CENTRE DE RECERCA MATEMÀTICA

Title:	*A new approach for sizing trials with composite binary endpoints using anticipated marginal values and accounting for the correlation between components*
Journal Information:	*Statistics in Medicine*
Author(s):	Marta Bofill Roig and Guadalupe Gómez Melis.
Volume, pages:	*1935-1956*, DOI:[www.doi.org/10.1002/sim.8092]

A new approach for sizing trials with composite binary endpoints using anticipated marginal values and accounting for the correlation between components

Marta Bofill Roig, Guadalupe Gómez Melis

Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Barcelona, Spain

Abstract

Composite binary endpoints are increasingly used as primary endpoints in clinical trials. When designing a trial, it is crucial to determine the appropriate sample size for testing the statistical differences between treatment groups for the primary endpoint. As shown in this work, when using a composite binary endpoint to size a trial, one needs to specify the event rates and the effect sizes of the composite components as well as the correlation between them. In practice, the marginal parameters of the components can be obtained from previous studies or pilot trials, however, the correlation is often not previously reported and thus usually unknown. We first show that the sample size for composite binary endpoints is strongly dependent on the correlation and, second, that slight deviations in the prior information on the marginal parameters may result in underpowered trials for achieving the study objectives at a pre-specified significance level.

We propose a general strategy for calculating the required sample size when the correlation is not specified, and accounting for uncertainty in the marginal parameter values. We present the web platform CompARE to characterize composite endpoints and to calculate the sample size just as we propose in this paper. We evaluate the performance of the proposal with a simulation study, and illustrate it by means of a real case study using CompARE.

Keywords: Composite Binary Endpoints, Correlated Endpoints, Sample Size.

1. Introduction

Many trials are designed to evaluate more than one endpoint with the aim of providing a wider picture of the intervention effects [1, 2]. When the rate of occurrence of an event is expected to be low, it is common to consider the composite event defined as the occurrence of any of a set of pre-specified events. This composite event is usually chosen as the primary efficacy endpoint for comparing two treatment groups, either by comparing proportions between groups at the end of the study or by using time-to-event analysis. In this paper, we focus on composite binary endpoints.

Power analysis and its subsequent sample size calculation have been widely discussed in the literature on comparing two proportions in the univariate case [3, 4, 5, 6]. These standard sample size formulae are based on the effect size and the frequency of occurrence of primary endpoint, and they could be applied in a straightforward way to a composite endpoint if its effect size and frequency are known prior to the initiation of the study. However, the effect size and frequency of observing the composite endpoint depend on the corresponding effect and frequency of the composite components, which are often quite dissimilar and thus make the composite parameters very difficult to anticipate.

The TACTICS-TIMI 18 trial [7] illustrates some problems that might arise when determining the sample size for a primary composite binary endpoint. TACTICS-TIMI 18 was an international, multicenter, randomized trial that evaluated the efficacy of invasive and conservative treatment strategies in patients with unstable angina or non-Q-wave

Email addresses: marta.bofill.roig@upc.edu (Marta Bofill Roig), lupe.gomez@upc.edu (Guadalupe Gómez Melis)

acute myocardial infarction treated with tirofiban, heparin, and aspirin. The primary hypothesis of the TACTICS-TIMI 18 trial was that an early invasive strategy would reduce the combined incidence of death, acute myocardial infarction, and rehospitalization for acute coronary syndromes at six months when compared with an early conservative strategy. The primary endpoint was the composite endpoint formed by death, non-fatal myocardial infarction, and rehospitalization for acute coronary syndrome at 6 months.

Similar research questions such as those in TACTICS-TIMI 18 were previously investigated in the TIMI IIIB and VANQWISH trials[8]. The TIMI IIIB trial [9] considered the primary composite endpoint of death, post-randomization myocardial infarction, and a positive exercise test at 6 weeks; whereas the primary endpoint in the VANQWISH trial [10] was the combination of death and non-fatal myocardial infarction at 12 months of follow-up. The initial planning of TACTICS-TIMI 18 was based on those trials expecting 22% events of the primary composite endpoint in the conservative-strategy group, to detect a relative difference of 25% between the two groups for a 80% power. Those anticipated values resulted in the need to recruit at least 1720 patients. However, TACTICS-TIMI 18 yielded a 19% frequency of observing the combination of death, acute myocardial infarction and rehospitalization at six months, which was remarkably lower than expected and delivered a relative difference of 20% between groups, a figure that is seriously lower than the anticipated 25%. Note that if the anticipated frequency of observing the composite endpoint had been closer to the observed results, at least 2000 patients rather than 1720 would have been required and the sample size needed would have been larger than the one initially planned.

In this paper, we present sample size formulations for detecting a hypothesized difference between treatments in a primary composite binary endpoint based on the event rates and effect sizes of the composite components. The motivation for this is mainly because prior information on the marginal effects and event rates is commonly available from previous or pivotal studies, as illustrated in the TACTICS-TIMI 18 trial. Moreover, the major findings in a trial with a primary composite endpoint should be well supported by its components [1, 11], since the trial could be considered negative if the components are not in line with the result [12, 13]. Nevertheless, as shown in this paper, the sample size calculation for composite endpoints relies not only on the anticipation of the effect size and the event rates of the composite components, but also on the correlation between them. However, even though the marginal parameters could be obtained previously, the correlation is usually not reported in practice and, thus, is frequently unknown and difficult to anticipate.

Several authors have addressed the correlation's influence on sample size determination when more than one endpoint is used as the primary endpoint. Sozu et al. [14] discuss several methods for calculating power and sample size for multiple co-primary binary endpoints, and they study the impact on the sample size, specifically regarding the association among endpoints. Senn and Bretz[15] examine sample size for trials under different power definitions for multiple testing problems. Rauch and Kieser [16] and Sander et al. [17] define a multiple test procedure focused on a composite binary endpoint and a pre-specified main component, and propose an internal pilot study for estimating the unknown parameters and revising the sample size. However, to the best of our knowledge, methodologies are limited in regard to handling the sample size calculation for composite binary endpoints when the correlation is unknown.

The aim of this paper is two-fold. First, we focus on providing a general procedure for sizing trials with composite binary endpoints, doing so on the basis of anticipated information of the composite components even if the correlation is unknown. We show that the sample size for composite binary endpoints is strongly dependent on the correlation, and that slight deviations in the prior information on the marginal parameters may result in trials being too underpowered for achieving the study objectives at the pre-specified significance level. We propose a sample size strategy to calculate the minimum sample size that guarantees the planned power while accounting for, on the one hand, the uncertainty of the correlation value and, on the other, plausible deviations in the marginal parameter values. Second, we present CompARE, a freely available web-based tool for characterizing binary composite endpoints and computing the needed sample size under several settings. CompARE provides aids to help understand the role played by each one of the components of the composite endpoint, as well as their consequences on the required sample size. The methodologies presented in this paper have been implemented in CompARE.

This paper is structured as follows. In Section 2, we introduce the settings of the problem and the CompARE web tool. In Section 3, we review sample size planning when evaluating risk difference. In Section 4, we present sample size formulae for composite binary endpoints based on the parameters of the components plus the correlation. We further describe the performance of these formulae according to the parameters and propose a strategy for sizing trials when the correlation is unknown. In Section 5, we exemplify the proposal by the TACTICS-TIMI 18 trial using CompARE, and in Section 6 we extend the proposal to those trials for which the treatment effect is measured

by the relative risk or odds ratio. In Section 7, we investigate the performance of the power and significance level under misspecification of the correlation and evaluate the proposed sample size strategy with a simulation study. We conclude the paper with the Discussion.

2. Notation, assumptions and CompARE

We consider a randomized clinical trial comparing two treatment groups: the control group ($i = 0$) and treatment group ($i = 1$), each one composed of $n^{(i)}$ patients who are followed for a pre-specified time τ. For simplicity, we consider only two events of potential interest, ε_1 and ε_2. Let X_{ijk} denote the response of the k-th binary endpoint for the j-th patient in the i-th group of treatment ($i = 0, 1$, $j = 1, ..., n^{(i)}$, $k = 1, 2$). The response X_{ijk} is defined by 1 if the event, ε_k, has occurred during the follow-up and 0 otherwise.

We define the binary composite endpoint as the event that occurs whenever one of the endpoints is observed, that is, $\varepsilon_* = \varepsilon_1 \cup \varepsilon_2$. At this point we assume that the composite endpoint is well-defined, that is, both composite components are important enough to be considered; and we include those adverse clinical outcomes that are relevant to the clinical setting. We denote by X_{ij*} the composite response defined as a Bernoulli random variable with probability of observing the event
$p_*^{(i)} = \mathrm{P}(X_{ij*} = 1) = 1 - q_*^{(i)}$, where:

$$X_{ij*} = \begin{cases} 1, \text{ if } X_{ij1} + X_{ij2} \geq 1 \\ 0, \text{ if else } X_{ij1} + X_{ij2} = 0 \end{cases} \tag{1}$$

To evaluate whether there is a risk reduction in the treatment group compared with the control group, we set a hypothesis test where the null hypothesis states that there is no difference between the control and the treatment groups; whereas the alternative hypothesis assumes a risk reduction in the treatment group. The usual measures to evaluate the treatment effect when comparing two groups are the difference in proportions (also called risk difference), denoted by δ_*; the relative risk (or risk ratio), R_*; and the odds ratio, OR_*. The relationship between these measures and the probabilities of observing the binary composite endpoint in each group are given in Table 1, together with the null and alternative hypothesis that should be set in each case. The following sections will be developed in terms of the risk difference $\delta_* = p_*^{(1)} - p_*^{(0)}$ of the composite binary endpoint. Section 6 extends the results to the relative risk and odds ratio.

Table 1: Parameter to anticipate the effect, and set of hypotheses.

	Parameter Effect	**Null hypothesis**	**Alternative hypothesis**
Risk difference	$\delta_* = p_*^{(1)} - p_*^{(0)}$	$\delta_* = 0$	$\delta_* < 0$
Relative risk	$\mathrm{R}_* = p_*^{(1)} / p_*^{(0)}$	$\log(\mathrm{R}_*) = 0$	$\log(\mathrm{R}_*) < 0$
Odds ratio	$\mathrm{OR}_* = \frac{p_*^{(1)}/q_*^{(1)}}{p_*^{(0)}/q_*^{(0)}}$	$\log(\mathrm{OR}_*) = 0$	$\log(\mathrm{OR}_*) < 0$

2.1. An insight into the parameters of the composite endpoint

Let $p_k^{(i)}$ and $q_k^{(i)}$ represent the probabilities that ε_k occurs or not, respectively, for a patient belonging to the i-th group. Let $\rho^{(i)}$ denote Pearson's correlation coefficient between the components in the i-th group. The probability of observing the composite event ε_* is in terms of the probabilities of ε_1 and ε_2 and the correlation, as follows:

$$p_*^{(i)} = 1 - q_1^{(i)} q_2^{(i)} - \rho^{(i)} \sqrt{p_1^{(i)} p_2^{(i)} q_1^{(i)} q_2^{(i)}}, \qquad i = 0, 1 \tag{2}$$

Note here that the probability of observing the composite endpoint becomes smaller as the correlation between the components of the composite increases.

The effect size in the composite endpoint in terms of the risk difference, δ_*, is given by:

$$\delta_* = \delta_1 q_2^{(0)} + \delta_2 q_1^{(0)} - \delta_1 \delta_2 + \rho^{(0)} \sqrt{p_1^{(0)} p_2^{(0)} q_1^{(0)} q_2^{(0)}} - \rho^{(1)} \sqrt{(p_1^{(0)} + \delta_1)(p_2^{(0)} + \delta_2)(q_1^{(0)} - \delta_1)(q_2^{(0)} - \delta_2)} \tag{3}$$

where δ_k ($k = 1, 2$) corresponds to the risk difference for each of its components.

From now on the correlation is assumed equal for both groups and denoted by ρ, that is, $\rho = \rho^{(0)} = \rho^{(1)}$. Let θ denote the vector of event rates of the composite components in the control group, that is, $\theta = (p_1^{(0)}, p_2^{(0)})$, and let λ represent the vector of marginal effect sizes, that is, $\lambda = (\delta_1, \delta_2)$. We will denote the risk difference as a function of the marginal parameters (θ, λ) and the correlation ρ by $\delta_*(\theta, \lambda, \rho)$; and the probability of observing ε_* under the control group by $p_*^{(0)}(\theta, \rho)$. We remark here that when λ and θ are fixed such that $p_k^{(0)} < 0.5$ and $\delta_k < 0$ ($k = 1, 2$), the risk difference $\delta_*(\theta, \lambda, \rho)$ increases with respect to the correlation ρ (see Appendix Appendix A).

2.2. CompARE

We present CompARE[1], an open-source and completely free web platform that can be used as a tool for clinicians, medical researchers and statisticians to compute the sample size according to the procedure proposed in this paper. Furthermore, CompARE can be used to:

1. Determine the sample size for different situations, among them, when the correlation is not known.
2. Specify the treatment effect for the composite endpoint based on the marginal information of the composite components, and to study the performance of the composite parameters according to them.
3. Calculate and interpret the measures of association among the composite components, then investigate their characteristics.
4. Choose the best primary endpoint to lead the trial. CompARE computes the Asymptotic Relative Efficiency method[18, 19], which quantifies differences in the efficiency of using – as the primary endpoint – a composite endpoint over one of its components.

Figure 1 summarizes all the capabilities of CompARE. To use CompARE, the least you should provide is the effect size and event rates of the composite components as well as the correlation.

3. Sample Size when the parameters of the composite endpoint can be anticipated

In this section we summarize the statistics and sample size formulae to test for a risk difference when the probability of occurrence in the control group of the composite binary endpoint can be anticipated and for a given expected risk difference. Since the composite endpoint is an univariate outcome, a single statistical test is performed and, consequently, no multiplicity problem occurs and no statistical adjustment is needed. Therefore, as we will see, the formulas follow the univariate case and are straightforward but to make the paper comprehensive and the following sections meaningful, we displayed them in terms of the composite endpoint parameters.

Herein we assume a clinical trial where, first, patients are randomized to one of two treatment arms following a balanced design and, second, where the primary endpoint is a binary composite endpoint. The aim is to detect a hypothesized risk reduction in the primary composite endpoint at the significance level of α and with desired power equal to $1 - \beta$. Let n be the total sample size required, with $n^{(i)} = n/2$ patients per group ($i = 0, 1$); and let us denote by z_α and z_β the values of standardized normal deviates corresponding to α and β.

The null hypothesis is stated as $H_0^* : p_*^{(1)} - p_*^{(0)} = 0$ and is compared against the alternative hypothesis $H_1^* : p_*^{(1)} - p_*^{(0)} < 0$. To test H_0^* against H_1^* we use the statistic:

$$T_{*,n} = \frac{\widehat{p}_*^{(1)} - \widehat{p}_*^{(0)}}{\sqrt{\widehat{Var}(\widehat{p}_*^{(1)} - \widehat{p}_*^{(0)})}} \tag{4}$$

[1]Link to CompARE: `https://cinna.upc.edu/compare/`, the open-source code for CompARE is available at: `https://github.com/MartaBofillRoig/CompARE`

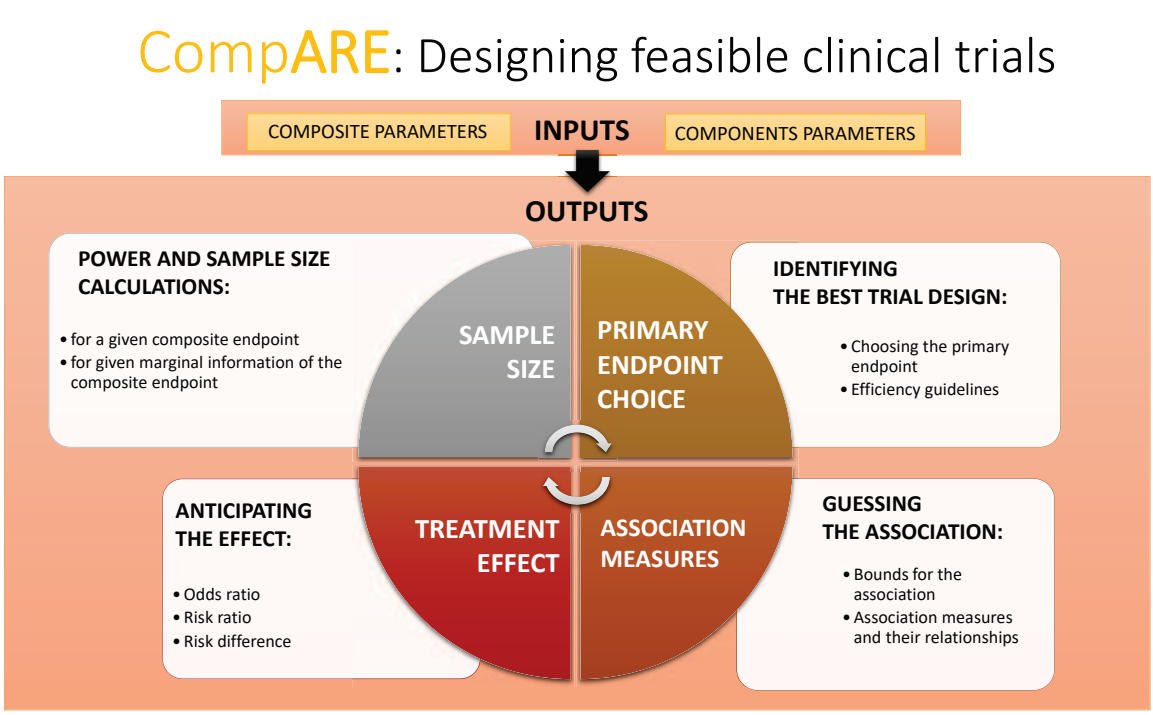

Figure 1: Scheme of the main lines of action in CompARE. Link to the CompARE homepage: `http://cinna.upc.edu/compare/`. Open-source code is available at: `https://github.com/MartaBofillRoig/CompARE`.

where $\widehat{p}_*^{(i)} = \frac{1}{n^{(i)}} \sum_{j=1}^{n^{(i)}} X_{ij*}$. Under H_0^*, $T_{*,n}$ follows, asymptotically, the standard normal distribution. We will reject the null hypothesis at the α level of significance if $T_{*,n} < -z_\alpha$. The variance $Var(\widehat{p}_*^{(1)} - \widehat{p}_*^{(0)})$ in equation (4) can be estimated under H_0^* using the pooled variance estimate[4]:

$$\widehat{Var}_{H_0}(\widehat{p}_*^{(1)} - \widehat{p}_*^{(0)}) = \frac{1}{2n^{(0)}} \cdot \left(\widehat{p}_*^{(0)} + \widehat{p}_*^{(1)}\right) \cdot \left(\widehat{q}_*^{(0)} + \widehat{q}_*^{(1)}\right)$$

or under H_1^* using the unpooled variance estimate:

$$\widehat{Var}_{H_1}(\widehat{p}_*^{(1)} - \widehat{p}_*^{(0)}) = \frac{1}{n^{(0)}} \left(\widehat{p}_*^{(0)}\widehat{q}_*^{(0)} + \widehat{p}_*^{(1)}\widehat{q}_*^{(1)}\right)$$

For a given probability under control group $p_*^{(0)}$, the required sample size using the pooled estimate to have power $1-\beta$ in order to detect an effect size of δ_* at a significance level α is given by [3, 5]:

$$n = 2 \cdot \left(z_\alpha \cdot \sqrt{(2p_*^{(0)} + \delta_*)(2q_*^{(0)} - \delta_*)} + z_\beta \cdot \sqrt{p_*^{(0)}q_*^{(0)} + (p_*^{(0)} + \delta_*)(q_*^{(0)} - \delta_*)} \right)^2 \Big/ \delta_*^2 \quad (5)$$

Note that in (5) we have replaced $p_*^{(1)}$ with $p_*^{(0)} + \delta_*$.

Similarly, the corresponding sample size using the unpooled variance estimate is given by:

$$n = 2 \cdot \left(\frac{z_\alpha + z_\beta}{\delta_*} \right)^2 \cdot \left(p_*^{(0)}q_*^{(0)} + (p_*^{(0)} + \delta_*)(q_*^{(0)} - \delta_*) \right) \quad (6)$$

Note that, under the null hypothesis $H_0^* : p_*^{(1)} - p_*^{(0)} = 0$, expressions (5) and (6) coincide.

4. Sample Size based on anticipated values of the composite components

Sample size formulae underlined in Section 3 are based on the parameters of the composite endpoint, that is, the event rate under the control group, $p_*^{(0)}$, and the treatment effect, δ_*. In this section, we derive the sample size based on the anticipated information on the marginal parameter values and the correlation, even if the correlation value is not fully specified and/or the event rates values are not accurately anticipated.

4.1. Sample size based on composite components

Given the event rates in the control group $\theta = (p_1^{(0)}, p_2^{(0)})$, the expected effect size for each component $\lambda = (\delta_1, \delta_2)$, and the correlation between the occurrence of both components ρ, we will denote by $n(\theta, \lambda, \rho)$ the needed sample size, which is computed by using the unpooled variance estimate, to detect a risk difference $\delta_*(\theta, \lambda, \rho)$ (see equation (3)) at significance level α with $1 - \beta$ power.

The expression for $n(\theta, \lambda, \rho)$ is obtained after direct substitution into formula (6) and is as follows:

$$n(\theta, \lambda, \rho) = \frac{2 \cdot \left(z_\alpha + z_\beta\right)^2 \cdot \left(p_*^{(0)}(\theta,\rho)\left(1 - p_*^{(0)}(\theta,\rho)\right) + \left(p_*^{(0)}(\theta,\rho) + \delta_*(\theta,\lambda,\rho)\right)\left(1 - p_*^{(0)}(\theta,\rho) - \delta_*(\theta,\lambda,\rho)\right)\right)}{\delta_*(\theta,\lambda,\rho)^2} \quad (7)$$

where $p_*^{(0)}(\theta, \rho)$ is given in (2). Note that the sample size also relies on the significance level α and the power $1-\beta$, but these are omitted for ease of notation. The corresponding sample size under the pooled estimate can be analogously calculated by using θ, λ and ρ and its expression can be found in the online support material.

4.2. Sample size bounds

Assuming that the correlation is the same in the two treatment groups, it follows that the correlation takes values between the lower bound, $B_L(\cdot)$, and the upper bound, $B_U(\cdot)$, which are functions of θ and λ, and are defined as:

$$B_L(\theta,\lambda) = \max\left\{-\sqrt{\frac{p_1^{(0)} \cdot p_2^{(0)}}{q_1^{(0)} \cdot q_2^{(0)}}},\ -\sqrt{\frac{q_1^{(0)} \cdot q_2^{(0)}}{p_1^{(0)} \cdot p_2^{(0)}}},\ -\sqrt{\frac{(p_1^{(0)}+\delta_1)\cdot(p_2^{(0)}+\delta_2)}{(q_1^{(0)}-\delta_1)\cdot(q_2^{(0)}-\delta_2)}},\ -\sqrt{\frac{(q_1^{(0)}-\delta_1)\cdot(q_2^{(0)}-\delta_2)}{(p_1^{(0)}+\delta_1)\cdot(p_2^{(0)}+\delta_2)}}\right\} \quad (8)$$

$$B_U(\theta,\lambda) = \min\left\{+\sqrt{\frac{p_1^{(0)} \cdot q_2^{(0)}}{p_2^{(0)} \cdot q_1^{(0)}}},\ +\sqrt{\frac{p_2^{(0)} \cdot q_1^{(0)}}{p_1^{(0)} \cdot q_2^{(0)}}},\ +\sqrt{\frac{(p_1^{(0)}+\delta_1)\cdot(q_2^{(0)}-\delta_2)}{(p_2^{(0)}+\delta_2)\cdot(q_1^{(0)}-\delta_1)}},\ +\sqrt{\frac{(p_2^{(0)}+\delta_2)\cdot(q_1^{(0)}-\delta_1)}{(p_1^{(0)}+\delta_1)\cdot(q_2^{(0)}-\delta_2)}}\right\}$$

Note that when at least one of the event rates is very close to 0, the lower bound $B_L(\lambda, \theta)$ will also be close to 0 and the plausible correlation values will be always positive. We also notice that, in clinical trials the probabilities of observing the events are often quite low and commonly smaller than 0.5. In this case, the expressions for $B_L(\lambda, \theta)$ and $B_U(\lambda, \theta)$ can be simplified. See the online supplementary material for more details.

Considering such bounds for a given marginal parameters θ and λ, the sample size $n(\theta, \lambda, \rho)$ is an increasing function of the correlation ρ, and it is bounded below and above by $n(\theta, \lambda, B_L(\theta, \lambda))$ and $n(\theta, \lambda, B_U(\theta, \lambda))$, respectively. As a consequence, the more correlated the single endpoints are, the larger will be the necessary sample size for detecting the differences between groups in the composite endpoint. Details for this derivation are provided in Appendix Appendix B (see Theorem 1).

4.3. Sample size with uncertain correlation value

Since the correlation plays an important role in calculating the sample size, we propose a strategy for deriving the sample size when the parameters that correspond to the composite components are known and the correlation value is not specified in advance.

Prior knowledge about the effect of the treatment being investigated can lead to scientists foreseeing whether the two events of interest, ε_1 and ε_2, are weakly, moderately or strongly correlated. We allow for prior information by splitting the rank of the correlation into three equal-sized intervals, and we consider three correlations categories: weak for the interval whose correlation values are lower; moderate for those intermediate correlation values; and strong for those correlation values that are higher. If any information exists, we will take it into account and will proceed as follows:

(i) *Correlation bounds for each category:*
Considering the categories weak/moderate/strong for the correlation, the plausible correlation values for a given (θ, λ) are in this situation those between the lower and upper values within each category. If the events are weakly correlated, the correlation is between $B_L(\theta, \lambda)$ and $(B_U(\theta, \lambda) - B_L(\theta, \lambda))/3$; if they are moderately correlated, its value lies between $(B_U(\theta, \lambda) - B_L(\theta, \lambda))/3$ and $2 \cdot (B_U(\theta, \lambda) - B_L(\theta, \lambda))/3$; and if they are strongly correlated, it is between $2 \cdot (B_U(\theta, \lambda) - B_L(\theta, \lambda))/3$ and $B_U(\theta, \lambda)$.
If we cannot place the correlation in any of the above categories, we use the most severe case within its plausible values, then, $B_U(\theta, \lambda)$. (See Table 2).

(ii) *Calculate the sample size in each category:*
For the sample size, we advocate using the maximum sample size across all its possible values. That is, $n(\theta, \lambda, (B_U(\theta, \lambda) - B_L(\theta, \lambda))/3)$, $n(\theta, \lambda, 2\,(B_U(\theta, \lambda) - B_L(\theta, \lambda))/3)$, and $n(\theta, \lambda, B_U(\theta, \lambda))$ for weak, moderate or strong correlations, respectively. Note that since we are assuming the correlation value that maximizes the sample size across its plausible values, we are guaranteeing that the pre-specified power $1 - \beta$ is attained.
If the correlation value can not be ascribed to any category, then, we propose a conservative sample size strategy of using the overall possible maximum sample size, that is, $n(\theta, \lambda, B_U(\theta, \lambda))$. Table 2 outlines the range of correlations and sample sizes values, together with the proposed sample size for each category.

Table 2: Correlation category and its subsequent correlation bounds, $B_L(\cdot)$ and $B_U(\cdot)$ (given in (8)) for event rates of the composite components $\theta = (p_1^{(0)}, p_2^{(0)})$, and marginal effect sizes $\lambda = (\delta_1, \delta_2)$. Sample size bounds for each correlation category and proposed sample size strategy calculated by (7) according to the margins (θ, λ) and for given significance level α and power $1 - \beta$.

Category	Correlation Bounds	Sample Size Bounds	Sample Size
Weak	$\left[B_L(\theta,\lambda),\ \frac{(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}\right)$	$\left[n(\theta,\lambda,B_L(\theta,\lambda)),\ \ n\left(\theta,\lambda,\ \frac{(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}\right)\right]$	$n\left(\theta,\lambda,\ \frac{(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}\right)$
Moderate	$\left[\frac{(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3},\ \frac{2(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}\right)$	$\left[n\left(\theta,\lambda,\ \frac{(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}\right),\ \ n\left(\theta,\lambda,\ \frac{2(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}\right)\right]$	$n\left(\theta,\lambda,\ \frac{2(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}\right)$
Strong	$\left[\frac{2(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3},\ B_U(\theta,\lambda)\right)$	$\left[n\left(\theta,\lambda,\ \frac{2(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}\right),\ \ n(\theta,\lambda,B_U(\theta,\lambda))\right]$	$n(\theta,\lambda,B_U(\theta,\lambda))$
No prior information	$[B_L(\theta,\lambda), B_U(\theta,\lambda))$	$\left[n(\theta,\lambda,B_L(\theta,\lambda)),\ \ n\,(\theta,\lambda,B_U(\theta,\lambda))\right]$	$n\,(\theta,\lambda,B_U(\theta,\lambda))$

4.4. Sample size accounting for departures from the anticipated event rates

The marginal parameters are often estimated through previous studies or pivotal trials with a limited number of patients and whose patient populations or concomitant drugs could differ from the current ones. Because of that, there is great uncertainty in the values that need to be anticipated for computing the sample size. In this section, we consider that the event rates $p_1^{(0)}$ and $p_2^{(0)}$ have been previously estimated and their corresponding standard errors of the point estimate are provided.

Let $I_k = \left[\underline{p}_k^{(0)}, \bar{p}_k^{(0)}\right]$ denote a set of plausible values for the true value of $p_k^{(0)}$. For instance, for those previous trials in which we have the standard deviations for the event rates, we can use the set of plausible values for $p_k^{(0)}$ that a 95% confidence interval would yield. We address the issue of sizing a trial for a significance level α and power $1 - \beta$ based on the intervals I_1 and I_2, and for fixed effects δ_1 and δ_2 when the correlation value is not known.

We state that, for given δ_1 and δ_2 and at fixed $\rho = r$, the sample size $n(p_1^{(0)}, p_2^{(0)}, \lambda, r)$ (see equation (7)) that is needed for power $1 - \beta$ at a significance level α, falls into the interval:

$$\mathcal{I}(r, I_1, I_2, \lambda) = [\ n(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, \lambda, r),\ \ n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, r)\ ] \tag{9}$$

This interval is such that it contains the sample size required to attain power $1 - \beta$, which is necessary for detecting an effect size equal to $\delta_* = p_*^{(1)} - p_*^{(0)}$ at a significance level α according to the marginal effects δ_1 and δ_2, the correlation r, and the event rates $p_k^{(0)}$ within I_k $(k = 1, 2)$. Note that the interval $\mathcal{I}(r, I_1, I_2, \lambda)$ gives us the plausible

sample size values by taking into account the uncertainty of the marginal parameter values, and it provides us the maximum sample size that we would need even though the anticipated event rates are not accurate.

Considering $\Theta = (I_1, I_2, \lambda)$ the set of values for the marginal parameters, and denoting by $\rho_L(\Theta) = \max_{(\pi_1,\pi_2)\in I_1\times I_2} B_L(\pi_1, \pi_2, \lambda)$ and $\rho_U(\Theta) = \min_{(\pi_1,\pi_2)\in I_1\times I_2} B_U(\pi_1, \pi_2, \lambda)$ the lower and upper bounds of the correlation within the set Θ. Then, for all $(\pi_1, \pi_2) \in I_1 \times I_2$, and $\rho \in (\rho_L(\Theta), \rho_U(\Theta))$, we have that:

$$n(\pi_1, \pi_2, \lambda, \rho) \leq U(\Theta) = n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho_U(\Theta)) \tag{10}$$

Furthermore, for given $\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda$, the sample size $n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho)$ is an increasing function of the correlation ρ.

The sample size given by $n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho)$ delimits the values that the sample size could have in terms of the correlation accounting for plausible deviations in the anticipated event rates. If there is no prior information on the correlation, we can use $U(\Theta)$ as the needed sample size. If otherwise, we have some prior information on the correlation value, the rationale used in 4.3 using correlation categories can be as well applied here to the function $n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho)$. Table 3 provides the sample size strategy under this circumstance. We lay out the performance of the sample size when varying the event rates in the intervals I_1 and I_2 and the subsequent sample size behavior according to the correlation in Propositions 2 and 3 in the supplementary material.

Table 3: Correlation category and its subsequent correlation bounds, $\rho_L(\cdot)$ and $\rho_U(\cdot)$ for the intervals of plausible values for event rates, $I_1 = \left[\underline{p}_1^{(0)}, \bar{p}_1^{(0)}\right]$ and $I_2 = \left[\underline{p}_2^{(0)}, \bar{p}_2^{(0)}\right]$, and marginal effect sizes $\lambda = (\delta_1, \delta_2)$, and where $\Theta = (I_1, I_2, \lambda)$ denotes the set of values for the marginal components. Sample size bounds for each correlation category and proposed sample size strategy calculated by (7) according to the intervals I_1 and I_2, the marginal effect sizes λ, for given significance level α and power $1 - \beta$.

Category	Correlation Bounds	Sample Size Bounds	Chosen Sample Size
Weak	$\left[\rho_L(\Theta), \frac{\rho_U(\Theta)-\rho_L(\Theta)}{3}\right)$	$\left[n(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, \lambda, \rho_L(\Theta)),\ \ n\left(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \frac{\rho_U(\Theta)-\rho_L(\Theta)}{3}\right)\right]$	$n\left(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \frac{\rho_U(\Theta)-\rho_L(\Theta)}{3}\right)$
Moderate	$\left[\frac{\rho_U(\Theta)-\rho_L(\Theta)}{3}, \frac{2(\rho_U(\Theta)-\rho_L(\Theta))}{3}\right)$	$\left[n\left(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, \lambda, \frac{\rho_U(\Theta)-\rho_L(\Theta)}{3}\right),\ \ n\left(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \frac{2(\rho_U(\Theta)-\rho_L(\Theta))}{3}\right)\right]$	$n\left(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \frac{2(\rho_U(\Theta)-\rho_L(\Theta))}{3}\right)$
Strong	$\left[\frac{2(\rho_U(\Theta)-\rho_L(\Theta))}{3}, \rho_U(\Theta)\right)$	$\left[n\left(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, \lambda, \frac{2(\rho_U(\Theta)-\rho_L(\Theta))}{3}\right),\ \ n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho_U(\Theta))\right]$	$n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho_U(\Theta))$
No prior information	$[\rho_L(\Theta), \rho_U(\Theta))$	$\left[n\left(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, \lambda, \rho_L(\Theta)\right),\ \ n\left(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho_U(\Theta)\right)\right]$	$n\left(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho_U(\Theta)\right)$

4.5. Power performance of the proposed strategies

Given (θ, λ, ρ) and for a fixed sample size N, the power function using the unpooled variance estimate is defined as:

$$\psi(\theta, \lambda, \rho, N) \quad = \quad \Phi\left(\frac{\sqrt{N}\cdot\delta_*(\theta,\lambda,\rho)}{\sqrt{p_*^{(0)}(\theta,\rho)\left(1-p_*^{(0)}(\theta,\rho)\right)+\left(p_*^{(0)}(\theta,\rho)+\delta_*(\theta,\lambda,\rho)\right)\left(1-p_*^{(0)}(\theta,\rho)-\delta_*(\theta,\lambda,\rho)\right)}} - z_\alpha\right) \tag{11}$$

where $\Phi(\cdot)$ denotes the cumulative distribution of the standard normal distribution. The power function for the pooled variance estimator can be found in the online support material.

In what follows, we show that the planned power $1-\beta$ is achieved with any of the previous strategies in Subsections 4.3 and 4.4.

- If θ and λ are fixed and the correlation value is not known, we have $n(\theta, \lambda, \rho) \leq n(\theta, \lambda, B_U(\theta, \lambda))$ and the proposed sample size becomes $N = n(\theta, \lambda, B_U(\theta, \lambda))$. The resulting power is then such that:

$$\psi\left(\theta, \lambda, \rho, n(\theta, \lambda, B_U(\theta, \lambda))\right) \leq \psi\left(\theta, \lambda, \rho, n(\theta, \lambda, \rho)\right).$$

  The power attained using the upper bound of the correlation is equal to the pre-specified power value $(1 - \beta)$ when the correlation ρ is the maximum value within its range, that is, $B_U(\theta, \lambda)$. Otherwise, if the correlation is less than $B_U(\theta, \lambda)$, the power will be always higher than the pre-specified power. Table S1 in the online supplementary material details the power performance when the correlation categories are taken into account.

- If the event rate value $p_k^{(0)}$ is within the interval I_k for $k = 1, 2$ and the effect sizes λ are fixed, then $n(p_1^{(0)}, p_2^{(0)}, \lambda, \rho) \leq n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho)$. If in addition we have no prior information on the correlation value, then since the sample size increases with respect to the correlation, it follows that $n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho) \leq n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho_U(\Theta))$, and then the proposed sample size turns into $N = n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho_U(\Theta))$. The corresponding power then satisfies:

$$\psi\left(\theta, \lambda, \rho, n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho_U(\Theta))\right) \leq \psi\left(\theta, \lambda, \rho, n(p_1^{(0)}, p_2^{(0)}, \lambda, \rho)\right).$$

  The power attained is equal to the pre-specified power value when the event rates $p_k^{(0)}$ take the upper values $\bar{p}_k^{(0)}$ and the correlation ρ is equal to $\rho_U(\Theta)$. If that is not the case, the power obtained will be larger than the pre-specified $1 - \beta$.

5. Motivating example: TACTICS-TIMI 18 trial

In managing the syndrome of unstable angina and non-Q-wave acute myocardial infarction, there is controversy over whether using an invasive strategy rather than a conservative strategy offers any advantage. TACTICS-TIMI 18 was a randomized trial that evaluated the efficacy of invasive and conservative treatment strategies in patients with unstable angina and non-Q-wave AMI treated with tirofiban, heparin, and aspirin [7].

Patients were randomly assigned to either an early invasive strategy or an early conservative strategy. The primary hypothesis of the TACTICS-TIMI 18 trial was that an early invasive strategy would reduce the combined incidence of death, acute myocardial infarction, and rehospitalization for acute coronary syndromes at six months when compared with an early conservative strategy. The primary endpoint was the composite endpoint formed by a combination of incidence of death or non-fatal myocardial infarction (ε_1), and rehospitalization for acute coronary syndrome (ε_2) at six months.

For illustrative purposes, we assume that a trial will be planned for a similar setting and that the results of TACTICS-TIMI 18 are to be used. Since previous studies to TACTICS-TIMI 18 also considered the events death and non-fatal myocardial infarction altogether, we presume that the event rate and effect size on the endpoint ε_1 can be anticipated despite being composed by two events. The estimated values for the frequency of death or non-fatal myocardial infarction (ε_1) in the conservative strategy group was $\hat{p}_1^{(0)} = 0.095$ with a standard deviation of 0.009; whereas the frequency of rehospitalization for acute coronary syndrome (ε_2) was $\hat{p}_2^{(0)} = 0.137$ with a standard deviation of 0.010. Based on the standard deviations of the estimated event rates, we use the 95% confidence intervals as a set of plausible values among which the true values $p_1^{(0)}$, $p_2^{(0)}$ take values, that is, $I_1 = [0.078, 0.112]$ and $I_2 = [0.117, 0.157]$. The observed effects on TACTICS-TIMI 18 were $\delta_1 = -0.022$ and $\delta_2 = -0.027$, and we will use these as the expected effects on the new experimental trial.

We consider these parameters to construct the correlation bounds outlined in equation (8). The effects δ_1 and δ_2 and the values $\hat{p}_1^{(0)}$ and $\hat{p}_2^{(0)}$ imply that the eligible values for ρ lie in the interval $(-0.10, 0.80)$. Using the intervals I_1 and I_2, the correlation bounds are such that the considered values are plausible for any event rate within I_1 and I_2. This gives us the correlation bounds $(-0.08, 0.77)$. Table 4 and its accompanying figure show the correlation bound according to δ_1 and δ_2 with varying values of the event rates. Observe that the upper bound takes the value 1 when both event rates are equal, and the lower bound tends to 0 when at least one of the event rates becomes smaller.

We illustrate the aspects of calculating power and sample size using the platform CompARE. CompARE calculates the sample size by anticipating the marginal information in terms of either risk difference, relative risk, or odds ratio. In this particular case, we use the statistical test for risk difference under pooled variance in order to ascertain the treatment differences in the composite endpoint at a significance level of $\alpha = 0.025$ and target power of $1 - \beta = 0.80$. The results obtained from CompARE are presented in the form of summary tables and plots.

Figure 2 (left panel) depicts the performance of the sample size in terms of the correlation for given marginal parameters $\theta = (\hat{p}_1^{(0)}, \hat{p}_2^{(0)})$ and $\lambda = (\delta_1, \delta_2)$; and it illustrates the recommended sample size for each correlation category (weak, moderate, and strong). The solid line represents the sample size as a function of the correlation computed for the anticipated values θ, and the shaded areas represent the region of values, constructed by I_1, I_2, δ_1 and δ_2, within which interval the sample size falls. Based on I_1 and I_2 the proposed sample size (in dotted lines) is the upper value of the shaded area within the correlation category.

Event rate values	Correlation Bounds
$\hat{p}_1^{(0)} = 0.095,\ \hat{p}_2^{(0)} = 0.137$	$-0.10 \leq \rho \leq 0.80$
$\bar{p}_1^{(0)} = 0.112,\ \bar{p}_2^{(0)} = 0.157$	$-0.12 \leq \rho \leq 0.81$
$\underline{p}_1^{(0)} = 0.078,\ \underline{p}_2^{(0)} = 0.117$	$-0.08 \leq \rho \leq 0.77$

Table 4: Lower bound, $B_L(\theta, \lambda)$, and upper bound, $B_U(\theta, \lambda)$, for the correlation according to the effect sizes $\delta_1 = -0.022$, $\delta_2 = -0.027$ and for different values of the event rates.

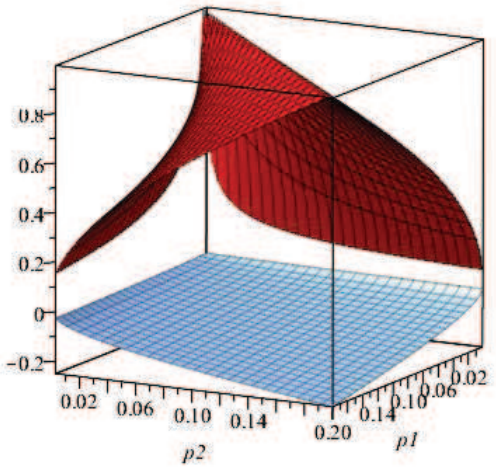

FIGURE Lower bound (surface in blue) and upper bound (in red) for the correlation according to the effect sizes $\delta_1 = -0.022$, $\delta_2 = -0.027$ and where the marginal event rates take values between 0 and 0.2.

Note that the sample size is highly sensitive to the anticipated parameters. For instance, for $\rho = 0.3$, using $\hat{p}_1^{(0)}$ and $\hat{p}_2^{(0)}$, the required sample size is $n = 3030$. This sample size, however, can differ substantially from that calculated using other reasonable values, such as the upper or lower limits for the intervals I_1 and I_2, which would imply $n = 2511$ and $n = 3540$, respectively.

Figure 2 (right panel) describes the statistical power achieved under the proposed method. Assuming that we have correctly anticipated the correlation category, observe that in all cases the achieved power is larger than the planned power, $1 - \beta$. Then, the method guarantees the desired power. If we could correctly anticipate the values of the event rates, then the achieved power would lie between 0.80 and 0.87, in accordance with the plausible correlation values. If we base the sample size calculation on the intervals I_1 and I_2, we will be overestimating the statistical power more than in the previous case, thus obtaining a power between 0.80 and 0.95.

Table 5 describes the proposed sample size for each correlation category and reports the possible values for the statistical power, assuming that we have correctly anticipated the correlation category.

Table 5: Recommended sample size for testing differences between the invasive strategy as compared with the conservative strategy. Underlying marginal parameters are as follows: $p_1^{(0)} = 0.095$, $p_2^{(0)} = 0.137$, $\delta_1 = -0.022$, $\delta_2 = -0.027$. Both sample size and power were calculated based on the statistic (4) under the pooled variance for a one-sided test at the significance level of $\alpha = 0.025$. The given sample size was calculated to detect the effect on the composite endpoint with the desired overall power of $1 - \beta = 0.80$. For calculating the power of the test, three sample size situations were considered, depending on the strength of the correlation: i) weak correlation; ii) moderate correlation; iii) strong correlation.

Based on point values $p_1^{(0)} = 0.095$, $p_2^{(0)} = 0.137$ for the event rates:
Correlation bounds: $B_L(\theta, \lambda) = -0.10$, $B_U(\theta, \lambda) = 0.80$.

Association strength	Correlation	Sample size	Achieved power
Weak	$-0.10 \leq \rho \leq 0.20$	2860	(0.80, 0.86)
Moderate	$0.20 < \rho \leq 0.50$	3425	(0.80, 0.87)
Strong	$0.50 < \rho \leq 0.80$	4201	(0.80,0.87)

Based on intervals $I_1 = [0.078, 0.112]$ and $I_2 = [0.117, 0.157]$ for the event rates:
Correlation bounds: $\rho_L(\Theta) = -0.08$, $\rho_U(\Theta) = 0.77$.

Association strength	Correlation	Sample size	Achieved power
Weak	$-0.08 \leq \rho \leq 0.21$	3355	(0.80, 0.95)
Moderate	$0.21 < \rho \leq 0.49$	3970	(0.80, 0.95)
Strong	$0.49 < \rho \leq 0.77$	4782	(0.80, 0.95)

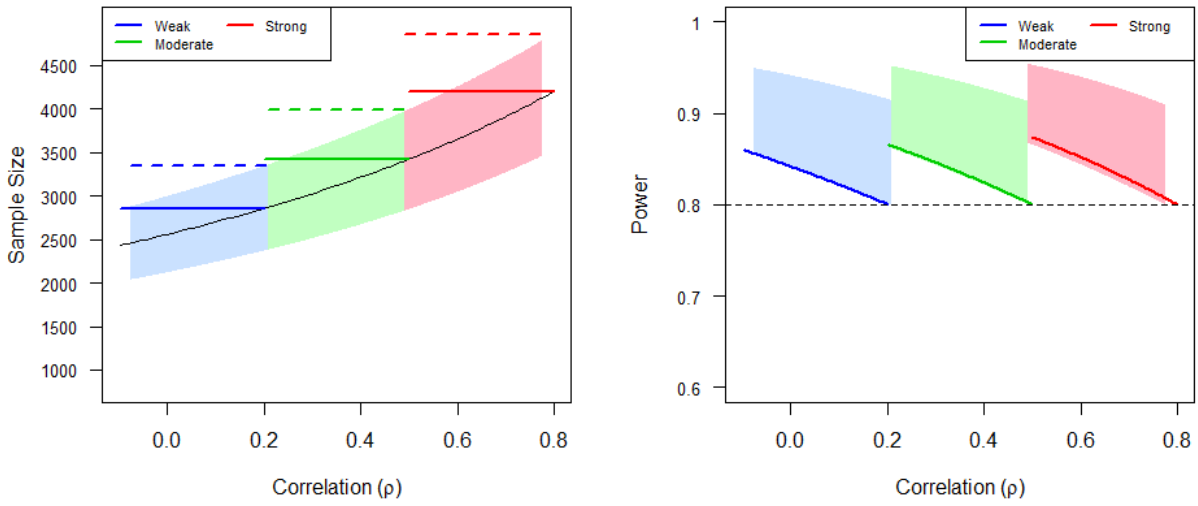

Figure 2: Sample size (left panel) and power (right panel) as a function of the correlation according to the marginal effect sizes $\delta_1 = -0.022$ and $\delta_2 = -0.027$; either based on the point values $\hat{p}_1^{(0)} = 0.095$, $\hat{p}_2^{(0)} = 0.137$ for the event rates (solid line) or based on the interval of plausible values for the event rates $I_1 = [0.078, 0.112]$ and $I_2 = [0.117, 0.157]$ (shaded areas). The proposed sample size for each correlation category is highlighted in solid and dotted lines for, respectively, the point values and the interval values for the event rates.

6. An extension for risk ratio and odds ratio

In this Section, we show that the risk ratio and odds ratio for the composite endpoint can also be expressed in terms of its margins plus the correlation, and we extend the sample size derivation given in Section 4 for evaluating the risk and odds ratio.

6.1. Composite effect expressed in terms of the risk ratio or the odds ratio

Let R_k and OR_k denote the risk ratio and odds ratio, respectively, for the k-th event. The risk ratio for the composite endpoint, R_*, is expressed in terms of the risk ratio of its components R_1 and R_2, the event rates under control group, $p_1^{(0)}$ and $p_2^{(0)}$, and the correlation between them, ρ, as follows:

$$\mathrm{R}_* = \frac{p_1^{(0)}\mathrm{R}_1 + p_2^{(0)}\mathrm{R}_2 - p_1^{(0)}p_2^{(0)}\mathrm{R}_1\mathrm{R}_2 - \rho\sqrt{p_1^{(0)}\mathrm{R}_1 p_2^{(0)}\mathrm{R}_2(1-p_1^{(0)}\mathrm{R}_1)(1-p_2^{(0)}\mathrm{R}_2)}}{1 - q_1^{(0)}q_2^{(0)} - \rho\sqrt{p_1^{(0)}p_2^{(0)}q_1^{(0)}q_2^{(0)}}} \tag{12}$$

Analogously, the odds ratio for the composite endpoint OR_* is defined according to its margins and the correlation is given by:

$$\mathrm{OR}_* = \frac{\left(\left(1+\frac{\mathrm{OR}_1 p_1^{(0)}}{1-p_1^{(0)}}\right)\left(1+\frac{\mathrm{OR}_2 p_2^{(0)}}{1-p_2^{(0)}}\right) - 1 - \rho\sqrt{\frac{\mathrm{OR}_1\mathrm{OR}_2 p_1^{(0)}p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}\right)\cdot\left(1+\rho\sqrt{\frac{p_1^{(0)}p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}\right)}{\left(\left(1+\frac{p_1^{(0)}}{(1-p_1^{(0)})}\right)\cdot\left(1+\frac{p_2^{(0)}}{(1-p_2^{(0)})}\right) - 1 - \rho\sqrt{\frac{p_1^{(0)}p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}\right)\cdot\left(1+\rho\sqrt{\frac{\mathrm{OR}_1\mathrm{OR}_2 p_1^{(0)}p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}\right)} \tag{13}$$

The derivations of equations (12) and (13) are postponed to Appendix Appendix A. By inspection of (3), (12), and (13), we observe that if there is no effect on the components, that is, $\delta_1 = \delta_2 = 1$, $\mathrm{R}_1 = \mathrm{R}_2 = 1$ or $\mathrm{OR}_1 = \mathrm{OR}_2 = 1$, then there is no effect on the composite endpoint, $\delta_* = \mathrm{R}_* = \mathrm{OR}_* = 1$. However, the reciprocal does not follow: no effect on the composite endpoint is compatible with some effect on the components. Therefore, it is important to remark, as other authors have warned before [21, 22, 23], that not finding a beneficial effect on composite endpoint is not a guarantee of not having some effect on the components, hence the effect on the composite endpoint cannot be treated as if it were an indicator of some specific effect on its components.

6.2. Sample size calculations in terms of risk ratio and odds ratio

The null hypothesis in terms of the risk ratio is stated as $H_0^* : \log(R_*) = 0$ and the alternative hypothesis assuming a risk reduction is $H_1^* : \log(R_*) < 0$. The statistic that we use for testing the significance of the relative risk R_* is:

$$Z_{*,n} \quad = \quad \log(\widehat{R}_*)/\sqrt{\widehat{Var}(\log(\widehat{R}_*))}$$

where $\widehat{R}_* = \widehat{p}_*^{(1)}/\widehat{p}_*^{(0)}$. Under H_0^*, $Z_{*,n}$ asymptotically follows the standard normal distribution; thus, we will reject H_0^* at the α significance level if $Z_{*,n} < -z_\alpha$. As in Section 3, we estimate the variance $Var(\widehat{R}_*)$ using the pooled variance by means of $\widehat{Var}_{H_0}(\log(\widehat{R}_*)) = \frac{2}{n^{(0)}} \cdot \frac{\widehat{q}_*^{(0)}+\widehat{q}_*^{(1)}}{\widehat{p}_*^{(0)}+\widehat{p}_*^{(1)}}$ or by using the unpooled variance, $\widehat{Var}_{H_1}(\log(\widehat{R}_*)) = \frac{1}{n^{(0)}}\left(\frac{1-\widehat{R}_*\widehat{p}_*^{(0)}}{\widehat{R}_*\widehat{p}_*^{(0)}} + \frac{\widehat{q}_*^{(0)}}{\widehat{p}_*^{(0)}}\right)$.

For a given probability under control group $p_*^{(0)}$, and a significance level α, the needed sample size for detecting a risk ratio $\Gamma_* = p_*^{(1)}/p_*^{(0)}$ with power $1-\beta$ is given by:

$$n \quad = \quad 2 \cdot \left(z_\alpha + z_\beta\right)^2 \cdot \left(\frac{1-\Gamma_* p_*^{(0)}}{\Gamma_* p_*^{(0)}} + \frac{q_*^{(0)}}{p_*^{(0)}}\right) \Big/ \log(\Gamma_*)^2 \tag{14}$$

The corresponding sample size when the pooled variance is used can be seen in Table 6.

When measuring the effect of treatment with the odds ratio, the null hypothesis $H_0^* : \log(OR_*) = 0$ is compared with the alternative hypothesis $H_1^* : \log(OR_*) < 0$. To test the above hypotheses we use the statistic:

$$W_{*,n} \quad = \quad \log(\widehat{OR}_*)/\sqrt{\widehat{Var}(\log(\widehat{OR}_*))}$$

where $\widehat{OR}_* = \frac{\widehat{p}_*^{(1)}/\widehat{q}_*^{(1)}}{\widehat{p}_*^{(0)}/\widehat{q}_*^{(0)}}$ and where the pooled and unpooled variance estimates are given, respectively, by $\widehat{Var}_{H_0}(\log(\widehat{OR}_*)) = \frac{8}{n^{(0)}(\widehat{p}_*^{(0)}+\widehat{p}_*^{(1)})(\widehat{q}_*^{(0)}+\widehat{q}_*^{(1)})}$ and $\widehat{Var}_{H_1}(\log(\widehat{OR}_*)) = \frac{1}{n^{(0)}}\left(\frac{1}{\hat{p}_*^{(0)}\hat{q}_*^{(0)}} + \frac{1}{\hat{p}_*^{(1)}\hat{q}_*^{(1)}}\right)$.

Then the needed sample size is calculated using the unpooled variance, for detecting a treatment difference of $OR_* = \Delta_*$ in order to have power $1-\beta$ at level α for given $p_*^{(0)}$, and it is given by:

$$n \quad = \quad 2 \cdot \left(\frac{z_\alpha + z_\beta}{\log(\Delta_*)}\right)^2 \cdot \left(\frac{(q_*^{(0)} + p_*^{(0)}\Delta_*)^2}{p_*^{(0)} q_*^{(0)} \Delta_*} + \frac{1}{p_*^{(0)} q_*^{(0)}}\right) \tag{15}$$

The sample size expression when using the pooled variance can be found in Table 6.

6.3. Sample size derivation based on its margins

Analogously to Section 4 and following the notation in Section 4.4,, we obtain the sample size based on the risk ratio as a function of the marginal effects R_1 and R_2, the event rates θ, and the correlation ρ. To do so, we take the event rate and risk ratio of the composite endpoint for their expressions (which are defined according to θ, R_1, R_2 and ρ, see equations (2) and (12)), and then substitute these into the sample size formula in (14). We denote by $n(\theta, R_1, R_2, \rho)$ the needed sample size for evaluating the risk ratio computed for specific values θ, R_1, R_2, and ρ. We will analogously proceed with sample size in terms of the odds ratio using the effects OR_1 and OR_2, then denote by $n(\theta, OR_1, OR_2, \rho)$ the corresponding sample size.

In what follows, we describe the performance of the sample size when the effect is measured by odds ratio or risk ratio. Further details of these properties and their empirical proof are to be found in the web supplementary material.

- For fixed (θ, R_1, R_2) or (θ, OR_1, OR_2), the sample size for testing the effect measured by the risk ratio, $n(\theta, R_1, R_2, \rho)$, and the sample size for testing the odds ratio, $n(\theta, OR_1, OR_2, \rho)$, are increasing functions of the correlation ρ.
- For given R_1 and R_2 at fixed $\rho = r$, the needed sample size $n(p_1^{(0)}, p_2^{(0)}, R_1, R_2, r)$ to have power $1-\beta$ at a significance level α falls into the interval:

$$\mathbf{I}(r, I_1, I_2, R_1, R_2) = [\ n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, R_1, R_2, r),\ \ n(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, R_1, R_2, r)\ ] \tag{16}$$

  Analogously, for given OR_1 and OR_2, the needed sample size $n(p_1^{(0)}, p_2^{(0)}, OR_1, OR_2, r)$ lies within the interval:

$$\mathbf{I}(r, I_1, I_2, OR_1, OR_2) = [\ n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, OR_1, OR_2, r),\ \ n(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, OR_1, OR_2, r)\ ]$$

Table 6: Formulae for sample size determination when comparing two treatments with respect to difference proportions, relative risks or odds ratio contrasts in a balanced design; where n and $n^{(i)}$ denote the total sample size and sample size per group ($i = 0, 1$) needed for testing the effect δ_*, Γ_* or Δ_* for a given event rate within control group $p_*^{(0)}$ at significance level α with $1-\beta$ power.

	Variance estimator	**Sample Size formula**
Risk difference		
Pooled variance	$\frac{\left(\widehat{p}_*^{(0)}+\widehat{p}_*^{(1)}\right)\left(\widehat{q}_*^{(0)}+\widehat{q}_*^{(1)}\right)}{2n^{(0)}}$	$n = 2\cdot\left(z_\alpha\cdot\sqrt{2\bar{p}_*\bar{q}_*}+z_\beta\cdot\sqrt{p_*^{(0)}q_*^{(0)}+(p_*^{(0)}+\delta_*)(q_*^{(0)}-\delta_*)}\right)^2\Big/\delta_*^2$
Unpooled variance	$\frac{\left(\widehat{p}_*^{(0)}\widehat{q}_*^{(0)}+\widehat{p}_*^{(1)}\widehat{q}_*^{(1)}\right)}{n^{(0)}}$	$n = 2\cdot\left(z_\alpha+z_\beta\right)^2\cdot\left(p_*^{(0)}q_*^{(0)}+(p_*^{(0)}+\delta_*)(q_*^{(0)}-\delta_*)\right)\Big/\delta_*^2$
Risk ratio		
Pooled variance	$\frac{2}{n^{(0)}}\cdot\frac{\widehat{q}_*^{(0)}+\widehat{q}_*^{(1)}}{\widehat{p}_*^{(0)}+\widehat{p}_*^{(1)}}$	$n = 2\cdot\left(z_\alpha\sqrt{\frac{2\bar{q}_*}{\bar{p}_*}}+z_\beta\sqrt{\frac{1-\Gamma_* p_*^{(0)}}{\Gamma_* p_*^{(0)}}+\frac{q_*^{(0)}}{p_*^{(0)}}}\right)^2\Big/\log(\Gamma_*)^2$
Unpooled variance	$\frac{1}{n^{(0)}}\left(\frac{1-\widehat{\mathrm{R}}_*\widehat{p}_*^{(0)}}{\widehat{\mathrm{R}}_*\widehat{p}_*^{(0)}}+\frac{\widehat{q}_*^{(0)}}{\widehat{p}_*^{(0)}}\right)$	$n = 2\cdot\left(z_\alpha+z_\beta\right)^2\cdot\left(\frac{1-\Gamma_* p_*^{(0)}}{\Gamma_* p_*^{(0)}}+\frac{q_*^{(0)}}{p_*^{(0)}}\right)\Big/\log(\Gamma_*)^2$
Odds ratio		
Pooled variance	$\frac{8}{n^{(0)}(\widehat{p}_*^{(0)}+\widehat{p}_*^{(1)})(\widehat{q}_*^{(0)}+\widehat{q}_*^{(1)})}$	$n = 2\cdot\left(z_\alpha\sqrt{\frac{2}{\bar{p}_*^{(1)}\bar{q}_*^{(1)}}}+z_\beta\cdot\sqrt{\frac{(q_*^{(0)}+p_*^{(0)}\Delta_*)^2}{p_*^{(0)}q_*^{(0)}\Delta_*}+\frac{1}{p_*^{(0)}q_*^{(0)}}}\right)^2\Big/\log(\Delta_*)^2$
Unpooled variance	$\frac{1}{n^{(0)}}\left(\frac{1}{\hat{p}_*^{(0)}\hat{q}_*^{(0)}}+\frac{1}{\hat{p}_*^{(1)}\hat{q}_*^{(1)}}\right)$	$n = 2\cdot\left(z_\alpha+z_\beta\right)^2\cdot\left(\frac{(q_*^{(0)}+p_*^{(0)}\Delta_*)^2}{p_*^{(0)}q_*^{(0)}\Delta_*}+\frac{1}{p_*^{(0)}q_*^{(0)}}\right)\Big/\log(\Delta_*)^2$

where: $\bar{p}_* = \frac{p_*^{(0)}+p_*^{(1)}}{2}$ and $\bar{q}_* = \frac{q_*^{(0)}+q_*^{(1)}}{2}$.

- For all $(\pi_1, \pi_2) \in I_1 \times I_2$ and $\rho \in (\rho_L(\Theta), \rho_U(\Theta))$, it follows that:

$$\begin{aligned} n(\pi_1, \pi_2, \mathrm{R}_1, \mathrm{R}_2, \rho) &\leq \mathcal{U}_R(\Theta) = n(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, \mathrm{R}_1, \mathrm{R}_2, \rho_U(\Theta)) \\ n(\pi_1, \pi_2, \mathrm{OR}_1, \mathrm{OR}_2, \rho) &\leq \mathcal{U}_{OR}(\Theta) = n(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, \mathrm{OR}_1, \mathrm{OR}_2, \rho_U(\Theta)) \end{aligned}$$

  Furthermore, for given $(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, \mathrm{R}_1, \mathrm{R}_2)$ or $(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, \mathrm{OR}_1, \mathrm{OR}_2)$, the sample size functions $n(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, \mathrm{R}_1, \mathrm{R}_2, \rho)$ and $n(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, \mathrm{OR}_1, \mathrm{OR}_2, \rho)$ increase with respect to the correlation ρ.

Note that, unlike when using risk differences, the sample size has its maximum value when both event rates take their lower interval values $\underline{p}_1^{(0)}, \underline{p}_2^{(0)}$ (see equations (9) and (16)).

Also note that if the marginal parameters $(\theta, \mathrm{R}_1, \mathrm{R}_2)$ or $(\theta, \mathrm{OR}_1, \mathrm{OR}_2)$ are anticipated and the correlation is not known, the sample size strategy described in Section 4.3 can be extended to the risk and odds ratio and analogously applied. For fixed effects $(\mathrm{R}_1, \mathrm{R}_2)$ or $(\mathrm{OR}_1, \mathrm{OR}_2)$, and given intervals I_1 and I_2 for the event rates, we can follow the same reasoning as for risk differences in Section 4.4, and use $\mathcal{U}_R(\Theta)$ (analogously $\mathcal{U}_{OR}(\Theta)$) to calculate the required sample size that guarantees the planned power while accounting for the unknown correlation value and uncertainty of the marginal parameter values.

7. A simulation study

We conduct a simulation study to evaluate the strategies proposed in Section 4 for computing the sample size.

7.1. Design

We simulate a two-arm trial with a composite primary endpoint composed of two events, ε_1 and ε_2, according to the following values (which are all summarized in Table 7): the marginal probabilities of observing ε_k ($k = 1, 2$) in the control group $\theta = (p_1^{(0)}, p_2^{(0)})$ take values between 0.01 and 0.2, and they are without loss of generality such that $p_1^{(0)} < p_2^{(0)}$; the risk ratios $\lambda = (\mathrm{R}_1, \mathrm{R}_2)$ are specified for beneficial effects and vary from 0.6 to 0.8; the true correlation between ε_1 and ε_2 is assumed to be common for both groups, and it covers the positive range between 0

and 1. The possible combinations add up to a total of 421 different scenarios which take into account that for given (θ, λ). simulations are performed only for those ρ_{true} between $B_L(\theta, \lambda)$ and $B_U(\theta, \lambda)$ (see (8)).

For each one of these 421 scenarios specified by $(\theta, \lambda, \rho_{true})$, we compute the required sample size $n(\theta, \lambda, \rho(\theta, \lambda))$ for a one-sided test with power $1-\beta = 0.80$ at the significance level $\alpha = 0.025$, which is done by following one of the six different formulations that are derived in Section 3 and Section 6 and, additionally, are all summarized in Table 6.

We distinguish 4 different situations according to the value we assume in $\rho(\theta, \lambda)$ to calculate $n(\theta, \lambda, \rho(\theta, \lambda))$:

1. For the weak correlation category, use $\rho(\theta, \lambda) = B_U(\theta, \lambda)/3$
2. For the moderate correlation category, use $\rho(\theta, \lambda) = 2B_U(\theta, \lambda)/3$
3. For the strong correlation category, use $\rho(\theta, \lambda) = B_U(\theta, \lambda)$
4. For guessing the true correlation, use $\rho(\theta, \lambda) = \rho_{true}$.

Given one scenario specified by $(\theta, \lambda = (R_1, R_2), \rho_{true})$, we evaluate the type I error first by calculating n based on $(\theta, \lambda = (R_1, R_2), \rho(\theta, \lambda))$ and simulating 100000 trials using $(\theta, \lambda = (1, 1), \rho_{true})$. To check the power, we start by calculating n as above, based on $(\theta, \lambda = (R_1, R_2), \rho(\theta, \lambda))$, and then we simulate 100000 trials using $(\theta, \lambda = (R_1, R_2), \rho_{true})$. Altogether, we have to analyze a total of 3368 scenarios.

The above steps have to be reproduced six times according to the different sample size formulae used to compute $n(\theta, \lambda, \rho(\theta, \lambda))$, that is, by stating the effect in terms of the difference in proportions, the risk ratios or the odds ratio, and using both the pooled and the unpooled estimates of the variance. We have performed all computations using the `R` software tool (Version 0.98.1087), and the time required to perform the considered simulations was 55.58h.

Table 7: **Simulation scenarios**: Values of marginal event rates in the control group: $\theta = (p_1^{(0)}, p_2^{(0)})$; treatment effects in terms of the risk ratio: $\lambda = (\mathrm{R}_1, \mathrm{R}_2)$; and correlation ρ_{true} between components. Note that not all the combinations are feasible because the correlation is between $B_L(\theta, \lambda)$ and $B_U(\theta, \lambda)$.

Parameter	**Values**
$p_1^{(0)}$	$0.01, 0.05, 0.10$
$p_2^{(0)}$	$0.01, 0.05, 0.10, 0.15, 0.20$
ρ_{true}	$0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$
Effects used to evaluate the power:	
$\mathrm{R}_1, \mathrm{R}_2$	$0.6, 0.7, 0.8$
Effect used to evaluate the type I error:	
$\mathrm{R}_1 = \mathrm{R}_2$	1

7.2. Power analysis of the proposed strategies for computing sample size

Let $n_{l,m}$ be the required sample size calculated using the formulae described in Table 6 where $l = p, u$, indicating whether the pooled or unpooled variance has been used; and $m = D, R, OR$, indicating the effect measure that has been tested. In other words, for the difference in proportions, $m = D$; for relative risk, $m = R$; and for odds ratio, $m = OR$. Let $\Psi_{l,m}$ denote the empirical power when the total number of participants is $n_{l,m}$ $(l = p, u;\ m = D, R, OR)$.

Whenever the correlation we are using to compute the sample size coincides with the one we have used to run the simulations $(\rho(\theta, \lambda) = \rho_{true})$, the empirical powers are always achieved whether we are using the pooled, $\Psi_{p,m}$, or unpooled, $\Psi_{u,m}$, estimator of the variance. Nevertheless, when testing the difference in proportions, the achieved powers do not substantially differ $(\Psi_{u,D} \cong \Psi_{p,D})$; when testing the treatment differences in terms of the risk ratio or the odds ratio, the power achieved if the unpooled variance estimator is used is slightly larger than the power achieved with the pooled estimator, $\Psi_{u,m} \leq \Psi_{p,m}$, $m = R, OR$ (see Table S3 in supplementary material for a comparison of the two approaches). The results presented herein refer to the unpooled variance estimator. The corresponding results for the pooled variance are summarized in the supplementary material (Table S4 and Figure S1).

When $\rho(\theta, \lambda) \neq \rho_{true}$, we distinguish two types of misspecification. Misspecification type I, ρ_{true} and $\rho(\theta, \lambda)$ pertain to the same correlation category; and Misspecification type II, ρ_{true} and $\rho(\theta, \lambda)$ do not belong to the same category.

Table 8 describes the empirical power in these two cases, which account for the correlation category for the three effect measures that we could use to test the difference between groups. If Misspecification I occurs, the pre-specified power is achieved and might exceed 7%.

For misspecification II, there are two possible scenarios. The first is for those cases where the correlation $\rho(\theta, \lambda)$ is assumed in a stronger correlation category than the one that ρ_{true} belongs to, for instance, if $\rho(\theta, \lambda)$ is assumed to be strong and ρ_{true} is moderate. Under this scenario, $\rho(\theta, \lambda) > \rho_{true}$, and then the planned power is always achieved. The second scenario is when the $\rho(\theta, \lambda)$ is assumed to be in a weaker correlation category than the one that ρ_{true} lies in. For instance, when $\rho(\theta, \lambda)$ is assumed weak and ρ_{true} is moderate. In those cases where $\rho(\theta, \lambda) < \rho_{true}$, the trial will be underpowered.

The empirical power in terms of the difference between the assumed and true correlations is illustrated in Figure 3. Observe that when the assumed correlation is greater than the true correlation, that is, $\rho(\theta, \lambda) > \rho_{true}$, the empirical power is equal to or greater than the pre-specified power. Note that in all cases under the strong correlation category we have $\rho_{true} \leq \rho(\theta, \lambda)$, the pre-specified power is assured even though we failed to anticipate the category. Also note that there are no differences in the achieved power, nor are there any in the method's performance in terms of the measure we are using to evaluate the effect.

Table 8: Median empirical power, given the sample size (under the unpooled variance), depending on the misspecification error and the assumed correlation. Values in parentheses indicate the maximum and minimum of the empirical power.

Assumption	**Misspecification I: Correlation within the category**	**Misspecification II: Correlation outside the category**
Risk Difference		
Weak	0.82 (0.80, 0.86)	0.78 (0.67, 0.80)
Moderate	0.82 (0.80, 0.87)	0.82 (0.74, 0.91)
Strong	0.82 (0.80, 0.87)	0.87 (0.81, 0.95)
Risk Ratio		
Weak	0.82 (0.80, 0.86)	0.78 (0.67, 0.81)
Moderate	0.82 (0.80, 0.87)	0.82 (0.74, 0.90)
Strong	0.82 (0.80, 0.87)	0.88 (0.81, 0.95)
Odds Ratio		
Weak	0.82 (0.80, 0.86)	0.78 (0.67, 0.81)
Moderate	0.82 (0.80, 0.87)	0.82 (0.74, 0.91)
Strong	0.82 (0.80, 0.87)	0.87 (0.81, 0.95)

7.3. Type I error analysis of the proposed strategies for computing sample size

The empirical results in the simulation study show that the type I error is not affected by the misspecification of the correlation. Nonetheless, the empirical type I error under the pooled estimator may be slightly superior to significance level 0.05, especially when the treatment is tested in terms of risk ratio and odds ratio (see Figure S3 in the online support material).

8. Discussion

Composite endpoints are increasingly used as primary endpoints to achieve greater incidence rates of observing the primary event, larger effect sizes and, hopefully, higher statistical power while avoiding multiplicity adjustment. Even so, their use creates challenges in both the design and interpretation of the studies.

It is well known that sample size determination plays a key role in trial design. We have shown that calculating the sample size for composite binary endpoints needs more than the anticipated effect size and event rates of the composite components; it also needs the correlation between them. Sizing clinical trials in which composite endpoints are involved often implies facing the challenge of dealing with the unknown value of the correlation. We have assessed how much the correlation impacts the sample size and, consequently, the attained power. Our conclusion is that

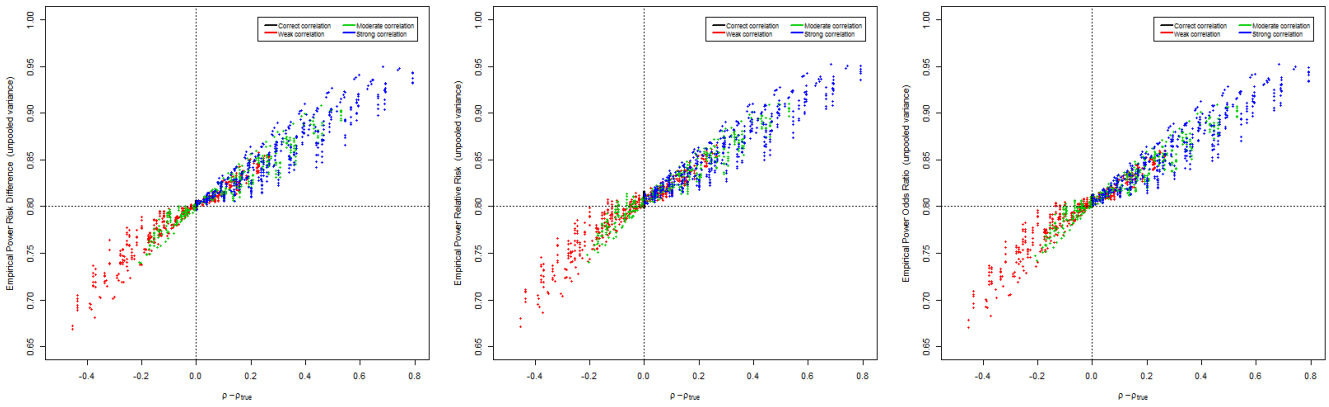

Figure 3: Scatterplot showing the relationship between empirical power versus the difference between the assumed and true correlations for each of the sample size formulas (under unpooled variance) that were used in the simulation study in section 7.

the sample size strongly depends on the correlation and that the more correlation between the components are, the more sample size is needed. Motivated by this concern, we have proposed some strategies for deriving the sample size when the correlation is not specified. The strategy, based on the stratification of the correlation into different categories, assures the pre-specified power even without previous knowledge on the correlation. In addition, if at least we could anticipate the category where the correlation falls into, the achieved power would slightly surpass the planned power (see Table 8). In those cases where not even the correlation category can be anticipated, the interval of plausible values for the sample size might be too wide and the proposed strategy might be extremely conservative. Further research is needed in such cases to obtain more accurate power.

We have illustrated our proposal using the platform CompARE. CompARE extends and includes a previous CompARE web page for composite time-to-event endpoints that was created for handling the relative efficiency of comparing treatment groups in terms of the composite endpoint versus one of its components as the primary endpoint. CompARE allows for the implementation of the Asymptotic Relative Efficiency method, which quantifies the gain in efficiency from using the composite endpoint over one of its components [18, 19], doing so specifically on the basis of information about the different endpoints and anticipated values.

Throughout this work, we have assumed that we are in the planning stage of a randomized clinical trial whose aim is to test the efficacy of a new treatment by comparing its performance with others that have already been approved. These trials are usually known to have much larger sample sizes. For this reason, we have restricted this work to sample size calculation based upon asymptotic approximations of the normal distribution. In previous trial phases devoted to obtain the optimal dose level or to study the toxicity of the new drug, the sample size is not as large as in efficacy trials. In those cases, it could be more appropriate to base the sample size calculations on an exact test. Unlike the tests based on asymptotic distributions, the power function of an exact test usually does not have an explicit form, and the sample size is obtained numerically by greedy search over the sample space. In practice, the applicability of such methods can come across difficulties because intensive computing is required[24]. There is controversy over whether or not to use exact tests, since when the sample size is not large enough, the asymptotic test may not preserve the test size, whereas exact tests could be conservative[25, 26, 27].

The sample size calculation in this work has been derived using the same correlations for both groups. Although this assumption is very often being used[28, 29, 30], it remains to be studied how plausible is in practice. We are working on an extension of our methods to account for different group correlations. Moreover, we are currently studying and implementing in CompARE other association measures for characterizing the strength of dependence between pairs of binary endpoints, such as the relative overlap [31], which in practice might be easier to anticipate.

Interpreting the results of a trial with a primary composite endpoint is particularly challenging. Composite endpoints comprise the information of its components and capture a more complex picture of the intervention's efficacy, however, they might oversimplify the evidence by looking only at the composite effect[12]. A proper study of the

contribution from each separate component should be conducted to ensure a clear understanding of the results. What is more, composite endpoints are in many cases formed by a set of endpoints among whom the clinical relevance highly differs. This could lead to misleading results about whether the treatment benefits only the less important endpoints. Moreover, as shown in Section 6, the effect for the composite does not necessarily reflect the effects for the components. CompARE computes the effect on the composite endpoint and gets constructive numerical and graphical results in order to investigate the role that each component plays.

Comparisons between two groups when using composite endpoints could be based either on the comparison of the corresponding proportions by means, for instance, of a two-proportion z-test or based on the comparison of survival summaries by means, for instance, of a log-rank test. Although this paper focuses on composite binary endpoints, we want to emphasize that time-to-first-event endpoints are very often used and that comparisons based on proportions test could be less powerful than those based on full survival information [32, 33, 34].

Different strategies such as the win ratio[35, 36] and the weighted combined approach[37] have been developed to take into consideration the order of clinical priorities for the composite components when analyzing composite endpoints. Extending this work to more than two components and by incorporating weights remains open for future research.

Acknowledgments

We would like to thank the referees for all their comments which have resulted in a clear improvement of the manuscript. We would also like to thank Dr A. Martín Andrés and the Spanish Network of Biostatistics *Biostatnet* (MTM2015-69068-REDT and MTM2017-90568-REDT).

This work is partially supported by grants MTM2015-64465-C2-1-R (MINECO/FEDER) from the Ministerio de Economía y Competitividad (Spain), and 2017 SGR 622 (GRBIO) from the Departament d'Economia i Coneixement de la Generalitat de Catalunya (Spain). M. Bofill Roig acknowledges financial support from the Spanish Ministry of Economy and Competitiveness, through the María de Maeztu Programme for Units of Excellence in R&D (MDM-2014-0445).

Financial disclosure

None reported.

Conflict of interest

The authors declare no potential conflict of interests.

Supporting information

Additional figures and tables may be found online in the supplementary document. Source code for implementing and reproducing the procedures discussed in this article is available at `https://github.com/MartaBofillRoig/CompARE`.

Appendix

Let X_{ijk} denote the response of the k-th binary endpoint for the j-th patient in the i-th group of treatment ($i = 0, 1$, $j = 1, ..., n$, $k = 1, 2$). We denote by X_{ij*} the composite response defined as

$$X_{ij*} = \begin{cases} 1, \text{ if } X_{ij1} + X_{ij2} \geq 1 \\ 0, \text{ if else } X_{ij1} + X_{ij2} = 0 \end{cases} \quad (.1)$$

W denote by $p_1^{(i)} = \mathrm{P}(X_{ij1} = 1) = 1 - q_1^{(i)}$, $p_2^{(i)} = \mathrm{P}(X_{ij2} = 1) = 1 - q_2^{(i)}$ and $p_*^{(i)} = \mathrm{P}(X_{ij*} = 1) = 1 - q_*^{(i)}$ the probabilities of observing each endpoint in the i-th group. Let $O_k^{(0)}, \delta_k, \mathrm{R}_k, \mathrm{OR}_k$ be the odds under the control group, the risk difference, risk ratio and odds ratio, respectively, for the k-th endpoint, that is, $O_k^{(0)} = \frac{p_k^{(0)}}{q_k^{(0)}}$, $\delta_k = p_k^{(1)} - p_k^{(0)}$,

$\mathrm{R}_k = \frac{p_k^{(1)}}{p_k^{(0)}}$, and $\mathrm{OR}_k = \frac{p_k^{(1)}/q_k^{(1)}}{p_k^{(0)}/q_k^{(0)}}$. We denote by $\theta = (p_1^{(0)}, p_2^{(0)})$ the vector of marginal event rates, and $\lambda = (\delta_1, \delta_2)$ the vector of effect sizes.

Let $\rho^{(i)}$ represent the correlation between X_{ij1} and X_{ij2} in the i-th treatment group, and ρ refer to the correlation when it is assumed to be equal in both groups, i.e., $\rho = \rho^{(0)} = \rho^{(1)}$.

Appendix A. Derivation of the composite effect from the margins

We derive the expression for the composite treatment effect in terms of the marginal component and the correlation described in Sections 2 and 6, and we prove the monotone performance of the risk difference with respect to the correlation ρ.

Theorem Appendix A.1 (Composite effect from margins)**.** *The composite effect for the composite endpoint can be expressed in terms of the component parameters as follows:*

(i) The risk difference for the composite endpoint, δ_*, is determined by the six parameters $p_1^{(0)}$, $p_2^{(0)}$, δ_1, δ_2, $\rho^{(0)}$, $\rho^{(1)}$ and has the following expression:

$$\delta_* = \delta_1 q_2^{(0)} + \delta_2 q_1^{(0)} - \delta_1\delta_2 + \rho^{(0)}\sqrt{p_1^{(0)}p_2^{(0)}q_1^{(0)}q_2^{(0)}} - \rho^{(1)}\sqrt{(p_1^{(0)}+\delta_1)(p_2^{(0)}+\delta_2)(q_1^{(0)}-\delta_1)(q_2^{(0)}-\delta_2)} \tag{A.1}$$

(ii) The risk ratio for the composite endpoint, R_*, is determined by the six parameters $p_1^{(0)}$, $p_2^{(0)}$, R_1, R_2, $\rho^{(0)}$, $\rho^{(1)}$ and has the following expression:

$$\mathrm{R}_* = \frac{p_1^{(0)}\mathrm{R}_1 + p_2^{(0)}\mathrm{R}_2 - p_1^{(0)}p_2^{(0)}\mathrm{R}_1\mathrm{R}_2 - \rho^{(1)}\sqrt{p_1^{(0)}\mathrm{R}_1 p_2^{(0)}\mathrm{R}_2(1-p_1^{(0)}\mathrm{R}_1)(1-p_2^{(0)}\mathrm{R}_2)}}{1 - q_1^{(0)}q_2^{(0)} - \rho^{(0)}\sqrt{p_1^{(0)}p_2^{(0)}q_1^{(0)}q_2^{(0)}}} \tag{A.2}$$

(iii) The odds ratio for the composite endpoint, OR_*, is determined by the six parameters $p_1^{(0)}$, $p_2^{(0)}$, OR_1, OR_2, $\rho^{(0)}$, $\rho^{(1)}$ and has the following expression:

$$\mathrm{OR}_* = \frac{\left(\left(1+\frac{\mathrm{OR}_1 p_1^{(0)}}{1-p_1^{(0)}}\right)\left(1+\frac{\mathrm{OR}_2 p_2^{(0)}}{1-p_2^{(0)}}\right) - 1 - \rho^{(1)}\sqrt{\frac{\mathrm{OR}_1\mathrm{OR}_2 p_1^{(0)}p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}\right)\cdot\left(1+\rho^{(0)}\sqrt{\frac{p_1^{(0)}p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}\right)}{\left(\left(1+\frac{p_1^{(0)}}{(1-p_1^{(0)})}\right)\cdot\left(1+\frac{p_2^{(0)}}{(1-p_2^{(0)})}\right) - 1 - \rho^{(0)}\sqrt{\frac{p_1^{(0)}p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}\right)\cdot\left(1+\rho^{(1)}\sqrt{\frac{\mathrm{OR}_1\mathrm{OR}_2 p_1^{(0)}p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}\right)} \tag{A.3}$$

Proof of Theorem Appendix A.1. (i), (ii) The two expressions (A.1) and (A.2) follow in a straightforward manner after noting that:

$$p_*^{(i)} \;=\; 1 - q_1^{(i)}q_2^{(i)} - \rho^{(i)}\sqrt{p_1^{(i)}p_2^{(i)}q_1^{(i)}q_2^{(i)}} = p_1^{(i)} + p_2^{(i)} - p_1^{(i)}p_2^{(i)} - \rho^{(i)}\sqrt{p_1^{(i)}p_2^{(i)}q_1^{(i)}q_2^{(i)}} \tag{A.4}$$

and taking into account $p_k^{(1)} = \delta_k + p_k^{(0)}$ and $p_k^{(1)} = p_k^{(0)}\mathrm{R}_1$.

(iii) Replacing the probabilities of the composite endpoint with its expression in terms of the marginal parameters plus the correlation (A.4), we have:

$$\mathrm{OR}_* \;=\; \left(\frac{1 - q_1^{(1)}q_2^{(1)} - \rho^{(1)}\sqrt{\frac{p_1^{(1)}p_2^{(1)}}{q_1^{(1)}q_2^{(1)}}}}{q_1^{(1)}q_2^{(1)} + \rho^{(1)}\sqrt{\frac{p_1^{(1)}p_2^{(1)}}{q_1^{(1)}q_2^{(1)}}}}\right)\cdot\left(\frac{1 - q_1^{(0)}q_2^{(0)} - \rho^{(0)}\sqrt{\frac{p_1^{(0)}p_2^{(0)}}{q_1^{(0)}q_2^{(0)}}}}{q_1^{(0)}q_2^{(0)} + \rho^{(0)}\sqrt{\frac{p_1^{(0)}p_2^{(0)}}{q_1^{(0)}q_2^{(0)}}}}\right)^{-1} = \left(\frac{\frac{1}{q_1^{(1)}q_2^{(1)}} - 1 - \rho^{(1)}\sqrt{\frac{p_1^{(1)}p_2^{(1)}}{q_1^{(1)}q_2^{(1)}}}}{1 + \rho^{(1)}\sqrt{\frac{p_1^{(1)}p_2^{(1)}}{q_1^{(1)}q_2^{(1)}}}}\right)\left(\frac{\frac{1}{q_1^{(0)}q_2^{(0)}} - 1 - \rho^{(0)}\sqrt{\frac{p_1^{(0)}p_2^{(0)}}{q_1^{(0)}q_2^{(0)}}}}{1 + \rho^{(0)}\sqrt{\frac{p_1^{(0)}p_2^{(0)}}{q_1^{(0)}q_2^{(0)}}}}\right)^{-1}$$

Notice that:

$$\frac{1}{q_1^{(i)}q_2^{(i)}} \;=\; (1+O_1^{(i)})(1+O_2^{(i)}), \qquad \frac{1}{q_1^{(1)}q_2^{(1)}} \;=\; (1+\mathrm{OR}_1 O_1^{(0)})(1+\mathrm{OR}_2 O_2^{(0)}), \qquad \frac{p_1^{(1)}p_2^{(1)}}{q_1^{(1)}q_2^{(1)}} \;=\; \mathrm{OR}_1\mathrm{OR}_2 O_1^{(0)}O_2^{(0)}$$

Hence:

$$\mathrm{OR}_* = \left(\frac{(1+\mathrm{OR}_1 O_1^{(0)})(1+\mathrm{OR}_2 O_2^{(0)}) - 1 - \rho^{(1)}\sqrt{\mathrm{OR}_1\mathrm{OR}_2 O_1^{(0)} O_2^{(0)}}}{1+\rho^{(1)}\sqrt{\mathrm{OR}_1\mathrm{OR}_2 O_1^{(0)} O_2^{(0)}}}\right)\cdot\left(\frac{(1+O_1^{(0)})(1+O_2^{(0)}) - 1 - \rho^{(0)}\sqrt{O_1^{(0)} O_2^{(0)}}}{1+\rho^{(0)}\sqrt{O_1^{(0)} O_2^{(0)}}}\right)^{-1}$$

By replacing $O_k^{(0)}$ by $\frac{p_k^{(0)}}{1-p_k^{(0)}}$, we obtain (A.3). □

Theorem Appendix A.2 (Risk difference performance)**.** *Assume that $p_k^{(0)} < 1/2$ and $\delta_k < 0$ ($k = 1, 2$). We denote by $\delta_*(\rho, \theta, \lambda)$ the risk difference for the composite endpoint function described in* (A.1)*, specifically in terms of the vector of event rates θ, the marginal effects λ and the correlation ρ. Then, the risk difference for the composite endpoint for a given θ and λ is an increasing function with respect to ρ.*

Proof of Theorem Appendix A.2. Observe that the difference in proportions (A.1) can be written as:

$$\delta_*(\rho, \theta, \lambda) = x(\theta, \lambda) + \rho \cdot y(\theta, \lambda).$$

where: $x(\theta, \lambda) = \delta_1 q_2^{(0)} + \delta_2 q_1^{(0)} - \delta_1\delta_2$, and $y(\theta, \lambda) = \sqrt{p_1^{(0)} p_2^{(0)} q_1^{(0)} q_2^{(0)}} - \sqrt{p_1^{(1)} p_2^{(1)} q_1^{(1)} q_2^{(1)}}$. Then: $\delta_*(\rho+\epsilon;\theta) - \delta_*(\rho;\theta) = \epsilon \cdot y(\theta, \lambda)$. Therefore, $\delta_*(\rho;\theta)$ is an increasing function if and only if $y(\theta, \lambda) > 0$, $\forall\lambda, \theta$, which is equivalent to showing that:

$$\frac{p_1^{(0)} p_2^{(0)} q_1^{(0)} q_2^{(0)}}{p_1^{(1)} p_2^{(1)} q_1^{(1)} q_2^{(1)}} > 1$$

It is enough to prove that for $k = 1, 2$,

$$\frac{p_k^{(1)} q_k^{(1)}}{p_k^{(0)} q_k^{(0)}} < 1$$

To ease the notation call $p_k^{(1)} = a$ and $p_k^{(0)} = b$; and, by assuming $a < b < 1/2$, that implies $a - b < 0$ and $a + b < 1$. We need to prove that:

$$\frac{a(1-a)}{b(1-b)} = \frac{a-a^2}{b-b^2} < 1 \Leftrightarrow b - a < b^2 - a^2 = (b+a)(b-a)$$

Since $b - a > 0$ and $a + b < 1$, then we have $(a + b)(b - a) < (b - a)$. As a consequence $y(\theta, \lambda) > 0$ and the risk difference of the composite endpoint is an increasing function with respect to the correlation. □

Appendix B. Derivation of the sample size for the composite binary endpoint

We establish the sample size formulae for the composite endpoint in terms of the margins and derive its properties, as outlined in Sections 4 and 6.

Appendix B.1. Sample size performance according to the correlation

Lemma Appendix B.1. *Let $N(p, d)$ denote the sample size function for testing the difference in proportions under the unpooled variance estimate, where p denotes the probability under the control group and d the relevant difference to be detected, that is:*

$$N(p, d) = \left(\frac{z_\alpha + z_\beta}{d}\right)^2 \cdot (p \cdot (1-p) + (p+d) \cdot (1-p-d)) \tag{B.1}$$

It follows that $N(p, d)$ is an increasing function with respect to p and with respect to d.

Proof. Observe that:

$$\frac{\partial}{\partial p}N\left(p,d\right)=\frac{\left(z_{\alpha}+z_{\beta}\right)^{2}\left(2-4\,p-2\,d\right)}{d^{2}}$$

Assuming $p<0.5$, then $1-2p>0$ and $2-4p0-2d>0$. Therefore $\frac{\partial}{\partial p}N\left(p,d\right)>0$, the sample size is increasing with respect to p. Moreover,

$$\frac{\partial}{\partial d}N(p,d)=-2\,\frac{\left(z_{\alpha}+z_{\beta}\right)^{2}\left(p\left(1-p\right)+\left(d+p\right)\left(1-p-d\right)\right)}{d^{3}}+\frac{\left(z_{\alpha}+z_{\beta}\right)^{2}\left(1-2p-2\,d\right)}{d^{2}}$$

Note that $1-2p-2d>0$ and therefore, $\frac{\partial}{\partial d}N(p,d)>0$; thus, the sample size is increasing with respect to d. □

Theorem 1. Let θ and λ be the vectors of, respectively, marginal event rates and effect sizes for the composite components, and we denote by ρ the correlation between both components. Then, the sample size $n(\theta,\lambda,\rho)$, for a given θ and λ is an increasing function of the correlation ρ.

Proof of Theorem 1. Since the probability of observing the composite event is given by θ and ρ (see equation (A.4)), and the risk difference for the composite endpoint is given by λ, θ and ρ (see equation (A.1)), then the sample size for the composite endpoint computed by $n(\theta,\lambda,\rho)=N\left(p_{*}(\theta,\rho),\delta_{*}(\lambda,\theta,\rho)\right)$ is a function of λ, θ and ρ.

To prove that the sample size for the composite endpoint $N\left(p_{*}(\theta,\rho),\delta_{*}(\lambda,\theta,\rho)\right)$ increases with ρ, we will show that:

$$\frac{\partial N(p_{*}(\theta,\rho),\delta_{*}(\lambda,\theta,\rho))}{\partial\rho}=\frac{\partial N(p_{*}(\theta,\rho),\delta_{*}(\lambda,\theta,\rho))}{\partial p_{*}(\theta,\rho)}\cdot\frac{\partial p_{*}(\theta,\rho)}{\partial\rho}+\frac{\partial N(p_{*}(\theta,\rho),\delta_{*}(\lambda,\theta,\rho))}{\partial\delta_{*}(\lambda,\theta,\rho)}\cdot\frac{\partial\delta_{*}(\lambda,\theta,\rho)}{\partial\rho}>0 \quad \text{(B.2)}$$

From now on we will omit θ, λ and ρ and use p_{*} and δ_{*} instead of $p_{*}(\theta,\rho)$ and $\delta_{*}(\lambda,\theta,\rho)$. From Lemma Appendix B.1, the sample size $N(p,d)$ in (B.1) is increasing with respect to the treatment effect, d, and with respect to the probability of observing the event under control group, p, hence:

$$\frac{\partial N(p_{*},\delta_{*})}{\partial p_{*}}>0\quad\text{and}\quad\frac{\partial N(p_{*},\delta_{*})}{\partial\delta_{*}}>0$$

We denote by:

$$\frac{\partial p_{*}(\theta,\rho)}{\partial\rho}=-a\quad\text{and}\quad\frac{\partial\delta_{*}(\rho)}{\partial\rho}=a-b$$

where $a,b>0$ and, from Theorem Appendix A.2, $a-b>0$. Then we have:

$$\frac{\partial N(p_{*},\delta_{*})}{\partial\rho}=(a-b)\cdot\left(-2\,\frac{\left(z_{\alpha}+z_{\beta}\right)^{2}\left(p_{*}\,\left(1-p_{*}(\theta,\rho)\right)+\left(\delta_{*}+p_{*}\right)\left(1-p_{*}-\delta_{*}\right)\right)}{{\delta_{*}}^{3}}+\frac{\left(z_{\alpha}+z_{\beta}\right)^{2}\left(1-2\,p_{*}-2\delta_{*}\right)}{{\delta_{*}}^{2}}\right)-a\cdot\frac{\left(z_{\alpha}+z_{\beta}\right)^{2}\left(2-4\,p_{*}-2\delta_{*}\right)}{{\delta_{*}}^{2}}$$

and this is positive if and only if:

$$(a-b)\cdot\left(-2\frac{p_{*}\,\left(1-p_{*}\right)+\left(\delta_{*}+p_{*}\right)\left(1-p_{*}-\delta_{*}\right)}{\delta_{*}}+1-2p_{*}-2\delta_{*}\right)-\left(2-4\,p_{*}-2\delta_{*}\right)\cdot a \quad \text{(B.3)}$$

(B.3) is positive. Then we have:

$$\begin{aligned}&-2\frac{(a-b)}{\delta_{*}}\left(p_{*}\left(1-p_{*}\right)+\left(\delta_{*}+p_{*}\right)\left(1-p_{*}-d\right)\right)-b\left(1-2\,p_{*}-2\delta_{*}\right)+a\left(-1+2\,p_{*}\right)\\&>-2\frac{(a-b)}{\delta_{*}}p_{*}\left(1-p_{*}\right)-2\frac{(a-b)}{\delta_{*}}\left(\delta_{*}+p_{*}\right)\left(1-p_{*}-\delta_{*}\right)-2a\left(1-p_{*}-\delta_{*}\right)+2ap_{*}\end{aligned} \quad \text{(B.4)}$$

Then (B.3) ¿ (B.4), because $a > b$. Note that the first and forth terms are positive, so we end if we see that the second plus third are also positive. This follows from the fact that:

$$-2\frac{(a-b)}{\delta_*}(\delta_* + p_*) - 2a > 0 \quad \Leftrightarrow \quad a\left(1 + \frac{\delta_* + p_*}{\delta_*}\right) < b\left(\frac{\delta_* + p_*}{\delta_*}\right) \Leftrightarrow \frac{a}{b} > \frac{\frac{\delta_* + p_*}{\delta_*}}{1 + \frac{\delta_* + p_*}{\delta_*}}$$

Since $a, b > 0$ and $a - b > 0$, we have $\frac{a}{b} > 1$; and since $\left(\frac{\delta_* + p_*}{\delta_*}\right)$, $\left(1 + \frac{\delta_* + p_*}{\delta_*}\right) < 0$, we have $\left(\frac{\delta_* + p_*}{\delta_*}\right) / \left(1 + \frac{\delta_* + p_*}{\delta_*}\right) \in (0, 1)$. Therefore (B.2) is positive, as we intended to prove. □

References

[1] FDA. Multiple Endpoints in Clinical Trials Guidance for Industry. 2017.

[2] Rosenblatt M. The Large Pharmaceutical Company Perspective. *New England Journal of Medicine*. 2017;376(1):52–60.

[3] Lachin JM. Introduction to Sample Size determination and Power analysis for Clinical Trials. *Controlled Clinical Trials*. 1981;2:92–113.

[4] Donner A. Approaches to sample size estimation in the design of clinical trials-a review. *Statistics in Medicine*. 1984;3(3),199–214.

[5] Fleiss JL. Statistical Methods for Rates and Proportions, Wiley, New York, 1981. ISBN: 978-0-471-52629-2

[6] Friedman LM, Furberg CD, DeMets DL. Fundamentals of Clinical Trials, John Wright, Boston, 1981.

[7] Cannon CP, Weintraub WS, Demopoulos LA, Vicari R, Frey MJ, Lakkis N, TACTICS (Treat Angina with Aggrastat and Determine Cost of Therapy with an Invasive or Conservative Strategy)–Thrombolysis in Myocardial Infarction 18 Investigators. Comparison of Early Invasive and Conservative Strategies in Patients with Unstable Coronary Syndromes Treated with the Glycoprotein IIb/IIIa Inhibitor Tirofiban. *New England Journal of Medicine*. 2001;344(25), 1879–1887.

[8] Cannon CP, Weintraub WS, Demopoulos LA, Robertson DH, Gormley GJ, Braunwald, E. Invasive versus conservative strategies in unstable angina and non-Q-wave myocardial infarction following treatment with tirofiban: rationale and study design of the international TACTICS-TIMI 18 Trial. *The American Journal of Cardiology*. 1998;82(6), 731–6.

[9] Anderson HV, Cannon CP, Stone PH, Williams DO, McCabe CH, Knatterud GL, Thompson B, Willerson JT, Braunwald E, for the TIMI-IIIB Investigators. One-year results of the Thrombolysis in Myocardial Infarction (TIMI) IIIB clinical trial. A randomized comparison of tissue-type plasminogen activator versus placebo and early invasive versus early conservative strategies in unstable angina and non-Q-wave myocardial infarction. *J Am Coll Cardiol*. 1995;26:1643– 1650.

[10] Boden WE, O'Rourke RA, Crawford MH, Blaustein AS, Deedwania PC, Zoble RG, Wexler LF, Pepine CJ, Ferry DR, Chow BK, Lavori PW, for the Veterans Affairs Non-Q-Wave Infaction Strategies in Hospital (VANQWISH) Trial Investigators. Outcomes in patients with acute non-Q-wave myocardial infarction randomly assigned to an invasive as compared with a conservative strategy. *N Engl J Med*. 1998;338:1785–1792.

[11] European Medicines Agengy Commitee For Proprietary Medicinal Products (CPMP). Guideline on multiplicity issues in clinical trials. 2016. `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2017/03/WC500224998.pdf`.

[12] Pocock SJ, McMurray JJV, Collier TJ. Statistical Controversies in Reporting of Clinical Trials Part 2 of a 4-Part Series on Statistics for Clinical Trials. *Journal of the American College of Cardiology*. 2015;66(23), 2648–2662.

[13] ICH guideline. Statistical principles for clinical trials (E9). 1999. `http://www.fda.gov/downloads/drugs/guidancecomplianceregulatory/informati`

[14] Sozu T, Sugimoto T, Hamasaki T. Sample size determination in clinical trials with multiple co-primary binary endpoints. *Statistics in Medicine*. 2010;29(21):2169–2179. doi:10.1002/sim.3972.

[15] Senn S, Bretz F. Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics*. 2007;6:161–170. doi:10.1002/pst.

[16] Rauch G, Kieser M. Multiplicity adjustment for composite binary endpoints. *Methods of Information in Medicine*. 2012;51(4):309–317.

[17] Sander, A, Rauch, G, Kieser, M. (2016). Blinded sample size recalculation in clinical trials with binary composite endpoints. *Journal of Biopharmaceutical Statistics*.

[18] Gómez G, Lagakos SW. Statistical considerations when using a composite endpoint for comparing treatment groups. *Statistics in Medicine*. 2013;32(5):719–738. doi:10.1002/sim.5547.

[19] Bofill Roig M, Gómez Melis G. Selection of composite binary endpoints in clinical trials. *Biometrical Journal*. 2017;00:1–16.

[20] Whitt, W. Bivariate distributions with given marginals. *Annals of Statistics*. 1976;4(6), 1280–1289.

[21] Ferreira-González I, Permanyer-Miralda G, Busse JW, et al. Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. *Journal of Clinical Epidemiology*. 2007;60(7):651-657.

[22] Ferreira-González I, Permanyer-Miralda G, Busse JW, et al. Composite endpoints in clinical trials: the trees and the forest. *Journal of Clinical Epidemiology*. 2007;60(7):660-661.

[23] Tomlinson, G, Detsky, AS. Composite End Points in Randomized Trials. *JAMA*. 2010;303(3), 267.

[24] Chow, SC, Shao, J, Wang, H. Sample size calculations in clinical research, Chapman & Hall/CRC, Boca Raton, FL, 2008 (2nd edn). ISBN: 1-58488-982-9.

[25] Fagerland MW, Lydersen S, Laake P. Recommended confidence intervals for two independent binomial proportions. *Statistical Methods in Medical Research*. 2015;24(2):224-254.

[26] Crans GD, Shuster JJ. How conservative is Fisher's exact test? A quantitative evaluation of the two-sample comparative binomial trial. *Statistics In Medicine*. 2008;27:3598-3611. doi:10.1002/sim.

[27] Andrés AM, Tejedor H. Comments on 'How conservative is Fisher's exact test? A quantitative evaluation of the two-sample comparative binomial trial'. *Statistics In Medicine*. 2009;(28):173-179.

[28] Sugimoto T, Hamasaki T, Evans SR, Sozu T. Sizing clinical trials when comparing bivariate time-to-event outcomes. *Statistics in Medicine*. 2017;36(9):1363-1382.

[29] Asakura K, Hamasaki T, Evans SR. Interim evaluation of efficacy or futility in group-sequential trials with multiple co-primary endpoints. *Biometrical Journal*. 2017;59(4):703-731.

[30] Ando Y, Hamasaki T, Evans SR, et al. Sample Size Considerations in Clinical Trials when Comparing Two Interventions using Multiple Co-Primary Binary Relative Risk Contrasts. *Stat Biopharm res*. 2015;7(2):81-94.

[31] Marsal JR, Ferreira-González I, Bertran S, Ribera A, Permanyer-Miralda G, García-Dorado D, Gómez G. The Use of a Binary Composite Endpoint and Sample Size Requirement: Influence of Endpoints Overlap. *American Journal of Epidemiology*. 2017;185(9):832-841.

[32] Ryan LM. Efficiency of Age-Adjusted Tests in Animal Carcinogenicity Experiments. *Biometrics*. 1985;41(2):525–531.

[33] Buyse M, Ryan LM. Issues of efficiency in combining proportions of deaths from several clinical trials. *Statistics in Medicine*. 1987;6(5):565–576.

[34] Cuzick J. The Efficiency of the Proportions Test and the Logrank Test for Censored Survival Data. *Biometrics*. 1982;38(4):1033–1039.

[35] Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: A new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal*. 2012;33(2):176-182.

[36] Luo X, Tian H, Mohanty S, Tsai WY. An alternative approach to confidence interval estimation for the win ratio statistic. *Biometrics*. 2015;71(1):139-145.

[37] Rauch G, Kunzmann K, Kieser M, Wegscheider K, König J, Eulenburg C. A weighted combined effect measure for the analysis of a composite time-to-first-event endpoint with components of different clinical relevance. *Statistics in Medicine*. 2017;1-19.