Impact of correlation structure on sample size requirements of statistical methods for multiple binary outcomes: A simulation study

Running head:
Sample Size and Outcome Correlations

Kanako Fuyama[1], Kentaro Sakamaki[2], Kohei Uemura[3] and Isao Yokota[1]

[1]Department of Biostatistics, Graduate School of Medicine, Hokkaido University, Sapporo, Japan
[2]Faculty of Health Data Science, Juntendo University, Tokyo, Japan
[3]Interfaculty Initiative in Information Studies, The University of Tokyo, Tokyo, Japan

Corresponding author:
Kanako Fuyama, Department of Biostatistics, Graduate School of Medicine, Hokkaido University, N15 W7, Kita-ku, Sapporo, Japan.
Email: kfuyama@pop.med.hokudai.ac.jp

Abstract

Background: In randomized clinical trials, multiple-testing procedures, composite endpoints, and prioritized outcome approaches are increasingly used to analyze multiple binary outcomes. Previous studies have shown that correlations between outcomes influence their sample size requirements. Although sample size is an important factor affecting the choice of statistical methods, the power and required sample sizes of methods for analyzing multiple binary outcomes have yet to be compared under the influence of outcome correlations.

Methods: We conducted simulations to evaluate the power of co-primary and multiple primary endpoints, composite endpoints, and prioritized outcome approaches based on generalized pairwise comparisons with varying correlations, marginal proportions, treatment effects, and number of outcomes. We then conducted a case study on sample size using a clinical trial of a migraine treatment as an example.

Results: The correlations significantly affected the statistical power and sample size of composite endpoints. The power and sample size of co-primary endpoints remained relatively stable across different correlations, though their power declined substantially when treatment effects were opposite on some components or more than two components were present. While the correlations influenced the power and sample size of all methods assessed, their direction and degree of influence varied between methods. Notably, the method with the greatest power and smallest sample size also differed depending on the correlations. When the correlations were the same between arms, prioritized outcome approaches usually had higher power and smaller sample sizes than other methods.

Conclusions: Anticipated correlations and their uncertainty should be considered when selecting statistical methods. Overall, co-primary endpoints remain a reliable option for evaluating the superiority of all components, although they are unsuitable for assessing the balance between treatment effects pointing in different directions. Generalized pairwise comparisons offer a useful alternative to deal with multiple prioritized outcomes, often providing the smallest sample sizes when the correlation structures are shared between the arms.

Keywords

binary endpoints, correlated endpoints, composite endpoints, multiple-testing procedures, prioritized outcomes, net benefit

Abstract: 306 words

Main text: 4073 words

Introduction

Migraine is a common neurological disease characterized by headache, nausea, and sensitivity to external stimuli.[1] Because pain control is essential for migraine treatment,[2,3] the proportion of patients free of pain at prespecified time points is often used as the single primary endpoint in randomized clinical trials.[4] However, studies have suggested that relieving other migraine-associated symptoms without introducing side effects further enhances patient satisfaction.[3,5]

Several frameworks are currently used to incorporate multiple binary components into the statistical evaluation. Multiple-testing procedures are increasingly used to prove effects on pain and associated symptoms,[1,6] and various composite endpoints have been proposed (e.g., freedom from pain and associated symptoms, sustained pain-free with no adverse events).[7,8] Nonetheless, recent investigations have suggested that these multiple outcomes have different degrees of clinical importance.[3] A relatively new method for handling multiple prioritized outcomes is called generalized pairwise comparisons.[9]

When selecting statistical methods, power and required sample sizes are crucial.[10] Previous studies on methods for multiple binary outcomes have found their sample sizes to be affected by correlations between the outcomes.[11–15] Statistical tools have been developed to account for correlations and their uncertainties when calculating sample sizes for multiple testing procedures[11] and composite endpoints.[13] As outcomes of interest are often correlated in clinical trials,[16–18] ignoring the influence of correlations could lead to excessively large or small sample sizes when designing trials. Because the degree and direction of the influence vary between methods,[14] the one that provides the greatest power may also differ based on the correlations. However, the power and required sample sizes of different methods have not been compared under the influence of various correlations.

This study compared the power and required sample sizes of multiple-testing, composite outcome, and prioritized outcome approaches in clinical trial settings involving correlated binary outcomes. After briefly summarizing these methods and the impact of correlations, we conduct a simulation study to compare their power under various correlations. We then present a case study on sample size estimation accounting for anticipated outcome correlations and their uncertainty using a clinical trial example involving migraine treatment.[19]

Methods

Statistical methods for multiple binary outcomes

We consider a randomized controlled trial with an experimental arm (e) and a control arm (c). K denotes the number of binary components, and $Y_{k,e}$ and $Y_{k,c}$ represent the outcomes (1 for a response and 0 for no response) observed in the respective arms for $k = 1, \ldots, K$. The trial aims to show the superiority of the experimental treatment with a one-sided significance level of α.

With multiple-testing procedures, statistical tests are conducted on individual components,

$$H_{0k}: \pi_{k,e} - \pi_{k,c} \leq 0 \text{ and } H_{1k}: \pi_{k,e} - \pi_{k,c} > 0$$

for $k = 1, \ldots, K$, where $\pi_{k,e} = \Pr[Y_{k,e} = 1]$ and $\pi_{k,c} = \Pr[Y_{k,c} = 1]$.[14,20] If K components are co-primary endpoints, effects must be shown for all components with level α. If they are multiple primary endpoints, superiority can be concluded if the effects are shown in at least one component and the significance levels are adjusted to avoid Type I error inflation.

When using a composite outcome approach, investigators must first define the outcome. Typically, composite response outcomes are defined as the occurrence of a response in all of the components: 1 if positive in all elements and 0 otherwise (hereafter referred to as composite[all]).[21,22] Composite outcomes can also be defined as the occurrence of a response in a certain number of components: 1 if positive in r or more elements and 0 otherwise (hereafter referred to as composite[$\geq r$]).[23,24] Statistical tests are based on the composite response proportions,

$$H_0: \pi_e - \pi_c \leq 0 \text{ and } H_1: \pi_e - \pi_c > 0,$$

where π_e and π_c denote Pr[composite outcome = 1] in the respective arms.

In generalized pairwise comparisons, all combinations of experimental and control subjects are evaluated on each component based on their predefined priority.[9] For each component, a pair is considered a win if the experimental subject has a better outcome than the control subject ($Y_{k,e} = 1$ and $Y_{k,c} = 0$), or a loss if the experimental subject has a worse outcome ($Y_{k,e} = 0$ and $Y_{k,c} = 1$). The pairs are evaluated sequentially until they are classified as a win/loss or a tie at the final component. Statistical tests can be conducted on the difference between proportions in favor of each treatment (net benefit):

$$H_0: \Pr[\text{pair is classified as win}] - \Pr[\text{pair is classified as loss}] \leq 0 \text{ and}$$
$$H_1: \Pr[\text{pair is classified as win}] - \Pr[\text{pair is classified as loss}] > 0.$$

Statistical tests can also be constructed for relative measures (win ratio and win odds).[25]

Impact of outcome correlations on treatment effect measures

Correlations between outcomes are fundamental characteristics of treatments and often differ depending on treatment options.[18] Treatment efficacy and adverse events, for example, may be positively associated if they share a common biological mechanism; conversely, if the adverse events prevent efficacy, they may be negatively associated.[17] When statistical tests are conducted on individual proportions ($\pi_{k,e} - \pi_{k,c}$) with multiple-testing procedures, the individual test statistics would be correlated according to the outcome correlations, affecting the overall power and sample size.[11,14]

When multiple components are reduced to one-dimensional endpoints, such as composite outcomes or the net benefit, correlations directly impact their point estimates.[12,13,18] The impact of the correlation in each arm can be formulated for simple cases with two binary components. In each arm *a* (*a* = *e*, *c*), we denote the degree of outcome overlap by $\pi_{1\&2,a} = \Pr[Y_{1,a} = 1, Y_{2,a} = 1]$. If $\pi_{1\&2,a} > \pi_{1,a}\pi_{2,a}$, Y_1 and Y_2 are positively correlated in arm *a*. Note that the range of $\pi_{1\&2,a}$ is determined by $\pi_{1,a}$ and $\pi_{2,a}$:

$$\max(0, \pi_{1,a} + \pi_{2,a} - 1) \leq \pi_{1\&2,a} \leq \min(\pi_{1,a}, \pi_{2,a}).$$

Therefore, the correlation has a smaller impact on the value of $\pi_{1\&2,a}$ when $\pi_{1,a}$ or $\pi_{2,a}$ has an extreme value (close to 0% or 100%).

With composite[all] (positive in both elements), the improvement in the experimental arm compared to the control arm is expressed as

$$\pi_{1\&2,e} - \pi_{1\&2,c}.$$

With fixed marginal probabilities, this probability is higher when the correlation is positive in the experimental arm and negative in the control arm.

Probabilities for composite[≥1] (positive in at least one element) can be expressed as $\pi_{1,a} + \pi_{2,a} - \pi_{1\&2,a}$ in each arm, as previously derived.[12] The improvement of the experimental arm over the control arm is

$$(\pi_{1,e} + \pi_{2,e} - \pi_{1\&2,e}) - (\pi_{1,c} + \pi_{2,c} - \pi_{1\&2,c}),$$

and is higher when the correlation is negative in the experimental arm and positive in the control arm.

The net benefit (priority 1: Y_1; priority 2: Y_2) is expressed as

$$\pi_{1\&2,e}(\pi_{1,c} - 0.5) - \pi_{1\&2,c}(\pi_{1,e} - 0.5) + C(\pi_{1,e}, \pi_{1,c}, \pi_{2,e}, \pi_{2,c}),$$

where $C(\pi_{1,e}, \pi_{1,c}, \pi_{2,e}, \pi_{2,c})$ is a term of marginal probabilities.[18] This equation shows that the direction of the impact of the correlations on the net benefit depends on whether the response probabilities at the first-priority component are higher than 50%, as reported previously.[18]

Simulation

We conducted simulations to compare the power of statistical methods for multiple binary outcomes. We considered a randomized trial with a moderate sample size of 200 per arm, aiming to prove the superiority of the experimental treatment with a one-sided significance level of 2.5%. First, we assessed the Type I error rate and power of each method with varying numbers of components (K = 2, …, 10), treatment effects, marginal proportions, and outcome correlations. The scenarios with common treatment effects of 0% on all components were used to assess the Type I error rate (Web Appendix: Table S1A), while those with 5%, 10%, and 15% were used to assess the power (Web Appendix: Table S1B). We considered relatively rare outcomes (common marginal proportions in the control arm of 20%) and prevalent outcomes (60%). Common pairwise correlations between the outcomes $\rho_e = \rho_c$ were also varied (0, 0.45, 0.9) using the Gaussian copula-based model (details are provided in the Web Appendix: Table S1). For each dataset, we estimated the one-sided Wald P values based on the differences in proportions at individual components, composite[all], and composite[≥K/2] (positive in at least half of the components) and the net benefit (priority was given in the order of k = 1, …, K). The *BuyseTest* (version 2.4.0) and *stats* packages of the R statistical software (version 4.3.1) were used to estimate the net benefit and other proportions, respectively. These procedures were repeated 100,000 times. The proportion of null hypothesis rejections was calculated for co-primary endpoints, multiple primary endpoints (Bonferroni-corrected significance level: 2.5% / K),[20] composite[all], composite[≥K/2], and the net benefit.

Additionally, using the case with two components (K = 2), we evaluated the power of different statistical methods under varying combinations of correlations. Outcome correlations ρ_e and ρ_c were varied systematically (−0.9, −0.45, 0, 0.45, 0.9). Scenarios also varied according to treatment effects $\pi_{1,e} - \pi_{1,c}$ and $\pi_{2,e} - \pi_{2,c}$ (10% and 10%, 10% and 0%, 0% and 0%, 15% and −5%, −5% and 15%) and marginal proportions in the control arm $\pi_{1,c}$ and $\pi_{2,c}$ (20% and 40%, 60% and 40%) (Web Appendix: Table S1C). For each dataset, we estimated the point estimates and one-sided Wald P values based on the individual components, composite[all], composite[≥1], and the net benefit (1: Y_1; 2: Y_2). We then computed the mean of the estimates and proportion of null hypothesis rejections (power) for each scenario and method. We also computed the power for co-primary endpoints as the

proportion of rejecting the null hypotheses of both components and the power for multiple primary endpoints as the proportion of rejecting at least one null hypothesis with a Bonferroni-corrected significance level of 1.25%.[20]

Application

Accurate sample size calculation is essential for appropriately conducting clinical trials. Previous studies have suggested that accounting for anticipated correlations is critical when determining sample size.[11–15] When outcome correlations are known prior to a trial, the sample size can be calculated using those values, either through simulations or sample size formulas. Correlations between outcomes are rarely reported, however, making it difficult for investigators to make an exact assumption about the values.

When an experimental treatment has an innovative mechanism for treating a disease, making assumptions about specific correlations can be challenging. Nevertheless, investigators can often apply certain constraints to these correlations based on early-phase trial results or known treatment characteristics. It may be justified, for instance, to assume that correlations are common in both arms if the two treatments share similar biological properties. Also, if a treatment impacts multiple outcomes through the same biological pathways, assuming positive outcome correlations could be plausible. When considering a range of potential correlation values, the standard practice is to select the largest sample size from those calculated across all scenarios, as previously described.[13,15] Here, we present a case study on sample size estimation that accounts for uncertainty in correlations, focusing on a clinical trial for migraine treatment.[19]

In this trial, 365 migraine patients were randomly assigned to receive 1.0, 1.9, or 3.8 mg of adhesive dermally applied microarray zolmitriptan or a placebo. The patients declared their most bothersome symptom (MBS) other than pain (nausea, photophobia, or phonophobia) during the initial stage of the trial. The co-primary endpoints were freedom from pain and MBS, measured two hours post-dose, and only the 3.8 mg dose of zolmitriptan was superior to the placebo in both endpoints (two-sided significance level: 5%). Pain-free status was observed in 41.5% of the 3.8 mg group and 14.3% of the placebo group, while MBS-free status was observed in 68.3% of the 3.8 mg group and 42.9% of the placebo group. Treatment-emergent adverse events (TEAEs) were observed in 42.5% of the 3.8 mg group and 18.1% of the placebo group.

Using the case of 3.8 mg of zolmitriptan versus placebo, we estimated the minimum sample size per arm (1:1 allocation) to achieve a power of 90% to demonstrate the superiority of 3.8 mg of

zolmitriptan with a one-sided significance level of 2.5% for various methods and correlations. The endpoint priorities were determined based on previous investigations on patient expectations:[3] (i) freedom from pain (*3.8 mg*: 41.5%, *placebo*: 14.3%), (ii) freedom from MBS (*3.8 mg*: 68.3%, *placebo*: 42.9%), and (iii) no TEAEs (*3.8 mg*: 57.5%, *placebo*: 81.9%).

First, we used the first two components, Pain and MBS, which were the co-primary endpoints in the actual trial. We considered co-primary endpoints, multiple primary endpoints, composite[all] (freedom from pain and MBS), composite[≥1] (freedom from at least one of the two), and the net benefit (1: pain-free; 2: MBS-free) with each combination of correlations (−0.9, −0.45, 0, 0.45, 0.9), following the same data generation process as the simulation. Second, using all three components, we considered co-primary endpoints, multiple primary endpoints, composite[all] (freedom from pain, MBS, and TEAEs), composite[≥2] (freedom from at least two of the three), and the net benefit (1: pain-free; 2: MBS-free; 3: no TEAEs). We set the common pairwise correlations between freedom from pain, freedom from MBS, and the *presence* of TEAEs as 0, 0.45, and 0.9, assuming no or positive correlations between the treatment-related symptom changes. Sample sizes were explored in the range below 10,000.

Additionally, we considered smaller treatment effects by halving the between-arm differences and maintaining the proportions for the control: (i) freedom from pain (27.9% [*3.8 mg*], 14.3% [*placebo*]), (ii) freedom from MBS (55.6%, 42.9%), and (iii) no TEAEs (69.7%, 81.9%). For the three-component case, we also considered another likely scenario where the presence of TEAEs is negatively correlated with freedom from pain and MBS. We set the correlations between freedom from pain, freedom from MBS, and the *absence* of TEAEs to 0.45 and 0.9.

Note that some of the analytical approaches described above are less likely to be used in actual trials. Since pain control is essential for migraine treatment,[2,3] methods that may determine overall responders or superior treatments without considering the evaluation of pain would be unsuitable (composite[≥1 or ≥2] and multiple primary endpoints). Furthermore, demonstrating a decrease in adverse events compared to a placebo is generally unnecessary, excluding the use of co-primary endpoints with the three components. These results are presented solely for illustration.

Results

Simulation results

Type I error rate. The empirical Type I error rate in the simulation for each scenario and method is presented in the Web Appendix: Table S2. In most cases, the Type I error rate was approximately equal to or below the nominal level.

Power. The power with treatment effects of 10% for 2–10 components is presented in Figure 1. Tabular results are provided in the Web Appendix: Table S3. The power of co-primary endpoints decreased as the number of components increased; conversely, the power of multiple primary endpoints increased when the correlations were weak (0 or 0.45). With a control proportion of 20%, the power of composite[all] decreased as the number of components increased; however, this trend was unclear with a control proportion of 60%. The composite[$\geq K/2$] with a control proportion of 20% and the net benefit were less susceptible in terms of power to increases in the number of components. Similar trends were observed, albeit less obvious, with treatment effects of 5% and 15% (Web Appendix: Figure S1, S2).

Mean estimate with two components. Figure 2 presents the mean differences in individual/composite response proportions and the means of the net benefit estimates for each scenario with two components. The marginal proportions or correlations did not impact the differences in proportions on individual components. In contrast, the mean estimates for composite outcomes and net benefit differed depending on the marginal proportions and correlations. Regardless of the marginal proportion, the mean estimate with composite[all] was higher with positive correlations in the experimental arm and negative correlations in the control arm, whereas the mean estimate with composite[≥1] was higher with negative correlations in the experimental arm and positive correlations in the control arm. The net benefit estimate was higher with negative correlations in the experimental arm and positive correlations in the control arm when $\pi_{1,c} = 20\%$ (**A**), whereas it was higher with positive correlations in the experimental arm and negative correlations in the control arm when $\pi_{1,c} = 60\%$ (**B**). These trends were consistent with the known results in the simple cases presented above.

Power with two components. The power for each scenario with two components is presented in Figure 3. The results with $\rho_e = \rho_c$ are presented in tabular form in the Web Appendix: Table S4. Both components 1 and 2 were moderately powered (51.8–64.6%) when treatment effects were 10% and 10% (row 1). In this scenario, the power of co-primary endpoints was higher when both arms had positive correlations, whereas the power of multiple primary endpoints was higher with negative correlations. The impact of correlations was more evident in the composite outcomes and the net benefit. Regardless of the marginal proportion, the power of composite[all] was higher with

positive correlations in the experimental arm and negative correlations in the control arm, whereas that of composite[≥1] was higher with negative correlations in the experimental arm and positive correlations in the control arm. The power of the net benefit was higher with negative correlations in the experimental arm and positive correlations in the control arm when $\pi_{1,c} = 20\%$ (**A**), whereas it was higher with positive correlations in the experimental arm and negative correlations in the control arm when $\pi_{1,c} = 60\%$ (**B**). These trends were consistent with the mean estimate. Even when component 2 had a null marginal effect (row 2), the composite outcomes and net benefit presented higher power than with component 1 individually, depending on the correlations; for example, when $\pi_{1,c} = 60\%$, with $\rho_c < 0$ for composite[all] and net benefit, and $\rho_c > 0$ for composite[≥1]. When both had null marginal effects (row 3), composite outcomes and net benefit demonstrated high power, particularly when compared to the specified Type I error rate (note: in these cases, mean point estimates were >0%). When either component had a negative effect, the power of net benefit was higher when the first component's effect was positive (row 4) rather than the second (row 5), despite the fact that only the order of treatment effects differed. When there was a discrepancy in power between the components, the power of co-primary endpoints was close to the lower of the two, while that of multiple primary endpoints was close to the higher (rows 2, 4, 5).

Application results

Pain and MBS. The sample sizes with two components (Pain and MBS) are presented in Table 1. The sample size for co-primary endpoints was the least affected by the correlations among the five methods, while that for composite outcomes varied greatly. While both composite outcomes had smaller sample sizes than the co-primary in over half of the scenarios, the maximum sample size across all combinations of correlations was smaller with co-primary endpoints. The sample sizes for composite[all] were extremely large when the correlations were negative in the experimental arm and positive in the control arm, and those for composite[≥1] were high when the correlation was positive in the experimental arm and negative in the control arm. The correlations also affected the sample sizes for multiple primary endpoints and the net benefit, although they remained smaller than co-primary endpoints. When $\rho_e = \rho_c$, the sample size was the smallest for the net benefit. Similar trends were observed with reduced treatment effects, though required sample sizes were generally larger and the sample sizes for the net benefit exceeded those of co-primary when $\rho_e = 0.9$ and $\rho_c \leq 0$ (Web Appendix: Table S5)

Pain, MBS, and TEAE. The sample sizes with three components (Pain, MBS, and TEAE) are presented in Table 2. With co-primary endpoints, the sample sizes were not found below 10,000. As was the case with two components, the correlations greatly affected the sample sizes for composite

outcomes. The net benefit had the smallest sample sizes in all scenarios. Similar trends were observed with reduced treatment effects (Web Appendix: Table S6). When TEAEs were negatively associated with freedom from Pain and MBS, the correlation worked in different directions for composite outcomes, while the results for the other methods remained unchanged (Web Appendix: Table S7).

Discussion

This study compared the power and sample sizes of statistical methods for multiple binary outcomes. Overall, the marginal proportions, treatment effects, number of components, and correlations affected the power and sample size of the methods assessed. While the correlations are less likely to be emphasized when designing actual trials, they greatly influenced the power and sample size of composite outcomes and net benefit, demonstrating their dependence not only on marginal effects but also on the joint distribution of outcomes. The power and sample size of co-primary endpoints were the most stable across various correlations, though their power declined substantially when the treatment effects were opposite on some components or more than two components were present.

Correlations influenced the power and sample size of multiple-testing procedures, composite endpoints, and generalized pairwise comparisons, consistent with previous results.[11–15] However, our results showed that the direction and degree of their influence depended on the method used. The method with the greatest power and smallest sample size also differed depending on the correlations. Therefore, anticipated correlations and their degree of uncertainty should be taken into account when selecting statistical methods. Given the maximum required sample size across all correlation scenarios, employing composite outcomes would be impracticable when no constraints can be set for correlations.

In the migraine treatment example, actual correlations between the outcomes are unknown. Freedom from MBS could be positively associated with freedom from pain if the treatment works for both symptoms via the same mechanism, while it could be negatively associated if the MBS becomes apparent once the pain disappears. As a placebo was the control in the trial, it would be sensible to assume that correlations differ from those in the active treatment. In this trial, choosing co-primary endpoints was appropriate because they were robust to correlation differences and, unlike other methods, ensured the superiority of both efficacy outcomes. However, it may be reasonable to assume that correlations are common if two active treatments have similar biological properties. When correlations were the same in both arms, the net benefit often had the greatest

power and smallest sample size. Generalized pairwise comparisons could be useful in such situations, allowing for comparisons that consider factors not included in the co-primary outcomes of the trial (e.g., adverse events, symptoms beyond the most bothersome ones).

Although we considered one-sided testing, many trials employ bilateral hypotheses. Our findings can be seamlessly applied to such cases. When calculating the sample size for a hypothetical treatment effect, a two-sided significance level is divided by 2, thus becoming equivalent to the calculation for a one-sided test. This study used simulations to estimate sample sizes. While various sample size formulas have been proposed for multiple binary outcomes,[11,26,27] our study focused on the actual sample sizes required to achieve certain powers rather than on the validity of those formulas.[11,28,29]

Additionally, while our study focused on the power and sample size, the suitability of each method should also be evaluated from a clinical perspective.[30] For example, if an analysis aims to measure the balance between a treatment's benefits and harms, methods yielding one-dimensional endpoints based on multiple components could be helpful, even when they require larger sample sizes (e.g., composite outcomes and generalized pairwise comparisons based on response and adverse events).[12,17,31] Moreover, when some components are more relevant than others, multiple primary endpoints and composite outcomes that require only some of the components to be positive would be inappropriate, as illustrated by our migraine treatment example.

Several limitations should be noted when interpreting our findings. First, while we defined composite outcomes as positive responses in some or all components, composite outcomes are sometimes defined as the occurrence of at least one event or disease of interest.[12,13] Our results with composite outcomes requiring all components to be positive can be extrapolated to those cases, as statistical tests on an increase in the proportion observing positive outcomes in all components are equivalent to those on a decrease in the proportion observing adverse outcomes in at least one component. Second, we only considered scenarios using common correlations in the estimation of power and sample size in the presence of three or more components. As the number of components increases, the lack of restrictions would greatly increase and complicate correlation scenarios, potentially resulting in unachievable sample sizes. Although it may sometimes be justified to assume common positive correlations between treatment-related symptom changes, it is generally advisable to choose statistical methods that are less affected by correlations when incorporating many outcomes. Third, there are further interesting adaptations of multiple testing procedures that

were not evaluated in our study. For example, 'ordered p-values' and 'rejecting at least k out of m hypotheses' may be considered depending on the nature of individual clinical trials. [32]

In conclusion, this study offers guidance on the use of different statistical methods for multiple possibly correlated binary outcomes. Overall, co-primary endpoints were found to be a reliable option for evaluating the superiority of all components; however, they were unsuitable for assessing the balance between treatment effects pointing in different directions. Generalized pairwise comparisons offered a useful alternative to handle multiple prioritized outcomes, often providing the smallest sample sizes when the correlations were shared between the arms.

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

1. Center for Drug Evaluation and Research. *Migraine: Developing Drugs for Acute Treatment*. Food and Drug Administration, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/migraine-developing-drugs-acute-treatment (February 2018, accessed 4 October 2023).

2. Davies GM, Santanello N, Lipton R. Determinants of patient satisfaction with migraine therapy. *Cephalalgia* 2000; 20: 554–560.

3. Mangrum R, Gerstein MT, Hall CJ 3rd, et al. Priority acute and preventive migraine treatment benefits: Results of the Migraine Clinical Outcome Assessment System (MiCOAS) qualitative study of people living with migraine. *Headache* 2023; 63: 953–964.

4. Houts CR, McGinley JS, Nishida TK, et al. Systematic review of outcomes and endpoints in acute migraine clinical trials. *Headache* 2021; 61: 263–275.

5. Gallagher RM, Kunkel R. Migraine medication attributes important for patient compliance: concerns about side effects may delay treatment. *Headache* 2003; 43: 36–43.

6. Tassorelli C, Diener H-C, Dodick DW, et al. Guidelines of the International Headache Society for controlled trials of preventive treatment of chronic migraine in adults. *Cephalalgia* 2018; 38: 815–832.

7. Edmeads J. Defining response in migraine: which endpoints are important? *Eur Neurol* 2005; 53 Suppl 1: 22–28.

8. Rodgers AJ, Hustad CM, Cady RK, et al. Total migraine freedom, a potential primary endpoint to assess acute treatment in migraine: comparison to the current FDA requirement using the complete rizatriptan study database. *Headache* 2011; 51: 356–368.

9. Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Stat Med* 2010; 29: 3245–3257.

10. Sakamaki K, Morita Y, Iba K, et al. Multiplicity adjustment and sample size calculation in clinical trials with multiple endpoints: an industry survey of current practices in Japan. *Ther Innov Regul Sci* 2020; 54: 1097–1105.

11. Sozu T, Sugimoto T, Hamasaki T. Sample size determination in clinical trials with multiple co-primary binary endpoints. *Stat Med* 2010; 29: 2169–2179.

12. Marsal J-R, Ferreira-González I, Bertran S, et al. The use of a binary composite endpoint and sample size requirement: influence of endpoints overlap. *Am J Epidemiol* 2017; 185: 832–841.

13. Bofill Roig M, Gómez Melis G. A new approach for sizing trials with composite binary endpoints using anticipated marginal values and accounting for the correlation between components. *Stat Med* 2019; 38: 1935–1956.

14. McMenamin ME, Barrett JK, Berglind A, et al. Sample size estimation using a latent variable model for mixed outcome co-primary, multiple primary and composite endpoints. *Stat Med* 2022; 41: 2303–2316.

15. Deltuvaite-Thomas V, Burzykowski T. Operational characteristics of generalized pairwise comparisons for hierarchically ordered endpoints. *Pharm Stat* 2022; 21: 122–132.

16. Evans SR, Follmann D. Using outcomes to analyze patients rather than patients to analyze outcomes: a step toward pragmatism in benefit:risk evaluation. *Stat Biopharm Res* 2016; 8: 386–393.

17. Buyse M, Saad ED, Peron J, et al. The Net Benefit of a treatment should take the correlation between benefits and harms into account. *J Clin Epidemiol* 2021; 137: 148–158.

18. Fuyama K, Ogawa M, Mizusawa J, et al. Impact of correlations between prioritized outcomes on the net benefit and its estimate by generalized pairwise comparisons. *Stat Med* 2023; 42: 1606–1624.

19. Spierings EL, Brandes JL, Kudrow DB, et al. Randomized, double-blind, placebo-controlled, parallel-group, multi-center study of the safety and efficacy of ADAM zolmitriptan for the acute treatment of migraine. *Cephalalgia* 2018; 38: 215–224.

20. Dmitrienko A, D'Agostino R Sr. Traditional multiplicity adjustment methods in clinical trials. *Stat Med* 2013; 32: 5172–5218.

21. Luijten KMAC, Tekstra J, Bijlsma JWJ, et al. The Systemic Lupus Erythematosus Responder Index (SRI); a new SLE disease activity assessment. *Autoimmun Rev* 2012; 11: 326–329.

22. Elkind MSV, Veltkamp R, Montaner J, et al. Natalizumab in acute ischemic stroke (ACTION II): A randomized, placebo-controlled trial. *Neurology* 2020; 95: e1091–e1104.

23. Ginsberg MD, Palesch YY, Hill MD, et al. High-dose albumin treatment for acute ischaemic stroke (ALIAS) Part 2: a randomised, double-blind, phase 3, placebo-controlled trial. *Lancet Neurol* 2013; 12: 1049–1058.

24. Arends S, de Wolff L, van Nimwegen JF, et al. Composite of Relevant Endpoints for Sjögren's Syndrome (CRESS): development and validation of a novel outcome measure. *The Lancet Rheumatology* 2021; 3: e553–e562.

25. Dong G, Huang B, Verbeeck J, et al. Win statistics (win ratio, win odds, and net benefit) can complement one another to show the strength of the treatment effect on time-to-event outcomes. *Pharm Stat* 2023; 22: 20–33.

26. Mao L, Kim K, Miao X. Sample size formula for general win ratio analysis. *Biometrics* 2022; 78: 1257–1268.

27. Yu RX, Ganju J. Sample size formula for a win ratio endpoint. *Stat Med* 2022; 41: 950–963.

28. Gasparyan SB, Kowalewski EK, Koch GG. Comments on 'Sample size formula for a win ratio endpoint' by R. X. Yu and J. Ganju. *Statistics in medicine* 2022; 41: 2688–2690.

29. Yu RX, Ganju J. Sample size formula for a win ratio endpoint (reply). *Statistics in medicine* 2022; 41: 2691–2692.

30. Center for Drug Evaluation and Research. *Multiple Endpoints in Clinical Trials: Guidance for Industry*. Food and Drug Administration, https://www.fda.gov/media/162416/download (October 2022, accessed 5 June 2024).

31. Shaw PA. Use of composite outcomes to assess risk-benefit in clinical trials. *Clin Trials* 2018; 15: 352–358.

32. Maurer W, Bretz F. Memory and other properties of multiple test procedures generated by entangled graphs. *Stat Med* 2013; 32: 1739–1753.

Table 1. Sample size per arm to achieve 90% power in each scenario with Pain and MBS.

Method	Correlation in the control arm	Correlation in the experimental arm				
		-0.9	-0.45	0	0.45	0.9
Co-primary	-0.9	***84***	84	84	84	83
	-0.45	84	***84***	83	84	84
	0	84	84	***84***	83	83
	0.45	84	83	83	***83***	83
	0.9	84	83	83	83	***82***
Multiple primary	-0.9	***37***	39	43	44	47
	-0.45	37	***42***	44	45	48
	0	39	43	***45***	47	51
	0.45	41	45	47	***48***	53
	0.9	44	47	48	53	***55***
Composite[all]	-0.9	***51***	29	22	18	15
	-0.45	118	***47***	32	24	19
	0	402	96	***55***	37	26
	0.45	3527	210	96	***57***	39
	0.9	> 9999	532	171	91	***56***
Composite[≥1]	-0.9	***19***	41	74	142	364
	-0.45	17	***36***	63	112	255
	0	15	30	***48***	83	160
	0.45	14	25	39	***60***	106
	0.9	12	21	32	48	***78***
Net benefit	-0.9	***19***	27	36	47	65
	-0.45	20	***27***	35	46	62
	0	21	27	***34***	44	59
	0.45	22	27	34	***42***	55
	0.9	23	28	33	40	***52***

Bold Italic: results with the same correlation values for both arms

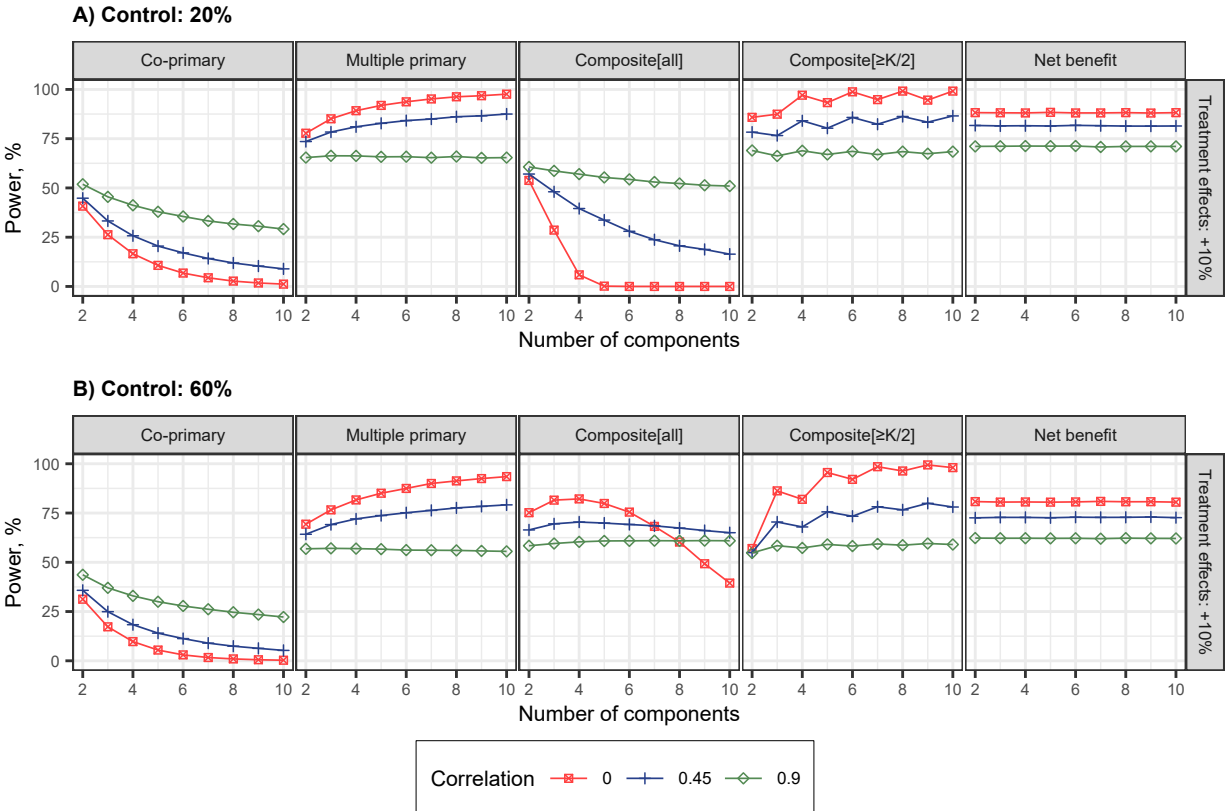
Table 2. Sample size per arm to achieve 90% power in each scenario with Pain, MBS, and TEAE.

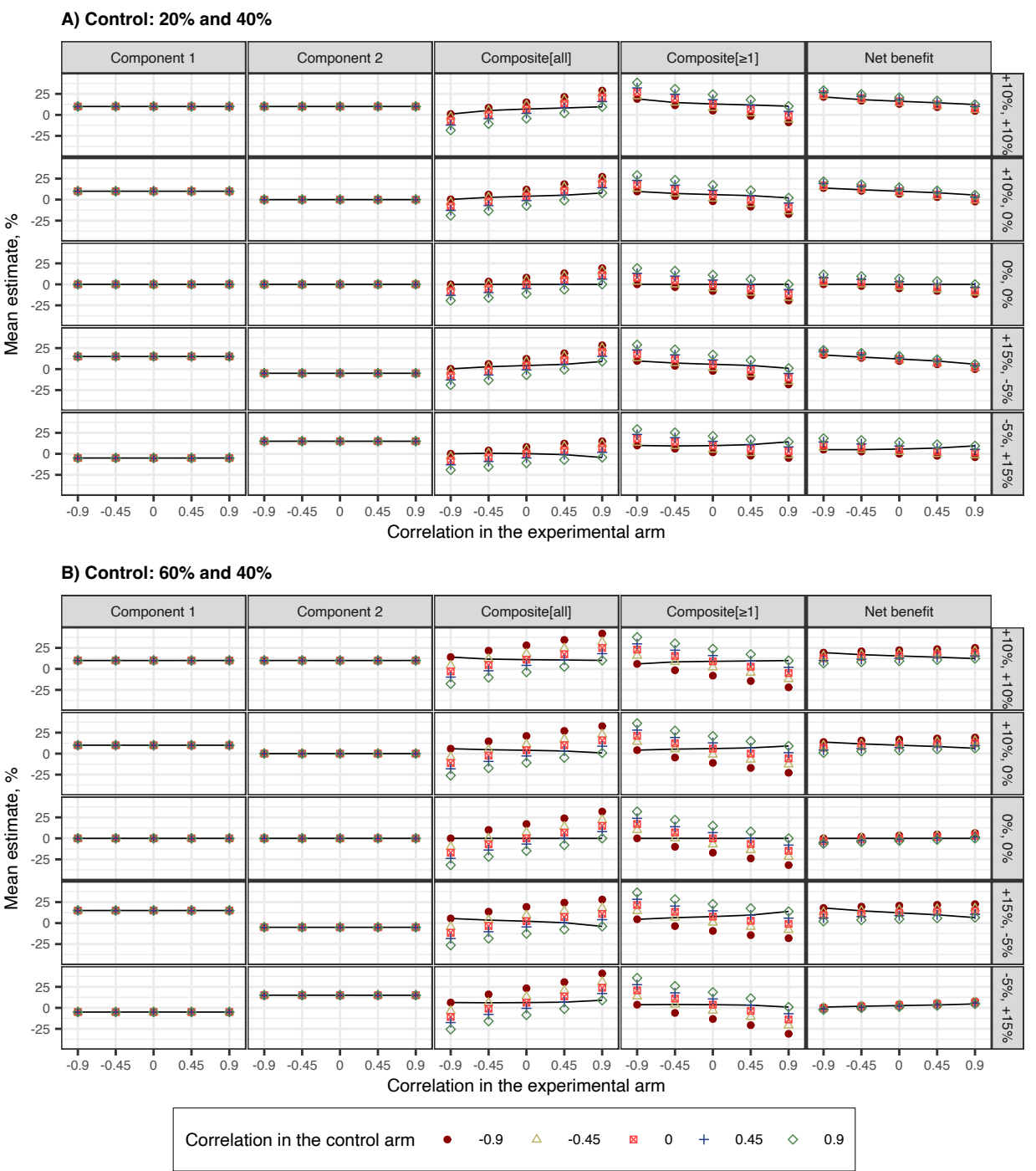
	Correlation		
Method	0	0.45	0.9
Co-primary	> 9999	> 9999	> 9999
Multiple primary	50	56	61
Composite[all]	148	338	671
Composite[≥2]	203	113	83
Net benefit	34	42	52

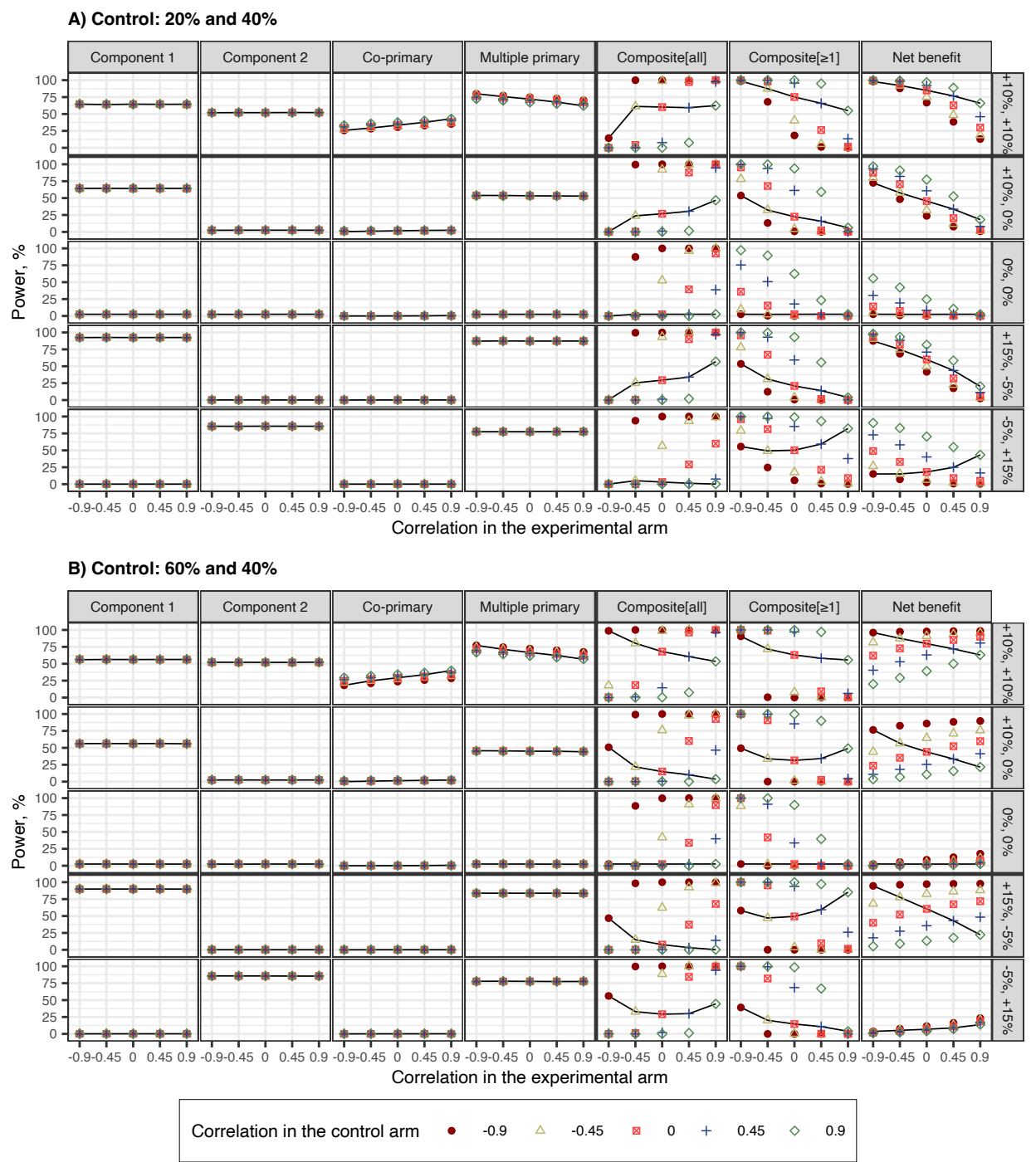
Figure 1. Power in the simulation with two to ten components and treatment effects of 10% according to the marginal proportions in the control arm (**A** 20%; **B** 60%).

Figure 2. Mean estimate in the simulation according to the marginal proportions in the control arm (**A** 20% and 40%; **B** 60% and 40%), methods (columns), and treatment effects (rows).
The lines connect the results with the same correlations for both arms.

Figure 3. Power in the simulation according to the marginal proportions in the control arm (**A** 20% and 40%; **B** 60% and 40%), methods (columns), and treatment effects (rows).
The lines connect the results with the same correlations for both arms.


A) Control: 20%
Co-primary
Multiple primary
Composite[all]
Composite[≥K/2]
Net benefit
Power, %
Number of components
Treatment effects: +10%
B) Control: 60%
Correlation
0
0.45
0.9


A) Control: 20% and 40%
Component 1
Component 2
Composite[all]
Composite[≥1]
Net benefit
+10%, +10%
+10%, 0%
0%, 0%
+15%, -5%
-5%, +15%
Mean estimate, %
25
0
-25
-0.9
-0.45
0
0.45
0.9
Correlation in the experimental arm
B) Control: 60% and 40%
Component 1
Component 2
Composite[all]
Composite[≥1]
Net benefit
+10%, +10%
+10%, 0%
0%, 0%
+15%, -5%
-5%, +15%
Mean estimate, %
25
0
-25
-0.9
-0.45
0
0.45
0.9
Correlation in the experimental arm
Correlation in the control arm
-0.9
-0.45
0
0.45
0.9


A) Control: 20% and 40%
Component 1
Component 2
Co-primary
Multiple primary
Composite[all]
Composite[≥1]
Net benefit
+10%, +10%
+10%, 0%
0%, 0%
+15%, -5%
-5%, +15%
Power, %
100
75
50
25
0
-0.9 -0.45 0 0.45 0.9
Correlation in the experimental arm
B) Control: 60% and 40%
Correlation in the control arm
-0.9
-0.45
0
0.45
0.9