \mathrm{HR}_{2,n}(t)\to 1}} \frac{n}{n_*}$, as we wanted to prove. □

Note that (7) implies

$$\frac{\left(\int_0^1 \log\left\{\lambda_*^{(1)}(t)/\lambda_*^{(0)}(t)\right\} f_*^{(0)}(t)dt\right)^2}{(\log(\mathrm{HR}_1))^2 \left(\int_0^1 f_*^{(0)}(t)dt\right)^2} = \lim_{\substack{\mathrm{HR}_{1,n}(t)\to 1\\ \mathrm{HR}_{2,n}(t)\to 1}} \frac{n\left(\int_0^1 f_1^{(0)}(t)dt\right)}{n_*\left(\int_0^1 f_*^{(0)}(t)dt\right)} \approx \frac{\text{expected number } \mathscr{E}_1}{\text{expected number } \mathscr{E}_*}$$

and whenever $\lambda_*^{(1)}(t)/\lambda_*^{(0)}(t)$ is approximately constant and equal to HR_*, we would have

$$\frac{\left(\frac{1}{\log(\mathrm{HR}_1)}\right)^2}{\left(\frac{1}{\log(\mathrm{HR}_*)}\right)^2} = \lim_{\substack{\mathrm{HR}_{1,n}(t)\to 1\\ \mathrm{HR}_{2,n}(t)\to 1}} \frac{n\left(\int_0^1 f_1^{(0)}(t)dt\right)}{n_*\left(\int_0^1 f_*^{(0)}(t)dt\right)} \approx \frac{\text{expected number } \mathscr{E}_1}{\text{expected number } \mathscr{E}_*}$$

4. Simulation

4.1. Simulation

Our next aim is to simulate data to empirically check how close we are to the limiting relationship $n/n_* = \mathrm{ARE}(Z_*, Z)$ when $\Pi_1 = \Pi_*$ for different finite sample sizes. To

conduct the simulations we will assume, as Gómez and Lagakos did, that $T_1^{(j)}$ and $T_2^{(j)}$ follow Weibull distributions. Weibull distributions are chosen for their wide use in the field of survival analysis due to its flexibility, allowing decreasing, constant and increasing hazard rates. The corresponding shape and scale parameters are denoted by β_k and $b_k^{(j)}$ ($j = 0,1$, $k = 1,2$) (shape parameters for both groups are taken equal so that the assumption of the proportionality of the hazard ratios holds). To establish the bivariate distribution of $(T_1^{(0)}, T_2^{(0)})$ we consider Frank's Archimedean survival copula, again as Gómez and Lagakos did. Other choices of copulas would be possible, although main conclusions and recommendations will not differ (Plana-Ripoll and Gómez, 2014). Frank's copula depends on an association parameter θ between $T_1^{(0)}$ and $T_2^{(0)}$ which is biunivocally related to Spearman's rank correlation ρ. Different scenarios will be simulated according to several choices of $(\beta_1, \beta_2, p_1^{(0)}, p_2^{(0)}, \mathrm{HR}_1, \mathrm{HR}_2, \rho)$ where $p_1^{(0)}$ and $p_2^{(0)}$ are the probability of observing events $\mathscr{E}_1$ and $\mathscr{E}_2$, respectively, for treatment group 0, HR_1 and HR_2 are relative treatment hazard ratios for $T_j^{(1)}$ versus $T_j^{(0)}$ ($j = 1,2$, respectively) and ρ is Spearman's rank correlation between $T_1^{(0)}$ and $T_2^{(0)}$.

Given a set of values for $(\beta_1, \beta_2, p_1^{(0)}, p_2^{(0)}, \mathrm{HR}_1, \mathrm{HR}_2, \rho)$, for a given power Π and a significance level α, the simulation steps are the following:

1. **Computations for the relevant endpoint $\mathscr{E}_1$.** The scale parameters $b_1^{(0)}$ and $b_1^{(1)}$ and the probability $p_1^{(1)}$ of observing the relevant endpoint in group 1 are derived as:

$$b_1^{(0)} = \frac{1}{(-\log(1-p_1^{(0)}))^{1/\beta_1}}$$

$$b_1^{(1)} = \frac{b_1^{(0)}}{\mathrm{HR}_1^{(1/\beta_1)}}$$

$$p_1^{(1)} = 1 - e^{-(1/b_1^{(1)})^{\beta_1}}$$

2. **Computations for the additional endpoint $\mathscr{E}_2$.** The scale parameters $b_2^{(0)}$ and $b_2^{(1)}$ and the probability $p_2^{(1)}$ of observing the additional endpoint in group 1 are derived as:

$$b_2^{(0)} = \begin{cases} \frac{1}{(-\log(1-p_2^{(0)}))^{1/\beta_2}} & \text{for Case 1} \\ * & \text{for Case 3} \end{cases}$$

$$b_2^{(1)} = \frac{b_2^{(0)}}{\mathrm{HR}_2^{(1/\beta_2)}}$$

$$p_2^{(1)} = 1 - e^{-(1/b_2^{(1)})^{\beta_2}}$$

$*$ For Case 3, $b_2^{(0)}$ is found as the solution of equation $p_2^{(1)} = \int_0^1 \int_u^1 f_{(1,2)}^{(0)}(u,v;\rho)dvdu$, where $f_{(1,2)}^{(0)}(\cdot,\cdot;\rho)$ is the joint density between $T_1^{(0)}$ and $T_2^{(0)}$ and ρ is Spearman's ρ coefficient between $T_1^{(0)}$ and $T_2^{(0)}$.

3. **Computation of sample sizes n and n_***

   (a) Compute n (per group) following Freedman (1982) formulas as follows

$$n = \frac{E}{p_1^{(0)} + p_1^{(1)}} \tag{8}$$

   where

$$E = \frac{(HR_1+1)^2(z_{1-\alpha}+z_\Pi)^2}{(HR_1-1)^2} \tag{9}$$

   (b) Compute $\text{ARE}(Z_*,Z)$ based on $(\beta_1,\beta_2,p_1^{(0)},p_2^{(0)},\text{HR}_1,\text{HR}_2,\rho)$.

   (c) Compute $n_* = n/\text{ARE}(Z_*,Z)$.

   (d) Compute $N = \max\{n,n_*\}$.

4. **Simulation of $T_1^{(0)},T_1^{(1)},T_2^{(0)},T_2^{(1)},T_*^{(0)},T_*^{(1)}$**
   Simulate 1000 samples of size N for the 4 endpoints $T_k^{(j)}$ from Weibull $(b_k^{(j)},\beta_k)$ $(j=0,1,\, k=1,2)$. Compute $T_*^{(j)} = min\{T_1^{(j)},T_2^{(j)}\}$.

5. **Computation of empirical powers $\hat{\Pi}_1$ and $\hat{\Pi}_*$**
   For each sample of size n (n_*), compute the logrank statistic Z (Z_*) to compare the treatment effect between $T_1^{(0)}$ and $T_1^{(1)}$ ($T_*^{(0)}$ and $T_*^{(1)}$). For a given significance level α, the rejection region comprises all observed Z (Z_*) such that $Z < z_{1-\alpha}$ ($Z_* < z_{1-\alpha}$) where $z_{1-\alpha}$ is the standard normal quantile corresponding to the left tail probability α. The empirical powers, denoted by $\hat{\Pi}_1$ ($\hat{\Pi}_*$), are calculated as the proportion of samples for which $Z < z_{1-\alpha}$ ($Z_* < z_{1-\alpha}$).
   We note here that whenever $n_* < n$, we only use, for each sample, the first n_* simulated values to compute $\hat{\Pi}_*$, while when $n < n_*$, we only use the first n simulated values to compute $\hat{\Pi}_1$.

6. **Comparison between $\hat{\Pi}_1$ and $\hat{\Pi}_*$**
   For each scenario $(\beta_1,\beta_2,p_1^{(0)},p_2^{(0)},\text{HR}_1,\text{HR}_2,\rho)$, we compare the differences between the two empirical powers $\hat{\Pi}_1$ and $\hat{\Pi}_*$ obtained from the 1000 simulations.

Table 1: *Values of parameters β_1, β_2, p_1, p_2, HR_1, HR_2 and ρ used for the simulations. There are 624 different configurations, excluding those yielding sample sizes larger than 1100 and* $\mathrm{ARE}(Z_*, Z) > 10$.

Parameters						
$\beta_1 = \beta_2$	0.5	1	2			
(p_1, p_2)	(0.05, 0.01)	(0.05, 0.15)	(0.05,0.35)	(0.1, 0.01)	(0.1, 0.15)	(0.1,0.35)
(p_1, p_2)	(0.15, 0.01)	(0.15, 0.15)	(0.15,0.35)	(0.35, 0.01)	(0.35, 0.15)	(0.35,0.35)
ρ	0.15	0.45	0.75			
$(\mathrm{HR}_1, \mathrm{HR}_2)$	(0.5, 0.3)	(0.5, 0.7)	(0.5, 0.9)	(0.6, 0.3)	(0.6, 0.7)	(0.6, 0.9)
$(\mathrm{HR}_1, \mathrm{HR}_2)$	(0.7, 0.3)	(0.7, 0.7)	(0.7, 0.9)	(0.8, 0.3)	(0.8, 0.7)	
Total number of cases	624					

4.2. Results

We have set $\Pi = 0.9$ and $\alpha = 0.05$ (other values would not provide additional information). We have chosen meaningful values for $(\beta_1, \beta_2, p_1^{(0)}, p_2^{(0)}, \mathrm{HR}_1, \mathrm{HR}_2, \rho)$, based on those arising in cardiovascular clinical trials (Gómez, Gómez-Mateu, Dafni, 2014) (see Table 1). We restrict our simulation study to 624 scenarios corresponding to $\mathrm{ARE}(Z_*, Z) \leq 10$ and sample sizes smaller than 1100 patients per group. These scenarios yield $\mathrm{ARE}(Z_*, Z)$ values between 0.20 and 9.93, sample sizes, n, for the relevant endpoint between 142 and 1081, and, n_*, for the composite endpoint between 53 and 1077 (see Table 2). Similar results were obtained for Case 1, when neither the relevant nor the additional endpoint includes a terminating event, and for Case 3 when the relevant endpoint includes a terminating event, and we only discuss here Case 1.

Table 2: *Computed values of n, n_* and* $\mathrm{ARE}(Z_*, Z)$ *in step 3 of the simulation based on the parameter values given in Table 1.*

	min	median	max
n	142	509	1081
n_*	53	398	1077
$\mathrm{ARE}(Z_*, Z)$	0.2	1.04	9.93

The empirical powers $\hat{\Pi}_1$ in our simulation study resulted in powers between 0.87 and 0.94, with a median of 0.91. A slightly higher median was found for scenarios with low hazard ratios. This finding is acknowledged as well by Freedman (1982).

Table 3 provides the percentiles for the absolute value differences between $\hat{\Pi}_*$ and $\hat{\Pi}_1$. We observe that in 75% of the cases the difference is smaller than 2.3%, and among cases with ARE as large as 3 the difference shrinks to 1.9%. There are, however, few instances, where this difference can be as large as 6%, and they deserve a closer look.

Table 3: *Percentiles of* $|\hat{\Pi}_* - \hat{\Pi}_1|$ *as a function of ARE values, where* w_i *indicates the corresponding percentile.*

	min	$w_{0.1}$	$w_{0.25}$	$w_{0.5}$	$w_{0.75}$	$w_{0.9}$	max
For all ARE	0	0.002	0.004	0.010	0.023	0.036	0.062
$\text{ARE}(Z_*, Z) \leq 3$	0	0.002	0.004	0.008	0.019	0.033	0.062
$\text{ARE}(Z_*, Z) > 3$	0.001	0.009	0.016	0.026	0.038	0.046	0.062

Figure 2 plots the differences $\hat{\Pi}_* - \hat{\Pi}_1$ as a function of the $\text{ARE}(Z_*, Z)$ values. The behaviour is remarkably different when $\text{ARE}(Z_*, Z) \leq 3$ or $\text{ARE}(Z_*, Z) > 3$. Whenever $\text{ARE}(Z_*, Z) \leq 3$, $\hat{\Pi}_*$ fluctuates around $\hat{\Pi}_1$, within a range of 4%. However, when $\text{ARE}(Z_*, Z) > 3$, corresponding mostly to scenarios where treatment has an stronger effect on the additional endpoint than on the relevant endpoint ($\text{HR}_2 \leq \text{HR}_1 - 0.2$) and the anticipated number of events in the control group is larger for the additional endpoint than for the relevant ($p_2^{(0)} \geq p_1^{(0)}$), the empirical power $\hat{\Pi}_*$ of the logrank test based on the CE never achieves the same power as the logrank test for the relevant endpoint would get. In these cases the interpretation of the $\text{ARE}(Z_*, Z)$ as the ratio of the sample sizes, n/n_*, is not as straightforward. Nevertheless, this does not mean that the recommendation of using the CE does not have to be followed since larger values for n_* needed to attain the same power as n does, would reduce the ARE value but not as much as to cross the "1" border that would imply to use the relevant endpoint instead of the CE.

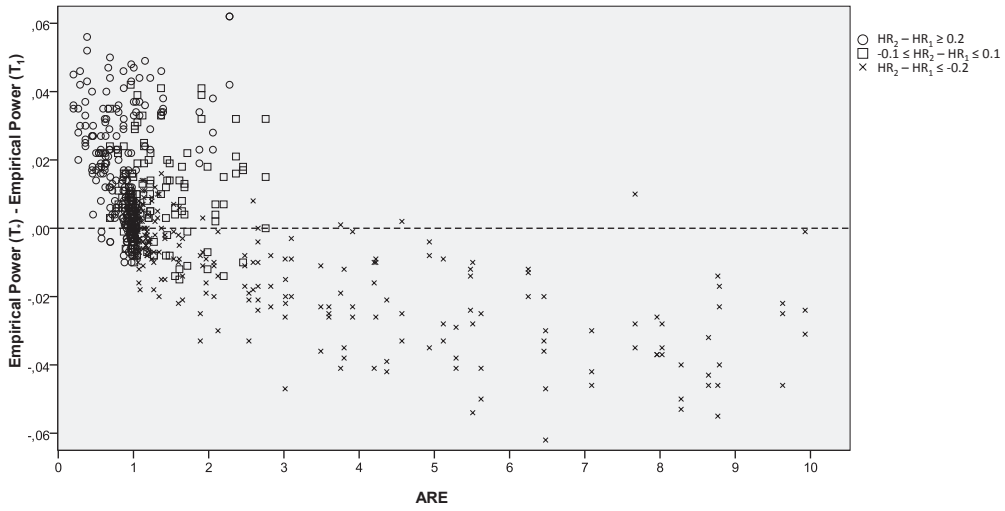

Figure 2: *Differences between empirical powers* $\hat{\Pi}_* - \hat{\Pi}_1$ *as function of* $\text{ARE}(Z_*, Z)$ *and in terms of* $\text{HR}_2 - \text{HR}_1$.

If we analyze the differences between $\hat{\Pi}_*$ and $\hat{\Pi}_1$ as a function of the differences between the two hazard ratios ($\mathrm{HR}_2 - \mathrm{HR}_1$), we observe that when the two hazard ratios are very close, the two empirical powers are as well very close. Whenever $\mathrm{HR}_2 - \mathrm{HR}_1 \leq -0.2$, not only $\mathrm{ARE}(Z_*, Z)$ values tend to be higher, but also $\hat{\Pi}_* < \hat{\Pi}_1$. (see Figure 2).

Taking into account that absolute differences between powers smaller than 5% could be considered irrelevant, we conclude that the asymptotic relationship $\mathrm{ARE}(Z_*, Z) = n/n_*$ is valid in the majority of scenarios.

All computations in this paper have been implemented in R and are available on request to either author.

5. Discussion

Pitman's relative efficiency is defined as the limiting ratio of sample sizes to give the same asymptotic power under sequences of local alternatives. Given two asymptotically standard normal tests S_n and T_m under the same null and alternative hypotheses, the alternative definition $\mathrm{ARE} = (\mu_S/\mu_T)^2$ where $\sqrt{n}\mu_S$ and $\sqrt{m}\mu_T$ are the respective means under local alternatives, can be used because the equality of the powers holds if $\frac{m}{n} = (\frac{\mu_S}{\mu_T})^2$.

Gómez and Lagakos' method uses the alternative definition of ARE to develop all the computations for the two corresponding logrank tests. Our goal has been to check that the relationship between $(\mu_S/\mu_T)^2$ and the ratio of sample sizes still held when the two hypotheses under test were not the same (H_0 versus H_a and H_0^* versus H_a^*).

It is important to keep in mind that these two hypotheses tests are by no means equivalent, for instance, to check whether treatment has a beneficial effect, we might use $\mathscr{E}_1$ or we might add endpoint $\mathscr{E}_2$ and use $\mathscr{E}_*$. As it is shown in Gómez (2011), even if we assume that the times to $\mathscr{E}_1$ and to $\mathscr{E}_2$ are independent, a beneficial effect on $\mathscr{E}_*$ can occur simultaneously with a beneficial effect on $\mathscr{E}_1$ and a harmful effect on $\mathscr{E}_2$ and not finding a beneficial effect on the composite event $\mathscr{E}_*$ is no guarantee of not having some effect on the individual events $\mathscr{E}_1$ or $\mathscr{E}_2$.

The main result of this paper proves that $\mathrm{ARE}(Z_*, Z)$ coincides with n/n_*, being n and n_* the sample sizes needed to detect specific alternatives HR_1 and HR_2 to attain power Π and for the same significance level α. Therefore, we can use and interpret ARE in its usual way.

The simulation study has been conducted in such a way that for fixed values n and $\mathrm{ARE}(Z_*, Z)$, the sample size n_* is calculated as $n_* = n/\mathrm{ARE}(Z_*, Z)$. Hence an approximate equality of the empirical powers $\hat{\Pi}_1$, of logrank test Z for H_0 versus $H_{a,n}$, and of $\hat{\Pi}_*$ of logrank test Z_* for H_0^* versus $H_{a,n}^*$, indicates that the relationship $\mathrm{ARE}(Z_*, Z) = n/n_*$ holds. Main results from our simulations show that the absolute differences between $\hat{\Pi}_1$ and $\hat{\Pi}_*$ are most of the times less than 2.5%, hence the usual interpretation between (n, n_*) and $ARE(Z_*, Z)$ holds for finite sample sizes.

For those scenarios under which $ARE(Z_*, Z) > 3$, we observe that the empirical power of the test based on $\mathscr{E}_*$ never achieves the empirical power that the logrank test based on $\mathscr{E}_1$ would get. Consequently, larger values of n_* would be needed to attain the same power as n does. In these instances, even though the relationship $\mathrm{ARE}(Z_*, Z) = n/n_*$ is not necessarily true, the recommendation to use the composite endpoint $\mathscr{E}_*$ instead of the relevant endpoint $\mathscr{E}_1$ will still be valid because very rarely a value of $ARE(Z_*, Z) > 3$ would go down to less than 1. However, caution will be needed if one wants to use the relationship $\mathrm{ARE}(Z_*, Z) = n/n_*$ to compute the required sample size n_* if $ARE(Z_*, Z) > 3$. In these cases, a different formulation should be seek.

References

Ferreira-González, I., Permanyer-Miralda, G., Busse, J.W., Bryant, D.M., Montori, V.M., Alonso-Coello, P., Walter, S.D. and Guyatt, G.H. (2007). Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. *Journal of Clinical Epidemiology*, 60, 651–657.

Freedman, L.S. (1982). Tables of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine*, 1, 121–129.

Freemantle, N., Calvert, M., Wood, J., Eastaugh, J. and Griffin, C. (2003). Composite outcomes in Randomized Trials. Greater precision but with greater uncertainty? *Journal of the American Medical Association*, 289, 2554–2559.

Gómez, G. (2011). Some theoretical thoughts when using a composite endpoint to prove the efficacy of a treatment. *International Workshop on Statistical Modelling. Proceedings of the 26th International Workshop on Statistical Modelling*, València 14–21. http://hdl.handle.net/2117/22571. Last accessed 19 May 2014.

Gómez, G., Gómez-Mateu, M. and Dafni, U. (2014). Informed Choice of Composite End Points in Cardiovascular Trials. *Circulation. Cardiovascular Quality and Outcomes*, 7, 170–178.

Gómez, G. and Lagakos, S.W. (2013). Statistical considerations when using a composite endpoint for comparing treatment groups. *Statistics in Medicine*, 32, 719–738.

Lehmann, E.L. and Romano, J.P. (2005). *Testing Statistical Hypotheses*, 3rd Ed. Springer.

Lindner, A.M. and Szimayer, A. (2005). A limit theorem for copulas. *Sonderforschungsbereich 386, Paper 433*. http://epub.ub.uni-muenchen.de/1802. Last accessed 19 May 2014.

Montori, V.M., Permanyer-Miralda, G., Ferreira-González, I., Busse, J.W., Pacheco-Huergo, V., Bryant, D., Alonso, J., Akl, E.A., Domingo-Salvany, A., Mills, E., Wu, P., Schnemann, H.J., Jaeschke, R. and Guyatt, G.H. (2005). Validity of composite end points in clinical trials. *British Medical Journal*, 330, 594–596.

Plana-Ripoll, O. and Gómez, G. (2014). Extension of the ARE method to select the Primary Endpoint in a Randomized Clinical Trial. *Submitted*.

Schoenfeld, D. (1981). The Asymptotic Properties of Nonparametric Tests for Comparing Survival Distributions. *Biometrika*, 68, 316–319.

Wittkop, L., Smith, C., Fox, Z., Sabin, C., Richert, L., Aboulker, J.P., Phillips, A., Chêne, G., Babiker, A. and Thiébaut, R. on behalf of NEAT-WP4. (2010). Methodological issues in the use of composite endpoints in clinical trials: examples from the HIV field. *Clinical Trials*, 7, 19–35.