Decision-making with multiple correlated binary outcomes in clinical trials

Xynthia Kavelaars[1] , **Joris Mulder[1,2]** and **Maurits Kaptein[2]**

Abstract

Clinical trials often evaluate multiple outcome variables to form a comprehensive picture of the effects of a new treatment. The resulting multidimensional insight contributes to clinically relevant and efficient decision-making about treatment superiority. Common statistical procedures to make these superiority decisions with multiple outcomes have two important shortcomings, however: (1) Outcome variables are often modeled individually, and consequently fail to consider the relation between outcomes; and (2) superiority is often defined as a relevant difference on a single, on any, or on all outcome(s); and lacks a compensatory mechanism that allows large positive effects on one or multiple outcome (s) to outweigh small negative effects on other outcomes. To address these shortcomings, this paper proposes (1) a Bayesian model for the analysis of correlated binary outcomes based on the multivariate Bernoulli distribution; and (2) a flexible decision criterion with a compensatory mechanism that captures the relative importance of the outcomes. A simulation study demonstrates that efficient and unbiased decisions can be made while Type I error rates are properly controlled. The performance of the framework is illustrated for (1) fixed, group sequential, and adaptive designs; and (2) non-informative and informative prior distributions.

Keywords

Multiple outcomes, compensatory decision rules, multivariate Bernoulli model, efficiency, Bayesian analysis

1 Introduction

Clinical trials often aim to compare the effects of two treatments. To ensure clinical relevance of these comparisons, trials are typically designed to form a comprehensive picture of the treatments by including multiple outcome variables. Collected data about efficacy (e.g. reduction of disease symptoms), safety (e.g. side effects), and other relevant aspects of new treatments are combined into a single, coherent decision regarding treatment superiority. An example of a trial with multiple outcomes is the CAR-B (Cognitive Outcome after WBRT or SRS in Patients with Brain Metastases) study, which investigated an experimental treatment for cancer patients with multiple metastatic brain tumors.[1] Historically, these patients have been treated with radiation of the whole brain (Whole Brain Radiation Therapy; WBRT). This treatment is known to damage healthy brain tissue and to increase the risk of (cognitive) side effects. More recently, local radiation of the individual metastases (stereotactic surgery; SRS) has been proposed as a promising alternative that saves healthy brain tissue and could therefore reduce side effects. The CAR-B study compared these two treatments based on cognitive functioning, fatigue, and several other outcome variables.[1]

[1]Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands
[2]Jheronimus Academy of Data Science, 's-Hertogenbosch, The Netherlands

Corresponding author:
Xynthia Kavelaars, Department of Methodology and Statistics, Tilburg University, PO Box 90153, Tilburg 5000LE, The Netherlands.
Email: x.m.kavelaars@tilburguniversity.edu

Statistical procedures to arrive at a superiority decision have two components: (1) A statistical model for the collected data; and (2) a decision rule to evaluate the treatment in terms of superiority based on the modelled data. Ideally, the combination of these components forms a decision procedure that satisfies two criteria: Decisions should be clinically relevant and efficient. Clinical relevance ensures that the statistical decision rule corresponds to a meaningful superiority definition, given the clinical context of the treatment. Commonly used decision rules define superiority as one or multiple treatment difference(s) on the most important outcome, on any of the outcomes, or on all of the outcomes.[2–5] Efficiency refers to achieving acceptable error rates while minimizing the number of patients in the trial. The emphasis on efficiency is motivated by several considerations, such as small patient populations, ethical concerns, limited access to participants, and other difficulties to enroll a sufficient number of participants.[6] In the current paper, we address clinical relevance and efficiency in the context of multiple binary outcomes and propose a framework for statistical decision-making.

In trials with multiple outcomes, it is common to use a univariate modeling procedure for each individual outcome and combine these with one of the aforementioned decision rules.[2,3] Such decision procedures can be inefficient since they ignore the relationships between outcomes. Incorporating these relations in the modeling procedure is crucial as they directly influence the amount of evidence for a treatment difference as well as the sample size required to achieve satisfactory error rates. A multivariate modeling procedure takes relations between outcomes into account and can therefore be a more efficient and accurate alternative when outcomes are correlated.

Another interesting feature of multivariate models is that they facilitate the use of decision rules that combine multiple outcomes in a flexible way, for example via a compensatory mechanism. Such a mechanism is characterized by the property that beneficial effects are given the opportunity to compensate adverse effects. The flexibility of compensatory decision-making is appealing, since a compensatory mechanism can be naturally extended with impact weights that explicitly take the clinical importance of individual outcome variables into account.[3] With impact weights, outcome variables of different importances can be combined into a single decision in a straightforward way.

Compensatory rules do not only contribute to clinical relevance, but also have the potential to increase trial efficiency. Effects on individual outcomes may be small (and seemingly unimportant) while the combined treatment effect may be large (and important),[7–9] as visualized in Figure 1 for fictive data of the CAR-B study. The two displayed bivariate distributions reflect the effects and their uncertainties on cognitive functioning and fatigue for SRS and WBRT. The univariate distributions of both outcomes overlap too much to clearly distinguish the two treatments on individual outcome variables or a combination of them. The bivariate distributions, however, clearly distinguish between the two treatments. Consequently, modeling a compensatory treatment effect with equal weights (visualized as the diagonal dashed line) would provide sufficient evidence to consider SRS superior in the presented situation.
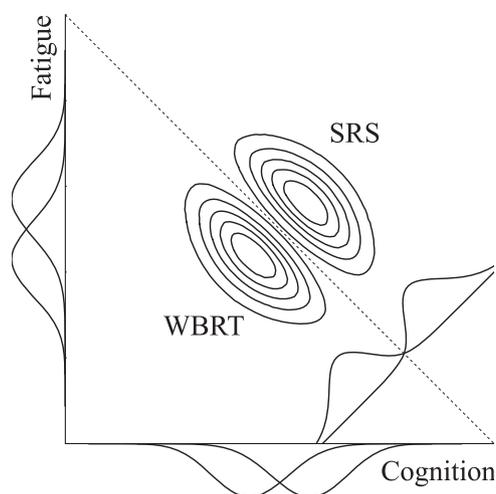

Figure 1. Separation of two bivariate distributions (diagonally) versus separation of their univariate distributions (horizontally/vertically) for the CAR-B study. The dashed diagonal line represents a Compensatory decision rule with equal weights. Each distribution reflects the plausibility of the treatment effects on cognitive functioning and fatigue after observing fictive data..

In the current paper, we propose a decision procedure for multivariate decision-making with multiple (correlated) binary outcomes. The procedure consists of two components. First, we model the data with a multivariate Bernoulli distribution, which is a multivariate generalization of the univariate Bernoulli distribution. The model is exact and does not rely on numerical approximations, making it appropriate for small samples. Second, we extend multivariate analysis with a compensatory decision rule to include more comprehensive and flexible definitions of superiority.

The decision procedure is based on a Bayesian multivariate Bernoulli model with a conjugate prior distribution. The motivation for this model is twofold. First, the multivariate Bernoulli model is a natural generalization of the univariate Bernoulli model, which intuitively parametrizes success probabilities per outcome variable. Second, a conjugate prior distribution can greatly facilitate computational procedures for inference. Conjugacy ensures that the form of the posterior distribution is known, making sampling from the posterior distribution straightforward.

Although Bayesian analysis is well known to allow for inclusion of information external to the trial by means of prior information,[10] researchers who wish not to include prior information can obtain results similar to frequentist analysis. The use of a non-informative prior distribution essentially results in a decision based on the likelihood of the data, such that (1) Bayesian and frequentist (point) estimates are equivalent; and (2) the frequentist p-value equals the Bayesian posterior probability of the null hypothesis in one-sided testing.[11] Since a (combined) p-value may be difficult to compute for the multivariate Bernoulli model, Bayesian computational procedures can exploit this equivalence and facilitate computations involved in Type I error control.[12,13]

The remainder of the paper is structured as follows. In section 2, we present a multivariate approach to the analysis of multiple binary outcomes. Subsequently, we discuss various decision rules to evaluate treatment differences on multiple outcomes in section 3. The framework is evaluated in section 4, and we discuss limitations and extensions in the section 5.

2 A model for multivariate analysis of multiple binary outcomes

2.1 Notation

We start the introduction of our framework with some notation. The joint response for patient i in treatment j on K outcomes will be denoted by $x_{j,i} = (x_{j,i,1}, \ldots, x_{j,i,K})$, where $i \in \{1, \ldots, n_j\}$ and $j \in \{E, C\}$ (i.e. Experimental and Control). The response on outcome k $x_{j,i,k} \in \{0, 1\}$ (0 = failure, 1 = success), such that $x_{j,i}$ can take on $Q = 2^K$ different combinations $\{1 \ldots 11\}, \{1 \ldots 10\}, \ldots, \{0 \ldots 01\}, \{0 \ldots 00\}$. The observed frequencies of each possible response combination for treatment j in a dataset of n_j patients are denoted by vector s_j of length Q. The elements of s_j add up to n_j, $\sum_{q=1}^{Q} s_{j,q} = n_j$.

Vector $\theta_j = (\theta_{j,1}, \ldots, \theta_{j,K})$ reflects success probabilities of K outcomes for treatment j in the population. Vector $\delta = (\delta_1, \ldots, \delta_K)$ then denotes the treatment differences on K outcomes, where $\delta_k = \theta_{E,k} - \theta_{C,k}$. We use $\phi_j = (\phi_{j,1\ldots11}, \phi_{j,1\ldots10}, \ldots, \phi_{j,0\ldots01}, \phi_{j,0\ldots00})$ to refer to probabilities of joint responses in the population, where $\phi_{j,q}$ denotes the probability of joint response combination $x_{j,i}$ with configuration q. Vector ϕ_j has Q elements, and sums to unity, $\sum_{q=1}^{Q} \phi_{j,q} = 1$. Information about the relation between outcomes k and l is reflected by $\phi_{j,kl}$, which is defined as the sum of those elements of ϕ_j that have the kth and lth elements of q equal to 1, e.g. $\phi_{j,11}$ for $K = 2$. Similarly, marginal probability $\theta_{j,k}$ follows from summing all elements of ϕ_j with the kth element of q equal to 1. For example, with three outcomes, the success probability of the first outcome is equal to $\theta_{j,1} = \phi_{j,111} + \phi_{j,110} + \phi_{j,101} + \phi_{j,100}$.

2.2 Likelihood

The likelihood of joint response $x_{j,i}$ follows a K-variate Bernoulli distribution[14]

$$
\begin{aligned}
p(x_{j,i}|\phi_j) &= \text{multivariate Bernoulli}(x_{j,i}|\phi_j) \\
&= \phi_{j,1\ldots11}^{x_{j,1}\times\ldots\times x_{j,K}} \phi_{j,1\ldots10}^{x_{j,1}\times\ldots\times x_{j,K-1}(1-x_{j,K})} \times \cdots \\
&\quad \times \phi_{j,0\ldots01}^{(1-x_{j,1})\times\ldots\times(1-x_{j,K-1})x_{j,K}} \phi_{j,0\ldots00}^{(1-x_{j,1}\times\ldots\times 1-x_{j,K})}
\end{aligned}
\tag{1}
$$

The multivariate Bernoulli distribution in equation (1) is a specific parametrization of the multinomial distribution. The likelihood of n_j joint responses summarized by cell frequencies in s_j follows a Q-variate multinomial distribution with parameters ϕ_j

$$p(s_j|\phi_j) = \text{multinomial}(s_j|\phi_j)$$
$$\propto \phi_{j,1\ldots11}^{s_{j,1\ldots11}}\phi_{j,1\ldots10}^{s_{j,1\ldots10}} \times \cdots \times \phi_{j,0\ldots01}^{s_{j,0\ldots01}}\phi_{j,0\ldots00}^{s_{j,0\ldots00}} \tag{2}$$

Conveniently, the multivariate Bernoulli distribution is consistent under marginalization. That is, marginalizing a $K-$variate Bernoulli distribution with respect to p variables results in a $(K-p)$-variate Bernoulli distribution.[14] Hence, the univariate Bernoulli distribution is directly related and results from marginalizing $(K-1)$ variables.

The pairwise correlation between variables $x_{j,k}$ and $x_{j,l}$ is reflected by $\rho_{x_{j,k},x_{j,l}}$[14]

$$\rho_{x_{j,k}x_{j,l}} = \frac{\theta_{j,kl} - \theta_{j,k}\theta_{j,l}}{\sqrt{\theta_{j,k}(1-\theta_{j,k})\theta_{j,l}(1-\theta_{j,l})}} \tag{3}$$

This correlation is over the full range, i.e. $-1 \leq \rho_{x_{j,k},x_{j,l}} \leq 1$.[15]

2.3 Prior and posterior distribution

A natural choice to model prior information about response probabilities ϕ_j is the Dirichlet distribution, since a Dirichlet prior and multinomial likelihood form a conjugate combination. The Q-variate prior Dirichlet distribution has hyperparameters $\alpha_j^0 = (\alpha_{j,11\ldots11}^0, \alpha_{j,11\ldots10}^0, \ldots, \alpha_{j,00\ldots01}^0, \alpha_{j,00\ldots00}^0)$

$$p(\phi_j) = \text{Dirichlet}\left(\phi_j|\alpha_j^0\right)$$
$$\propto \phi_{j,1\ldots11}^{\alpha_{j,1\ldots11}^0-1}\phi_{j,1\ldots10}^{\alpha_{j,1\ldots10}^0-1} \times \cdots \times \phi_{j,0\ldots01}^{\alpha_{j,0\ldots01}^0-1}\phi_{j,0\ldots00}^{\alpha_{j,0\ldots00}^0-1} \tag{4}$$

where each of the prior hyperparameters α_j^0 should be larger than zero to ensure a proper prior distribution.

The posterior distribution of ϕ_j results from multiplying the likelihood and the prior distribution and follows a Dirichlet distribution with parameters $\alpha_j^n = \alpha_j^0 + s_j$

$$p(\phi_j|s_j) = \text{Dirichlet}\left(\phi_j|\alpha_j^0 + s_j\right)$$
$$\propto \phi_{j,1\ldots11}^{s_{j,1\ldots11}}\phi_{j,1\ldots10}^{s_{j,1\ldots10}} \times \cdots \times \phi_{j,0\ldots01}^{s_{j,0\ldots01}}\phi_{j,0\ldots00}^{s_{j,0\ldots00}}$$
$$\times \phi_{j,1\ldots11}^{\alpha_{j,1\ldots11}^0-1}\phi_{j,1\ldots10}^{\alpha_{j,1\ldots10}^0-1} \times \cdots \times \phi_{j,0\ldots01}^{\alpha_{j,0\ldots01}^0-1}\phi_{j,0\ldots00}^{\alpha_{j,0\ldots00}^0-1}$$
$$\propto \phi_{j,1\ldots11}^{\alpha_{j,1\ldots11}^n-1}\phi_{j,1\ldots10}^{\alpha_{j,1\ldots10}^n-1} \times \cdots \times \phi_{j,0\ldots01}^{\alpha_{j,0\ldots01}^n-1}\phi_{j,0\ldots00}^{\alpha_{j,0\ldots00}^n-1} \tag{5}$$

Since prior hyperparameters α_j^0 impact the posterior distribution of treatment difference δ, specifying them carefully is important. Each of the hyperparameters contains information about one of the observed frequencies s_j and can be considered a prior frequency that reflects the strength of prior beliefs. Equation (5) shows that the influence of prior information depends on prior frequencies α_j^0 relative to observed frequencies s_j. When all elements of α_j^0 are set to zero, $\alpha_j^n = s_j$. This (improper) prior specification results in a posterior mean of $\phi_{j,q}|s_{j,q} = \frac{\alpha_{j,q}^n}{\sum_{p=1}^{Q}\alpha_{j,p}^n}$, which is equivalent to the frequentist maximum likelihood estimate of $\phi_{j,q} = \frac{s_{j,q}}{\sum_{p=1}^{Q}s_{j,p}}$. To take advantage of this property with a proper non-informative prior, one could specify hyperparameters slightly larger than zero such that the posterior distribution is essentially completely based on the information in the data (i.e. $\alpha_j^n \approx s_j$).

To include prior information – when available – in the decision, α_j^0 can be set to specific prior frequencies to increase the influence on the decision. These prior frequencies may, for example, be based on results from related historical trials. We provide more technical details on prior specification in Supplementary Appendix *Specification of prior hyperparameters*. There we also highlight the relation between the Dirichlet distribution and the multivariate beta distribution, and demonstrate that the prior and posterior distributions of θ_j are multivariate beta distributions.

The final superiority decision relies on the posterior distribution of treatment difference δ. Although this distribution does not belong to a known family of distributions, we can approach the distribution of δ via a two-step transformation of the posterior samples of ϕ_j. First, a sample of ϕ_j is drawn from its known Dirichlet distribution. Next, these draws can be transformed to a sample of θ_j using the property that joint response frequencies sum to the marginal probabilities. Finally, these samples from the posterior distributions of θ_E and θ_C can then be transformed to obtain the posterior distribution of joint treatment difference δ, by subtracting draws of θ_C from draws of θ_E, i.e. $\delta = \theta_E - \theta_C$. Algorithm 1 in section 3.3 includes pseudocode with the steps required to obtain a sample from the posterior distribution of δ.

3 Decision rules for multiple binary outcomes

The current section discusses how the model from the previous section can be used to make treatment superiority decisions. Treatment superiority is defined by the posterior mass in a specific subset of the multivariate parameter space of $\delta = (\delta_1, \ldots, \delta_K)$. The complete parameter space will be denoted by $S \subset (-1, 1)^K$, and the superiority space will be denoted by $\mathcal{S}_{Sup} \subset S$. Superiority is concluded when a sufficiently large part of the posterior distribution of δ falls in superiority region $\mathcal{S}_{Sup}$

$$P\big(\delta \in \mathcal{S}_{sup}|s_E, s_C\big) > p_{cut} \tag{6}$$

where p_{cut} reflects the decision threshold to conclude superiority. The value of this threshold should be chosen to control the Type I error rate α.

3.1 Four different decision rules

Different partitions of the parameter space define different superiority criteria to distinguish two treatments. The following decision rules conclude superiority when there is sufficient evidence that:

1. *Single rule:* an a priori specified primary outcome k has a treatment difference larger than zero. The superiority region is denoted by

$$\mathcal{S}_{Single(k)} = \{\delta|\delta_k > 0\} \tag{7}$$

Superiority is concluded when

$$P\big(\delta \in \mathcal{S}_{Single(k)}|s_E, s_C\big) > p_{cut} \tag{8}$$

2. *Any rule:* at least one of the outcomes has a treatment difference larger than zero. The superiority region is a combination of K superiority regions of the Single rule

$$\mathcal{S}_{Any} = \{\mathcal{S}_{Single_1} \cup \ldots \cup \mathcal{S}_{Single_K}\}$$

Superiority is concluded when

$$\max_k P\big(\delta \in \mathcal{S}_{Single(k)}|s_E, s_C\big) > p_{cut} \tag{9}$$

3. *All rule:* all outcomes have a treatment difference larger than zero. Similar to the Any rule, the superiority region is a combination of K superiority regions of the Single rule: The superiority region is denoted by

$$\mathcal{S}_{All} = \{\mathcal{S}_{Single_1} \cap \ldots \cap \mathcal{S}_{Single_K}\}$$

Superiority is concluded when

$$\min_k P(\boldsymbol{\delta} \in \mathcal{S}_{Single(k)} | \boldsymbol{s}_E, \boldsymbol{s}_C) > p_{cut} \tag{10}$$

Next to facilitating these common decision rules, our framework allows for a Compensatory decision rule:

4. *Compensatory rule:* the weighted sum of treatment differences is larger than zero. The superiority region is denoted by

$$\mathcal{S}_{Compensatory}(\boldsymbol{w}) = \left\{ \boldsymbol{\delta} | \sum_{k=1}^{K} w_k \delta_k > 0 \right\} \tag{11}$$

where $\boldsymbol{w} = (w_1, \ldots, w_K)$ reflect the weights for outcomes $1, \ldots, K$,
$0 \leq w_k \leq 1$ and $\sum_{k=1}^{K} w_k = 1$
Superiority is then concluded when

$$P(\boldsymbol{\delta} \in \mathcal{S}_{Compensatory}(\boldsymbol{w}) | \boldsymbol{s}_E, \boldsymbol{s}_C) > p_{cut} \tag{12}$$

Figure 2 visualizes these four decision rules.

From our discussion of the different decision rules, a number of relationships between them can be identified. First, mathematically the Single rule can be considered a special case of the Compensatory rule with weight $w_k = 1$ for primary outcome k and $w_l = 0$ for all other outcomes. Second, the superiority region of the All rule is a subset of the superiority regions of the other rules, i.e.

$$\mathcal{S}_{All} \subset \mathcal{S}_{Single}, \mathcal{S}_{Compensatory}, \mathcal{S}_{Any} \tag{13}$$

The Single rule is in turn a subset of the superiority region of the Any rule, such that

$$\mathcal{S}_{Single} \subset \mathcal{S}_{Any} \tag{14}$$

These properties can be observed in Figure 2 and translate directly to the amount of evidence provided by data $\boldsymbol{s}_E$ and $\boldsymbol{s}_C$. The posterior probability of the All rule is always smallest, while the posterior probability of the Any rule is at least as large as the posterior probability of the Single rule

$$P(\mathcal{S}_{Any} | \boldsymbol{s}_E, \boldsymbol{s}_C) \geq P(\mathcal{S}_{Single} | \boldsymbol{s}_E, \boldsymbol{s}_C) > P(\mathcal{S}_{All} | \boldsymbol{s}_E, \boldsymbol{s}_C)$$
$$P(\mathcal{S}_{Compensatory} | \boldsymbol{s}_E, \boldsymbol{s}_C) > P(\mathcal{S}_{All} | \boldsymbol{s}_E, \boldsymbol{s}_C) \tag{15}$$

The ordering of the posterior probabilities of different decision rules (equation (15) implies that superiority decisions are most conservative under the All rule and most liberal under the Any rule. In practice, this difference has two consequences. First, to properly control Type I error probabilities for these different decision rules, one needs to set a larger decision threshold p_{cut} for the Any rule than for the All rule. Second, the All rule typically requires the largest sample size to obtain sufficient evidence for a superiority decision.

Additionally, the correlation between treatment differences, $\rho_{\delta_k, \delta_l}$, influences the posterior probability to conclude superiority. The correlation influences the overlap with the superiority region, as visualized in Figure 3. Consequently, the Single rule is not sensitive to the correlation. A negative correlation requires a smaller sample size than a positive correlation under the Any and Compensatory rules, and vice versa for the All rule.
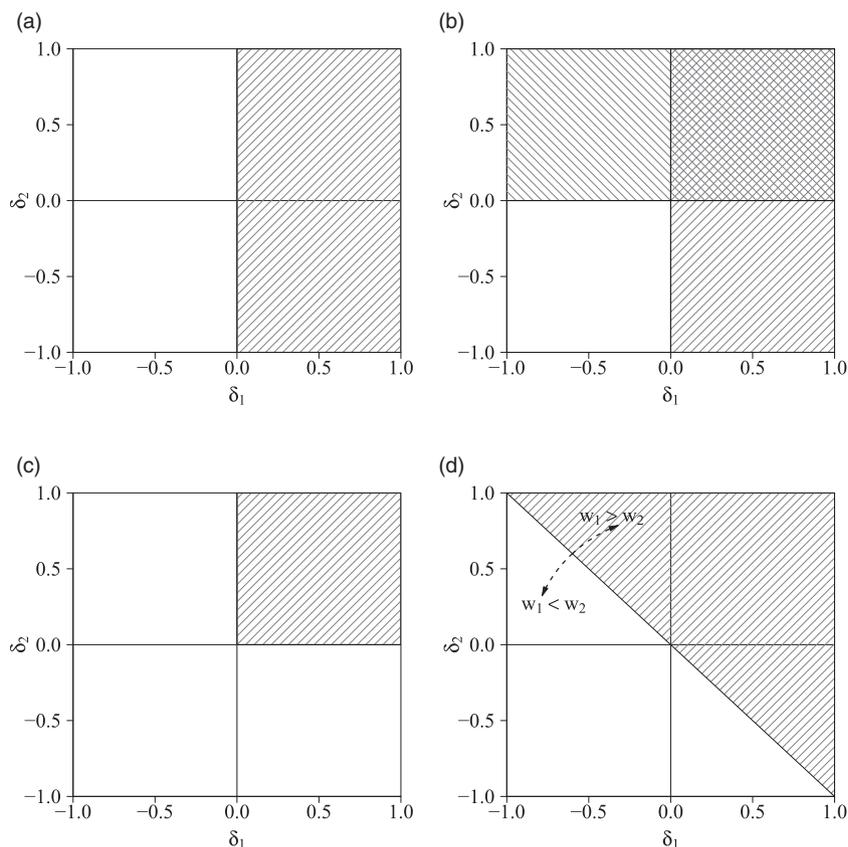
Figure 2. Superiority regions of various decision rules for two outcome variables ($K = 2$). The Any rule is a combination of the two Single rules. The Compensatory rule reflects $w = (0.5, 0.5)$.
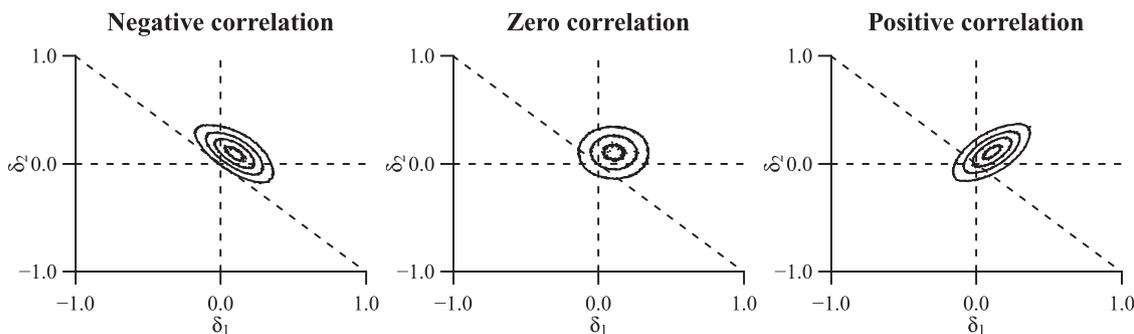

Figure 3. Influence of the correlation between two treatment differences on the proportion of overlap between the bivariate distribution of treatment differences δ and the superiority regions.

3.2 Specification of weights of the Compensatory decision rule

To utilize the flexibility of the Compensatory rule, researchers may wish to specify weights w. The current subsection discusses two ways to choose these weights: Specification can be based on the impact of outcome variables or on efficiency of the decision.

Specification of impact weights is guided by substantive considerations to reflect the relative importance of outcomes. When $w = \left(\frac{1}{K}, \ldots, \frac{1}{K}\right)$, all outcomes are equally important and all success probabilities in θ_j exert an identical influence on the weighted success probability. Any other specification of w that satisfies $\sum_{k=1}^{K} w_k = 1$

implies unequal importance of outcomes. To make the implications of importance weight specification more concrete, let us reconsider the two potential side effects of brain cancer treatment in the CAR-B study: cognitive functioning and fatigue.[1] When setting $(w_{cognition}, w_{fatigue}) = (0.50, 0.50)$, both outcomes would be considered equally important and a decrease of (say) 0.10 in fatigue could be compensated by an increase on cognitive functioning of at least 0.10. When $w_{cognition} > 0.50$, cognitive functioning is more influential than fatigue; and vice versa when $w_{cognition} < 0.50$. If $w_{cognition} = 0.75$ and $w_{fatigue} = 0.25$ for example, the treatment difference of cognitive functioning has three times as much impact on the decision as the treatment difference of fatigue.

Efficiency weights are specified with the aim of optimizing the required sample size. As the weights directly affect the amount of evidence for a treatment difference, the efficiency of the Compensatory decision rule can be optimized with values of w that are a priori expected to maximize the probability of falling in the superiority region. This strategy could be used when efficiency is of major concern, while researchers do not have a strong preference for the substantive priority of specific outcomes. The technical details required to find efficient weights are presented in Supplementary Appendix *Specification of efficiency weights*.

3.3 Implementation of the framework

The procedure to arrive at a decision using the multivariate analysis procedure proposed in the previous sections is presented in Algorithm 1 for a design with fixed sample size n_j of treatment j. We present the algorithm for designs with interim analyses in Algorithm 2 in Supplementary Appendix *Implementation of the framework in group sequential and adaptive designs*.

Algorithm 1 Decision procedure for a fixed design

1. Initialize
   a Choose decision rule
     if Compensatory **then** specify weights w
     if Single **then** specify k
     end if
   for each treatment $j \in \{E, C\}$ **do**
   b Choose prior hyperparameters $\boldsymbol{\alpha}_j^0$
   end for
   c Choose Type I error rate α and power $1 - \beta$
   d Determine decision threshold p_{cut}
     if Any rule **then** $1 - \frac{1}{2}\alpha$
     else $1 - \alpha$
     end if
   e Determine sample size n_j based on anticipated treatment differences $\boldsymbol{\delta}^n$
2. Collect data and compute evidence
   for each treatment $j \in \{E, C\}$
   a Collect n_j joint responses $\boldsymbol{x}_{j,i}$
   b Compute joint response frequencies $\boldsymbol{s}_j$
   c Compute posterior parameters $\boldsymbol{\alpha}_j^n = \boldsymbol{s}_j + \boldsymbol{\alpha}_j^0$
   d Sample L posterior draws, $\boldsymbol{\phi}_j^l$, $\boldsymbol{\phi}_j | \boldsymbol{\alpha}_j^n \sim Dirichlet(\boldsymbol{\phi}_j | \boldsymbol{\alpha}_j^n)$
   e Sum draws $\boldsymbol{\phi}_j^l$ to $\boldsymbol{\theta}_j^l$
   end for
   f Transform draws $\boldsymbol{\theta}_j^l$ to $\boldsymbol{\delta}^l$ via $\delta_k^l = \theta_{E,k}^l - \theta_{C,k}^l$
   g Compute posterior probability of treatment superiority $P(\boldsymbol{\delta} \in \mathcal{S}_{Sup} | \boldsymbol{s}_E, \boldsymbol{s}_C)$ as the proportion of posterior draws in superiority region $\mathcal{S}_{Sup}$
3. Make final decision
   if $P(\boldsymbol{\delta} \in \mathcal{S}_{Sup} | \boldsymbol{s}_E, \boldsymbol{s}_C) > p_{cut}$ **then** conclude superiority
   else conclude non-superiority
   end if

---

4 Numerical evaluation

The current section evaluates the performance of the presented multivariate decision framework by means of simulation in the context of two outcomes ($K = 2$). We seek to demonstrate (1) how often the decision procedure results in an (in)correct superiority conclusion to learn about decision error rates; (2) how many observations are required to conclude superiority with satisfactory error rates to investigate the efficiency of different decision rules, and (3) how well the average estimated treatment difference corresponds to the true treatment difference to examine bias. The current section is structured as follows. We first introduce the simulation conditions, the procedure to compute sample sizes for each of these conditions, and the procedure to generate and evaluate data. We then discuss the results of the simulation.

4.1 Conditions

The performance of the framework is examined as a function of the following factors:

1. *Data generating mechanisms:* We generated data of eight treatment difference combinations δ^T and three correlations between outcomes $\rho_{\theta_{j,1},\theta_{j,2}}$. An overview of these $8 \times 3 = 24$ data generating mechanisms is given in Table 1. In the remainder of this section, we refer to these data generating mechanisms with numbered combinations (e.g. 1.2), where the first number reflects treatment difference δ^T and the second number refers to correlation $\rho_{\theta_{j,1},\theta_{j,2}}^T$.

2. *Decision rules:* The generated data were evaluated with six different decision rules. We used the Single (for outcome $k = 1$), Any, and All rules, as well as three different Compensatory rules: One with equal weights $w = (0.50, 0.50)$ and two with unequal weights $w = (0.76, 0.24)$ and $w = (0.64, 0.36)$. The weight combinations of the latter two Compensatory rules optimize the efficiency of data generating mechanisms with uncorrelated (i.e. 8.2) and correlated (i.e. 8.1) treatment differences, respectively, following the procedure in Supplementary Appendix *Specification of efficiency weights*. We refer to these three Compensatory rules as Compensatory-Equal (C-E), Compensatory-Unequal Uncorrelated (C-UU) and Compensatory-Unequal Correlated (C-UC), respectively.

Table 1. Data generating mechanisms (DGM) used in numerical evaluation of the framework.

DGM	δ_1^T	δ_2^T	$\rho_{\theta_{j,1},\theta_{j,2}}^T$	$\theta_{E,1}^T$	$\theta_{E,2}^T$	$\phi_{E,11}^T$	$\theta_{C,1}^T$	$\theta_{C,2}^T$	$\phi_{C,11}^T$
1.1	−0.20	−0.20	−0.30	0.40	0.40	0.09	0.60	0.60	0.29
1.2			0.00			0.16			0.36
1.3			0.30			0.23			0.43
2.1	0.00	0.00	−0.30	0.50	0.50	0.17	0.50	0.50	0.17
2.2			0.00			0.25			0.25
2.3			0.30			0.32			0.32
3.1	0.10	0.10	−0.30	0.55	0.55	0.23	0.45	0.45	0.13
3.2			0.00			0.30			0.20
3.3			0.30			0.38			0.28
4.1	0.20	0.20	−0.30	0.60	0.60	0.29	0.40	0.40	0.09
4.2			0.00			0.36			0.16
4.3			0.30			0.43			0.23
5.1	0.40	0.40	−0.30	0.70	0.70	0.43	0.30	0.30	0.03
5.2			0.00			0.49			0.09
5.3			0.30			0.55			0.15
6.1	0.40	0.00	−0.30	0.70	0.50	0.28	0.30	0.50	0.08
6.2			0.00			0.35			0.15
6.3			0.30			0.42			0.22
7.1	0.20	−0.40	−0.30	0.60	0.30	0.11	0.40	0.70	0.21
7.2			0.00			0.18			0.28
7.3			0.30			0.25			0.35
8.1	0.24	0.08	−0.30	0.62	0.54	0.26	0.38	0.46	0.10
8.2			0.00			0.33			0.17
8.3			0.30			0.41			0.25

4.2 Sample size computations

To properly control Type I error and power, each of the 24×6 conditions requires a specific sample size. These sample sizes n_j are based on anticipated treatment differences $\boldsymbol{\delta}^n$, that corresponded to the true parameters of each data generating mechanism in Table 1 (i.e. $\boldsymbol{\delta}^n = \boldsymbol{\delta}^T$ and $\rho^n_{\theta_{j,1}, \theta_{j,2}} = \rho^T_{\theta_{j,1}, \theta_{j,2}}$). Procedures to compute sample sizes per treatment group for the different decision rules were the following:

1. For the Single rule, we used a two-proportion $z-$ test, where we plugged in the anticipated treatment difference on the first outcome variable (i.e. δ^n_1).
2. Following Sozu et al.,[5,16] we used multivariate normal approximations of correlated binary outcomes for the All and Any rules.
3. For the Compensatory rule, we used a continuous normal approximation with mean $\sum_{k=1}^{K} w_k \theta_{j,k}$ and variance $\sum_{k=1}^{K} w_k^2 \sigma_{j,k}^2 + 2 \sum \sum_{k<l} w_k w_l \sigma_{j,kl}$. Here, $\sigma_{j,k}^2 = \theta_{j,k}(1 - \theta_{j,k})$ and $\sigma_{j,kl} = \phi_{j,kl} - \theta_{j,k}\theta_{j,l}$.

The computed sample sizes are presented in Table 3. Conditions that should not result in superiority were evaluated at sample size $n_j = 1000$.

4.3 Data generation and evaluation

Of each data generating mechanism presented in Table 1, we generated 5000 samples of size $2 \times n_j$. These data were combined with a proper uninformative prior distribution with hyperparameters $\boldsymbol{\alpha}_j^0 = (0.01, \dots, 0.01)$ to satisfy $\boldsymbol{\alpha}_j^n \approx \boldsymbol{s}_j$, as discussed in Section 2. We aimed for Type I error rate $\alpha = .05$ and power $1 - \beta = .80$, which corresponds to a decision threshold p_{cut} of $1 - \alpha = 0.95$ (Single, Compensatory, All rules) and $1 - \frac{1}{2}\alpha = 0.975$ (Any rule).[4,5,11] The generated datasets were evaluated using the procedure in steps 2 and 3 of Algorithm 1.

The proportion of samples that concluded superiority reflects Type I error rates (when false) and power (when correct). We assessed the Type I error rate under the data generating mechanism with the least favorable population values of $\boldsymbol{\delta}^T$ under the null hypothesis in frequentist one-sided significance testing. These are values of $\boldsymbol{\delta}^T$ outside $\mathcal{S}_{Sup}$ that are most difficult to distinguish from values of $\boldsymbol{\delta}^T$ inside $\mathcal{S}_{Sup}$. Adequate Type I error rates for the least favorable treatment differences imply that the Type I error rates of all values of $\boldsymbol{\delta}^T$ outside $\mathcal{S}_{Sup}$ are properly controlled. The least favorable values of $\boldsymbol{\delta}^T$ were reflected by treatment difference 2 for the Single, Any, and Compensatory rules, and treatment difference 6 for the All rule. Bias was computed as the difference between the observed treatment difference at sample size n_j and the true treatment difference $\boldsymbol{\delta}^T$.

4.4 Results

The proportion of samples that concluded superiority and the required sample size are presented in Tables 2 and 3, respectively. Type I error rates were properly controlled around $\alpha = .05$ for each decision rule under its least favorable data generating mechanism. The power was around .80 in all scenarios with true superiority. Moreover, average treatment differences were estimated without bias (smaller than 0.01 in all conditions).

Given these satisfactory error rates, a comparison of sample sizes provides insight in the efficiency of the approach. We remark here that a comparison of sample sizes is only relevant when the decision rules under consideration have a meaningful definition of superiority. Further, in this discussion of results we primarily focus on the newly introduced Compensatory rule in comparison to the other decision rules. The results demonstrate that the Compensatory rule consistently requires fewer observations than the All rule, and often – in particular when treatment differences are equal (i.e. treatment differences $3 - 5$) – than the Any and the Single rule. Similarly, the Any rule consistently requires fewer observations than the All rule and could be considered an attractive option in terms of sample sizes. Note, however, that the more lenient Any rule may not result in a meaningful decision for all trials, since the rule would also conclude superiority when the treatment has a small positive treatment effect and large negative treatment effect (i.e. treatment difference 7); a scenario that may not be clinically relevant.

The influence of the relation between outcomes is also apparent: Negative correlations require fewer observations than positive correlations. The variation due to the correlation is considerable: The average sample size almost doubles in scenarios with equal treatment differences (e.g. data generating mechanisms 3.1 vs. 3.3 and 4.1 vs. 4.3).

Table 2. P(Conclude superiority) for different data generating mechanisms (DGM) and decision rules.

DGM	Single	Any	All	C-E	C-UU	C-UC
1.1	0.000	0.000	0.000	0.000	0.000	0.000
1.2	0.000	0.000	0.000	0.000	0.000	0.000
1.3	0.000	0.000	0.000	0.000	0.000	0.000
2.1	**0.051**	**0.048**	0.000	0.049	0.052	0.051
2.2	0.046	0.045	0.003	0.056	0.048	0.054
2.3	0.051	0.045	0.008	**0.049**	**0.049**	**0.049**
3.1	0.810	0.796	0.801	0.807	0.804	0.790
3.2	0.799	0.801	0.804	0.806	0.788	0.791
3.3	0.799	0.807	0.809	0.800	0.797	0.803
4.1	0.794	0.784	0.806	0.811	0.789	0.784
4.2	0.808	0.802	0.814	0.813	0.804	0.803
4.3	0.804	0.801	0.816	0.804	0.796	0.800
5.1	0.807	0.806	0.830	0.881	0.817	0.857
5.2	0.807	0.814	0.838	0.831	0.813	0.813
5.3	0.809	0.847	0.822	0.809	0.798	0.802
6.1	0.811	0.779	0.053	0.824	0.798	0.819
6.2	0.813	0.777	0.045	0.805	0.808	0.820
6.3	0.803	0.758	**0.051**	0.801	0.788	0.803
7.1	0.799	0.789	0.000	0.000	0.863	0.002
7.2	0.804	0.792	0.000	0.000	0.857	0.003
7.3	0.807	0.794	0.000	0.000	0.867	0.005
8.1	0.787	0.782	0.789	0.808	0.804	0.805
8.2	0.777	0.797	0.807	0.804	0.799	0.804
8.3	0.785	0.811	0.807	0.805	0.805	0.806

Note: Bold-faced values indicate the conditions with least favorable values.

Table 3. Average sample size to correctly conclude superiority for different data generating mechanisms (DGM) and decision rules.

DGM	Single	Any	All	C-E	C-UU	C-UC
1.1	–	–	–	–	–	–
1.2	–	–	–	–	–	–
1.3	–	–	–	–	–	–
2.1	–	–	–	–	–	–
2.2	–	–	–	–	–	–
2.3	–	–	–	–	–	–
3.1	307	191	424	**108**	157	119
3.2	307	217	418	**154**	192	162
3.3	307	247	406	**199**	226	206
4.1	75	47	105	**26**	39	29
4.2	75	53	103	**38**	47	40
4.3	75	60	101	**49**	55	50
5.1	17	11	25	**6**	9	7
5.2	17	12	25	**9**	11	**9**
5.3	17	14	24	**11**	12	**11**
6.1	17	21	–	25	**15**	17
6.2	**17**	21	–	36	19	24
6.3	**17**	21	–	47	22	30
7.1	**75**	95	–	–	608	–
7.2	**75**	95	–	–	733	–
7.3	**75**	95	–	–	858	–
8.1	51	56	482	41	38	**36**
8.2	51	60	482	59	**46**	49
8.3	**51**	63	482	76	55	62

Note: Bold-faced values indicate the lowest sample size per data generating mechanism. Conditions with a hyphen should not result in treatment superiority.

Comparison of the three different Compensatory rules further highlights the influence of weights w and illustrates that a Compensatory rule is most efficient when weights have been optimized with respect to the treatment differences and the correlation between them. The Compensatory rule with equal weights (C-E) is most efficient when treatment differences on both outcomes are equally large (treatment differences $3-5$), while the Compensatory rule with unequal weights for uncorrelated outcomes (C-UU) is most efficient under data generating mechanism 8.2. The Compensatory rule with unequal weights, optimized for negatively correlated outcomes (C-UC) is most efficient in data generating mechanism 8.1. The Compensatory is less efficient than the Single rule in the scenario with an effect on one outcome only (treatment difference 6). Effectively, in this situation the Single rule is the Compensatory rule with optimal weights for this specific scenario $w = (1, 0)$. Utilizing the flexibility of the Compensatory rule to tailor weights to anticipated treatment differences and their correlations thus pays off in terms of efficiency.

Note that in practice it may be difficult to accurately estimate treatment differences and correlations in advance. This uncertainty may result in inaccurate sample size estimates, as demonstrated in Supplementary Appendix *Numerical evaluation: Comparison of trial designs*. The simulations in this appendix also show that the approach can be implemented in designs with interim analyses as well, which is particularly useful under uncertainty about anticipated treatment differences. Specifically, we demonstrate that (1) both Type I and Type II error rates increase, while efficiency decreases in a fixed design when the anticipated treatment difference does not correspond to the true treatment difference; and (2) designs with interim analyses could compensate for this uncertainty in terms of error rates and efficiency, albeit at the expense of upward bias.

Further, Supplementary Appendix *Numerical evaluation: Comparison of prior specifications* shows how prior information influences the properties of decision-making. Informative priors support efficient decision-making when the prior treatment difference corresponds to the treatment difference in the data. In contrast, evidence is influenced by dissimilarity between prior hyperparameters and data, and may either increase or decrease (1) the required sample size; and (2) the average posterior treatment effect, depending on the nature of the non-correspondence.

5 Discussion

The current paper presented a Bayesian framework to efficiently combine multiple binary outcomes into a clinically relevant superiority decision. We highlight two characteristics of the approach.

First, the multivariate Bernoulli model has shown to capture relations properly and support multivariate decision-making. The influence of the correlation between outcomes on the amount of evidence in favor of a specific treatment highlights the urgency to carefully consider these relations in trial design and analysis in practice.

Second, multivariate analysis facilitates comprehensive decision rules such as the Compensatory rule. More specific criteria for superiority can be defined to ensure clinical relevance, while relaxing conditions that are not strictly needed for clinical relevance lowers the sample size required for error control; a fact that researchers may take advantage of in practice where sample size limitations are common.[6]

Several other modeling procedures have been proposed for the multivariate analysis of multiple binary outcomes. The majority of these alternatives assume a (latent) normally distributed continuous variable. When these models rely on large sample approximations for decision-making (such as methods presented by Whitehead et al.,[17] Sozu et al.,[5,16] and Su et al.[18]; see for an exception Murray et al.[3]), their applicability is limited, since the validity of z-tests for small samples may be inaccurate. A second class of alternatives uses copula models, which is a flexible approach to model dependencies between multiple univariate marginal distributions. The use of copula structures in discrete data can be challenging, however.[19] Future research might provide insight in the applicability of copula models for multivariate decision-making in clinical trials.

Two additional remarks concerning the number of outcomes should be made. First, the modeling procedure becomes more complex when the number of outcomes increases, since the number of cells increases exponentially. Second, the proposed Compensatory rule has a linear compensatory mechanism. With two outcomes, the outcomes compensate each other directly and the size of a negative effect is maximized by the size of the positive effect. A decision based on more than two outcomes might have the – potentially undesirable – consequence of compensating a single large negative effect by two or more positive effects. Researchers are encouraged to carefully think about a suitable superiority definition and might consider additional restrictions to the Compensatory rule, such as a maximum size of individual negative effects.

Acknowledgements

Declaration of conflicting interests

Funding

ORCID iDs

Xynthia Kavelaars https://orcid.org/0000-0003-1600-3153
Maurits Kaptein https://orcid.org/0000-0002-6316-7524

Supplemental material

Supplemental material for this article is available online. The R code used to generate results in Section Numerical evaluation, Appendix Numerical evaluation: Comparison of trial designs, and Appendix Numerical evaluation: Comparison of prior specifications can be found on https://github.com/XynthiaKavelaars/Decision-making-with-multiple-correlated-binary-outcomes-in-clinical-trials

References

1. Schimmel WC, Verhaak E, Hanssens PE, et al. A randomised trial to compare cognitive outcome after gamma knife radiosurgery versus whole brain radiation therapy in patients with multiple brain metastases: research protocol car-study b. *BMC cancer* 2018; **18**: 218.
2. Food and Drug Administration. *Multiple endpoints in clinical trials guidance for industry*. Center for Biologics Evaluation and Research (CBER), 2017.
3. Murray TA, Thall PF and Yuan Y. Utility-based designs for randomized comparative trials with categorical outcomes. *Stat Med* 2016; **35**: 4285–4305.
4. Sozu T, Sugimoto T and Hamasaki T. Sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables. *Biometric J* 2012; **54**: 716–729.
5. Sozu T, Sugimoto T and Hamasaki T. Reducing unnecessary measurements in clinical trials with multiple primary endpoints. *J Biopharmaceut Stat* 2016; **26**: 631–643.
6. Van de Schoot R and Miocevic M (eds) *Small sample size solutions*. London: Routledge, 2020.
7. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics* 1984; 40: 1079–1087.
8. Tang DI, Gnecco C and Geller NL. Design of group sequential clinical trials with multiple endpoints. *J Am Stat Assoc* 1989; **84**: 775–779.
9. Pocock SJ, Geller NL and Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987; **43**: 487–498.
10. Gelman A, Carlin JB, Stern HS, et al. *Bayesian data analysis*. London: Chapman and Hall/CRC, 2013.
11. Marsman M and Wagenmakers EJ. Three insights from a bayesian interpretation of the one-sided p value. *Educ Psychol Measure* 2017; **77**: 529–539.
12. Food and Drug Administration. *Guidance for industry adaptive design clinical trials for drugs and biologics*. Washington, DC: Food and Drug Administration, 2010.
13. Wilson DJ. The harmonic mean p-value for combining dependent tests. *Proc Natl Acad Sci* 2019; **116**: 1195–1200.
14. Dai B, Ding S, Wahba G, et al. Multivariate Bernoulli distribution. *Bernoulli* 2013; **19**: 1465–1483.
15. Olkin I and Trikalinos TA. Constructions for a bivariate beta distribution. *Stat Probabil Lett* 2015; **96**: 54–60.
16. Sozu T, Sugimoto T and 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.
17. Whitehead J, Branson M and Todd S. A combined score test for binary and ordinal endpoints from clinical trials. *Stat Med* 2010; **29**: 521–532.
18. Su TL, Glimm E, Whitehead J, et al. An evaluation of methods for testing hypotheses relating to two endpoints in a single clinical trial. *Pharmaceut Stat* 2012; **11**: 107–117.
19. Panagiotelis A, Czado C and Joe H. Pair copula constructions for multivariate discrete data. *J Am Stat Assoc* 2012; **107**: 1063–1072.