Journal: American Journal of Epidemiology
Article doi: kww105
Article title: The Use of a Binary Composite Endpoint and Sample Size Requirement: Influence of Endpoints Overlap
First Author: Josep-Ramon Marsal
Corr. Author: Ignacio Ferreira-González

INSTRUCTIONS

We encourage you to use Adobe's editing tools (please see the next page for instructions). If this is not possible, please print out the proof, mark your corrections clearly in black ink, scan it, and upload the file. Please do not send corrections as track changed Word documents.

Changes should be corrections of typographical errors only. Changes that contradict journal style will not be made.

These proofs are for checking purposes only. They should not be considered as the final publication format. The proof must not be used for any other purpose. In particular, we request that you do not post them on your personal/institutional website and do not print and distribute multiple copies (please use the attached offprint order form). Neither excerpts nor all of the article should be included in other publications written or edited by yourself until the final version has been published and the full citation details are available. You will be sent these details when the article is published.

1. **Licence to Publish:** If you have not already done so, please also complete and return the Licence to Publish form online.

2. **Author groups:** Please check that all names have been spelled correctly and appear in the correct order. Please also check that all initials are present. Please check that the author surnames (family name) have been correctly identified by a pink background. If any of this information is incorrect, please identify the full surname of the relevant authors. Occasionally, the distinction between surnames and forenames can be ambiguous; this process is to ensure that the authors' full surnames and forenames are tagged correctly for accurate indexing online. Please also check all author affiliations.

3. **Figures:** If applicable, figures have been placed as close as possible to their first citation. Please check that they are complete and that the correct figure legend is present. Figures in the proof are low-resolution versions that will be replaced with high-resolution versions when the journal is printed.

4. **Missing elements:** Please check that the text is complete and that all figures, tables, and legends are included.

5. **Special characters:** Please check that special characters, equations, dosages, and units, if applicable, have been reproduced accurately.

6. **Funding:** We use the institution names to tag NIH-funded articles so they are deposited at PMC. If we already have this information, it will appear as colored text in the funding paragraph (after the affiliations). Please check that the information is correct. If funding information has not been received, please provide a funding statement that details the funding received. Remember that any funding used while completing this work should be highlighted in a separate paragraph. Please ensure that you use the full official name of the funding body, and if your paper has received funding from any institution, such as the NIH, please inform us of the grant number.

MAKING CORRECTIONS TO YOUR PROOF

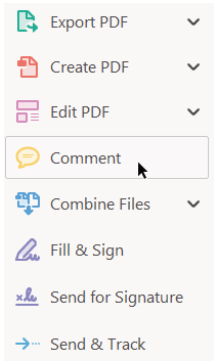
These instructions show you how to mark changes or add notes to the document using Adobe Acrobat Reader or Professional, version 10 or 11 or Adobe Reader DC. To check which version you are using, go to **Help** and then **About**. The latest version of Adobe Reader is available for free at get.adobe.com/reader. **We suggest using version 10 or higher.**

For additional help, please use the **Help** function or go to https://helpx.adobe.com/support.html**.**

Displaying the Toolbars: To edit the proofs, open the Comment pane and use the tools under Annotations.

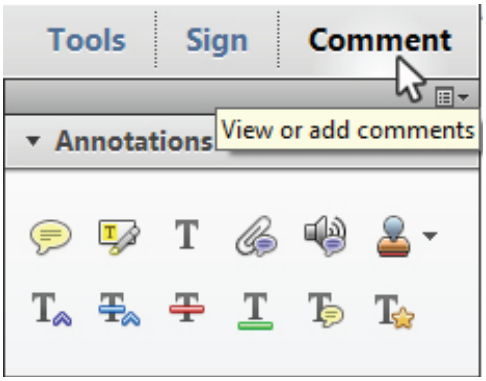
Adobe Reader 10 or 11

Adobe Reader DC



Using Text Edits

This is the quickest and easiest way to make corrections and for your corrections to be transferred and checked.

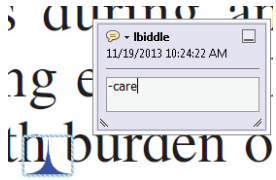


1. Open the **Comment** pane.
2. Select the text to be annotated or place your cursor at the insertion point.
3. Select the correct tool from the box. You can hover over each thumbnail for a description of the tool.

   You can also right click on selected text for a range of commenting options.

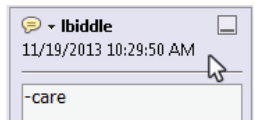
Pop-up Notes

Using the tools, it is possible to add notes. In some cases (e.g., when inserting or replacing text), a pop-up note is displayed automatically.



To **display** the pop-up note for other markup tools, right click on the annotation in the document and select **Open Pop-Up Note**.

To **move** a note, click on the title area and drag.



To **resize** the note, click on and drag the bottom right corner.

To **close** the note, click on the underscore in the top right corner.

To **delete** an edit, right click on it and select **Delete**. The edit and associated note will be removed. You can also delete by selecting the comment from the list to the right and pressing the delete key.

SAVING COMMENTS

To save your comments and notes, you must save the file (**File**, then **Save**) before you close the document.
A full list of the comments and edits you have made can be viewed by opening the Comment pane by clicking the button in the upper right corner (Adobe 10 or higher).

American Journal of Epidemiology

DOI: 10.1093/aje/kww105

Practice of Epidemiology

The Use of a Binary Composite Endpoint and Sample Size Requirement: Influence of Endpoints Overlap

Josep-Ramon Marsal, Ignacio Ferreira-González*, Sandra Bertran, Aida Ribera, Gaietà Permanyer-Miralda, David García-Dorado, and Guadalupe Gómez

* Correspondence to Dr. Ignacio Ferreira-Gonzalez, Epidemiology Unit, Department of Cardiology, Vall d'Hebron Hospital, Passeig de la Vall d'Hebron, 119-129, 08035 Barcelona, Spain (e-mail: nachoferreira@secardiologia.es).

Initially submitted November 13, 2015; accepted for publication April 4, 2016.

Although composite endpoints (CE) are common in clinical trials, the impact of the relationship between the components of a binary CE on the sample size requirement (SSR) has not been addressed. We performed a computational study considering 2 treatments and a CE with 2 components: the relevant endpoint (RE) and the additional endpoint (AE). We assessed the strength of the components' interrelation by the degree of relative overlap between them, which was stratified into 5 groups. Within each stratum, SSR was computed for multiple scenarios by varying the events proportion and the effect of the therapy. A lower SSR using CE was defined as the best scenario for using the CE. In 25 of 66 scenarios the degree of relative overlap determined the benefit of using CE instead of the RE. Adding an AE with greater effect than the RE leads to lower SSR using the CE regardless of the AE proportion and the relative overlap. The influence of overlapping decreases when the effect on RE increases. Adding an AE with lower effect than the RE constitutes the most uncertain situation. In summary, the interrelationship between CE components, assessed by the relative overlap, can help to define the SSR in specific situations and it should be considered for SSR computation.

association measures; binary endpoints; composite endpoint; correlated endpoints; outcome assessment; sample size

Abbreviations: AE, additional endpoint; AMI, acute myocardial infarction; CE, composite endpoint; RCT, randomized clinical trial; RE, relevant endpoint; RR, relative risk; SR, Sample Ratio; SSR, sample size requirement.

Use of a composite endpoint (CE) is a common strategy to reduce sample size in randomized clinical trials (RCT) (1–5). In addition to other potential benefits (6–8), combining 2 or more endpoints in a single CE might increase the number of events, thus reducing sample size requirements (SSRs). This strategy has been widely employed in cardiovascular clinical trials. For instance, the endpoint acute myocardial infarction (AMI) is usually combined with the endpoint "mortality." It is biologically plausible that reducing AMI could also lead to mortality reductions, and both are important outcomes for patients. The fact that both components are important for clinicians and patients provides a rationale for using CEs (4, 5, 9–12). Although the effect of a new treatment on the CE clearly depends on the effect on each component, the impact of the relationship between components on the statistical power and sample size is not usually considered. Thus, the increased probability of mortality in patients with AMI, which determines the correlation between both components, could be an important issue in sample size computation. In other words, the presence of a strong or a weak relationship between components of a CE may modify sample size computation in a RCT.

The benefits, risks, and influence on sample size computation of using CEs have been broadly studied (1, 2, 6, 7, 9, 13, 14). However, to our knowledge, the impact of the strength of the relationship between CE components on such computation has not been addressed in depth (3, 4, 15–18). In this sense, it is well known that the degree of relationship between CE components can affect both type I and type II errors when computing the sample size of the CE. However, a numerical approach to quantifying their impact on the power of the CE has not, to our knowledge,

been explored. Although in the setting of 2 or more coprimary binary endpoints, the correlation between endpoints and impact on power and sample size has been explored (16, 17, 19, 20), this is not the case regarding CE components. In the setting of survival analysis, Gómez and Lagakos (21) have developed a method to quantify the efficiency of adding an endpoint to an outcome expected to capture the main effect of the treatment. Their method is valid for time-to-event primary endpoints and is based on the asymptotic relative efficiency between 2 log-rank tests. The method is built from the marginal laws of the times to each of these 2 outcomes and uses Spearman's rank correlation between these 2 times. Other papers discuss the relationship between asymptotic relative efficiency and needed sample sizes (22) and provide recommendations in cardiovascular studies (23).

We propose a method to quantify the strength of the relationship between components of a binary CE. It is based on the quantification of the overlap between them (i.e., the probability of both events happening together) and their relative contribution to the probability of the CE. We explore, based on this method, the impact of the strength of the interrelation between the components on the SSR using the CE, which, ultimately, can be useful when choosing between a binary CE and one of its components.

METHODS

Definitions and assumptions

For simplicity we consider an RCT with only 2 treatment arms. We define a relevant endpoint (RE) as the outcome that is expected to drive the main effect of the treatment. For example, if the treatment is expected to reduce the rate of AMI, then AMI is the RE. The additional endpoint (AE) is the outcome that the researcher considers combining with the RE in a CE to reduce the SSR. For instance, mortality and AMI could be combined in a CE attempting to decrease sample size. The statistical test is based on the difference in events proportion between groups. Sample size estimates are computed, assuming a normal distribution, for the CE and the RE. If using the CE requires a smaller sample than the single RE then researchers could prefer to use the CE (22) instead of the RE alone.

Let (X_{ij1}, X_{ij2}) be the vector of responses where X_{ij1} and X_{ij2} denote the responses of the RE and the AE endpoints respectively for the jth subject in the ith treatment group (i = 1, 2). We assume that the responses are independently distributed as a bivariate Bernoulli distribution with $E(X_{ijk}) = \pi_{ik}$ and $V(X_{ijk}) = \pi_{ik}(1 - \pi_{ik})$, where k only takes the values 1 or 2. Define the response of the CE by X^*_{ij}; note that it is equal to 1 whenever X_{ij1} *or* X_{ij2} are equal 1, and it is 0 otherwise. X^*_{ij} is a Bernouilli random variable with expectation:

$$\begin{aligned} E(X^*_{ij}) &= \text{Prob}(X^*_{ij} = 1) = \text{Prob}((X_{ij1} = 1) \cup (X_{ij2} = 1)) \\ &= \text{Prob}(X_{ij1} = 1) + \text{Prob}(X_{ij2} = 1) - \text{Prob}((X_{ij1} = 1) \\ &\quad \cap (X_{ij2} = 1)) = \pi_{i1} + \pi_{i2} - \pi_{i12}, \end{aligned} \tag{1}$$

where π_{i12} corresponds to the probability of both outcomes (RE and AE) happening together. Note that with these 6 parameters $(\pi_{i=1,1}, \pi_{i=1,2}, \pi_{i=1,12}, \pi_{i=2,1}, \pi_{i=2,2}, \pi_{i=2,12})$, the sample size for RE and CE can be computed. Usually π_{i1}, π_{i2} are known, but this is not the case for π_{i12}.

The overall possible scenarios concerning the probability of the intersection of both outcomes in an RCT can be represented as in Figure 1.

Figure 1A depicts a scenario without overlapping between endpoints (i.e., no patient who had AMI died). Figure 1B, Figure 1C, and Figure 1D represent weak, moderate, and strong overlap, respectively. Figure 1E represents the maximum possible overlap (i.e., all patients who died had had AMI). Each scenario has different impact on the computation of the sample size.

We define relative overlap as the conditional probability of experiencing the 2 outcomes given that the patient has experienced one of them, which is evaluated as the ratio between the probability of the intersection and the probability of the CE as computed in equation (1). This statistic is well known in ecology and genetics as Jaccard's index of similarity (24–26). It is used to estimate the similarity of sample sets. Using our notation, the relative overlap in group i (for an arbitrary subject j from treatment group i):

$$RO_i = \text{Prob}(X_{i1} = X_{i2} = 1 | X_{i1} + X_{i2} \geq 1) = \frac{\pi_{i12}}{\pi_{i1} + \pi_{i2} - \pi_{i12}} \tag{2}$$

Note that equation (2) corresponds to $P(A \cap B) \div P(A \cup B)$ when A and B stand, respectively, for the events $[X_{ij1} = 1]$

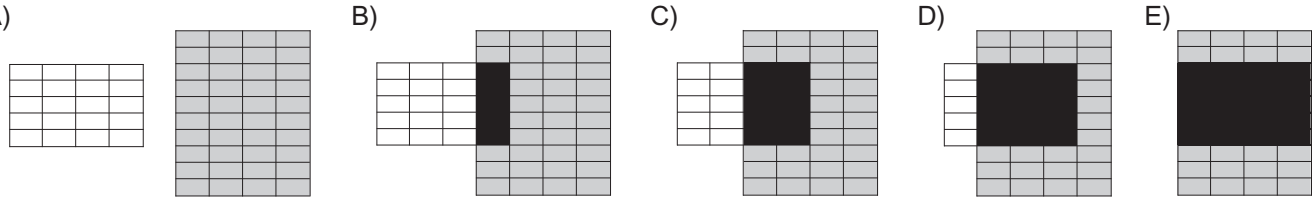


Figure 1. Example of possible scenarios concerning the probability of the intersection of both outcomes in a randomized clinical trial. Each square represents 1% probability. The white square represent the probability of death ($\pi_{i1} = 0.20$), and gray square represents the probability of acute myocardial infarction ($\pi_{i1} = 0.40$) for the ith treatment. The π_{i12} in each scenario are: 0.00 (A), 0.05 (B), 0.10 (C), 0.15 (D), and 0.20 (E), and the probability of the composite endpoint in each scenario was assessed as in equation (1), at 0.60, 0.55, 0.50, 0.45, and 0.40, respectively.

and $[X_{ij2} = 1]$. Following basic relationships between the probability of the union of two events and the probability of their intersection it follows that the relative overlap in group i, RO_i, is bounded by $\max\{0,\pi_{i1} + \pi_{i2} - 1\} \leq RO_i \leq \min\left\{\frac{\pi_{i1}}{\pi_{i2}}, \frac{\pi_{i2}}{\pi_{i1}}\right\}$.

The π_{i12} value ranges from 0, in the case of 2 disjoint sets (Figure 1A), to the value of the less prevalent endpoint probability, in the case of completely overlapping sets (Figure 1E); then:

$$\max\{0; \pi_{1j} + \pi_{2j} - 1\} \leq \pi_{i12} \leq \min\{\pi_{i1}, \pi_{i2}\}. \quad (3)$$

From equations (1) and (3), the range of possible values of $E(X^*_{ij})$ is given by

$$[\{\max\{\pi_{i1}, \pi_{i2}\} \leq \text{Prob}(X^*_{ij} = 1) \leq \min\{\pi_{i1} + \pi_{i2}, 1\}\}]. \quad (4)$$

And from equations (1) to (4), the range of values of RO_i is bounded by:

$$\left[\max\{0, \pi_{i1} + \pi_{i2} - 1\}, \min\left\{\frac{\pi_{i1}}{\pi_{i2}}, \frac{\pi_{i2}}{\pi_{i1}}\right\}\right]. \quad (5)$$

We additionally assume that the magnitude of the relationship between endpoints is the same in treatment (i = 1) and control groups (i = 2); the conditional probability of having the AE as well as having had the RE (and vice versa) is not affected by the treatment. Note that this does not imply $\pi_{1,12} = \pi_{2,12}$. For example, consider that the proportion of the RE and the AE in the control group (i = 2) are 0.20 and 0.40 respectively. From equation (3) it easily follows that the range of values of $\pi_{2,12}$ can fall from 0 to 0.2. Assuming that the effect of the treatment is quantified by the relative risk (RR) and is equal to 0.9 for the RE and 0.5 for the AE, the proportion for the treatment group (i = 1) will be 0.18 and 0.20 for the RE and the AE, respectively. Again, from equation (3), $\pi_{1,12}$ can take values from 0 to 0.18. Considering, for instance, the maximum possible overlap (Figure 1E), then $\pi_{2,12} = 0.20 \neq \pi_{1,12} = 0.18$.

Sample size and statistical power calculation

The analysis aimed at quantifying to what extent using a binary CE instead of a single RE can decrease the SSR for a given significance level and for a given power to detect a given proportion difference, considering different overlapping scenarios.

SSRs using the RE or using the CE are calculated to detect differences in the proportion of events between both groups, considering a type I error of 0.05 (note that other hypothesis tests such as superiority could be used). We address 2 different, not equivalent, hypothesis tests, depending on whether we base our primary endpoint on RE or CE. Specifically:

$$\begin{cases} H_0 : \pi_{2,RE} = \pi_{1,RE} \\ H_1 : \pi_{2,RE} \neq \pi_{1,RE} \end{cases} \qquad \begin{cases} H^*_0 : \pi_{2,CE} = \pi_{1,CE} \\ H^*_1 : \pi_{2,CE} \neq \pi_{1,CE}. \end{cases}$$

Sample size is calculated using the normal approximation to the binomial test. It can be computed for both the RE and CE as it is presented in equation (6) (27, 28), where π represents the probability of the RE or the probability of the CE as calculated in equation (1), $\pi_{i=1}$ the probability in one group, and $\pi_{i=2}$ the probability in the other group (estimated assuming certain effect). The α and β are the type I and II errors respectively:

$$n \geq \left(\frac{Z_{\alpha/2}\sqrt{2\pi_1(1-\pi_1)} + Z_\beta\sqrt{\pi_{i=1}(1-\pi_{i=1}) + \pi_{i=2}(1-\pi_{i=2})}}{\pi_{i=1} - \pi_{i=2}}\right)^2,$$
$$\pi_1 = \frac{\pi_{i=1} + \pi_{i=2}}{2}. \quad (6)$$

A CE will be preferred if the number of subjects required (n^*) to detect an effect using the CE is lower than using the single RE (22) (n). We define the sample ratio (SR) as the ratio between sample sizes: ($SR = n/n^*$). Thus the CE will be preferred in all situations where SR > 1.

Finally, given a fixed number of patients, the statistical power (pw) using the CE and the RE can be calculated as follows:

$$\text{pw} = 1 - \beta = P\left(X \leq \frac{\sqrt{n}(\pi_{i=1} - \pi_{i=2}) - Z_{\alpha/2}\sqrt{2\pi_1(1-\pi_1)}}{\sqrt{\pi_{i=1}(1-\pi_{i=1}) + \pi_{i=2}(1-\pi_{i=2})}}\right), \quad (7)$$

where X is a standardized normal distribution.

Numerical examples

We simulated several scenarios (Table 1) assuming different values of the event proportion for both the RE and the

Table 1. Prevalence and Range of Effects Simulated for the Relevant and Additional Endpoints in Different Scenarios

P(RE)[a]	RR(RE)	P(AE)	RR(AE)
0.10	0.70 to 0.90	0.10	0.70 to 1.00
0.10	0.70 to 0.90	0.20	0.70 to 1.00
0.20	0.70 to 0.90	0.10	0.70 to 1.00
0.20	0.70 to 0.90	0.20	0.70 to 1.00
0.03	0.70 to 0.90	0.10	0.70 to 1.00
0.03	0.70 to 0.90	0.01	0.70 to 1.00

Abbreviations: P(AE), proportion of additional endpoint; P(RE), proportion of relevant endpoint; RR(AE), relative risk of additional endpoint; RR(RE), relative risk of relevant endpoint.

[a]The results of the scenarios with P(RE) = 0.10 are presented in Web Table 1, for P(RE) = 0.20 in Web Table 2, and for P(RE) = 0.03 in Web Table 3.

AE, the effect of therapy on each endpoint, and the degree of overlapping:

- Scenario 1: RE with a fixed proportion $\pi_{i=1,1} = 0.10$ and increasing treatment effect (RR_1 from 0.9 to 0.7). An AE with 2 proportions $\pi_{i=1,2} = 0.10$ and $\pi_{i=1,2} = 0.20$ was added, without effect of the therapy on the AE ($RR_2 = 1.0$).
- Scenario 2: RE with a fixed proportion $\pi_{i=1,1} = 0.10$ and fixed and low treatment effect ($RR_1 = 0.9$). An AE with 2 proportions $\pi_{i=1,2}$=0.10 and $\pi_{i=1,2}$=0.20 was added. Progressively higher therapy effect on AE (RR_2= 0.9 to 0.7).
- Scenario 3: RE with a fixed proportion $\pi_{i=1,1} = 0.10$ and fixed and moderate treatment effect ($RR_1 = 0.8$). An AE with 2 proportions $\pi_{i=1,2} = 0.10$ and $\pi_{i=1,2} = 0.20$ was added. Progressively higher treatment effect on AE ($RR_2 = 0.9$ to 0.7).
- Scenario 4: RE with a fixed proportion $\pi_{i=1,1} = 0.10$and fixed and high treatment effect ($RR_1 = 0.7$). An AE with 2 proportions $\pi_{i=1,2} = 0.1$ and $\pi_{i=1,2} = 0.2$ was added. Progressively higher treatment effect ($RR_2 = 0.9$ to 0.7) on AE.

Then, for each scenario, 5 situations of overlapping were considered, from absence (Figure 1A) to maximum overlapping (Figure 1E). The whole range of possible values of relative overlap in each scenario was divided into 5 strata. The Web Material (available at http://aje.oxfordjournals.org/) includes all the scenarios showed in Table 1.

For each scenario we calculated sample sizes of the RE and the CE, the SR, and the RR on the CE. The SRs generated in each scenario were plotted to highlight the situations in which the use of CE is more "efficient" than the use of a single RE in terms of sample size. Additionally, we compared the power of using a single RE with that using the CE across different sample sizes and considered different situations regarding the effect of the treatment and the degree of overlapping.

Calculations for the 4 scenarios were replicated, increasing the proportion of the RE to 0.20 (see Web Table 1), decreasing the proportion of the RE to 0.03 (see Web Table 2), and for 2 different scenarios of the proportion of the AE (0.01 and 0.1). The scenarios have been chosen considering a general framework of credible clinical situations in research. Other scenarios could be easily reproduced using the formulas (1–7).

RESULTS

Web Table 3 shows the sample size calculated for the RE and the CE, the SR, and the RR of the CE across all scenarios considered. In scenario 1 (S1.1 – S1.4), adding an AE without effect (RR = 1) results in a higher SSR, with all SRs being lower than 1; thus the CE requires a higher sample size than the RE. In addition, the higher the overlap, the lower the effect on the CE and the greater the additional sample size required. In general, an increase in the proportion of the AE leads to either an increase in SSRs when the overlap is very low or low or a reduction when overlap is moderate or high. Then adding components without effect always results in an increase of the sample size, even in the most favorable situation when the components are disjoint.

In scenario 2 (S2.1 – S2.6), adding an AE with effect results in a lower SSR except for maximum overlap, when for sample size computation a single RE has equivalent effects. Again, the higher the overlap, the lower the SR. However, the impact of overlapping is greater when the proportion of the AE is low (i.e., 0.1). The most favorable situation corresponds to the absence of overlapping, which results in an exponential increase of SRs as long as the RR on the AE increases. In all cases, increasing the AE proportion implies a lower SSR with the CE. The impact of the increase in AE proportion is greater as the effect on the AE increases and the overlap grows.

In scenario 3 (S3.1–S3.6) the effect of the RE was incremented with respect to scenario 2 (RR = 0.9 for scenario 3 against RR = 0.8 in scenario 2) while maintaining all the other parameters. The comparison of these 2 scenarios suggests that a greater effect on the RE implies lower SSRs, lower SRs, and less marked changes in SRs when increasing the overlap. However, it is to be noted that, when the effect on the AE is lower than the effect on the RE (in both cases the probabilities of events are restricted to $\pi_{i=c,2} = 0.10$ and $\pi_{i=c,2}$=0.20), overlap can determine a lower (SRs > 1; overlap very low or low) or higher (SRs < 1; overlap moderate to very high) SSR when using the CE. This leads to a higher number of instances—9 instances in scenario 3 and 3 in scenario 2—in which the use of the CE implies the same or greater SSRs than the use of the isolated RE. Thus, the relative overlapping net impact on efficiency may be greater when RR = 0.9 (scenario 3) than when RR = 0.8 (scenario 2). Also, the impact of the increase of the proportion in AE is greater as the effect on the AE and the overlap increases. In scenario 4, when RR = 0.7 (S4.1–S4.6), the effect on the RE is greater than in scenario 3 and scenario 2 (RR = 0.7 vs. 0.8 vs. 0.9), and the other conditions are equal.

Similarly, compared with scenario 2 and scenario 3, scenario 4 leads to a lower SSRs, lower SRs, and less marked changes in SRs when increasing overlap. The number of instances in which the use of the CE implies the same or greater SSRs than the use of the isolated RE is greater than in scenario 3 and scenario 2 (15 vs. 9 vs. 3).Therefore, the net impact of overlapping on efficiency may be greater as the effect on the RE increases (scenarios 1 through 4).

In general, in all situations a greater overlap implies higher SSR of the CE (all other conditions being equal). However, the qualitative impact of overlapping is higher when the effect on the AE is lower than the effect on the RE. This can be observed in scenario 3 and scenario 4. In these situations, adding an AE with lower effect than the RE may result in a lower or higher SSR depending on the relative overlap. Globally in 11 out of 22 scenarios explored, the relative overlap magnitude could be determinant for CE use.

In general, the relative benefit expected of adding an AE to an RE, assessed by the SRs, decreases when the proportion of the RE increases (Web Table 2). The exception is when both the proportion and the effect on the AE are greater and the effect on the RE is greater. In cases of low proportion of the RE (Web Table 3), the main impact of the

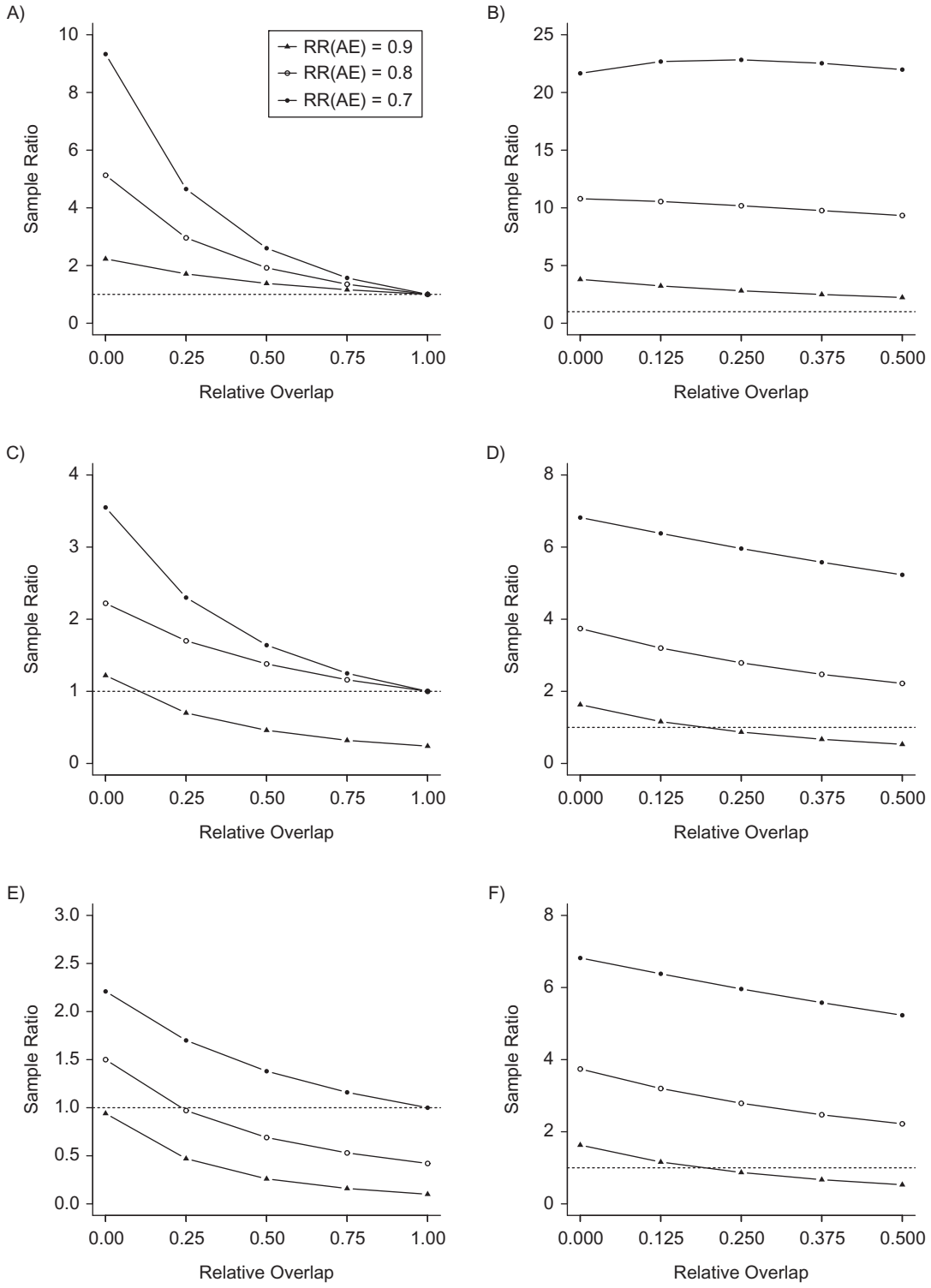


Figure 2. Sample ratio trends in scenarios 2–4. A) P(AE) = 0.10, P(RE) = 0.10, and RR(RE) = 0.90. B) P(AE) = 0.20, P(RE) = 0.10, and RR(RE) = 0.90. C) P(AE) = 0.10, P(RE) = 0.10, and RR(RE) = 0.80. D) P(AE) = 0.20, P(RE) = 0.10, and RR(RE) = 0.80. E) P(AE) = 0.10, P(RE) = 0.10, and RR(RE) = 0.70. F) P(AE) = 0.20, P(RE) = 0.10, and RR(RE) = 0.70. Note that the possible range of relative overlap values (*x*-axis) depends on the proportions of both the AE and the RE (i.e., lower bound: $\{0; \pi_{i1} + \pi_{i2} - 1\}$; upper bound: $\min\{\pi_{i1}/\pi_{i2}, \pi_{i2}/\pi_{i1}\}$). Dashed black lines denote equal sample size requirement (SSR) using the RE and the composite endpoint (CE); above the dashed black line implies lower SSR using the CE, and below the dashed black line implies higher SSR using the CE. P(AE), proportion of additional endpoint; P(RE), proportion of relevant endpoint; RR(AE), relative risk of additional endpoint; RR(RE), relative risk of relevant endpoint.

CE SSR rests on both the proportion and the effect on the AE. In 25 of the total of 66 scenarios explored, the degree of relative overlap could determine the benefit for SSR of using a CE.

Figure 2 summarizes the behavior of the SR across the spectrum of overlapping in scenarios 2–4 (scenario 1 was included because all SRs were lower than 1 in all situations). In general, the greater the effect on the RE, the lower the benefit of adding an AE, and the smaller the influence of overlapping. By contrast, the impact of overlapping on SSR is greater with lower proportions of the AE. In general, the more the added AE is prevalent, the greater the benefit of the CE on SSR (only in cases with a SR > 1). It must be noted that, as shown above (in Definitions and Assumptions), the possible range of relative overlap values depends on the proportion of both the AE and the RE.

Figures 3 and 4 show the power function, for both the RE (solid line) and the CE (different dotted lines), given different sample size intervals and considering different scenarios concerning the effect on endpoints and the degree of overlapping. As expected, the greater the effect on the AE, the greater the statistical power of the CE, especially when the overlap is small. In some scenarios the statistical power is clearly greater using the CE regardless of the overlap (Figure 4B, 4D, and 4F), whereas in other scenarios it is more efficient to use a single RE (Figures 3E and 4E). However there are situations in which the overlap is determinant to the selection of the single RE or the CE (Figures 3C and 4C).

DISCUSSION

We have addressed the issue of the statistical "efficiency" in sample size and statistical power terms using a binary CE in a conceptual and visual manner. We have employed the SR between the RE and the binary CE that results from adding an AE to evaluate to what extent the proportion of events, the effect on endpoints, and especially the overlap between them determines the SSR using a CE. Examination of different scenarios can be summarized as follows (Table 2):

A. Should the effect on the AE be superior to the effect on the RE, the use of CE will always be convenient to decrease sample size regardless of the event's proportion and the relative overlap. Even when adding a low-prevalence AE with similar effect to that on the RE, the CE will be a better solution than the RE. Influence of overlapping will decrease as long as the effect on RE increases.
B. The greater the degree of overlapping, the lower the potential benefit of the CE to reduce sample size. However, the benefit of combining endpoints with small overlap is variable, depending on other parameters, such as the treatment effect on each endpoint and the event proportion. Occasionally the degree of overlapping will be sufficient to recommend the use of CE.
C. Adding an AE with lower effect than that on the RE constitutes the most uncertain situation. It could result in either a greater or lower SSR when using the CE, depending on the degree of overlap (i.e., the more, the lower), the effect on the AE (i.e., the greater, the higher), and the proportion of the AE. In general, increased AE proportion will decrease the SSR using the CE.

The potential benefits of using CE in clinical trials have been widely assessed (1–8). It is well recognized that, among other benefits, CEs can be useful to decrease sample size. Moreover, many authors agree that this could be the main reason for using CEs. However, it is less recognized that, besides other problems, the use of CEs may lead to a statistical power reduction (21, 22). This is obvious when the treatment effect on the AE is in the opposite direction from the effect on the RE. Now we have described other scenarios in which endpoint combination could also lead to increased SSRs even when components have the same direction. Moreover, we have shown situations where, even with substantial treatment effect on the AE, the use of CE could imply a higher SSR to detect the true treatment differences. This is closely related to the degree of overlapping between components.

We have introduced "overlap between endpoints" as a conceptual approach to visualizing how relationships between correlated components of a binary CE may influence statistical power for the CE. However, is to be noted that overlap between components is neither equivalent to correlation nor the intersection probability. Correlation between endpoints is more related to the conditional probability of having the AE after having had the RE and vice versa, which is in relationship with the overlap but not equivalent. In this case, we used the relative overlap as an approach to quantify to what extent the relationship between components of the CE can affect the power. This ratio is also known as Jaccard's statistic in ecology and genetics (24, 25), where it is used to assess the similarity between different samples or communities. In the context of CE, we believe that the concept of relative overlap can be more intuitive for clinicians than the concept of correlation and the concept of the intersection probability.

Elements that have an impact on the statistical efficiency of the CE have been well described within time-to-event analysis. Along with the probabilities of observing the event components and the magnitude of treatment effect on each component via the hazard ratio, Spearman's rank correlation between the times-to-each-component is needed (23). Describing the relationship between the overlap of components of a binary CE and the correlation between them in the context of the time-to-event analysis is beyond the scope of this manuscript and should be addressed in future work.

Schriger et al. (18) have identified that the relationship between endpoints is not usually reported in publications of RCTs. Moreover, sample computations of RCTs do not usually include this parameter. Given its implications for CE construction, an anticipated estimation of the plausible relationship between components would be useful to increase the precision of sample size estimates. As a future line of outcomes research, it would be desirable to quantify the relationship between outcomes in specific populations from real RCT results. For example, knowing the proportion of stroke in patients who had an AMI could be of

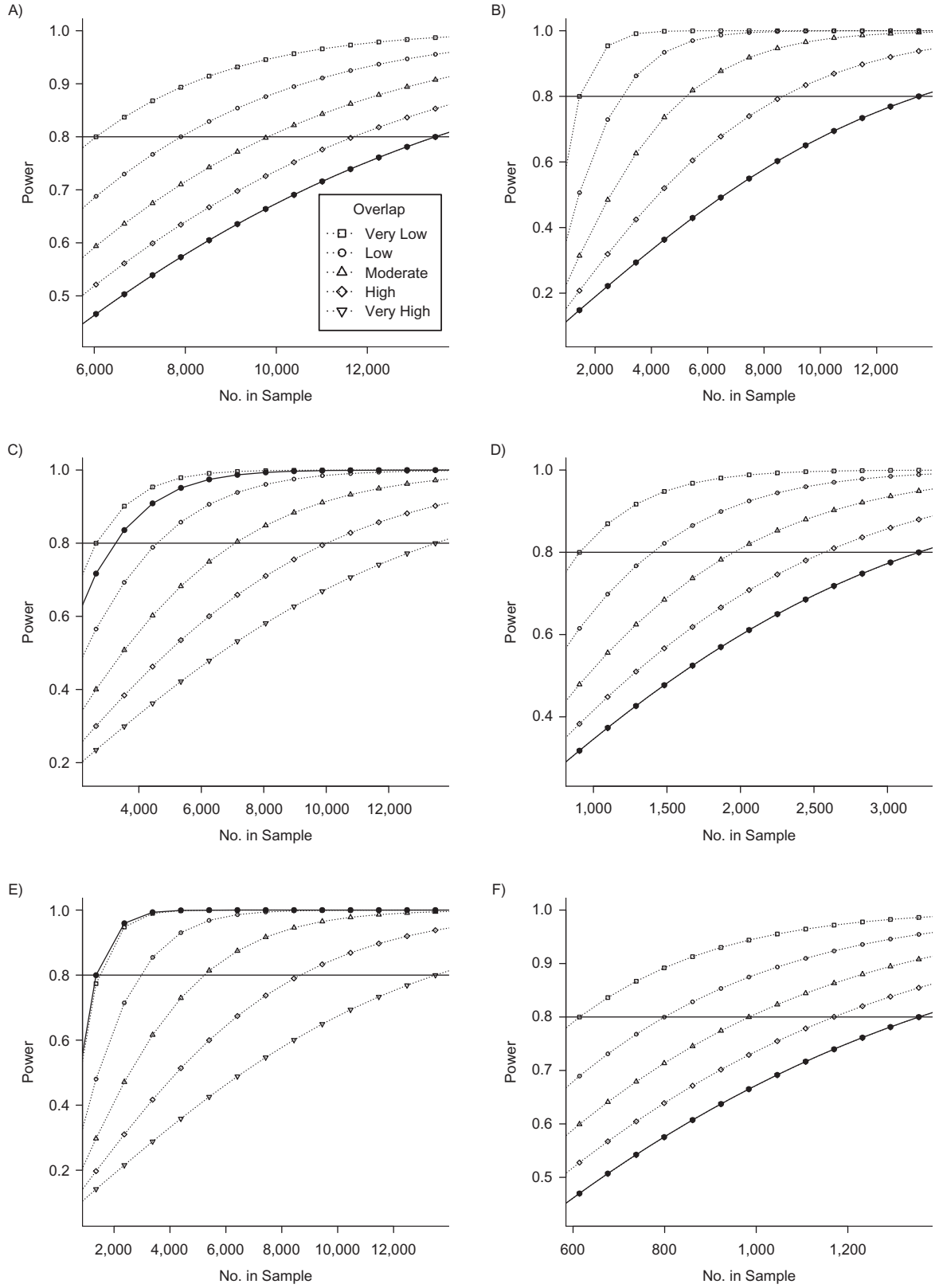


Figure 3. Power curves considering different sample sizes in several scenarios for a P(RE) = 0.10 and a P(AE) = 0.10. A) RR(RE) = 0.90 and RR(AE) = 0.90. B) RR(RE) = 0.90 and RR(AE) = 0.70. C) RR(RE) = 0.80 and RR(AE) = 0.90. D) RR(RE) = 0.80 and RR(AE) = 0.70. E) RR (RE) = 0.70 and RR(AE) = 0.90. F) RR(RE) = 0.70 and RR(AE) = 0.70. Solid black line with closed circles denotes the statistical power curves for the single RE. Dashed lines denote the statistical power curves for the composite endpoint for different relative overlap. Black horizontal lines denote the threshold for a statistical power of 0.80. P(AE), proportion of additional endpoint; P(RE), proportion of relevant endpoint; RR(AE), relative risk of additional endpoint; RR(RE), relative risk of relevant endpoint.

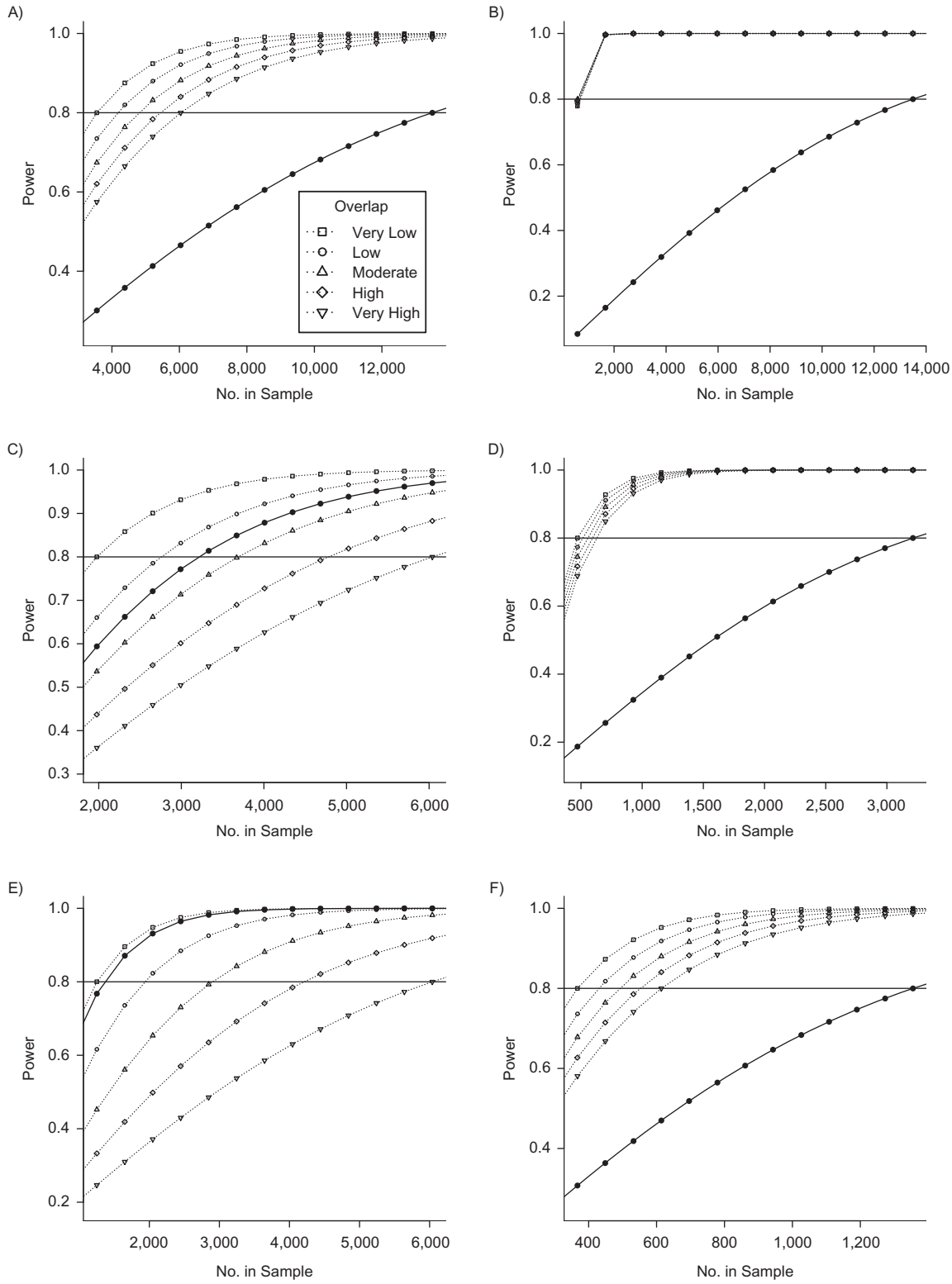


Figure 4. Power curves considering different sample sizes in several scenarios for a P(RE) = 0.10 and a P(AE) = 0.20. A) RR(RE) = 0.90 and RR(AE) = 0.90. B) RR(RE) = 0.90 and RR(AE) = 0.70. C) RR(RE) = 0.80 and RR(AE) = 0.90. D) RR(RE) = 0.80 and RR(AE) = 0.70. E) RR (RE) = 0.70 and RR(AE) = 0.90. F) RR(RE) = 0.70 and RR(AE) = 0.70. Solid black line with closed circles denotes the statistical power curves for the single RE. Dashed lines denote the statistical power curves for the composite endpoint for different relative overlap. Black horizontal lines denote the threshold for a statistical power of 0.80. P(AE), proportion of additional endpoint; P(RE), proportion of relevant endpoint; RR(AE), relative risk of additional endpoint; RR(RE), relative risk of relevant endpoint.

Table 2. Summary of the Qualitative Impact of Relative Overlapping Between Components on the Sample Size Requirement of the Composite Endpoint Considering Different Scenarios for the Relative Risks and Probabilities of the Components

Scenarios		Qualitative Impact of Overlapping[a]	Sample Size Requirement Using the CE[b]		
P(AE)	RR(AE)		Before Considering Overlap	After Considering Overlap	
				Low Overlap	High Overlap
≤P(RE)	1	++	i	i	ii
≥P(RE)	1	++	i	i	ii
≤P(RE)	<RR(RE)	++	d	dd	d
≤P(RE)	>RR(RE)	+++	i/d	d	i
≥P(RE)	<RR(RE)	+	dd	ddd	dd
≥P(RE)	>RR(RE)	+++	i/d	d	I

Abbreviations: CE, composite endpoint; P(AE), proportion of additional endpoint; P(RE), proportion of relevant endpoint; RR(AE), relative risk of additional endpoint; RR(RE), relative risk of relevant endpoint.

[a]The impact of components overlap in scenarios labeled + is lower than in scenarios labeled ++. Impact is highest in scenarios labeled +++.

[b]Sample size requirement using the CE in scenarios labeled i is always higher than using only the RE regardless the degree of the overlap. In scenarios labeled ii the increase is higher than scenarios labeled i. Sample size requirement using the CE in scenarios labeled d is always lower than using only the RE regardless the degree of the overlap. In scenarios labeled ddd the decrease is lower than the scenarios labeled dd and lower than scenarios labeled d. Sample size requirement using the CE in scenarios labeled i/d depends on the degree of the overlap.

Note: In all scenarios, probability for the CE was assumed to be lower than 0.5.

interest for sample size computations using a CE including these 2 outcomes.

We have limited our exploration to binary CEs with only 2 components. Quantifying the associations between 3 or more components of a CE and the resulting impact on the SSRs should be addressed in future studies. However, although computationally much more challenging, the essential concepts described here would be valid. In any case, their extension to more complex CEs will be necessary to simulate real-life examples.

We have focused exclusively on the study of the SSR using CEs. CEs may have other additional benefits that could be apparent even if the CE is not efficient. For example, the study of the net benefit of a therapy by a CE including efficacy and safety endpoints could be completely "inefficient," because the CE may include components with treatment effects in opposite directions. Thus a treatment could reduce the rate of AMI, increasing at the same time the rate of severe hemorrhage events. In this case, computation of sample size based on the net benefit of the therapy (i.e., the rate of the CE AMI or severe hemorrhage) could be inefficient in both sample size and statistical power terms. However, sometimes this approach will be pragmatically essential to depict the net usefulness of a treatment. Even in these extreme cases, the study of the overlap between endpoints will be useful to estimate the SSR.

ACKNOWLEDGMENTS

Author Affiliations: Epidemiology Unit, Department of Cardiology, Vall d'Hebron Hospital, Barcelona, Spain (Josep-Ramon Marsal, Ignacio Ferreira-González, Sandra Bertran, Aida Ribera, Gaietà Permanyer-Miralda, David García-Dorado); Consorcio de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain (Josep-Ramon Marsal, Ignacio Ferreira-González, Sandra Bertran, Aida Ribera, Gaietà Permanyer-Miralda, David García-Dorado); Department of Pediatrics, Obstetrics and Gynecology and Public Health of Universitat Autònoma de Barcelona, Cerdanyola del Vallés, Spain (Josep-Ramon Marsal); Unit of Research Support, Primary Care Research Institute (IDIAP) Jordi Gol, Catalan Institute of Health, Lleida, Spain (Josep-Ramon Marsal); and Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain (Sandra Bertran, Guadalupe Gómez).

G.G. is partially supported by the Ministerio de Economía y Competitividad del Gobierno de España (grants MTM2012-38067-C02-01 and MTM2015-64465_C02_1R) and by the Departament d'Economia i Coneixement de la Generalitat de Catalunya (grant 2014 SGR 464).

Conflict of interest: none declared.

REFERENCES

1. Ferreira-González I, Alonso-Coello P, Solà I, et al. Composite endpoints in clinical trials. *Rev Esp Cardiol*. 2008;61(3):283–290.
2. Chi GY. Some issues with composite endpoints in clinical trials. *Fundam Clin Pharmacol*. 2005;19(6):609–619.
3. Sankoh AJ, D'Agostino RB Sr, Huque MF. Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. *Stat Med*. 2003;22(20):3133–3150.

4. Rauch G, Rauch B, Schüler S, et al. Opportunities and challenges of clinical trials in cardiology using composite primary endpoints. *World J Cardiol*. 2015;7(1):1–5.
5. Cannon CP. Clinical perspectives on the use of composite endpoints. *Control Clin Trials*. 1997;18(6):517–529.
6. Cordoba G, Schwartz L, Woloshin S, et al. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. *BMJ*. 2010;341:c3920.
7. 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. *J Clin Epidemiol*. 2007;60(7):651–657.
8. Holmes DR Jr, Brennan JM, Rumsfeld JS, et al. Clinical outcomes at 1 year following transcatheter aortic valve replacement. *JAMA*. 2015;313(10):1019–1028.
9. Prieto-Merino D, Smeeth L, Staa TP, et al. Dangers of non-specific composite outcome measures in clinical trials. *BMJ*. 2013;347:f6782.
10. European Medicines Agency. *ICH Topic E 9: Statistical Principles for Clinical Trials*. London, UK: European Medicines Agency; 2006: 1–37.
11. Committee For Proprietary Medicinal Products, European Medicines Agency. *Points to Consider on Multiplicity Issues in Clinical Trials*; London, UK: EMEA (European Medicines Agency); 2006: 1–11.
12. Kip KE, Hollabaugh K, Marroquin OC, et al. The problem with composite end points in cardiovascular studies: the story of major adverse cardiac events and percutaneous coronary intervention. *J Am Coll Cardiol*. 2008;51(7):701–707.
13. Ferreira-González I, Busse JW, Heels-Ansdell D, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ*. 2007;334(7597):786.
14. Cannon CP, Sharis PJ, Schweiger MJ, et al. Prospective validation of a composite end point in thrombolytic trials of acute myocardial infarction (TIMI 4 and 5). Thrombosis In Myocardial Infarction. *Am J Cardiol*. 1997;80(6):696–699.
15. Lefkopoulou M, Ryan L. Global tests for multiple binary outcomes. *Biometrics*. 1993;49(4):975–988.
16. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics*. 1984;40(4):1079–1087.
17. Rauch G, Kieser M. Multiplicity adjustment for composite binary endpoints. *Methods Inf Med*. 2012; 51(4):309–317.
18. Schriger DL, Cooper RJ, Lopez-O'Sullivan A, et al. Is the relationship among outcome variables shown in randomized trials? *Trials*. 2015;16(1):57.
19. Sozu T, Sugimoto T, Hamasaki T. Sample size determination in clinical trials with multiple co-primary binary endpoints. *Stat Med*. 2010;29(21):2169–2179.
20. Xiong C, Yu K, Gao F, et al. Power and sample size for clinical trials when efficacy is required in multiple endpoints: application to an Alzheimer's treatment trial. *Clin Trials*. 2005;2(5):387–393.
21. Gómez G, Lagakos SW. Statistical considerations when using a composite endpoint for comparing treatment groups. *Stat Med*. 2013;32(5):719–738.
22. Gómez G, Gómez-Mateu M. The asymptotic relative efficiency and the ratio of sample sizes when testing two different null hypotheses. *SORT*. 2014;38(1):73–88.
23. Gómez G, Gómez-Mateu M, Dafni U. Informed choice of composite end points in cardiovascular trials. *Circ Cardiovasc Qual Outcomes*. 2014;7(1):170–178.
24. Miller JA, Miller JH, Pham DS, et al. Cyberdiversity: improving the informatic value of diverse tropical arthropod inventories. *PLoS One*. 2014;9(12):e115750.
25. Sousa FM, Mendes RM, Mendes RF, et al. Evaluation of genetic similarity between accessions of *Pityrocarpa moniliformis* (angico-de-bezerro) using RAPD markers. *Genet Mol Res*. 2014;13(3):5815–5821.
26. McCormick WP, Lyons NI, Hutcheson K. Distributional properties of Jaccard's index of similarity. *Commun Stat Theory Methods*. 1992;21(1):51–68.
27. Fleiss JL. *Statistical Methods for Rates and Proportions*. Hoboken, NJ: John Wiley & Sons; 1981:33–49.
28. Casagrande JT, Pike MC. An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics*. 1978;34(3):483–486.