Published in final edited form as:
Stat Med. 2022 June 15; 41(13): 2303–2316. doi:10.1002/sim.9356.

Sample size estimation using a latent variable model for mixed outcome co-primary, multiple primary and composite endpoints

Martina E. McMenamin[1,2], **Jessica K. Barrett**[1], **Anna Berglind**[3], **James M. S. Wason**[1,4]

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

[2]WHO Collaborating Centre for Infectious Disease Epidemiology and Control, School of Public Health, The University of Hong Kong, Hong Kong Special Administrative Region, China

[3]Late Respiratory & Immunology, Biometrics, BioPharmaceuticals R& D, AstraZeneca, Gothenburg, Sweden

[4]Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK

Abstract

Mixed outcome endpoints that combine multiple continuous and discrete components are often employed as primary outcome measures in clinical trials. These may be in the form of co-primary endpoints, which conclude effectiveness overall if an effect occurs in all of the components, or multiple primary end-points, which require an effect in at least one of the components. Alternatively, they may be combined to form composite endpoints, which reduce the outcomes to a one-dimensional endpoint. There are many advantages to joint modeling the individual outcomes, however in order to do this in practice we require techniques for sample size estimation. In this article we show how the latent variable model can be used to estimate the joint endpoints and propose hypotheses, power calculations and sample size estimation methods for each. We illustrate the techniques using a numerical example based on a four-dimensional end-point and find that the sample size required for the co-primary endpoint is larger than that required for the individual endpoint with the smallest effect size. Conversely, the sample size required in the multiple primary case is similar to that needed for the outcome with the largest effect size. We show that the empirical power is achieved for each endpoint and that the FWER can be sufficiently controlled using a Bonferroni correction if the correlations between endpoints are less than 0.5. Otherwise, less conservative adjustments may be needed. We further illustrate empirically the efficiency gains that may be achieved in the composite endpoint setting.

Keywords

latent variable modeling; mixed outcome endpoints; sample size estimation

Correspondence to: Martina E. McMenamin.

Correspondence Martina E. McMenamin, WHO Collaborating Centre for Infectious Disease Epidemiology and Control, School of Public Health, The University of Hong Kong, G/F, Patrick Manson Building, 7 Sassoon Road, Pokfulam, Hong Kong Special Administrative Region, China. mmcm@hku.hk.

Europe PMC Funders Author Manuscripts

1 Introduction

Sample size estimation plays an integral role in the design of a study. The objective is to determine the minimum sample size that is large enough to detect, with a specified power, a clinically meaningful treatment effect. Although it is crucial that investigators have enough patients enrolled to detect this effect, overestimating the sample size also has ethical and practical implications. Namely, in a placebo-controlled trial, more patients are subjected to a placebo arm than is necessary, therefore withholding access to potentially beneficial drugs from them and delaying access to future patients.[1–3] Furthermore it results in longer, more expensive trials, using resources that could be allocated elsewhere.

One vital aspect of sample size determination is the primary endpoint. Typically this is a single outcome, however in some instances there may be multiple outcomes of interest and so various combinations of these outcomes can be selected as the primary endpoint, depending on the hypothesis of interest. Assuming we have three outcomes of interest ν_1, ν_2 and ν_3, one option is a co-primary endpoint, which takes the form of the multivariate endpoint $\nu_1 \cap \nu_2 \cap \nu_3$. This means that an intervention is deemed to be effective overall if it is shown to be effective in each of ν_1, ν_2, and ν_3. Alternatively multiple primary endpoints may be of interest, which take the multivariate form $\nu_1 \cup \nu_2 \cup \nu_3$, where an intervention is deemed effective if it is shown to be effective in at least one of ν_1, ν_2, or ν_3. Another possibility is a composite endpoint, involving some function that maps the multivariate outcome to a univariate outcome for inference, for example $\nu_1 + \nu_2 + \nu_3$. In this case the outcomes within the composite may be assigned equal or differing degrees of relevance depending on clinical importance.[4] Alternatively, the composite endpoint may combine outcomes by labeling patients as ‘responders’ or ‘non-responders’ based on whether they exceed predefined thresholds in each of the outcomes. For instance, we let a response indicator $S = 1$ if $\nu_1 \leq \eta_1$, $\nu_2 \leq \eta_2$, and $\nu_3 \leq \eta_3$, where η denotes the response cutpoints. Note that the composite case is distinct from the others in that it combines the parameters and hence test statistics for each outcome into one, rather than these remaining separate for each outcome. This will have implications for sample size estimation.

For each of these endpoints, the individual outcomes may be a mix of multiple continuous, ordinal, and binary measures. One possible way to jointly model the outcomes is using a latent variable framework, arising in the graphical modeling literature, in which discrete outcomes are assumed to be latent continuous variables subject to estimable thresholds and modeled using a multivariate normal distribution.[5,6] By employing this framework we can take account of the correlation between the outcomes, improve the handling of missing data in individual components and potentially increase efficiency. Furthermore, in the case of multiple primary outcomes, it may reduce the severity of multiple testing corrections required by accounting for correlation between endpoints.

A barrier to adopting these techniques is a lack of consensus on sample size determination. A recent and comprehensive overview of the existing literature for sample size calculation in clinical trials with co-primary and multiple primary endpoints is provided by Sozu et al.[7] The review found many proposals for power and sample size calculations for multiple continuous outcomes. In the co-primary case, some of these were based on assuming that

the endpoints were bivariate normally distributed,[8,9] and extended for the case of more than two endpoints.[10,11] Other work focused on testing procedures[12,13] and controlling the type I error rate.[14–17] Similar ideas were investigated for multiple primary endpoints.[14,17–19]

Approaches to sample size estimation for composite endpoints have focused primarily on the case of multiple binary components.[20–25] In the case of binary co-primary endpoints, five methods of power and sample size calculation based on three association measures have been introduced.[26] Additionally, sample size calculation for trials using multiple risk ratios and odds ratios for treatment effect estimation is discussed by Hamasaki et al,[27] and Song[28] explores co-primary endpoints in non-inferiority clinical trials. Consideration has also been given to the case where two co-primary endpoints are both time-to-event measures where effects are required in both endpoints,[29–31] and at least one of the endpoints.[32,33] Furthermore, composites comprised of time-to-event measures are common, in which the composite reflects time-to-first-event variable.[34] Sample size estimation in this case has been considered by Sugimoto et al.[35]

The mixed outcome setting has received substantially less consideration. One paper considers overall power functions and sample size determinations for multiple co-primary endpoints that consist of mixed continuous and binary variables.[36] They assume that response variables follow a multivariate normal distribution, where binary variables are observed in a dichotomized normal distribution, and use Pearson's correlations for association. A modification was suggested by Wu and de Leon[37] which involved using latent-level tests and pairwise correlations, and provided increased power. Thesemethods focus on the co-primary endpoint case,where effects are required in all outcomes. The case ofmultiple primary or composite endpoints where the components are measured on different scales has not been considered, each of which will require distinct hypotheses. In practice, if a mixed outcome composite is selected as the primary endpoint in a trial then the sample size calculation may be based on an overall binary endpoint or collapsed to form multiple binary endpoints however this will result in a large loss in efficiency.[38]

In this article we build on the existing work for co-primary continuous and binary endpoints to include any combination of continuous, ordinal, and binary outcomes for co-primary, multiple primary, and composite endpoints. We propose a framework based on the same latent variable model and show how it may be tailored to each of the three endpoints to facilitate sample size estimation. The article will proceed as follows: in Section 2 we introduce the latent variable model, detailing how it can be used in each context, and specify hypothesis tests for each of the three combinations of mixed outcomes; in Section 3 we propose power calculations and sample size estimation techniques in each case; in Section 4 we illustrate the methods on a four dimensional endpoint consisting of two continuous, one ordinal and one binary outcome using a numerical example based on the MUSE trial;[39] and in Section 5 we simulate the empirical power for each test and the FWER for the union-intersection test. We conclude with a discussion and recommendations for practice in Section 6, and introduce user-friendly software and documentation for implementation in Section 7.

2 Endpoints and Hypothesis Testing

2.1 Latent variable framework

Let n_T and n_C represent the number of patients in the treatment group and the control group respectively and let K be the number of outcomes measured for each patient. Let $\mathbf{Y_{Ti}} = (Y_{Ti1}, \ldots, Y_{TiK})^T, i = 1, \ldots, n_T$ be vector of K responses for patient i on the treatment arm and $\mathbf{Y_{Ci}} = (Y_{Ci1}, \ldots, Y_{CiK})^T, i = 1, \ldots, n_C$ the vector of K responses for patient i on the control arm. Without loss of generality, the first $1 \leq k \leq k_m$ elements of $\mathbf{Y_{Ti}}$ and $\mathbf{Y_{Ci}}$ are observed as continuous variables, the next $k_m < k \leq k_o$ are observed as ordinal and the remaining $k_o < k \leq K$ are observed as binary. For instance, for a three dimensional endpoint with one continuous, one ordinal and one binary measure, $k_m = 1$, $k_o = 2$, and $K = 3$. We use the biserial model of association by Tate,[40] which is based on latent continuous measures manifesting as discrete variables. Formally, we say that $\mathbf{Y_{Ti}}$ and $\mathbf{Y_{Ci}}$ have latent variables $\mathbf{Y^*_{Ti}}$ and $\mathbf{Y^*_{Ci}}$ respectively, where $\mathbf{Y^*_{Ti}} \sim N_K(\mu_T, \sum_T)$ and $\mathbf{Y^*_{Ci}} \sim N_K(\mu_C, \sum_C)$, where $\mu_T = (\mu_{1T}, \ldots \mu_{KT}), \mu_{kT} = \mu_{kT0} + \mu_{kT1}x_{kT1} + \ldots + \mu_{kTp}x_{kTp}$ and $x_{kT1} \ldots x_{kTP}$ denotes the p covariates included in the model for outcome k. Likewise $\mu_C = (\mu_{1C}, \ldots \mu_{KC})$ are the corresponding quantities for the control arm. Then for $k \neq k' : 1 \leq k < k' \leq k_m$ let $Var(Y_{Tik}) = \sigma^2_{Tk}, Var(Y_{Cik}) = \sigma^2_{Ck}$ and $Corr(Y_{Tik}, Y_{Tik'}) = \rho_{Tkk'}, Corr(Y_{Cik}, Y_{Cik'}) = \rho_{Ckk'}$ where $\rho_{Tkk'}$ and $\rho_{Ckk'}$ are the associationmeasures between the endpoints. For $k_m < k \leq K$, $Var(Y^*_{Tik}) = Var(Y^*_{Cik}) = 1$ and $Corr(Y^*_{Tik}, Y^*_{Tik'}) = \rho^*_{Tkk'}, Corr(Y^*_{Cik}, Y^*_{Cik'}) = \rho^*_{Ckk'}$. The latent variables can be related to the observed variables by:

- $1 \leq k \leq k_m : Y_{Tik} = Y^*_{Tik}$ and $Y_{Cik} = Y^*_{Cik}$
- $$k_m < k \leq k_o: \; Y_{Tik} = \begin{cases} 0 & \text{if} \quad \tau_{k0} \leq Y^*_{Tik} < \tau_{k1}, \\ 1 & \text{if} \quad \tau_{k1} \leq Y^*_{Tik} < \tau_{k2}, \\ \vdots & \vdots \\ w_k & \text{if} \; \tau_{kw_k} \leq Y^*_{Tik} < \tau_{k(w_k+1)} \end{cases} \quad Y_{Cik}$$
$$= \begin{cases} 0 & \text{if} \quad \tau_{k0} \leq Y^*_{Cik} < \tau_{k1}, \\ 1 & \text{if} \quad \tau_{k1} \leq Y^*_{Cik} < \tau_{k2}, \\ \vdots & \vdots \\ w_k & \text{if} \; \tau_{kw_k} \leq Y^*_{Cik} < \tau_{k(w_k+1)} \end{cases}$$
- $$k_o < k \leq K : Y_{Tik} = \begin{cases} 0 \text{ if } \tau_{k0} \leq Y^*_{Tik} < \tau_{k1}, \\ 1 \text{ if } \tau_{k1} \leq Y^*_{Tik} < \tau_{k2} \end{cases} \quad Y_{Cik} = \begin{cases} 0 \text{ if } \tau_{k0} \leq Y^*_{Cik} < \tau_{k1}, \\ 1 \text{ if } \tau_{k1} \leq Y^*_{Cik} < \tau_{k2} \end{cases}$$

We set $\tau_{k\,0} = -\infty$, $\tau_k(w_k+1) = \infty$ and the intercepts μ_{kT0} and μ_{kC0} equal to zero for $k_m < k \leq k_o$ in order to estimate the cut-points. Additionally, $\tau_{k0} = -\infty$, $\tau_{k1} = 0$, $\tau_{k2} = \infty$ for $k_o < k \leq K$ so that the intercepts can be estimated for the outcomes observed as binary. The mixed outcomes are then combined as follows.

Europe PMC Funders Author Manuscripts

2.2 Co-primary endpoint

In this case, a treatment must be shown to be effective as measured by each of the outcomes in order to be deemed effective overall. We generalize previous work for mixed continuous and binary outcomes to include ordinal outcomes, as shown below.[36,37] In many clinical trials the hypothesis of interest is based on superiority, namely that the proposed treatment will perform better than the control treatment. The null hypothesis is that the difference in treatment effects for the treatment arm and control arm is less than or equal to zero. This is straightforward to formalize in the case of one endpoint but less so when there are multiple co-primary endpoints, particularly when they are measured on different scales. The hypothesis of interest is as shown in (1)

$$\begin{aligned} H_0&: \ \exists\, k\ s.t.\ \pi_{Tk} - \pi_{Ck} \leq 0 \\ H_1&: \ \pi_{Tk} - \pi_{Ck} > 0 \forall k, \end{aligned} \tag{1}$$

where $\boldsymbol{\pi}_{Tk}$ and $\boldsymbol{\pi}_{Ck}$ is the effect of the intervention in the treatment and control arm respectively. For $k_o < k \leq K$ we can specify $\pi_{Tik} = P(Y_{Tik} = 0) = P\left(Y^*_{Tik} < 0\right)$ and $\pi_{Cik} = P(Y_{cik} = 0) = P\left(Y^*_{Cik} < 0\right)$ for the treatment and control group.

We can generalize this assumption to account for the ordinal endpoints based on the fact that for $k_m < k \leq k_o \pi_{Tik} P(Y_{Tik} = w_k) = P\left(\tau_k w_k < Y^*_{Tik} < \tau_{k(w_k+1)}\right)$. The definition of treatment effect for ordinal outcomes may be modified to include multiple ordinal levels by selecting the appropriate $\boldsymbol{\tau}$ thresholds. For instance, $\pi_{Tik} = P(Y_{Tik} = 0) + P(Y_{Tik} = 1) + P(Y_{Tik} = 2) = P\left(-\infty < Y^*_{Tik} < \tau_{k3}\right)$. As the latent means are estimable by maximum likelihood, $\mu^*_{Ti1} = \Phi^{-1}(\pi_{Ti1}), \ldots, \mu^*_{Tik} = \Phi^{-1}(\pi_{Tik})$ in the treatment group and $\mu^*_{Ci1} = \Phi^{-1}(\pi_{Ci1}), \ldots, \mu^*_{Cik} = \Phi^{-1}(\pi_{CiK})$ in the control group.

We can proceed by specifying that the hypothesis in (1) holds if and only if the hypothesis

$$\begin{aligned} H^*_0&: \ \exists\, k\ s.t.\ \delta^*_k \leq 0 \\ H^*_1&: \ \delta^*_k > 0 \ \forall k, \end{aligned} \tag{2}$$

holds, where $\delta^*_k = \mu^*_{TK} - \mu^*_{CK}, \mu^*_{TK} = 1/n_T \Sigma_{i=1}^{nT} \mu^*_{Tik}$ and $\mu^*_{CK} = 1/n_C \Sigma_{i=1}^{nC} \mu^*_{Cik}$. The maximum likelihood estimates $\hat{\mu}^*_{TK}$ and $\hat{\mu}^*_{CK}$ can be used for a test of H^*_0 and the variance of this test statistic can be obtained using the inverse of the Fisher information matrix.

2.3 Multiple primary endpoint

Multiple primary endpoints conclude a treatment is effective if it is shown to work in at least one of the outcomes. We would expect the sample size required to be reduced compared with the co-primary endpoint case which would require power to detect treatments in all outcomes. We can allow for sample size estimation for multiple primary endpoints as follows.

The hypothesis of interest, accounting for the fact that a significant effect in only one outcome is required, is shown below.

$$\begin{aligned} H_0&: \ \pi_{Tk} - \pi_{Ck} \leq 0 \forall k \\ H_1&: \ \exists\, k\, s.t.\, \pi_{Tk} - \pi_{Ck} > 0. \end{aligned} \tag{3}$$

As before, $\boldsymbol{\pi}_{Tk}$ and $\boldsymbol{\pi}_{Ck}$ can be determined for $k_m < k \leq k_o$ using the relevant $\boldsymbol{\tau}$ thresholds.

$$\begin{aligned} H_0^*&: \ \delta_k^* \leq 0 \forall k \\ H_1&: \ \exists\, k\, s.t.\, \delta_k^* > 0. \end{aligned} \tag{4}$$

The difference in latent means $\delta_k^* = \mu_{Tk}^* - \mu_{Ck}^*$ and their variance are estimated using the maximum likelihood estimates and Fisher information matrix, as before.

2.4 Composite endpoint

A review conducted by Wason et al[41] showed that composite responder endpoints are widely used and identified many clinical areas in which they are common, such as oncology, rheumatology, cardiovascular, and circulation. The latent variable framework may be used to model the underlying structure of these mixed outcome composite endpoints to greatly improve efficiency.[38] The joint distribution of the continuous, ordinal, and binary outcomes is modeled using the latent variable structure as before. However, in this case the endpoint of interest is a composite responder endpoint and so the required quantity is some function of the probability of response in the treatment group p_T and in the control group p_C.

For instance, an overall responder index S_i can be formed for patient i, where $S_i = 1$ if $Y_{i1} \leq \eta_1, \ldots, Y_{ik}^* \leq \eta_k$ and 0 otherwise, where the quantities $(\eta_1, \ldots, \eta_K)$ are predefined responder thresholds. Generalizations where response only requires a certain number of the components to meet the thresholds are possible, but involve more complex sums. Note that this definition of response is distinct from that commonly found in composites formed from survival endpoints or binary composites typical in cardiovascular studies. We can specify p_{iT} and p_{iC}, the probability of response for patient i in the treatment arm and control arm respectively, as shown in (5),

$$\begin{aligned} p_{iT} &= P(S_i = 1 | T_i = 1) = \int_{-\infty}^{\eta 1} \ldots \int_{-\infty}^{\eta K} f_{Y_1, \ldots, Y_K}(y_{i1}, \ldots, y_{iK} | T_i = 1, \theta) dy_K \ldots dy1 \\ p_{iC} &= P(S_i = 1 | T_i = 0) = \int_{-\infty}^{\eta 1} \ldots \int_{-\infty}^{\eta K} f_{Y_1, \ldots, Y_K}(y_{i1}, \ldots, y_{iK} | T_i = 0, \theta) dy_K \ldots dy1, \end{aligned} \tag{5}$$

where θ is the vector of model parameters and we assume that $p_T \sim N\left(\delta_T, \sigma_{\delta_T}^2\right)$ and $p_C \sim N\left(\delta_C, \sigma_\delta^2\right)$. As in the case of co-primary and multiple primary endpoints, the assumptions allow us to estimate latent means $\left(\mu_{k_{m+1}}^*, \ldots, \mu_K^*\right)$ for the observed discrete components using the model parameters.

In the mixed outcome composite endpoint setting, note that although we are exploiting the latent multivariate Gaussian structure for efficiency gains we are ultimately still interested in a one dimensional endpoint, such as the difference in response probabilities between the treatment and control arms of the trial. This is distinct from the co-primary and multiple primary endpoints cases, where the overall hypothesis test must be based on some union or intersection of the hypotheses for the individual outcomes. For the composite endpoint we can formulate the hypothesis as shown in (6),

$$\begin{aligned} H_0&: \; p_T - p_C \leq 0 \\ H_1&: \; p_T - p_C > 0, \end{aligned} \tag{6}$$

where p_T and p_C are as in (5). For sample size estimation, we require the distribution of $\delta = p_T - p_C$ under H_1, which we can assume to be $\delta \sim N\left(\delta_T - \delta_C, \; \sigma_\delta^2\right)$. The hypothesis can therefore be stated as

$$\begin{aligned} H_0&: \; \delta^* \leq 0 \\ H_1&: \; \delta^* > 0, \end{aligned} \tag{7}$$

where $\delta^* = \delta_T^* - \delta_C^*, \delta_T^* = \; \Phi_K\left(\eta_1, \ldots, \eta_K; \; \boldsymbol{\mu}_T^*, \Sigma_T\right), \delta_C^* = \; \Phi_K\left(\eta_1, \ldots, \eta_K; \; \boldsymbol{\mu}_C^*, \Sigma_C\right)$ and $\Phi_K(.;\boldsymbol{\mu},\Sigma)$ is the K-dimensional multivariate normal distribution function, with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Estimates of the quantities can be obtained using the maximum likelihood estimates for the model parameters, as in the co-primary and multiple primary endpoint settings, so that $\hat{\delta}_T^* = \; \Phi_K\left(\eta_1, \ldots, \eta_K; \hat{\mu}_T^*, \hat{\Sigma}_T\right)$ and $\hat{\delta}_C^* = \; \Phi_K\left(\eta_1, \ldots, \eta_K; \hat{\mu}_C^*, \hat{\Sigma}_C\right)$, where $\boldsymbol{\mu}_T^*$ is the K-dimensional vector of mean values in the treatment arm and $\boldsymbol{\mu}_C^*$ is the corresponding vector for the control arm. Using a Taylor series expansion, we can obtain the quantity σ_δ^2 using the fact that $\text{var}\left(\hat{\delta}^*\right) \approx ({}''\delta)^T Cov\left(\hat{\theta}\right)({}'\delta)$. Then, $\hat{var}\left(\hat{\delta}^*\right) = ({}''\delta_T)^T \widehat{Cov}\left(\hat{\theta}\right)({}''\delta_T)$, where ${}''\delta$ is the vector of partial derivatives of δ* with respect to each of the parameter estimates. We can obtain $\hat{\theta}$ and covariance matrix $\widehat{Cov}\left(\hat{\theta}\right)$ by fitting the model to pilot trial data.

3 Sample Size Estimation

3.1 Co-primary endpoints

To construct the power function, we define the required quantities as follows. Let $\bar{Y}_{Tk} - \bar{Y}_{Ck}$ and $\hat{\mu}_{Tk}^* - \hat{\mu}_{Ck}^*$ denote the difference in sample means for the continuous and discrete outcomes respectively. We assume $\delta_k = \mu_{Tk} - \mu_{Ck}, \delta_k^* = \mu_{Tk}^* - \mu_{Ck}^*, \kappa = n_C/n_T$ and let z_a denote the $(1 - a)100$th standard normal percentile, where a is the prespecified significance level. We define the z score as $Z_k = \dfrac{\bar{Y}_{Tk} - \bar{Y}_{Ck}}{\sigma_k \sqrt{\dfrac{1+\kappa}{\kappa n_T}}}$ and $Z_k^* = \dfrac{\hat{\mu}_{Tk}^* - \hat{\mu}_{Ck}^*}{\sqrt{\dfrac{1+\kappa}{\kappa n_T}}}$ for the observed continuous and latent continuous measures respectively. The test statistic can then be defined as shown below.

$$Z_k^{\dagger} = \begin{cases} Z_k - \dfrac{\delta_k}{\sigma_k}\sqrt{\dfrac{\kappa n_T}{1+\kappa}} = \dfrac{\bar{Y}_{Tk} - \bar{Y}_{Ck} - \delta_k}{\sigma_k\sqrt{\dfrac{1+\kappa}{\kappa n_T}}},\ k = 1, \ldots, k_m \\ Z_k^* - \delta_k^*\sqrt{\dfrac{\kappa n_T}{1+\kappa}} = \dfrac{\hat{\mu}_{Tk}^* - \hat{\mu}_{Ck}^* - \delta_k^*}{\sqrt{\dfrac{1+\kappa}{\kappa n_T}}},\ k = k_{m+1}, \ldots, K \end{cases}, \tag{8}$$

$$z_k^{\dagger} = \begin{cases} z_\alpha - \dfrac{\delta_k}{\sigma_k}\sqrt{\dfrac{\kappa n_T}{1+\kappa}},\ k = 1, \ldots, k_m \\ z_\alpha - \delta_k^*\sqrt{\dfrac{\kappa n_T}{1+\kappa}},\ k = k_{m+1}, \ldots, K \end{cases}. \tag{9}$$

A useful property of $Z^{\dagger} = \left(Z_1^{\dagger}, \ldots Z_K^{\dagger}\right)^T$ is that it is asymptotically multivariate normal under regularity conditions.[11] The power function for the joint co-primary endpoints is as shown in (10) and hence can be approximated by (11).

$$1 - \beta = P\left(\bigcap_{k=1}^{k_m}\left\{Z_k > z_a\right\} \bigcap_{k_m+1}^{K}\left\{z_k^* > z_a\right\} \middle| \delta\right) \simeq P\left(\bigcap_{k=1}^{K}\left\{z_k^{\dagger} > z_k^{\dagger}\right\} \middle| \delta\right), \tag{10}$$

for $\delta = \left(\delta_1, \ldots, \delta_{k_m}, \ldots, \delta_{k_0}, \ldots, \delta_K\right)^T \neq \mathbf{0}$.

$$1 - \beta \simeq P\left(\bigcap_{k=1}^{K}\left\{z_k^{\dagger} > z_k^{\dagger}\right\} \middle| \delta\right) = \Phi_K\left(-z_1^{\dagger}, \ldots, -z_K^{\dagger}; \Gamma\right). \tag{11}$$

Assuming $n_T = n_C = n$ it is possible to rearrange (11) to obtain a sample size formula in terms of n as shown below.[7]

$$n = \frac{\left(C_K + z_\alpha\right)^2}{\delta_K^2}, \tag{12}$$

where the sample size depends on the number of outcomes and C_K is the solution of

$$1 - \beta = \int_{-\infty}^{\gamma_1 C_K + z_\alpha(\gamma_1 - 1)} \cdots \int_{-\infty}^{\gamma_{K-1} C_K + z_\alpha(\gamma_{K-1} - 1)} \int_{-\infty}^{C_K} f\left(z_1, \ldots, z_K^*; \mathbf{0}, \Gamma\right) dz_K^* \ldots dz_1. \tag{13}$$

Alternatively, we can input different values for n in (11) to achieve the required power.

3.2 Multiple primary endpoints

Using the $Z_k^{\dagger}$ and $Z_k^{\dagger}$ defined for co-primary endpoints and assuming $n_T = n_C = n$, we can define the overall power as in (14).

$$1-\beta = P\left(\bigcup_{k=1}^{k_m}\left\{Z_k > z_a\right\}\bigcup_{k_m+1}^{K}\left\{z_k^* > z_a\right\}\middle|\delta\right) \simeq P\left(\bigcup_{k=1}^{K}\left\{z_k^\dagger > z_k^\dagger\right\}\middle|\delta\right). \tag{14}$$

In order to obtain an appropriate power function we rely on the inclusion-exclusion principle as follows.

$$\begin{aligned} P\left(\bigcup_{k=1}^{K}\left\{Z_k^\dagger > z_k^\dagger\right\}\middle|\delta\right) = & \sum_{k=1}^{K} P\left(\left\{Z_k^\dagger > z_k^\dagger\right\}\middle|\delta\right) - \sum_{k<l} P\left(\left\{Z_k^\dagger > z_k^\dagger\right\}\cap\left\{Z_l^\dagger > z_l^\dagger\right\}\middle|\delta\right) \\ + & \sum_{k<l<m} P\left(\left\{Z_k^\dagger > z_k^\dagger\right\}\cap\left\{Z_l^\dagger > z_l^\dagger\right\}\cap\left\{Z_m^\dagger > z_m^\dagger\right\}\middle|\delta\right) \\ + & \cdots + (-1)^{K-1}\sum_{k<\cdots<K} P\left(\bigcap_{k=1}^{K}\left\{Z_k^\dagger > z_k^\dagger\right\}\middle|\delta\right). \end{aligned}$$

A closed form expression for the overall power is shown in (15)

$$P\left(\bigcup_{k=1}^{K}\left\{Z_k^\dagger > z_k^\dagger\right\}\middle|\delta\right) = \sum_{i=1}^{K}\left((-1)^{i-1}\sum_{I\subseteq\{1,\ldots,K\}} P\left(\bigcap_{k\in I}\left\{Z_k^\dagger > z_k^\dagger\right\}\middle|\delta\right)\right). \tag{15}$$

We then input different values for n to achieve the required power. Note when using the union-intersection test for multiple primary endpoints that a correction must be applied to control the family-wise error rate (FWER). Approaches used for multiple primary continuous endpoints, such as Bonferroni and Holm corrections, may also be implemented in this setting.

3.3 Composite endpoints

As the endpoint of interest is specified in terms of the overall one dimensional composite endpoint, we can use the formula assumed when employing the standard test of proportions technique. As $\sigma_\delta = \sqrt{\frac{\sigma_{\delta T}^2}{n_T} + \frac{\sigma_{\delta C}^2}{n_C}}$, we can assume that $\sigma_T = \sigma_C = \sigma$ and $n_T = n_C = n$, so that $\delta \sim N\left(\delta_T - \delta_C, 2\sigma^2/n\right)$. The power is deduced in the standard way, as demonstrated below.

$$\begin{aligned} 1-\beta &= P\left(\bar{P}_T - \bar{P}_C > z_\alpha\sqrt{2\sigma^2/n}\middle|H_1\right) \\ &= P\left(z > \frac{z_\alpha\sqrt{2\sigma^2/n} - \delta^*}{\sqrt{2\sigma^2/n}}\middle|H_1\right) \\ &= \Phi\left(\frac{\delta^*}{\sqrt{2\sigma^2/n}} - z_a\right). \end{aligned} \tag{16}$$

Note that $\sigma_\sigma^2 = \frac{2\sigma^2}{n}$, however to obtain a formula in terms of the required sample size we will need to separate n from the variance estimate. By fitting the model to pilot trial data we can obtain an estimate for σ^2, as the value of n will be known in this instance and n can be obtained using (17).

Europe PMC Funders Author Manuscripts

$$n = \frac{2\,\hat{\sigma}^2\left(z_{1-\beta} + z_{\alpha}\right)^2}{\hat{\sigma}^{*2}}. \tag{17}$$

This is similar to the sample size equation used for the binary method, however σ is not derived in the standard way and δ^* is obtained using latent means as opposed to provided directly.

4 Numerical Application

4.1 Muse trial

We illustrate the technique for sample size determination using the MUSE trial.[39] The trial was a phase IIb, randomized, double-blind, placebo-controlled study investigating the efficacy and safety of anifrolumab in adults with moderate to severe systemic lupus erythematosus (SLE). Patients (n=305) were randomized (1:1:1) to receive anifrolumab (300 or 1000 mg) or placebo, in addition to standard therapy every 4 weeks for 48 weeks. The primary endpoint in the study was the percentage of patients achieving an SLE Responder Index (SRI) response at week 24, with sustained reduction of oral corticosteroids (<10 mg/day and less than or equal to the dose at week 1 from week 12 through 24). SRI is comprised of a continuous Physician's Global Assessment (PGA) measure, a continuous SLE Disease Activity Index (SLEDAI) measure and an ordinal British Isles Lupus Assessment Group (BILAG) measure.[42] The study had a target sample size of 100 patients per group based on providing 88% power at the two-sided 0.10 alpha level, to detect at least 20% absolute improvement in SRI(4) response rate at week 24 for anifrolumab relative to placebo. The investigators assumed a 40% placebo response rate.

4.2 Model

In this case (Y_1, Y_2, Y_3, Y_4) are SLEDAI, PGA, BILAG and the corticosteroid tapering indicator respectively and $(Y_1, Y_2, Y_3^*, Y_4^*) \sim N_4(\boldsymbol{\mu}^*, \Sigma)$ where,

$$\boldsymbol{\mu}^* = \left(\mu_1, \mu_2, \mu_3^*, \mu_4^*\right)^T \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1 & \rho_{14}\sigma_1 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2 & \rho_{24}\sigma_2 \\ \rho_{13}\sigma_1 & \rho_{23}\sigma_2 & 1 & \rho_{34} \\ \rho_{14}\sigma_1 & \rho_{24}\sigma_2 & \rho_{34} & 1 \end{pmatrix}, \tag{18}$$

and the ordinal and binary components may be related to their latent variables as shown in (19). The thresholds (τ_{31}, τ_{32}, τ_{33}, τ_{34}) are estimated from the data.

Europe PMC Funders Author Manuscripts

$$Y_{i3} = \begin{cases} 0 \text{ if } -\infty < Y_{i3}^* < \tau_{31}, \\ 1 \text{ if } \tau_{31} \le Y_{i3}^* < \tau_{32}, \\ 2 \text{ if } \tau_{32} \le Y_{i3}^* < \tau_{33}, \\ 3 \text{ if } \tau_{33} \le Y_{i3}^* < \tau_{34}, \\ 4 \text{ if } \tau_{34} \le Y_{i3}^* < \infty, \end{cases} \quad Y_{i4} = \begin{cases} 0 \text{ if } -\infty < Y_{i4}^* < 0, \\ 1 \text{ if } 0 \le Y_{i4}^* < \infty \end{cases} \tag{19}$$

We can use the MUSE trial to design future studies where we assume that the endpoints of interest are co-primary, multiple primary and composite endpoints. The overall power functions for each are shown below.

$$\begin{aligned} Power_{co} &= \Phi_4\left(-z_1^\dagger, -z_2^\dagger, -z_3^\dagger, -z_4^\dagger; \Sigma\right) \\ Power_{mult} &= \sum_{i=1}^{4}\left((-1)^{i-1} \sum_{I \subseteq \{1,2,3,4\}} \Phi_{k \in I}\left(-z_k^\dagger; \Sigma\right)\right) \\ Power_{comp} &= \Phi(-z), \end{aligned}$$

where $z_k^\dagger = Z_\alpha - \frac{\delta_k}{\sqrt{2\sigma_k^2/n}}$ for $k = \{1, 2\}$ and $z_k^\dagger = Z_\alpha - \frac{\delta_k^*}{\sqrt{2/n}}$ for $k = \{3, 4\}$. In the composite setting $z = \frac{\delta^*}{\sqrt{2\,\sigma^2/n}} - z_\alpha$ where σ is estimated using the delta method. For the $Power_{mult}$ calculation we apply the Bonferroni correction, such that each outcome is assessed at the $\frac{\alpha}{4}$ level.

4.3 Computation

We have conducted the computations in R version 4.0.2. We define functions to evaluate the power for each of the endpoints using a combination of the pnorm and pmvnorm functions. Sample size is obtained by inserting values for n until the desired power is achieved. Details of our source code and a web app for implementation is included in the Software section. Code to obtain the results shown in this article can be obtained at https://github.com/martinamcm/mcmenamin_2021_multsamp. Considerations and instructions for fitting the latent variable model are discussed in detail McMenamin et al. (2021).[38]

4.4 Results

The power is largest for the multiple primary endpoint, where 80% is achieved for n=37 patients in each arm. The power for the composite endpoint is similar to that of PGA, the component with the highest effect size. As we would expect the power is considerably lower for co-primary endpoints, which would require n=325 for 80% power (Figure 1).

Table 1 shows the sample sizes required in each group, for the co-primary and multiple primary endpoints to obtain an overall power of at least 80% to detect a difference of 0.88 in SLEDAI, 0.38 in PGA, 0.24 in BILAG and 0.40 in the taper outcome based on the values observed in the trial. We allow for uncertainty in the variance of the continuous measures

by setting $\sigma_1^2 = 18, 19, 20$ and $\sigma_2^2 = 0.35, 0.45, 0.55, 0.65$. The sample sizes required for each individual endpoint are also shown, based on achieving a power of at least 80%. Allowing for uncertainty in the variance of the SLEDAI outcome varies the required sample size for the co-primary endpoint but not the multiple primary endpoint. The opposite is true when the assumed variance of the PGA outcome is changed, namely affecting the sample size required for the multiple primary endpoint but not the co-primary. This is intuitive given that the treatment effect observed in the SLEDAI outcome is smallest and is largest for the PGA outcome. For the co-primary and composite endpoints the power is largest when the correlation between the endpoints is high whereas for multiple primary endpoints the power is largest for zero correlation between endpoints (Figure 2).

We assume that a future trial in SLE is to be conducted using the composite responder endpoint, allowing for uncertainty in σ. The estimated variance for the risk difference from the trial dataset is $\sigma_\delta^2 = 0.048$ with correlation parameters $\rho_{12} = 0.448,\ \rho_{13} = 0.521,\ \rho_{14} = 0.003,\ \rho_{23} = 0.448,\ \rho_{24} = -0.031,\ \rho_{34} = 0.066$. For a risk difference of 0.14, the required sample size per group is 50, compared to 135 for 88% power in the standard binary method. If the method were to be employed for increased power, rather than a decrease in required sample size, the estimated power of the latent variable method is over 99.99% for sample sizes giving 88% power at the 0.05 one-sided alpha level in the binary method. The empirical power is shown for the latent variable method in 1000 simulated datasets, which is approximately 88% for each sample size, as required. Note that the sample size for composite endpoints are highly dependent on the responder threshold chosen, which will be predefined by clinicians.

5 Empirical Performance Of Sample Sizes

The behavior of the sample sizes obtained for each of the endpoints can be shown empirically. Assuming the four dimensional SLE endpoint, we calculate the empirical power by simulating 100 000 datasets from the multivariate normal distribution and applying the corresponding tests for both the observed and latent continuous outcomes. The key concern for the co-primary endpoints is that the method gives the appropriate power whereas for multiple primary endpoints we must ensure the family-wise error rate is controlled.

The sample sizes required for each of the three endpoints and the corresponding empirical power is shown for effect sizes observed in the MUSE trial with low, medium, and high correlation assumed between endpoints (Table 2). The empirical power derived is approximately equal to the desired power of 80% for all endpoints. As is well recognized in the multiple testing literature, the type I error rate must be controlled when multiple primary endpoints are tested using the union-intersection test. The degree to which the type I error rate is inflated depends on the number of outcomes and the correlation between outcomes, where lower correlation between outcomes and larger number of outcomes result in larger inflations (Figure 3). The performance of the Bonferroni correction in this setting is shown, where it is conservative in the case of high correlation between endpoints. As the maximum correlation between outcomes in the MUSE trial endpoint used in the numerical example is 0.5, we expect the sample sizes shown for this application to be a good estimate. If very

large positive correlations between the endpoints are expected the required sample size from this approach may be overestimated. The code to obtain these empirical results is provided in the ‘Software’ section.

6 Discussion

The work in this article demonstrated the various ways in which a latent variable framework may be employed for mixed continuous, ordinal, and binary outcomes. We illustrated sample size determination in the case of mixed continuous, ordinal, and binary co-primary outcomes. We extended this to allow for sample size determination in the case of mixed multiple primary endpoints and proposed a technique to estimate the sample size when using a latent variable model for the underlying structure of a mixed composite endpoint. For co-primary and multiple primary endpoints the resulting hypothesis is based on an intersection or union of the hypotheses for the individual outcomes and so is multivariate in nature. However, for composite responder endpoints the hypothesis of interest is stated in relation to the overall responder endpoint and so is univariate. Sample size estimation in this case can make use of the standard power and sample size functions but requires the distribution of the test statistic under the alternative hypothesis which we approximate using latent-level means and a Taylor series expansion.

We applied the methods to a numerical example based on a phase IIb study. For the correlation structure observed in the MUSE trial, the sample size required for the co-primary endpoint was greater than that required for the individual endpoint with the lowest effect size. Alternatively, the sample size required for the multiple primary endpoint changes based on the variance assumed for the outcome with the largest treatment effect, however is similar to that required by the individual endpoint. The sample size required for the composite endpoint was between that required for the individual outcome with the largest and second largest effect size. Given that in the composite case we are concerned with the overall binary response endpoint, we compared the sample sizes required for the endpoint using the latent variable model with the standard binary method which we showed offered a large gain in efficiency. Results of the simulated scenarios agree with previous findings that the inclusion of the ordinal component with five levels is only responsible for a very small proportion of the precision gains. Given that the inclusion of the ordinal component substantially increases complexity and computational demand, it may be sufficient to combine any ordinal components with the binary outcome if necessary. Detailed simulation results for the composite endpoint are shown in the Supplementary Material.

One practical consideration when calculating the sample size for a trial using the latent variable model is the need to specify a large number of parameters, even in the case of only a few outcomes. Estimates for the parameters could be obtained by fitting the model to pilot data however this is potentially challenging and restrictive for a number of reasons. First, it requires that a pilot or earlier phase trial must have already taken place. Furthermore, the pilot data could be fundamentally different to the future trial and observed effects may be imprecise. Therefore, placing too much emphasis on the existing data may lead to problems in the main trial. In theory, it is possible to specify the required covariance parameters without data however this would be difficult in practice. Additionally, in the case

Europe PMC Funders Author Manuscripts

of composite endpoints, we cannot define the variance in terms of the model parameters only, as the treatment effect is defined for the one-dimensional composite and so is a function of the parameters. This means that the full covariance matrix of the estimated parameters is required for the Taylor series derivation. An alternative when there is no data available is to apply the method using the sample size required to achieve 80% power for the binary method and avail of the large increase in power. Alternatively, we can directly specify σ_δ based on expert elicitation, as is sometimes the case in practice for standard one-dimensional endpoints. Allowing for uncertainty in the quantities and choosing conservative values should provide an appropriate sample size estimate.

It is possible to extend this approach to use adaptive sample size re-estimation, or an internal pilot to allow for reductions in the required sample size in the trial as we collect more information about the treatment effect variability.

Software

The code to obtain the results in this article is available at https://github.com/martinamcm/mcmenamin_2021_multsamp. A Shiny application for implementing the method is available at https://martinamcm.shinyapps.io/multsampsize/. Documentation and example data are available at https://github.com/martinamcm/MultSampSize.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014) and the Medical Research Council (MC_UU_00002/5). The funding bodies did not have any role in the design or analysis of the study, interpretation of data or writing the manuscript.

Funding information

Medical Research Council, Grant/Award Number: MC_UU_00002/5; NIHR Cambridge Biomedical Research Centre, Grant/Award Number: BRC-1215-20014

References

1. Altman D. Statistics and ethics in medical research: III how large a sample? Br Med J. 1980; 281: 1336–1338. [PubMed: 7437789]
2. Moher D, Dulberg C, Wells G. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA. 1994; 272: 122–124. [PubMed: 8015121]
3. Wittes J. Sample size calculations for randomized controlled trials. Epidemiol Rev. 2002; 24: 39–53. [PubMed: 12119854]
4. 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. Eur Heart J. 2012; 33 (2) 176–182. [PubMed: 21900289]
5. Ashford J, Sowden R. Multivariate probit analysis. Biometrics. 1970; 26: 535–546. DOI: 10.2307/2529107 [PubMed: 5480663]
6. Chib S, Greenberg E. Analysis of multivariate probit models. Biometrika. 1998; 85 (2) 347–361. DOI: 10.1093/biomet/85.2.347

7. Sozu, T, Sugimoto, T, Hamasaki, T, Evans, S. Sample Size Determination in Clinical Trials with Multiple Endpoints. New York, NY: Springer; 2015.
8. Sozu T, Kanou T, Hamada C, Yoshimura I. Power and sample size calculations in clinical trials with multiple primary variables. Jpn J Biometr. 2006; 27: 83–96.
9. Xiong C, Yu K, Gao F, Yan Y, Zhang Z. Power and sample size for clinical trials when efficacy is required in multiple endpoints: application to Alzheimer's treatment trial. Clin Trials. 2005; 2: 387–393. DOI: 10.1191/1740774505cn112oa [PubMed: 16317808]
10. Sozu T, Sugimoto T, Hamasaki T. Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. J Biopharm Stat. 2011; 21: 650–668. DOI: 10.1002/sim.3972 [PubMed: 21516562]
11. Sugimoto T, Sozu T, Hamasaki T. A convenient formula for sample size calculations in clinical trials with multiple co-primary continuous endpoints. Pharm Stat. 2012; 11: 118–128. DOI: 10.1002/pst.505 [PubMed: 22415870]
12. Eaton M, Muirhead R. On a multiple endpoints testing problem. J Stat Plan Infer. 2007; 137: 3416–3429. DOI: 10.1016/j.jspi.2007.03.021
13. Julious S, McIntyre N. Sample sizes for trials involving multiple correlated must-win comparisons. Pharm Stat. 2012; 11: 177–185. DOI: 10.1002/pst.515 [PubMed: 22383136]
14. Senn S. Disappointing dichotomies. Pharm Stat. 2003; 2: 239–240. DOI: 10.1002/pst.90
15. Chuang-Stein C, Stryszak P, Dmitrienko A, Offen W. Challenge of multiple co-primary endpoints: a new approach. Stat Med. 2007; 26: 1181–1192. DOI: 10.1002/sim.2604 [PubMed: 16927251]
16. Kordzakhia G, Siddiqui O, Huque M. Method of balanced adjustment in testing co-primary endpoints. Stat Med. 2010; 29: 2055–2066. DOI: 10.1002/sim.3950 [PubMed: 20683896]
17. Hung H, Wang S. Some controversial multiple testing problems in regulatory applications. J Biopharm Stat. 2009; 19: 1–11. DOI: 10.1080/10543400802541693 [PubMed: 19127460]
18. Dmitrienko, A, Tamhane, A, Bretz, F. Multiple Testing Problems in Pharmaceutical Statistics. Boca Raton, FL: Chapman & Hall/CRC Press; 2010.
19. Gong J, Pinheiro J, DeMets D. Estimating significance level and power comparisons for testing multiple endpoints in clinical trials. Control Clin Trials. 2000; 21: 323–329. DOI: 10.1016/S0197-2456(00)00049-0
20. Sander A, Rauch G, Kieser M. Blinded sample size recalculation in clinical trials with binary composite endpoints. J Biopharm Stat. 2017; 27 (4) 705–715. [PubMed: 27295402]
21. Rauch G, Kieser M. Multiplicity adjustment for composite binary endpoints. Methods Inf Med. 2012; 51 (4) 309–317. [PubMed: 22525969]
22. Ruse M, Ritz C, Hothorn L. Simultaneous inference of a binary composite endpoint and its components. J Biopharm Stat. 2016; 3406: 1–14.
23. Marsal J, Ferreira-González I, GPMDGGG SB. The use of a binary composite endpoint and sample size requirement: influence of endpoints overlap. Am J Epidemiol. 2017; 185 (9) 832–841. [PubMed: 28402501]
24. Roig MB, Gómez 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 (11) 1935–1956. [PubMed: 30637797]
25. Roig MB, Gómez G. Selection of composite binary endpoints in clinical trials. Biometr J. 2018; 60 (2) 246–261.
26. Sozu T, Sugimoto T, Hamasaki T. Sample size determination in clinical trials with multiple co-primary binary endpoints. Stat Med. 2010; 29: 2169–2179. DOI: 10.1002/sim.3972 [PubMed: 20687162]
27. Hamasaki, T, Evans, S, Sugimoto, T, Sozu, T. Power and sample size determination for clinical trials with two correlated binary relative risks. Technical report. ENAR Spring Meeting; Washington DC: 2012.
28. Song J. Sample size for simulataneous testing of rate differences in non-inferiority trials with multiple endpoints. Comput Stat Data Anal. 2009; 53: 1201–1207. DOI: 10.1016/j.csda.2008.10.028

29. Hamasaki T, Sugimoto T, Evans S, Sozu T. Sample size determination for clinical trials with co-primary outcomes: exponential event-times. Pharma Stat. 2013; 12: 28–34. DOI: 10.1002/pst.1545
30. Sugimoto, T; Hamasaki, T; Sozu, T. Sample size determination in clinical trials with two correlated co-primary time-to-event endpoints; Paper presented at: Proceedings of the 7th International Conference on Multiple Comparison Procedures; Washington DC. 2011.
31. Sugimoto T, Sozu T, Hamasaki T, Evans S. A logrank test-based method for sizing clinical trials with two co-primary time-to-events endpoints. Biostatistics. 2013; 14: 409–421. DOI: 10.1093/biostatistics/kxs057 [PubMed: 23307913]
32. Sugimoto, T; Hamasaki, T; Sozu, T; Evans, S. Sample size determination in clinical trials with two correlated time-to-event endpoints as primary contrast; Proceedings of the 6th FDA-DIA Forum; Washington, DC. 2012.
33. Sugimoto T, Hamasaki T, Evans S, Sozu T. Sizing clinical trials when comparing bivariate time-to-event outcomes. Stat Med. 2017; 36 (9) 1363–1382. [PubMed: 28120524]
34. Ferreira-González I, Permanyer-Miralda G, Domingo-Salvany A, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. BMJ. 2007; 334 (7597) 786. [PubMed: 17403713]
35. Gómez G, Lagakos S. Statistical considerations when using a composite endpoint for comparing treatment groups. Stat Med. 2013; 32: 719–738. [PubMed: 22855368]
36. Sozu T, Sugimoto T, Hamasaki T. Sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables. Biometr J. 2012; 54: 716–729. DOI: 10.1002/bimj.201100221
37. Wu B, de Leon A. Letter to the Editor re: ɛSample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variablesɛ Sozu T, Sugimoto T, Hamasaki T. Biometr J. 2013; 55 (5) 807–812. DOI: 10.1002/bimj.201200254
38. McMenamin M, Barrett JK, Berglind A, Wason JM. Employing a latent variable framework to improve efficiency in composite endpoint analysis. Stat Methods Med Res. 2021; 30 (3) 702–716. [PubMed: 33234028]
39. Furie R, Khamashta M, Merrill J, et al. Anifrolumab, an anti interferon alpha receptor monoclonal antibody, in moderate-to-severe systemic lupus erythematosus. Arthritis Rheumatol. 2017; 69 (2) 376–386. DOI: 10.1002/art.39962 [PubMed: 28130918]
40. Tate R. The theory of correlation between two continuous variables when one is dichotomized. Biometrika. 1955; 2 (1-2) 205–216. DOI: 10.2307/2333437
41. Wason J, McMenamin M, Dodd S. Analysis of responder-based endpoints: improving power through utilising continuous components. Trials. 2020; 21: 427. [PubMed: 32450909]
42. Luijten K, Tekstra J, Bijlsma J, Bijl M. The systemic lupus erythematosus responder index (SRI): a new SLE disease activity assessment. Autoimmun Rev. 2012; 11 (5) 326–329. DOI: 10.1016/j.autrev.2011.06.011 [PubMed: 21958603]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

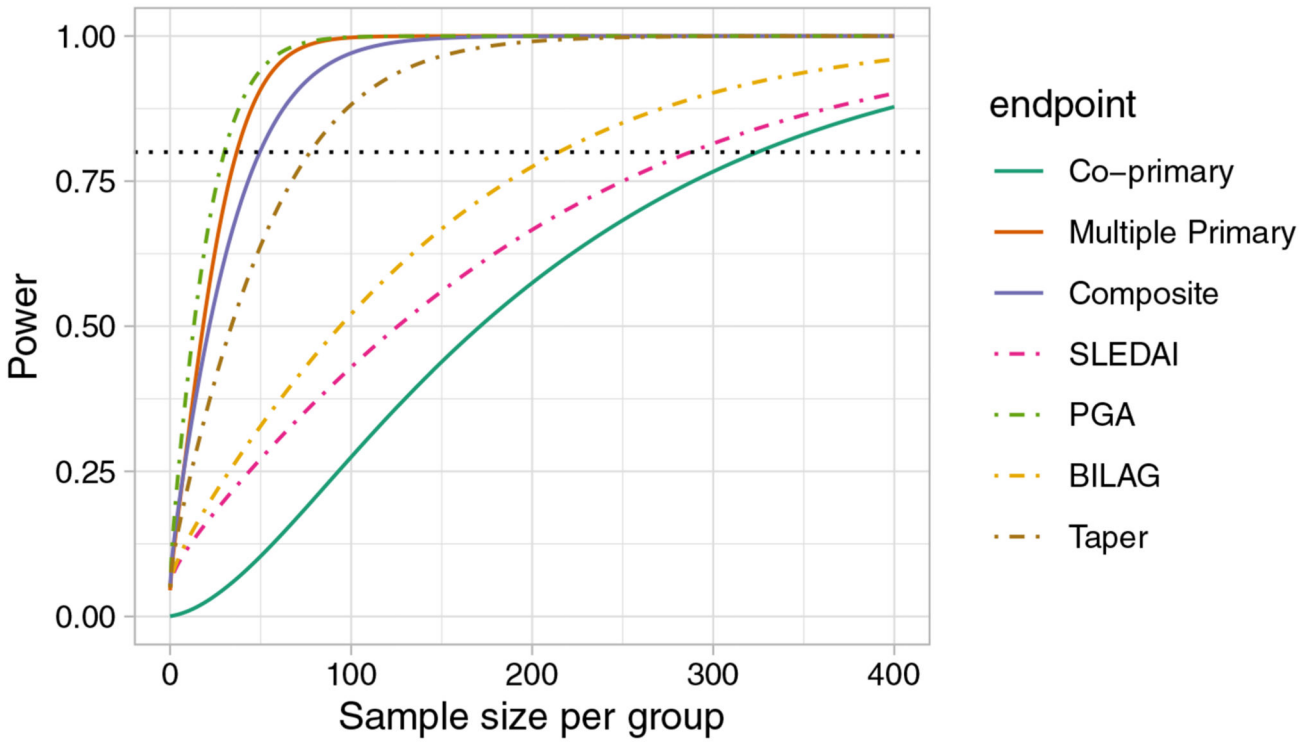

Figure 1.

Power function for individual SLEDAI (continuous), PGA (continuous), BILAG (ordinal), and Taper (binary) outcomes and the power functions with when they are treated as co-primary, multiple primary, and composite endpoints using data from the MUSE trial

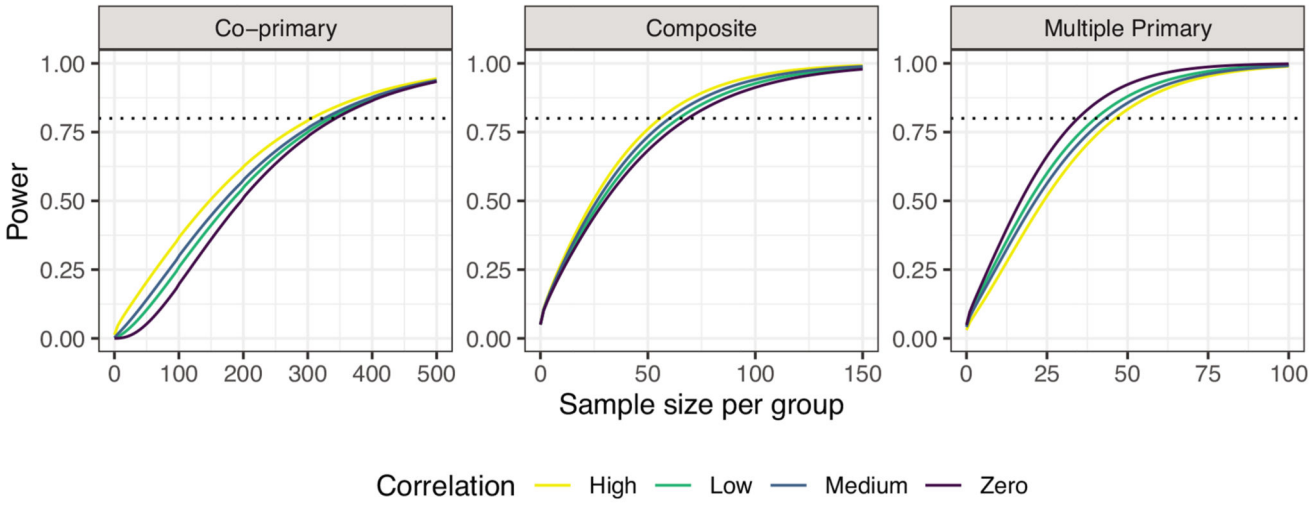

Figure 2.

Overall power $1 - \beta$ to detect the treatment effects assumed from the MUSE trial for the systemic lupus erythematosus co-primary, multiple primary, and composite endpoints for different sample sizes per group $n = n_C = n_T$ and differing correlations between outcomes, where Low = 0.3, Medium = 0.5, and High = 0.8

Europe PMC Funders Author Manuscripts

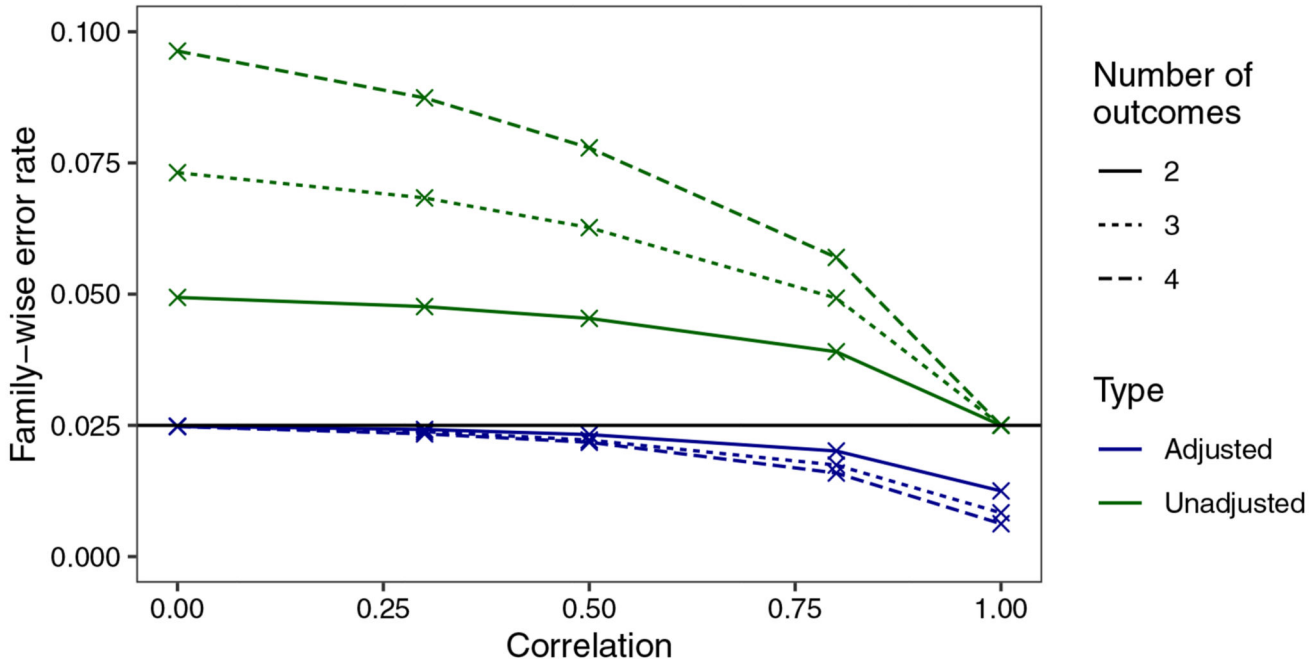

Figure 3.

Family-wise error rate (FWER) of the multiple primary endpoints shown both unadjusted and adjusted using the Bonferroni correction. FWERs are shown for $K = (2, 3, 4)$ outcomes and correlations are constrained to be equal between all outcomes

Europe PMC Funders Author Manuscripts

Table 1

Sample sizes $n = n_C = n_T$ for the co-primary and multiple primary endpoints for overall power $1 - \beta \approx 0.80$, $\alpha = 0.025$, $k_m = 2$, $K = 4$ using the MUSE trial data

SLEDAI		PGA		BILAG		Taper							
δ_1	σ_1^2	δ_2	σ_2^2	(π_{T3},π_{C3})	σ_3^*	(π_{T4},π_{C4})	σ_4^*	n_{co}	n_{mult}	SS_1	SS_2	SS_3	SS_4
0.88	18	0.38	0.35	(0.97,0.95)	0.24	(0.54,0.38)	0.40	403	46	365	39	273	99
0.88	19	0.38	0.35	(0.97,0.95)	0.24	(0.54,0.38)	0.40	419	46	386	39	273	99
0.88	20	0.38	0.35	(0.97,0.95)	0.24	(0.54,0.38)	0.40	435	46	406	39	273	99
0.88	18	0.38	0.45	(0.97,0.95)	0.24	(0.54,0.38)	0.40	403	55	365	49	273	99
0.88	18	0.38	0.55	(0.97,0.95)	0.24	(0.54,0.38)	0.40	403	63	365	60	273	99
0.88	18	0.38	0.65	(0.97,0.95)	0.24	(0.54,0.38)	0.40	403	70	365	71	273	99

Note: SS_1, SS_2, SS_3, SS_4 are sample sizes required per group for the individual endpoints for a power of at least $1 - \beta = 0.80$.

Europe PMC Funders Author Manuscripts

Table 2

Sample sizes and empirical power (%) for $n = n_C = n_T$ for the co-primary, multiple primary, and composite endpoints for overall power $1 - \beta \approx 0.80$, $\alpha = 0.025$, $k_m = 2$, $K = 4$ with observed and latent effect sizes δ_1 , δ_2 , δ_3^* , δ_4^* and correlation ρ equal to 0.3, 0.5, 0.8 where correlations are assumed to be equal between all endpoints

δ_1	δ_2	δ_3^*	δ_4^*	ρ	Co-primary	Multiple primary	Composite
0.12	0.12	0.12	0.12	0.0	1766 (80.1)	591 (80.0)	1031 (80.1)
				0.3	1692 (80.0)	744 (80.1)	883 (80.0)
				0.5	1617 (80.0)	867 (80.1)	687 (80.0)
				0.8	1439 (80.0)	1117 (80.0)	589 (79.9)
0.35	0.35	0.15	0.15	0.0	917 (79.9)	105 (80.1)	201 (79.9)
				0.3	894 (80.0)	122 (79.9)	156 (80.0)
				0.5	870 (80.0)	134 (80.0)	115 (80.1)
				0.8	815 (80.1)	153 (80.0)	92 (80.0)
0.12	0.35	0.55	0.10	0.0	1772 (80.2)	61 (80.3)	81 (80.0)
				0.3	1736 (80.2)	67 (80.5)	72 (80.1)
				0.5	1700 (80.0)	70 (79.9)	67 (80.2)
				0.8	1625 (80.2)	74 (80.6)	58 (80.1)