Simultaneous inference of a binary composite endpoint and its components

M. Große Ruse¹*, C. Ritz², and L.A. Hothorn³

¹Center for Mathematical Sciences, Lund University, Lund, Sweden; ²Department of Nutrition, Exercise and Sports, Faculty of Science, University of Copenhagen, Copenhagen, Denmark; ³Institute of Biostatistics, Leibniz University, Hannover, Germany

ABSTRACT

Binary composite endpoints offer some advantages as a way to succinctly combine evidence from a number of related binary endpoints recorded in the same clinical trial into a single outcome. However, as some concerns about the clinical relevance as well as the interpretation of such composite endpoints have been raised, it is recommended to evaluate the composite endpoint jointly with the involved components. We propose an approach for carrying out simultaneous inference based on separate model fits for each endpoint, yet controlling the familywise type I error rate asymptotically. The key idea is to stack parameter estimates from the different fits and derive their joint asymptotic distribution. Simulations show that the proposed approach comes closer to nominal levels and has comparable or higher power as compared to existing approaches, even for moderate sample sizes (around 100-200 observations). The method is compared to the gatekeeping approach and results are provided in the Supplementary Material. In two data examples we show how the procedure may be adapted to handle local significance levels specified through a priori given weights.

ARTICLE HISTORY

Received 25 September 2014
Accepted 10 January 2016

KEYWORDS

Adjusted p-values; asymptotic representation; correlated endpoints; familywise error rate; weighted local significance levels

1. Introduction

For a number of reasons, it may be convenient to use composite binary endpoints (i.e., the maximum of a number of binary endpoints) in clinical trials: Such endpoints may allow a parsimonious summary of treatments and also achieve increased statistical efficiency as compared to analyses of individual endpoints, in particular if some endpoints are associated with low probabilities of observing the event (Ross, 2007).

Unfortunately, the usage of binary composite endpoints also entails some shortcomings, which have been identified and discussed by a number of authors recently, e.g., Cordoba et al. (2010) and Tomlinson and Detsky (2010). A key problem is how to interpret the composite endpoint clinically, in particular in view of constituent endpoints with incomparable contributions due to varying degrees of relevance of the biological effects captured by these individual endpoints. Subsequent interpretation of results is also complicated as incomparable contributions are not easily disentangled from the composite endpoint (Lim et al., 2008; Mascha and Sessler, 2011). In short, clinical relevance and interpretation may be difficult for composite endpoints (Ross, 2007).

Hence both the composite endpoint and its constituents should be analyzed to ensure a satisfactory clinical evaluation (Turk et al., 2008). A confirmatory statistical analysis, which will require
simultaneous evaluation of the composite endpoint and its constituent, is therefore strongly recommended (Huque et al., 2011; Chi, 2005).

Simultaneous confirmatory analysis involves some kind of multiplicity adjustment for the significance levels or \( p \)-values considered. Bonferroni adjustment is readily available also in this setting, but due to the expected high correlation between the composite endpoints and its components it easily becomes much too conservative. A few approaches for simultaneous inference of binary composite endpoints have been proposed in the literature: Pogue et al. (2010) suggested using generalized estimating equations (GEE) or logistic mixed-effects models to carry out a heterogeneity test for evaluating how different the treatment differences were across endpoints as compared to a common effect. Mascha and Sessler (2011) suggested a similar multivariate approach using GEE. A different approach is taken by Rauch and Kieser (2012), who derive a multiplicity adjustment using an approximation argument for establishing correlations between test statistics by means of (raw) observations and the parameter estimates and subsequently exploiting asymptotic joint normality of the test statistics.

We present a novel approach for simultaneous assessment of a binary composite endpoint and its component endpoints. Specifically, we consider separate models for each binary endpoint but derive correlations between test statistics through an asymptotic argument that allows us to combine the individual model fits, exploiting the idea by Pipper et al. (2012).

Section 2 contains the theoretical background of the proposed method. In Section 3, the proposed approach is evaluated through an extensive simulation study where its performance is compared to standard Bonferroni adjustment and to the approach proposed by Rauch and Kieser (2012). We also compare the approaches in two data examples. In Section 4, we discuss the advantages and limitations of the proposed approach and also outline some possible extensions.

2. Methods

We consider a study where \( J \) binary endpoints were observed for \( n \) subjects: \( Y_{1}^{(j)}, \ldots, Y_{n}^{(j)} \) for endpoint \( j = 1, \ldots, J \). These endpoints were combined into a single composite endpoint, which is considered to be the primary outcome. The original binary outcomes will be treated as secondary outcomes, e.g., Turk et al. (2008). More specifically, we define the composite binary endpoint \( Y^{(J+1)} \) as the following maximum:

\[
Y_{i}^{(J+1)} = \max_{j=1,\ldots,J} \{ Y_{i}^{(j)} \} , \quad i = 1, \ldots, n.
\]

By definition the composite endpoint \( Y^{(J+1)} \) will also be a binary endpoint (Quan et al., 2007).

2.1. A novel approach

We will assume that observations for each endpoint may be described by a generalized linear model for binary data (arbitrary link); a logistic regression model would be a common choice. So we assume that \( Y_{i}^{(j)} \) is a binary random variable with probability \( \pi_{i}^{(j)} \) specified through a generalized linear model with link function \( g \):

\[
g(\pi_{i}^{(j)}) = \sum_{k=1}^{p_j} x_{ik}^{(j)} \beta_{k}^{(j)} ,
\]

(1)

where \( p_j \in \mathbb{N} \) and \( x_{i}^{(j)} = (x_{i1}^{(j)}, \ldots, x_{ip_j}^{(j)}) \) denotes the covariate vector for observation \( Y_{i}^{(j)} \) (McCullagh and Nelder, 1989, Chapter 4). This general model specification allows for inclusion of a treatment factor while at the same time additional covariates may be incorporated.

We assume that \( \beta_{k}^{(j)} \) denotes the parameter of interest for the \( j \)th endpoint \( (k_j \in \{1, \ldots, p_j\}) \), corresponding to the null hypotheses:

\[
H_{0}^{(j)} : \beta_{k_j}^{(j)} = \beta_{0,k_j}^{(j)} , \quad j = 1, \ldots, J + 1 .
\]

(2)
The intersection of the null hypotheses in equation (1) is referred to as the global null hypothesis. We denote the maximum likelihood estimator of the unknown parameter $\hat{\beta}^{(j)}_{k_j}$ by $\hat{\psi}^{(j)}$. Under the null hypothesis $H_0^{(j)}$ and assuming standard regularity conditions, the maximum likelihood estimator $\hat{\beta}^{(j)}_{k_j}$ satisfies the asymptotic property $(j=1, \ldots, J+1)$:

$$\sqrt{n} \left( \hat{\beta}^{(j)}_{k_j} - \beta_{0,k_j}^{(j)} \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ -I^{(j)} \right]^{-1} \hat{\psi}^{(j)}_i + o_p(1) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i^{(j)} + o_p(1),$$

(3)

where for simplicity we introduced the notation $\psi_i^{(j)} = \left[ -I^{(j)} \right]^{-1} \hat{\psi}^{(j)}_i$. Here $\hat{\psi}^{(j)}_i$ is the score function and Fisher information matrix are evaluated at the maximum likelihood estimator $\hat{\beta}^{(j)}_{k_j}$. Piiper et al. (2012) show that the asymptotic representations for the individual endpoints (both constituents and composite endpoints) carry over to the stacked asymptotic representation:

$$\sqrt{n} \left( \hat{\beta} - \beta_0 \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i^{(j)} + o_p(1)$$

with the following definitions $\hat{\beta} = \left( \hat{\beta}^{(1)}_{k_1}, \ldots, \hat{\beta}^{(J+1)}_{k_{J+1}} \right)$, $\beta_0 = \left( \beta^{(1)}_{0,k_1}, \ldots, \beta^{(J+1)}_{0,k_{J+1}} \right)$, and $\psi_i = \left( \psi^{(1)}_i, \ldots, \psi^{(J+1)}_i \right)$. Note that $\psi_1, \ldots, \psi_n$ are independent realizations of a zero-mean random variable with finite variance. The multivariate central limit theorem ensures weak convergence $\sqrt{n} \left( \hat{\beta} - \beta_0 \right) \rightarrow \mathcal{N}(0, \Sigma)$ as $n \rightarrow \infty$. The variance-covariance matrix $\Sigma$ is defined as the following limit in probability $\frac{1}{n} \sum_{i=1}^{n} \psi_i^T \psi_i \rightarrow \Sigma$, which may be estimated by $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \hat{\psi}_i^T \hat{\psi}_i$, where $\hat{\psi}_i$ is the empirical counterpart of $\psi_i$ with parameter estimates inserted. For additional details, we refer to Piiper et al. (2012).

The null hypotheses in equation (2) are evaluated in the usual way by means of Wald test statistics $T_1, \ldots, T_{J+1}$, defined as the estimate divided by its standard error. Each of the corresponding (marginal) $p$-values, which are obtained with reference to a standard normal distribution, may then be referred to a global multiplicity adjusted significance level $\alpha_{adj}$. Consequently, the global null hypothesis is rejected if and only if (at least) one $|T_j|$ exceeds the $(1 - \alpha_{adj}/2)$ percentile of the standard normal distribution (Dickhaus, 2014). The global multiplicity adjusted significance level may be deduced from the asymptotic multivariate normal distribution of $T = (T_1, \ldots, T_{J+1})$ (Hothorn et al., 2008; Hasler and Hothorn, 2011) in such a way that control at a given nominal FWER $\alpha_{nominal}$, which is often equal to 0.05, is ensured.

In some applications it can, however, be appropriate to prioritize certain hypotheses. Such a hierarchy among hypotheses can be induced by means of a vector of weights, $w = (w_1, \ldots, w_{J+1})$ (such that $\sum_{j=1}^{J+1} w_j = 1$), where higher weights are assigned to more important hypotheses. In this case the global, adjusted significance level $\alpha_{adj}$ would be replaced by local (i.e., hypothesis-specific) adjusted significance levels defined as $\alpha^{(j)}_{w,adj} = w_j \alpha_{w,adj}$, $j = 1, \ldots, J+1$. The reference significance level $\alpha_{w,adj}$ will naturally depend on the specified prioritization $w$ and is chosen in order for the $\alpha^{(j)}_{w,adj}$ to simultaneously control the FWER. Here one would compare the marginal $p$-value corresponding to hypothesis $H^{(j)}$ with the value of the respective weight-specific local adjusted significance level $\alpha^{(j)}_{w,adj}$. Consequently, the global null hypothesis would be rejected if and only if (at least) one $|T_j|$ exceeds the $\left(1 - \alpha^{(j)}_{w,adj}/2\right)$ percentile.
This idea was already proposed by Westfall and Soper (1998) and it may be extended to correlated endpoints considered in the present framework as follows: For a given weight vector \( w \) and variance-covariance matrix \( \Sigma \in \mathbb{R}^{(j+1) \times (j+1)} \), we define the function \( f_{w, \Sigma} \): \([0,1] \rightarrow [0,1] \) of tail probabilities in a multivariate normal distribution as:

\[
f_{w, \Sigma}(\alpha) = 1 - \int_{z_1 - w_j \alpha / 2}^{\infty} \ldots \int_{z_{j+1} - w_{j+1} \alpha / 2}^{\infty} \phi(s, 0, \Sigma) ds.
\] (4)

Here \( z_j - w_j \alpha / 2 \) denotes the \( 1 - w_j \alpha / 2 \) percentile in the standard normal distribution and \( \phi(\cdot, 0, \Sigma) \) is the density of the \((j + 1)\)-dimensional normal distribution with mean 0 and variance-covariance matrix \( \Sigma \). Given a prioritization represented by \( w \) and an estimated variance-covariance matrix \( \hat{\Sigma} \) of the vector \( T \) of test statistics, the function \( f_{w, \hat{\Sigma}} \) may then be used to derive the local adjusted weight-specific significance levels \( \alpha_{w, \text{adj}}^{(j)} \) from a prescribed FWER \( \alpha_{\text{nom}} \) as follows. As \( f \) is invertible, the reference level \( \alpha_{w, \text{adj}} \) may be calculated for any specified FWER \( \alpha_{\text{nom}} \) by:

\[
\alpha_{w, \text{adj}} = f_{w, \hat{\Sigma}}^{-1}(\alpha_{\text{nom}}).
\] (5)

Subsequently, the local multiplicity adjusted significance levels are obtained as \( \alpha_{w, \text{adj}}^{(j)} = w_j \alpha_{w, \text{adj}} \). If the test statistics are independent, their variance-covariance matrix is given by the identity matrix \( I \). With equal weighting \( w_j = \frac{1}{J+1} \) for \( j = 1, \ldots, J+1 \), (which corresponds to no prioritization) the above equation would yield \( \alpha_{w, \text{adj}} = f_{w, I}(\alpha_{\text{nom}}) = (J+1) \cdot \left[ 1 - (1 - \alpha_{\text{nom}})^{1/(J+1)} \right] \), implying that the local adjusted significance levels \( \alpha_{w, \text{adj}}^{(j)} \) simply recover the Slepian correction (Pipper et al., 2012).

In practice equation (5) may be solved by means of a modified bisection method. The method is implemented using \( R \) (R Core Team, 2014) and the extension package \texttt{multcomp} (Hothorn et al., 2008), an example code thereof can be found in Appendix A.

3. Results

A simulation study was performed to explore and compare the performance of the proposed method, the standard Bonferroni adjustment and the suggested approach by Rauch and Kieser (2012).

Within a logistic regression setting, we considered binary endpoints \( Y^{(1)}, Y^{(2)} \) and their composite defined as their maximum, i.e., \( Y^{(3)} = \max\{Y^{(1)}, Y^{(2)}\} \). We were interested in the simultaneous assessment of the null hypotheses:

\[
H_0^{(j)} : \pi_C^{(j)} = \pi_T^{(j)} \quad \text{versus} \quad H_0^{(j)} : \pi_C^{(j)} \neq \pi_T^{(j)} , \quad j = 1, 2, 3.
\] (6)

Here \( \pi_C^{(j)} \) and \( \pi_T^{(j)} \) denote the event probability for endpoint \( j \in \{1, 2, 3\} \) in control and treatment group, respectively.

Translated into the framework introduced in Section 2, this setting corresponds to choosing \( g \) as the logit function and \( J = p_1 = p_2 = 2 \). The first covariate is constant, \( x_{1i}^{(j)} = 1 \), and represents the intercept \( \beta_1^{(j)} \). The second covariate \( x_{12}^{(j)} \) is an indicator variable expressing affiliation of subject \( i \) to either treatment or control group, i.e., \( x_{12}^{(j)} = 1 \) if subject \( i \) belongs to the treatment group and \( x_{12}^{(j)} = 0 \) else. With these choices equation (1) reads:

\[
\ln \left( \frac{\pi_T^{(j)}}{1 - \pi_T^{(j)}} \right) = \beta_1^{(j)} + \beta_2^{(j)} x_{12}^{(j)}.
\]
The parameter of interest is $\beta_j$ and its value under the null hypothesis $H_{j0}$ of no treatment effect is $\beta_{j0} = 0$. As indicated in the previous section, Wald test statistics may be applied to investigate the hypotheses of interest. Note, however, that in this specific setting (inclusion of only a treatment factor without incorporation of additional covariates), it is possible and advantageous to apply the generally more powerful score test for hypotheses assessment.

### 3.1. Settings for the simulation study

The global two-sided significance level was set to $\alpha = 0.05$. Estimation of FWER and power was each based on 10,000 simulated data sets. Each data set contained $n$ independent observations of the two endpoints $Y_{(1)}, Y_{(2)}$ for the control group and additional $n$ independent observations for the treatment group such that the total sample size was $2n$. The group sample size $n$ varied between 50, 100, 200 and 500 and the component endpoints were correlated with predefined correlation $\rho$ (correlations were equal in control and treatment group).

For the simultaneous confirmatory assessment of treatment effect in composite and its constituents in the simulation study, we considered unequally weighted hypotheses: More weight was assigned to the null hypothesis $H_3$ related to the composite endpoint by choosing $w_3 = 0.5$, whereas its constituents $Y_{(1)}$ and $Y_{(2)}$ were weighted according to $w_1 = w_2 = 0.25$. As indicated above, this specific setting facilitates the use of the score test for hypotheses assessment, which results in a slight gain of power as compared to the Wald test and thus simulation results are based on the score test.

As mentioned in Appendix B, the correlation $\rho$ between the components $Y_{(1)}, Y_{(2)}$ is given by:

$$\rho = \frac{\pi_{(1)\cap(2)} - \pi_{(1)}\pi_{(2)}}{\sqrt{\pi_{(1)}(1 - \pi_{(1)})\pi_{(2)}(1 - \pi_{(2)})}},$$

where $\pi_{(1)\cap(2)} = \mathbb{P}(Y_{(1)} = 1, Y_{(2)} = 1)$ denotes the overlap rate of the components. In Rauch and Kieser (2012), the authors apply an approximation of $\rho$ by assuming a zero overlap rate, thereby introducing a bias in their estimation procedure. For the sake of completeness our simulation study includes both their proposed approximate approach with hypothesized zero overlap rate (in the sequel this approach will be denoted by RK1) and their approach without such approximation (henceforth denoted by RK2). This will en passant give a quantification of the effect the approximation has on the performance of their method.

#### 3.1.1. Specific settings for FWER estimation

The event probability $\pi_{(1)}^{(C)}$ for the first endpoint in the control group was chosen among $\{0.1, 0.4\}$. We assumed the second endpoint corresponding to the less severe and more frequent event; therefore, its event probability under control treatment was set to $\pi_{(2)}^{(C)} = \pi_{(1)}^{(C)} + 0.1$. Under the prevailing assumption of no treatment effect, one had $\pi_{(1)}^{(T)} = \pi_{(1)}^{(C)}$ and $\pi_{(2)}^{(T)} = \pi_{(2)}^{(C)}$ (which then naturally implied $\pi_{(3)}^{(T)} = \pi_{(3)}^{(C)}$). Given fixed values for $\pi_{(1)}^{(C)}$ and $\pi_{(2)}^{(C)}$, formula (7) and the fact that the overlap $\pi_{(1)\cap(2)}^{(C)}$ is within the interval $[0,1]$ naturally yield restrictions on the possible values of the components’ correlation $\rho$ (confer Appendix B). Choices of $\rho$ in line with the above settings were $\{0.2, 0.4, 0.6\}$.

#### 3.1.2. Specific settings for power estimation

The baseline value for the success probability for the first endpoint was set to $\pi_{(1)}^{(C)} = 0.4$ and as above $\pi_{(2)}^{(C)}$ was defined by $\pi_{(2)}^{(C)} = \pi_{(1)}^{(C)} + 0.1$. It was assumed that treatment causes a reduction in incidence of events, while reduction was supposed to be stronger in the less severe (and more frequent) endpoint:
\[ \pi_T^{(1)} = \pi_C^{(1)}(1 - r) \quad \text{and} \quad \pi_T^{(2)} = \pi_C^{(2)}(1 - (r + 0.03)). \]

The reduction \( r \) varied among values \{0.05, 0.2, 0.3, 0.4, 0.5, 0.8\}. Choices of \( \rho \) in line with the chosen settings for power estimation were \{0.2, 0.4, 0.6, 0.8\}.

### 3.2. Simulation results

Table 1 displays the estimated (mean) FWER for the proposed method as well as for the comparison methods RK1 (with zero overlap assumption), RK2 (using the true overlap) of Rauch and Kieser and the Bonferroni approach. The proposed method assures a strict FWER control for almost all parameter constellations asymptotically. At single constellations, the estimated FWER exceeds the value 0.05 slightly. However, throughout all parameter constellations the FWER level of \( \alpha = 0.05 \) lies within the respective 95% confidence interval. The same observation can be made for the methods RK1 and RK2. At the same time, all three approaches are notably less conservative than Bonferroni adjustment. Supplementary Table 1 shows the reference significance levels \( \alpha_{w,\text{adj}} \) for each method and the weight vector \( w = (0.25, 0.25, 0.5) \). The Bonferroni level is equal to the nominal alpha of 0.05 and depicts a lower bound for reference significance levels.

The results from power simulations for different parameter constellations are shown in Tables 2 and 3. Table 4 summarizes the chosen parameter constellations, where \( \pi_T^{(1)} \) and \( \pi_T^{(2)} \) are calculated according to formula (8). The overlap rates for control and treatment group are provided as well. The respective reference significance levels for each considered method can be inferred from Supplementary Table 2.

For the component endpoints and the composite the proposed method and RK2 show comparable performance for detecting a treatment effect, across all parameter constellations. The simplified method RK1 looses efficiency by ignoring present overlap and particularly drops power for high endpoint correlations. However, all three approaches show an improvement as compared to standard Bonferroni adjustment, and the gain in power increases for high correlations as well as for low treatment effects. For instance, for a correlation of 0.8 and treatment effect of \( r = 0.05 \) (in the first component endpoint) and in a regime in which asymptotics can be assumed to hold \((n = 200, 500)\), the proposed method achieves a gain in power of more than 50% (corresponding to an absolute gain of 0.027) for the CE, as compared to standard Bonferroni correction. In the same setting, but with weakly correlated endpoints there is still a gain of 12–16% in the composite endpoint.

These findings are also supported by Supplementary Table 2 showing the strong increase of the reference significance levels as compared to the nominal level of 0.05.

### Table 1. Estimated FWER for the method suggested by Rauch and Kieser with (RK1) and without (RK2) zero overlap assumption, for the proposed method (PrM) and for the Bonferroni approach (Bonf). The sample size \( n \) corresponds to the sample size in each of the two groups.

| \( n \) | \( \rho \) | \( \pi_T^{(1)} \) | RK1 | RK2 | PrM | Bonf |
|------|-------|-------|-----|-----|-----|------|
| 50   | 0.2   | 0.1   | 0.048 | 0.049 | 0.050 | 0.038 |
|      | 0.4   | 0.44  | 0.045 | 0.049 | 0.047 | 0.041 |
| 100  | 0.2   | 0.1   | 0.044 | 0.048 | 0.049 | 0.033 |
|      | 0.4   | 0.44  | 0.045 | 0.052 | 0.052 | 0.033 |
| 200  | 0.2   | 0.4   | 0.046 | 0.048 | 0.047 | 0.041 |
|      | 0.4   | 0.041 | 0.050 | 0.050 | 0.050 | 0.033 |
| 500  | 0.2   | 0.1   | 0.046 | 0.048 | 0.048 | 0.038 |
|      | 0.4   | 0.050 | 0.052 | 0.053 | 0.053 | 0.046 |
|      | 0.4   | 0.044 | 0.053 | 0.054 | 0.054 | 0.037 |

For the component endpoints and the composite the proposed method and RK2 show comparable performance for detecting a treatment effect, across all parameter constellations. The simplified method RK1 looses efficiency by ignoring present overlap and particularly drops power for high endpoint correlations. However, all three approaches show an improvement as compared to standard Bonferroni adjustment, and the gain in power increases for high correlations as well as for low treatment effects. For instance, for a correlation of 0.8 and treatment effect of \( r = 0.05 \) (in the first component endpoint) and in a regime in which asymptotics can be assumed to hold \((n = 200, 500)\), the proposed method achieves a gain in power of more than 50% (corresponding to an absolute gain of 0.027) for the CE, as compared to standard Bonferroni correction. In the same setting, but with weakly correlated endpoints there is still a gain of 12–16% in the composite endpoint.

These findings are also supported by Supplementary Table 2 showing the strong increase of the reference significance levels as compared to the nominal level of 0.05.
### Table 2. Estimated power for the method suggested by Rauch and Kieser with (RK1) and without (RK2) zero overlap assumption, for the proposed method (PrM) and for Bonferroni adjustment (Bonf). The sample size $n = 50,100$ corresponds to the sample size in each of the two groups.

| $n$ | $\rho$ | Endpoint $r$ | $Y^{(1)}$ | $Y^{(2)}$ | $Y^{(3)} = \max(Y^{(1)}, Y^{(2)})$ |
|-----|--------|--------------|----------|----------|-------------------|
| 50  | 0.2    | 0.05         | 0.016    | 0.018    | 0.018             |
|     |        |              | 0.202    | 0.205    | 0.205             |
|     |        |              | 0.019    | 0.020    | 0.020             |
|     |        |              | 0.018    | 0.018    | 0.018             |
|     |        |              | 0.037    | 0.037    | 0.037             |
| 0.2 | 0.05   | 0.019        | 0.019    | 0.020    | 0.020             |
|     |        |              | 0.019    | 0.020    | 0.020             |
|     |        |              | 0.019    | 0.020    | 0.020             |
|     |        |              | 0.037    | 0.037    | 0.037             |
| 0.2 | 0.2    | 0.023        | 0.019    | 0.020    | 0.020             |
|     |        |              | 0.019    | 0.020    | 0.020             |
|     |        |              | 0.019    | 0.020    | 0.020             |
|     |        |              | 0.037    | 0.037    | 0.037             |
| 0.4 | 0.05   | 0.015        | 0.015    | 0.016    | 0.016             |
|     |        |              | 0.016    | 0.017    | 0.017             |
|     |        |              | 0.016    | 0.017    | 0.017             |
|     |        |              | 0.033    | 0.033    | 0.033             |
| 0.6 | 0.05   | 0.017        | 0.017    | 0.018    | 0.018             |
|     |        |              | 0.018    | 0.019    | 0.019             |
|     |        |              | 0.018    | 0.019    | 0.019             |
|     |        |              | 0.033    | 0.033    | 0.033             |
| 0.8 | 0.05   | 0.022        | 0.022    | 0.023    | 0.023             |
|     |        |              | 0.023    | 0.024    | 0.024             |
|     |        |              | 0.023    | 0.024    | 0.024             |
|     |        |              | 0.035    | 0.035    | 0.035             |
| 100 | 0.2    | 0.019        | 0.019    | 0.020    | 0.020             |
|     |        |              | 0.020    | 0.021    | 0.021             |
|     |        |              | 0.020    | 0.021    | 0.021             |
|     |        |              | 0.034    | 0.034    | 0.034             |
| 0.4 | 0.05   | 0.018        | 0.018    | 0.020    | 0.020             |
|     |        |              | 0.019    | 0.021    | 0.021             |
|     |        |              | 0.019    | 0.021    | 0.021             |
|     |        |              | 0.034    | 0.034    | 0.034             |
| 0.8 | 0.05   | 0.023        | 0.023    | 0.024    | 0.024             |
|     |        |              | 0.024    | 0.025    | 0.025             |
|     |        |              | 0.024    | 0.025    | 0.025             |
|     |        |              | 0.036    | 0.036    | 0.036             |
| 0.6 | 0.05   | 0.018        | 0.018    | 0.020    | 0.020             |
|     |        |              | 0.019    | 0.021    | 0.021             |
|     |        |              | 0.019    | 0.021    | 0.021             |
|     |        |              | 0.034    | 0.034    | 0.034             |
| 0.8 | 0.05   | 0.023        | 0.023    | 0.024    | 0.024             |
|     |        |              | 0.024    | 0.025    | 0.025             |
|     |        |              | 0.024    | 0.025    | 0.025             |
|     |        |              | 0.036    | 0.036    | 0.036             |

### 4. Data examples

#### 4.1. Example 1: Testing and treating osteoporosis

In a randomized controlled trial Yuksel et al. (2010) investigated the effect of a screening program conducted by community pharmacists on testing and treatment of osteoporosis. Patients assigned to the control group received usual osteoporosis care. Intervention entailed tailored osteoporosis education by a community pharmacist, quantitative heel ultrasound measurement and subsequent discussion of results.
The study comprised a total number of 262 patients with 133 patients being allocated to the control group and the remaining 129 patients being assigned to the intervention group. The primary endpoint $Y(3)$ was the composite of the secondary endpoints $Y(1)$ (obtaining a bone mineral density (BMD) test) and $Y(2)$ (receiving osteoporosis treatment), i.e., $Y(3) = \max\{Y(1), Y(2)\}$. The empirical Pearson correlation between both component endpoints was $\hat{\rho} = 0.4$.

The study found a distinct effect of intervention on the number of patients tested or treated for osteoporosis (22% observed events in intervention group as opposed to 11% in control.

| $n$ | $\rho$ | $\gamma$ | Endpoint | $\gamma(1)$ | $\gamma(2)$ | $\gamma(3) = \max\{\gamma(1), \gamma(2)\}$ |
|-----|-------|---------|----------|----------|----------|-------------------|
| 200 | 0.2   | 0.05    |          | 0.023   | 0.025   | 0.019            |
| 0.2 |       |         |          | 0.021   | 0.025   | 0.020            |
| 0.3 |       |         |          | 0.054   | 0.057   | 0.050            |
| 0.4 |       |         |          | 0.845   | 0.849   | 0.831            |
| 0.5 |       |         |          | 0.978   | 0.979   | 0.974            |
| 0.6 |       |         |          | 1.000   | 1.000   | 1.000            |
| 0.7 |       |         |          | 1.000   | 1.000   | 1.000            |
| 0.8 |       |         |          | 1.000   | 1.000   | 1.000            |
| 0.9 |       |         |          | 1.000   | 1.000   | 1.000            |
| 1.0 |       |         |          | 1.000   | 1.000   | 1.000            |

### Table 3: Estimated power for the method suggested by Rauch and Kieser with (RK1) and without (RK2) zero overlap assumption, for the proposed method (PrM) and for Bonferroni adjustment (Bonf). The sample size $n = 200$, 500 corresponds to the sample size in each of the two groups.

| $n$ | $\rho$ | $\gamma$ | Endpoint | $\gamma(1)$ | $\gamma(2)$ | $\gamma(3) = \max\{\gamma(1), \gamma(2)\}$ |
|-----|-------|---------|----------|----------|----------|-------------------|
| 0.2 |       |         |          | 0.023   | 0.026   | 0.022            |
| 0.3 |       |         |          | 0.215   | 0.225   | 0.198            |
| 0.4 |       |         |          | 0.541   | 0.561   | 0.524            |
| 0.5 |       |         |          | 0.975   | 0.978   | 0.972            |
| 0.6 |       |         |          | 1.000   | 1.000   | 1.000            |
| 0.7 |       |         |          | 1.000   | 1.000   | 1.000            |
| 0.8 |       |         |          | 1.000   | 1.000   | 1.000            |
| 0.9 |       |         |          | 1.000   | 1.000   | 1.000            |
| 1.0 |       |         |          | 1.000   | 1.000   | 1.000            |

| $n$ | $\rho$ | $\gamma$ | Endpoint | $\gamma(1)$ | $\gamma(2)$ | $\gamma(3) = \max\{\gamma(1), \gamma(2)\}$ |
|-----|-------|---------|----------|----------|----------|-------------------|
| 0.2 |       |         |          | 0.024   | 0.028   | 0.020            |
| 0.3 |       |         |          | 0.216   | 0.238   | 0.200            |
| 0.4 |       |         |          | 0.531   | 0.564   | 0.512            |
| 0.5 |       |         |          | 0.846   | 0.863   | 0.833            |
| 0.6 |       |         |          | 0.977   | 0.981   | 0.973            |
| 0.7 |       |         |          | 1.000   | 1.000   | 1.000            |
| 0.8 |       |         |          | 1.000   | 1.000   | 1.000            |
| 0.9 |       |         |          | 1.000   | 1.000   | 1.000            |
| 1.0 |       |         |          | 1.000   | 1.000   | 1.000            |

The study comprised a total number of 262 patients with 133 patients being allocated to control group and the remaining 129 patients being assigned to the intervention group. The primary endpoint $Y(3)$ was the composite of the secondary endpoints $Y(1)$ (obtaining a bone mineral density (BMD) test) and $Y(2)$ (receiving osteoporosis treatment), i.e., $Y(3) = \max\{Y(1), Y(2)\}$. The empirical Pearson correlation between both component endpoints was $\hat{\rho} = 0.4$. The study found a distinct effect of intervention on the number of patients tested or treated for osteoporosis (22% observed events in intervention group as opposed to 11% in control.
group, \( p \)-value of \( p = 0.0216 \) using a \( \chi^2 \) distribution with 1 degree of freedom). However, the result was largely driven by the first endpoint describing the event of BMD testing (22% of the patients in treatment group received BMD testing while 10% in the control group were tested).

Using the proposed approach, we considered simultaneous assessment of the treatment effect based on the composite endpoint and its two constituents (assuming a FWER of \( \alpha = 0.05 \)).

The results of the simultaneous analysis are shown in Table 5: The second column displays the estimated odds ratios (OR), followed by the respective multiplicity adjusted 95% confidence intervals for both equally and unequally weighted endpoints. The subsequent column provides the unadjusted (marginal) \( p \)-values. The local significance levels for the proposed method and for the Bonferroni correction are presented in the latter columns. Here two sets of weights have been applied, \( w_1 = w_2 = w_3 = 1/3 \) (equal weights) and \( w_1 = w_2 = 1/4 \) and \( w_3 = 1/2 \) (unequal weights with double weight on the composite endpoint). Observe how prioritizing the composite (by assigning the weight \( w_3 = 1/2 \)) leads to a shortening of its OR confidence interval.

### Table 4. Overview of the chosen parameter constellations and the (thereby implicitly defined) overlap rates.

| \( \rho \) | \( r \) | \( n_{C}^{(1)} \) | \( n_{L}^{(1)} \) | \( n_{C}^{(2)} \) | \( n_{L}^{(2)} \) | \( n_{C}^{(1)} \cup n_{L}^{(2)} \) | \( n_{L}^{(1)} \cup n_{C}^{(2)} \) |
|---|---|---|---|---|---|---|---|
| 0.2 | 0.4 | 0.38 | 0.5 | 0.460 | 0.2490 | 0.2232 |
| 0.2 | 0.4 | 0.32 | 0.5 | 0.385 | 0.2490 | 0.1666 |
| 0.3 | 0.4 | 0.28 | 0.5 | 0.335 | 0.2490 | 0.1362 |
| 0.4 | 0.4 | 0.24 | 0.5 | 0.285 | 0.2490 | 0.1070 |
| 0.5 | 0.4 | 0.20 | 0.5 | 0.235 | 0.2490 | 0.0809 |
| 0.8 | 0.4 | 0.08 | 0.5 | 0.085 | 0.2490 | 0.0219 |
| 0.4 | 0.05 | 0.4 | 0.38 | 0.5 | 0.460 | 0.2490 | 0.2716 |
| 0.2 | 0.4 | 0.32 | 0.5 | 0.385 | 0.2490 | 0.2140 |
| 0.3 | 0.4 | 0.28 | 0.5 | 0.335 | 0.2490 | 0.1786 |
| 0.4 | 0.4 | 0.24 | 0.5 | 0.285 | 0.2490 | 0.1455 |
| 0.5 | 0.4 | 0.20 | 0.5 | 0.235 | 0.2490 | 0.1148 |
| 0.8 | 0.04 | 0.4 | 0.08 | 0.5 | 0.085 | 0.2490 | 0.0371 |
| 0.6 | 0.05 | 0.4 | 0.38 | 0.5 | 0.460 | 0.3470 | 0.3199 |
| 0.2 | 0.4 | 0.32 | 0.5 | 0.385 | 0.3470 | 0.2594 |
| 0.3 | 0.4 | 0.28 | 0.5 | 0.335 | 0.3470 | 0.2210 |
| 0.4 | 0.4 | 0.24 | 0.5 | 0.285 | 0.3470 | 0.1841 |
| 0.5 | 0.4 | 0.20 | 0.5 | 0.235 | 0.3470 | 0.1488 |
| 0.8 | 0.4 | 0.08 | 0.5 | 0.085 | 0.3470 | 0.0522 |
| 0.8 | 0.05 | 0.4 | 0.38 | 0.5 | 0.460 | 0.3960 | 0.3683 |
| 0.2 | 0.4 | 0.32 | 0.5 | 0.385 | 0.3960 | 0.3048 |
| 0.3 | 0.4 | 0.28 | 0.5 | 0.335 | 0.3960 | 0.2633 |
| 0.4 | 0.4 | 0.24 | 0.5 | 0.285 | 0.3960 | 0.2226 |
| 0.5 | 0.4 | 0.20 | 0.5 | 0.235 | 0.3960 | 0.1827 |
| 0.8 | 0.4 | 0.08 | 0.5 | 0.085 | 0.3960 | 0.0673 |

### Table 5. Simultaneous assessment of the composite endpoint and its constituents. The second column contains the estimated odds ratios, followed by the corresponding multiplicity adjusted 95% confidence intervals for the proposed method and both weightings. The subsequent column shows the raw \( p \)-values (marginal models) and the adjusted local significance levels \( \alpha_{j, \text{adj}} = \frac{w_j \alpha_j}{\sum w_j \alpha_j} \) for the proposed method and the Bonferroni correction are given in the right-most columns.

| Endpoint | OR | 95% confidence interval [\( \ell, \uparrow \)] | \( p_j \) | Proposed method | Proposed method | Bonferroni |
|---|---|---|---|---|---|---|
| Obtaining BMD test | 2.56 | [1.25, 5.25] | [1.19, 5.49] | 0.0094 | 0.0236 | 0.0173 | 0.0167 | 0.0125 |
| Receiving osteop. trt. | 2.11 | [0.51, 8.80] | [0.46, 9.63] | 0.2973 | 0.0236 | 0.0173 | 0.0167 | 0.0125 |
| Composite endpoint | 2.36 | [1.17, 4.76] | [1.24, 4.48] | 0.0155 | 0.0236 | 0.0345 | 0.0167 | 0.0250 |

Obtaining BMD test 2.56 [1.25, 5.25] [1.19, 5.49] 0.0094 0.0236 0.0173 0.0167 0.0125
Receiving osteop. trt. 2.11 [0.51, 8.80] [0.46, 9.63] 0.2973 0.0236 0.0173 0.0167 0.0125
Composite endpoint 2.36 [1.17, 4.76] [1.24, 4.48] 0.0155 0.0236 0.0345 0.0167 0.0250
4.2. Example 2: Testing and treating osteoporosis with many-to-one comparisons

Related to the previous example is the randomized controlled trial conducted by Solomon et al. (2007). The study targeted the assessment of three different educational strategies on osteoporosis management. Primary care physicians and their patients were randomly allocated to control group (Treatment A, usual care) or one of three intervention groups, patient intervention only (Treatment B), physician intervention only (Treatment C), or both interventions (Treatment D). Intervention comprised a thorough education in osteoporosis diagnosis, prevention, BMD testing and treatment. The primary endpoint \( Y^{(3)} \) was defined by means of the secondary endpoints \( Y^{(1)} \) (undergoing a BMD test) and \( Y^{(2)} \) (obtaining osteoporosis medication) in the same manner as in the first example. The empirical Pearson correlation between the component endpoints was \( \hat{\rho} = 0.4 \).

To investigate which type of intervention has a significant effect on osteoporosis management, many-to-one comparisons (Dunnett contrasts) to the control group (usual care) were calculated. The assessment of intervention effect is conducted simultaneously for the composite endpoint and its components.

For the analysis, both equal \( \alpha \) allocation and weighted allocation among the three endpoints were considered. For the latter, the two individual endpoints were assigned the weights \( w_1 = w_2 = 1/4 \) while double weight was allocated to the composite, \( w_3 = 1/2 \). Within a single endpoint, all comparisons to the control group were weighted equally. Such allocation resulted in weights \( w_{iAB} = w_{iAC} = w_{iAD} = \frac{1}{3} \cdot \frac{1}{4}, i \in \{1, 2\} \), for the first and second endpoint while the third endpoints was weighted according to \( w_{3AB} = w_{3AC} = w_{3AD} = \frac{1}{3} \cdot \frac{1}{2} \).

Table 6 shows the results of the simultaneous analysis. The estimated odds ratios with multiplicity adjusted 95% confidence intervals based on the proposed approach, for both equally and unequally weighted endpoints, are given in the first columns. Unadjusted \( p \)-values from the marginal models and each comparison to the control are listed in the subsequent column while the latter four columns display local significance levels based on equal and unequal weights, both for the proposed approach and for Bonferroni adjustment. Again observe that the chosen weighting of endpoints leads to shorter confidence intervals regarding all many-to-one comparisons for the composite.

5. Discussion

We have proposed a flexible and versatile approach for simultaneous inference of a binary composite endpoint and its individual binary components. The proposed method controls the FWER asymptotically and shows a marked increase in power for strongly correlated components as compared to standard Bonferroni adjustment and the approach proposed by Rauch and Kieser (2012), which is based on a related maximum test. The gain in efficiency is most pronounced for small sample sizes and small treatment effects. Moreover, the proposed approach gave results similar to what was obtained by using a common gatekeeping procedure, where the test based on the composite outcome serves as the gatekeeper, deciding whether or not to proceed with testing the individual outcomes, Dmitrienko et al. (2010). A comparison of the estimated power for composite endpoint and its constituents, for the gatekeeping approach and the proposed method, can be made on the basis of simulation results as shown in Supplementary Table 3 and Supplementary Table 4. A further interesting extension of our simulation study would be a comparison to exact, finite sample size methods, such as a permutation-based global gatekeeping approach.

The advantages of our method are (i) its simplicity, particularly an explicit formulation of the correlation matrix is not needed, (ii) the availability of convenient software through the function \texttt{mmm()} in the \textit{R} package mult-comp (Hothorn et al., 2008), (iii) flexibility by providing both multiplicity-adjusted \( p \)-values and simultaneous confidence intervals, (iv) full flexibility for the individual models (per endpoint) to include different covariates next to the treatment variable of
interest (e.g., adjustment for baseline imbalance) and (v) simultaneous assessment of differently scaled endpoints (e.g. proportions and continuous). Moreover, the overall significance level may be split according to pre-specified weights following the approach by Westfall and Soper (1998).

It is a limitation that the proposed method relies on asymptotic properties and, in particular, strong FWER control may not hold for small sample sizes as it is only guaranteed to hold asymptotically (Pipper et al., 2012). However, the simulation results indicate that the method may in practice still be useful for quite small sample sizes and for specific applications it may be recommendable to explore the possible inflation of FWER due to finite sample sizes via simulations. Although the proposed method does not involve any a priori assumptions or specifications of the covariance structure, using the proposed method for sample size calculations will require a priori specification of correlations. This may be a limitation in practice.

The proposed methodology may also be extended to composite endpoints used in cardiovascular trials but also increasingly seen in studies on rheumatoid arthritis, asthma and diabetes (Unnikrishnan et al., 2013). Cardiovascular death, nonfatal myocardial infarction (MI) and nonfatal stroke are widely accepted primary endpoints in cardiovascular studies. While the outcome death has typically a low incidence, nonfatal components such as non-fatal stroke and nonfatal myocardial infarction have substantially higher event rates. Therefore, using the time to first event of death, MI, and stroke as a composite endpoint would result in an increased event rate (Yusuf et al., 2000). Generally, we may consider composite end-points of the form \( \min_{j=1, \ldots, J} T_j \), where \( T_1, \ldots, T_J \) denote the individual event times (which of course may be subject to censoring). This proposed method extends in a straightforward way to such endpoints.

### References

Chi, G. Y. (2005). Some issues with composite endpoints in clinical trials. Fundamental and Clinical Pharmacology 19 (6):609–619.

Cordoba, G., Schwartz, L., Woloshin, S., Bae, H., Getzsche, P. C. (2010). Definition, reporting, and interpretation of composite outcomes in clinical trials: Systematic review. British Medical Journal 341:c3920.

Dickhaus, T. (2014). Simultaneous Statistical Inference: With Applications in the Life Sciences. Berlin: Springer.

Dmitrienko, A., Tamhane, A. C., Bretz, F. (2010). Multiple Testing Problems in Pharmaceutical Statistics. Boca Raton: Chapman & Hall/CRC.

Hasler, M., Hothorn, L. A. (2011). A Dunnett-type procedure for multiple endpoints. The International Journal of Biostatistics 7(1):1–15.

---

**Table 6.** Simultaneous assessment of the composite endpoint and its components. The first columns of numbers provide the estimated odds ratios with multiplicity adjusted 95% confidence intervals for the proposed method (in case of both equally and unequally weighted endpoints). The subsequent column contains the raw \( p \)-values. The adjusted local significance levels \( \alpha_{w,adj} = w_1 \alpha_{w,adj} \) for the proposed method and the Bonferroni correction in case of equally and unequally weighted endpoints are shown as well.

| Endpoint                  | Many-to-one comparison OR | 95% confidence interval [\( \hat{\theta}_{ij} \)] | \( p_{\text{unadj}} \) | Adjusted local significance levels \( \alpha_{w,adj} = w_1 \alpha_{w,adj} \) |
|---------------------------|---------------------------|-----------------------------------------------|--------------------------|-------------------------------------------------|
|                           |                           |                                               |                          | Proposed method Proposed method Bonferroni        |
|                           |                           |                                               |                          | Equal Unequal Equal Unequal Equal Unequal        |
| Proposed method           |                           |                                               |                          | 0.0693 0.0696 0.06 0.05                           |
|                           |                           |                                               |                          | Proposed method Proposed method Bonferroni        |
|                           |                           |                                               |                          | Equal Unequal Equal Unequal Equal Unequal        |
|                           |                           |                                               |                          | 0.0077 0.0058 0.0056 0.0042                       |
|                           |                           |                                               |                          | 0.0077 0.0058 0.0056 0.0042                       |
|                           |                           |                                               |                          | 0.0077 0.0058 0.0056 0.0042                       |
|                           |                           |                                               |                          | 0.0077 0.0058 0.0056 0.0042                       |
| Obtaining BMD test        | B-A                       | 0.77 [0.60, 0.98]                             | 0.0081                   | 0.0077 0.0058 0.0056 0.0042                       |
|                           | C-A                       | 0.89 [0.70, 1.13]                             | 0.2477                   | 0.0077 0.0058 0.0056 0.0042                       |
|                           | D-A                       | 1.00 [0.79, 1.26]                             | 0.9933                   | 0.0077 0.0058 0.0056 0.0042                       |
| Receiving osteop. trt.    | B-A                       | 0.73 [0.57, 0.94]                             | 0.0025                   | 0.0077 0.0058 0.0056 0.0042                       |
|                           | C-A                       | 1.12 [0.89, 1.41]                             | 0.2411                   | 0.0077 0.0058 0.0056 0.0042                       |
|                           | D-A                       | 0.97 [0.77, 1.23]                             | 0.7762                   | 0.0077 0.0058 0.0056 0.0042                       |
| Composite endpoint        | B-A                       | 0.78 [0.64, 0.95]                             | 0.0023                   | 0.0077 0.0117 0.0056 0.0083                       |
|                           | C-A                       | 1.02 [0.84, 1.24]                             | 0.8021                   | 0.0077 0.0117 0.0056 0.0042                       |
|                           | D-A                       | 1.00 [0.83, 1.21]                             | 0.9917                   | 0.0077 0.0117 0.0056 0.0083                       |
Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. Biometrical Journal, 50(3):346–363.

Huque, M. F., Alosh, M., Bhore, R. (2011). Addressing multiplicity issues of a composite endpoint and its components in clinical trials. Journal of Biopharmaceutical Statistics 21(4):610–634.

Lim, E., Brown, A., Helmy, A., Mussa, S., Altman, D. G. (2008). Composite outcomes in cardiovascular research: A survey of randomized trials. Annals of Internal Medicine 149(9):612–617.

Mascha, E. J., Sessler, D. I. (2011). Design and analysis of studies with binary-event composite endpoints: Guidelines for anesthesia research. Anesthesia and Analgesia 112(6):1461–1471.

McCullagh, P., Nelder, J. A. (1989). Generalized Linear Models, 2 ed. London: Chapman & Hall.

Pipper, C. B., Ritz, C., Bisgaard, H. (2012). A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. Journal of the Royal Statistical Society: Series C (Applied Statistics) 61(2):315–326.

Quan, H., Zhang, D., Zhang, J., Devlaminck, L. (2007). Analysis of a binary composite endpoint with missing data in components. Statistics in Medicine 26(26):4703–4718.

R Core Team. 2014 A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. Available at. http://www.r-project.org

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

Ross, S. (2007). Composite outcomes in randomized clinical trials: Arguments for and against. American, Journal of Obstetrics and Gynecology 196(2):119.e1-119.e6.

Solomon, D. H., Katz, J. N., Finkelstein, J. S., Polinski, J. M., Stedman, M., Brookhart, M. A., Arnold, M., Gauthier, S., Avorn, J. (2007). Osteoporosis improvement: A large-scale randomized controlled trial of patient and primary care physician education. Journal of Bone and Mineral Research 22(11):1808–1815.

Tomlinson, G., Detsky, A. S. (2010). Composite end points in randomized trials: There is no free lunch. JAMA 303(3):267–268.

Turk, D. C., Dworkin, R. H., McDermott, M. P., Bellamy, N., Burke, L. B., Chandler, J. M., Cleeland, C. S., Cowan, P., Dimitrova, R., Farrar, J. T., Hertz, S., Heyse, J. F., Iyengar, S., Jadad, A. R., Jay, G. W., Jermano, J. A., Katz, N. P., Manning, D. C., Martin, S., Max, M. B., McGrath, P., McQuay, H. J., Quesy, S., Rappaport, B. A., Revicki, D. A., Rothman, M., Stauffer, J. W., Svensson, O., White, R. E., & Witter, J. (2008). Analyzing multiple endpoints in clinical trials of pain treatments: Impact recommendations. Pain 139(3):485–493.

Unnikrishnan, A. G., Bhattacharyya, A., Baruah, M. P., Sinha, B., Dhar-malingam, M., Rao, P. V. (2013). Importance of achieving the composite endpoints in diabetes. Indian Journal of Endocrinology and Metabolism 17(5):835–843.

Westfall, P. Soper, K. (1998). Weighted multiplicity adjustments for animal carcinogenicity tests. Journal of Biopharmaceutical Statistics 8(1):23–44.

Yuksel, N., Majumdar, S., Biggs, C., Tsuyuki, R. (2010). Community pharmacist-initiated screening program for osteoporosis: Randomized controlled trial. Osteoporosis International, 21(3):391–398.

Yusuf, S., Sleight, P., Pogue, J., Bosch, J., Davies, R., Dagenais, G. (2000). Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. The heart outcomes prevention study investigators. The New England Journal of Medicine 342(3):145–153.

Appendix A: R code for example 1

```r
## Define functions needed for calculation of local significance levels
## Corr: correlation matrix
## wei: vector of weights

alphaFct <- function(alpha, Corr, wei) {
  q <- sapply(wei, function(x) qnorm(1-x*alpha/2))
  1 - as.numeric(pmvnorm(lower=-q, upper=q, mean=rep(0,length(q)),
                        corr=Corr))
}

adjAlphaFct <- function(Corr, wei = rep(1/nrow(Corr), nrow(Corr)), level = {
  # optimization on log scale for convenience
  objFct <- function(x) {log((alphaFct(x, Corr, wei) - level) + 1)}
  adjAlpha <- uniroot(objFct, interval = c(0.04, 1)) [["root"]]
  localAlphas <- adjAlpha * wei
  localAlphas
}
```
## Recover data from Yuksel et al. (2010)
## endpoint 1: BMD
## endpoint 2: Osteoporosis treatment

# control group
control <- matrix(0, 133, 4)
control[2:14, 2] <- 1
control[1:3, 3] <- 1
control[, 4] <- pmax(control[, 2], control[, 3])

# treatment group
treat <- matrix(0, 129, 4)
treat[, 1] <- 1
treat[1:28, 2] <- 1
treat[1:6, 3] <- 1
treat[, 4] <- pmax(treat[, 2], treat[, 3])
df <- as.data.frame(rbind(control, treat))
colnames(df) <- c("group", "BMD", "Osteo", "CE")
df$group <- as.factor(df$group)

## Logistic regression for each endpoint
m1 <- glm(BMD ~ group, data = df, family = binomial)
m2 <- glm(Osteo ~ group, data = df, family = binomial)
m3 <- glm(CE ~ group, data = df, family = binomial)

## Extracting raw p-values
p.raw <- round(c(coef(summary(m1))[2, 4], coef(summary(m2))[2, 4], coef(summary(m3))[2, 4]), 4)

## Simultaneous analysis: comparison of treatment group to control group library (multcomp)
fit <- glht(mmm(BMD=m1, Osteo=m2, CE=m3), mlf(mcp(group = "Dunnett")))

## Extract correlation matrix for parameters of interest
Sigma <- vcov(fit)
Corr <- cov2cor(Sigma)

## weighted alphas (weights: w1 = w2 = 0.25, w3 = 0.5)
level <- 0.05
wei <- c(0.25, 0.25, 0.5)
adj.alphas.unequal <- adjAlphaFct(Corr, wei)

## equally weighted alphas
wei <- rep(1/3, 3)
adj.alphas.equal <- adjAlphaFct(Corr, wei)

## calculate confidence intervals
betas <- coef(fit)
lower <- betas - qnorm(1 - adj.alphas.equal) * sqrt(diag(Sigma))
upper <- betas + qnorm(1 - adj.alphas.equal) * sqrt(diag(Sigma))
lower.unequal <- betas - qnorm(1 - adj.alphas.unequal) * sqrt(diag(Sigma))
upper.unequal <- betas + qnorm(1 - adj.alphas.unequal) * sqrt(diag(Sigma))

## transform confidence intervals to Odds Ratio
OR <- exp(betas)
lowerOR <- exp(lower)
upperOR <- exp(upper)
lowerOR.unequal <- exp(lower.unequal)
upperOR.unequal <- exp(upper.unequal)
Appendix B: Generation of two binary random variables with given correlation

The correlation \( \rho \) between two Bernoulli distributed random variables \( X, Y \) with success probabilities \( p_X \) and \( p_Y \), respectively, is given by:

\[
\rho = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}} = \frac{\mathbb{E}(XY) - p_X p_Y}{\sqrt{p_X(1-p_X)p_Y(1-p_Y)}} = \frac{p_{X|Y} p_Y - p_X p_Y}{\sqrt{p_X(1-p_X)p_Y(1-p_Y)}},
\]

with \( p_{X|Y} = \mathbb{P}(X = 1|Y = 1) \). This yields the following explicit expression of \( p_{X|Y} \) in terms of \( \rho, p_X, p_Y \):

\[
p_{X|Y} = \frac{\rho \sqrt{p_X(1-p_X)p_Y(1-p_Y)} + p_X p_Y}{p_Y}.
\]

Moreover, the success probability of \( X \) given that \( Y = 0 \) can be calculated as:

\[
p_{X|0} = \mathbb{P}(X = 1|Y = 0) = \frac{p_X - p_Y p_{X|Y}}{1 - p_Y}.
\]

Summing up, the conditional distribution of \( X \) given \( Y \) is \( X|Y \sim \text{Bin}(1, p_{X|Y}) \), where the conditional success probability is defined as:

\[
p_{X|Y} = p_{X|0}(1 - Y) + p_{X|1} Y.
\]

Simulation of two Bernoulli random variables with predefined success probabilities \( p_X, p_Y \) and a given correlation \( \rho \) can thus be performed as follows:

```r
simulate Y
simulate X from Bin(1, p_{X|Y})
end
```

One has to ensure that \( p_{X|0}, p_{X|1} \) are valid probabilities, in other words one has to guarantee that \( p_{X|0}, p_{X|1} \in [0,1] \). This condition translates to the following restriction on the triple \( (p_X, p_Y, \rho) \):

\[
\max \left\{ -\frac{p_X p_Y}{\sqrt{(1-p_X)(1-p_Y)}}, \sqrt{\frac{(1-p_X)(1-p_Y)}{p_X p_Y}} \right\} \leq \rho \leq \min \left\{ \sqrt{\frac{p_X(1-p_Y)}{(1-p_X)p_Y}}, \sqrt{\frac{p_Y(1-p_X)}{(1-p_Y)p_X}} \right\}.
\]