Multiplicity Adjustment for Composite Binary Endpoints

G. Rauch, M. Kieser
Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany

Keywords
Binary endpoints, clinical trials, composite endpoints, correlation matrix, multiple testing

Summary
Background: Binary composite outcome measures are increasingly used as primary endpoints in clinical trials. Composite endpoints combine several events of interest within a single variable. However, as the effect observed for the composite does not necessarily reflect the effects for the individual components, it is recommended in the literature to additionally evaluate each component separately.

Objectives: The task is to define an adequate multiple test procedure which focuses on the composite outcome measure but allows for a confirmatory interpretation of the components in case of large effects.

Methods: In this paper, we determine the correlation matrix for a multiple binary endpoint problem of a composite endpoint and its components based on the normal approximation test statistic for rates. Thereby, we assume multinomial distributed components. We use this correlation to calculate the adjusted local significance levels. We discuss how to use our approach for a more informative formulation of the test problem. Our work is illustrated by two clinical trial examples.

Results: By taking into account the special correlation structure between a binary composite outcome and its components, an adequate multiple test procedure to assess the composite and its components can be defined based on an approximate multivariate normal distribution without much loss in power compared to a test problem formulated exclusively for the composite.

Conclusions: By incorporating the correlation under the null hypotheses, the global power for the multiple test problem assessing both the composite and its components can be increased as compared to simple Bonferroni-adjustment. Thus, a confirmatory analysis of the composite and its components might be possible without a large increase in sample size as compared to a single endpoint problem formulated exclusively for the composite.

1. Introduction
Clinical trials using composite endpoints as primary outcome measures can be found in various medical fields. Composite endpoints are commonly used if the clinical effect of interest cannot directly be assessed by a single specific outcome measure but refers to several endpoints [1]. Moreover, using a composite measure instead of single event variables increases the number of expected events and is thus meant to increase the power or reduce sample size [2–5].

Usually, composite endpoints are defined rather as binary variables or as time-to-event outcomes. We focus on binary event data in this work, where the event is one out of several predefined events of interest, which have to be specified in advance.

If the underlying sample size is large enough, the comparison of two event rates can be done by the standard chi-square test or the equivalent normal approximation test for rates. Composite endpoints defined as event rates can be found in clinical studies which assess response or failure rates, respectively, were response or failure is defined as a combination of several binary criteria.

The interpretation of results for a composite outcome measure can be difficult as the observed effect for the composite does not necessarily reflect the effects for the components [1–7]. Even if a statistically significant and clinically relevant effect in the composite has been observed, it may happen that the effects of some components are of very different magnitude or even show an adverse effect. This is especially a problem, if the single endpoints forming a composite are of different clinical relevance.

Therefore, it is recommended in the corresponding guidelines not to combine hard and soft endpoints and moreover to evaluate each individual component separately, additionally to the composite [2, 5, 8, 9].

The statistical evaluation of the individual components can be done either by descriptive statistical methods or by including them in the primary confirmatory analysis which results in a multiple test problem. Thereby, it is not always necessary to include all components in the confirmatory analysis. In many clinical applications, some components are clearly more relevant than other. It might therefore suffice to assess only the most important components in the confirmatory analysis.

For composite endpoints defined as event rates, the corresponding test problem thus corresponds to a multiple binary endpoint situation. A confirmatory evaluation of the components clearly simplifies the interpretation of the composite effect and provides important supplementary information but yields the need for adjustment of the local significance levels in order to...
control the family wise type I error rate by \( \alpha \). However, as the component event rates and the composite event rates are usually highly correlated, the actually required adjustment is limited.

We will show how the correlation between the composite and its components can be used to define an adequate multiple test procedure which focuses mainly on the composite but additionally allows for a confirmatory interpretation of the components in case of large effects. Our method also implies a direct approach to determine the required sample size for the multiple test problem. Our results will be illustrated by two clinical trial examples.

2. Objectives

2.1 Hypotheses and Test Statistics

Consider a controlled clinical trial with a binary composite endpoint (CE) consisting of \( k \) components (EP\(_1\), \( i = 1, \ldots, k \)). Without loss of generality, assume that the first \( l \leq k \) components refer to the clinically most relevant ones which should be included in the confirmatory analysis. Let \( n \) and \( m \) denote the sample sizes in the control group (C) and in the intervention group (I), respectively. Let \( X^C \) and \( X^I \) denote the random vectors of the event frequencies for the individual components in the control and intervention group, respectively,

\[
X^C = (X^C_{EP1}, \ldots, X^C_{EPk}),
X^I = (X^I_{EP1}, \ldots, X^I_{EPk}),
\]

which are assumed to follow a multinomial distribution

\[
X^C \sim \text{Multi}(n, p^C_{EP1}, \ldots, p^C_{EPk}),
X^I \sim \text{Multi}(m, p^I_{EP1}, \ldots, p^I_{EPk}),
\]

where \( p^C_{EP}, p^I_{EP} \) denote the corresponding event probabilities. The multinomial distribution corresponds to the assumption that there exists no overlap between the individual components, thus the simultaneous occurrence of an event in two components is not possible. Although this assumption does not hold true in general, it will be demonstrated below, that assuming no overlap between the components yields to an underestimated the correlations between the components and is therefore a conservative approach.

Without loss of generality, we assume that lower event rates correspond to more favorite results. The null hypothesis to be assessed in confirmatory analysis thus states that the event rate for the composite in the intervention group (I) is higher or equal to the event rate in the control group (C)

\[
H_0 : p^{I, CE} - p^{I, CE} \leq 0.
\]

This hypothesis can be tested with the standard chi-square test or the equivalent normal approximation test for rates in case that the underlying sample sizes are large enough. Lydersen et al. [10] provide simulation results for power and type I error of the standard chi-square test for different sample sizes and event rate assumptions, which show that even for relatively small sample sizes of 34 patients per group the test preserves its level or is just slightly anti-conservative as long as the underlying event rates do not approach the boundaries 0 or 1. The normal approximation test for rates therefore covers most relevant situations of clinical trials. For more extreme situations with very small sample sizes or event rates close to the boundaries, unconditional tests can be used which have also been discussed by Lydersen et al. [10].

The test statistic of the normal approximation test for rates is given by

\[
T_{CE} = \frac{\hat{p}^{CE} - \hat{p}^{IE}}{\sqrt{\frac{n}{n+m} \cdot \frac{1}{n} \cdot \frac{m}{m+n}}},
\]

with

\[
\hat{p}^{CE} = \frac{n + \hat{p}^{CE} \cdot m}{n + m},
\]

\[
\hat{p}^{IE} = \frac{m + \hat{p}^{IE} \cdot n}{m + n},
\]

where \( \hat{p}^{CE}, \hat{p}^{IE} \) denote the estimates for the composite event rate in the control and in the intervention group, respectively, which are given by the corresponding relative frequencies

\[
\hat{p}^{CE} = \frac{x^{CE} \cdot n}{n}, \quad \hat{p}^{IE} = \frac{x^{IE} \cdot m}{m},
\]

where \( x^{CE}, x^{IE} \) denote the observed absolute event frequencies, which are determined by the random variables

\[
x^{CE} = \sum_{i=1}^{k} x^{CE}_{EPi}, \quad x^{IE} = \sum_{i=1}^{k} x^{IE}_{EPi}
\]

Note that the test statistic \( T_{CE} \) is the square root of the standard chi-square test statistic. Under the null hypothesis of no difference between \( p^{CE} \) and \( p^{IE} \), \( T_{CE} \) is approximately standard normally distributed \( T_{CE} \sim N(0,1) \).

Generally, an adequate interpretation of the treatment effect observed for the composite is only possible if the effects for the single components are also considered. Therefore, it is common practice to report the component rates in addition to the result obtained for the composite. The corresponding null hypotheses for the components

\[
H_0 : p^{CE} - p^{IE} \leq 0
\]

are usually tested simultaneously by the corresponding test statistics

\[
T_{EPi} = \frac{\hat{p}^{CE}_{EPi} - \hat{p}^{IE}_{EPi}}{\sqrt{\frac{n}{n+m} \cdot \frac{1}{n} \cdot \frac{m}{m+n}}},
\]

without taking into account the multiplicity issue. Therefore, the results can only be interpreted descriptively. A confirmatory interpretation of the component effect sizes would only be adequate after an appropriate adjustment for multiplicity.

2.1 Multiplicity Considerations

Adjustment for multiplicity that does not take into account the specific situation at hand and is based on ad hoc methods usually leads to very conservative local significance levels. However, composite endpoints and its components are often highly correlated which can be used to define a less conservative adjustment procedure.
Multiplicity adjustments for multiple endpoints have been widely discussed in the statistical literature. There exists a variety of sequentially rejective approaches based on weighted Bonferroni-type tests. A common example is the standard Bonferroni-Holm approach [11], extensions are given by gatekeeping procedures [12–15], fixed sequence tests [16] and fallback procedures [17]. Westfall [18] discusses a sequentially rejective method for binary outcomes which takes account of the discreteness and which is especially powerful for small sample sizes and low event frequencies.

Bretz et al. [19] proposed a graphical method for the illustration of sequentially rejective test procedures which allow to reallocate the local significance levels of hypotheses which have already been rejected. Applications of these methods to composite endpoints have been recently discussed by Huque et al. [20]. All of these methods intend to maximize power by an optimal exhaustion of the global significance level \( \alpha \). However, in the planning stage of a clinical trial, sample size calculation based on sequentially rejective methods is not straightforward as the local significance level assigned to an individual hypothesis depends on previous test decisions and can therefore only be realized by simulations. For planning purposes, approaches that assign predefined local significance levels to the individual hypotheses can be much easier applied.

James [21] proposed an approximate method to calculate the adjusted local significance levels based on an explicit formula incorporating the correlation between test statistics. Leon and Heo [22] supported the validity of James’ method in simulation studies. However, James’ method only provides an approximation to the true adjusted levels which can be obtained from the multivariate normal distribution. As software programs that allow calculation of the true adjusted levels can be easily obtained in our days, James’ explicit formula no longer provides a true benefit.

One main concern is that adjustment procedures based on the correlation between endpoints are not robust against wrong correlation specifications. In the special case of composite binary endpoints, however, the correlation is of special structure and depends exclusively on the event rates under the corresponding null hypotheses which have to be specified in the planning stage to calculate the required sample size. Thus, if there exists reliable knowledge on the event rate in the control group, the correlation matrix can be deduced directly.

### 3. Methods

#### 3.1 Defining the Test Procedure and Calculating Sample Size

Testing the composite and its components simultaneously seems contra intuitive, as the composite measure was chosen to assess several variables of clinical relevance within a single statistical test. Thus, the focus of the analysis clearly lays on the composite measure. However, in the case of huge effects in the composite and all components, it would be helpful if a confirmatory interpretation of the individual test results is possible. A possible approach would be to formulate a hierarchical test problem were the components are only assessed if the composite shows a significant result.

However, it might happen that the composite fails to be significant as one of the components refers to an adverse effect which reduces the effect in the composite. The remaining components, however, might show quite large effects so that a confirmatory interpretation is desirable. In this situation, a hierarchical approach does not solve the problem of a confirmatory analysis of the components. One might argue, that this situation only arises if the composite outcome measure was not adequately defined in the planning stage, as the treatment should generally positively affect all components. However, such wrong expectations about the outcome are often met in practice as the behavior of the components cannot always be correctly predicted in advance. A clinical study example is given by the CAPRICORN trial [13] where the composite of all-cause mortality and hospital admission failed to be significant but the component all-cause mortality showed a highly positive effect.

The question is thus how to define an adequate test procedure which focuses on the composite outcome but in addition allows for a confirmatory interpretation of the results for the components in case of large effects. The idea is to test the composite and all components simultaneously at different predefined local levels \( \alpha_{CE}^{local} \), \( \alpha_{EP}^{local} \), \( \alpha_{EP_1}^{local} \), \( \alpha_{EP_2}^{local} \) which are defined such that the family wise type I error rate is controlled by \( \alpha \) taking into account the correlation between the composite and its components. Therefore, the local significance level for the composite \( \alpha_{CE}^{local} \) is chosen such that it is close to the family wise level \( \alpha \) and the components are tested at the remaining level.

When the family wise significance level \( \alpha \) has been chosen, the sample size for a clinical trial with a binary composite outcome measure is calculated as follows: The assumed event rates for the single components \( p_{CE_1}^{EP}, \ldots, p_{CE_i}^{EP} \) in the control group yield the event rate assumption for the composite in the control group

\[
p_{CE}^{CE} = \sum_{i=1}^{l} p_{CE_i}^{EP}.
\]

A minimal clinically relevant treatment effect in the composite outcome measure \( \delta \) is defined and the sample size is calculated such that a reduction in the composite to \( p_{CE}^{CE} = p_{CE}^{CE} - \delta \) is detected with adequate power at the predefined significance level \( \alpha_{CE}^{local} \). Note that sample size calculation is exclusively based on the detection of a clinically relevant effect in the composite measure.

Under the assumption that all null hypotheses are true and under additional assumptions about the event rates in the control group for the composite and its components the correlation matrix can be calculated as shown in the following.

#### 3.2 Correlation Between the Test Statistics

Under the null hypotheses assumptions \( H_0: p_{CE}^{CE} - p_{CE}^{CE} \leq 0 \) and \( H_0: p_{CE}^{EP} - p_{CE}^{EP} \leq 0 \), respectively, the effect sizes for the composite and its components equal 0. However, no statement about the position of the effect is made, which is determined by the specifying the corresponding event rates.
Assuming no effect in the composite and all components and further assuming a given vector of event rates in the control group

\( p_C = (p_{CE1}, p_{CE2}, \ldots, p_{CEp}) \),

it can easily be shown [14] that the correlation between the different test statistics can be adequately approximated by:

\[
\text{Corr}(T_{CE1}, T_{CE2}, \ldots, T_{CEp}) = \text{Corr}(X_{CE1}, X_{CE2}, \ldots, X_{CEp})
\]

Thereby note, that the correlations between the test statistics corresponds to the covariances, as the variances of the test statistics equal 1 under the corresponding null hypothesis assumption.

We now consider the correlations between a composite endpoint and its components. The dependencies of event rates between the composite and the individual components are illustrated in Table 1.

As an event in a single component always corresponds to an event in the composite, the correlation is given as follows, compare for example Sozu [24]

\[
\text{Corr}(X_{CE1}, X_{CE2}, \ldots, X_{CEp}) = \frac{p_{CE1} (1 - p_{CE1})}{\sqrt{p_{CE1}(1 - p_{CE1}) p_{CE1}(1 - p_{CE1})}}
\]

Note that the estimated product-moment correlation between two binary endpoints corresponds to the well-known phi-coefficient.

In order to obtain the general correlation structure between two individual components that may overlap consider the dependencies of the event rates as shown in Table 1.

### Table 1 Dependencies of event rates between the composite and its components

<table>
<thead>
<tr>
<th>CE +</th>
<th>CE -</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPj +</td>
<td>( p_{CE1} - p_{CEj} )</td>
</tr>
<tr>
<td>EPj -</td>
<td>( p_{CEj} - p_{CE1} )</td>
</tr>
</tbody>
</table>

### Table 2 Dependencies of event rates between the components, general case

<table>
<thead>
<tr>
<th>EPj +</th>
<th>EPj -</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_{EPj} )</td>
<td>( 1 - p_{EPj} )</td>
</tr>
<tr>
<td>( 1 - p_{EPj} )</td>
<td>( p_{EPj} )</td>
</tr>
</tbody>
</table>

### Table 3 Dependencies of event rates between the components, no overlap

<table>
<thead>
<tr>
<th>EPj +</th>
<th>EPj -</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0 )</td>
<td>( p_{EPj} )</td>
</tr>
<tr>
<td>( p_{EPj} )</td>
<td>( 1 - p_{EPj} )</td>
</tr>
</tbody>
</table>

\[
\text{Corr}(X_{CE1}, X_{CE2}, \ldots, X_{CEp}) = \frac{p_{CE1} (1 - p_{CE1})}{\sqrt{p_{CE1}(1 - p_{CE1}) p_{CE1}(1 - p_{CE1})}}
\]

### 3.3 Calculation of the Adjusted Local Levels and the Global Power

Under the assumption that all null hypotheses (1) and (2) are true and a predefined vector of event rates \( p \) can be specified to determine the position of the effects, the random vector \( T_p = (T_{CE1}, T_{CE2}, \ldots, T_{CEp}) \) follows a multivariate normal distribution \( T_p \sim N(\mu, \Sigma) \), with mean vector \( \mu = (0, \ldots, 0) \) and a variance covariance matrix \( \Sigma \) where the diagonal elements equal 1 and the covariances are given in Formula 3 and 5 presented in Section 3.2.

If \( \alpha \) denotes the family wise significance level which is commonly given by 0.05 the corresponding local levels are chosen such that the following equality holds

\[
\alpha = 1 - \Theta(z_{1 - 0.05}, z_{1 - 0.05}, \ldots, z_{1 - 0.05})
\]

where \( z_{1 - \alpha} \) denotes the corresponding \((1 - \alpha)\)-quantile of the standard normal distribution and \( \Theta \) denotes the normal distribution function of the random vector \( T_{p'} \). Thereby, the local significance levels must not necessarily be equal such that...
```r
library(mvtnorm)
library(triangle)

# Please specify
# pce= composite event rate under the null hypothesis assumption  
# p01= component 1 event rate under the null hypothesis assumption  
# p02= component 2 event rate under the null hypothesis assumption  
# alpha= global one-sided significance level  
# alpha.ce= one-sided significance level assigned to the composite, must be smaller than alpha  
# weight= determines local level for component 1 by alpha_1=(alpha-alpha.ce)*weight

power<-function(pce,p01,p02,alpha, alpha_ce, weight){

corr_e1_e2<-(-p01*p02)/sqrt(p01*(1-p01)*p02*(1-p02))
corr_e1_ce<-(-p01*(1-pce))/sqrt(p01*(1-p01)*pce*(1-pce))
corr_e2_ce<-(-p02*(1-pce))/sqrt(p02*(1-p02)*pce*(1-pce))

corr<-diag(3)
corr[1,2]<-corr_e1_e2
corr[1,3]<-corr_e1_ce
corr[2,3]<-corr_e2_ce

corr[1,3]<-corr_e1_ce
corr[2,3]<-corr_e2_ce

corr[1,2]<-corr_e1_e2

corr[1,3]<-corr_e1_ce
 corr[2,3]<-corr_e2_ce

print("Correlation Matrix")
print(corr)

alpha_e1<-((alpha-alpha_ce)*weight
alpha_e2<-((alpha-alpha_ce)*(1-weight)

quantil1<-qnorm(1-alpha_e1)
quantil2<-qnorm(1-alpha_e2)
quantil3<-qnorm(1-alpha_ce)

x<-pnorm(lower=c(-Inf,-Inf,-Inf), upper=c(quantil1,quantil2,quantil3), mean=c(0,0,0), corr=corr)
print(1-x[1])

while ((1-x[1])<alpha){
 alpha_e1<-alpha_e1+0.000001*weight
 alpha_e2<-alpha_e2+0.000001*(1-weight)
 quantil1<-qnorm(1-alpha_e1)
 quantil2<-qnorm(1-alpha_e2)
 quantil3<-qnorm(1-alpha_ce)

print("Correlation Matrix")
print(corr)

x<-pnorm(lower=c(-Inf,-Inf,-Inf), upper=c(quantil1,quantil2,quantil3), mean=c(0,0,0), corr=corr)
#Global Level#
print(1-x[1])
#Lokal Levels#
print(c(alpha_ce, alpha_e1, alpha_e2))
}
```

**Fig. 2** R code to calculate adjusted local levels
Equation 6 has multiple solutions. It can be solved iteratively by standard statistical software, for example with the 'mvtnorm'-package in R Version 1.9.0 or higher. Exemplary program code is provided in Figure 2.

The global power to reject the null hypothesis for the composite and for the $l$ most important components is given in Figure 3, which can be directly calculated with the 'mvtnorm'-package in R Version 1.9.0 or higher once the adjusted significance levels have been determined. Note that the power exclusively depends on the assumptions on the event rates under the corresponding null hypotheses. The underlying correlations can directly be calculated for any given event rate assumption.

A potential argument against the inclusion of the correlation for the definition of the local levels is that the correlation depends on the event rates in the control group. Thus, if these rates are wrongly specified, the correlation is wrong, too, and the defined local levels do not necessarily preserve the family wise significance level $\alpha$. Therefore, we provide simulation results to demonstrate the robustness of the method under misspecifications of the event rates.

4. Results

4.1 Simulations

Consider a controlled, clinical trial with a composite primary endpoint consisting of two components which should both be assessed in the confirmatory analysis. The global one-sided significance level is given by 0.025. As the focus of the confirmatory analysis lays on the assessment of the composite, the local significance level of the test for the composite null hypothesis will be fixed at 0.02 which corresponds to a very small reduction. Therefore, the power to test exclusively the composite at level 0.02 will be only slightly reduced as compared to the power for the full level of 0.025. Moreover, we focus on the situation where both components are tested at the same local level. Table 4 provides simulation results for different parameter settings. We did not consider situations where the event rates of the composite or the components are too close to the boundaries of the interval $[0;1]$ as the normal approximation might not be valid in these cases. We first calculated the adjusted local levels for the components based on fixed parameter assumptions. We then varied these assumptions within a certain range but kept the local significance levels fixed. We simulated the global type I error rate based on 10,000 replications for each parameter constellation within these ranges using a step width of 0.005. Finally, the maximum over all type I error rates for these settings was determined.

For illustration, consider the first two rows of Table 4. Assuming a composite event rate of 0.3 under the null hypothesis of no treatment effect and event rates of 0.15 for both components, the local adjusted one-sided level to test both components is given by 0.0043. Simulated type I error rates for a sample size of 50 and 200 patients per group are provided based on 10,000 replications. It can be seen that the multiplicity adjustment is still slightly conservative. In row 2, the maximum over all type I error rates for parameters that deviate from these assumptions is shown. We assume here that the event rate for the composite lies within the interval $[0.25;0.4]$ and that the first component corresponds to a fraction between 0.4 and 0.6 of the composite rate. As we assume no overlap between the components, the event rate for the second component is determined by $p_{CE}^C - p_{EP1}^C.$

![Fig. 3](image-url)  
*Global power for the multiple test problem*
The results show that the maximal type I error rate under deviations from the originally assumed parameter setting can exceed the global one-sided level of 0.025. If the underlying sample size is small (50 patients per group), this anti-conservatism can become unacceptable for some parameter constellations. However, clinical trials with composite primary endpoints most often include much higher sample sizes of 200 patients per group or even more. For an underlying sample size of 200 patients per group, the anti-conservatism of our approach is limited which demonstrates the robustness of the method under deviations from the original parameter assumptions.

4.2 Clinical Trial Examples

In a randomized, controlled clinical trial including a total of 63 patients with cirrhosis and acute variceal bleeding, the treatment effect of an early use of a polytetrafluoroethylene-covered stent was compared to standard drug treatment alone [25]. The primary end point of the study was a composite outcome of failure to control for acute bleeding (EP₁) or failure to prevent clinically relevant variceal rebleeding within one year after enrollment (EP₂). Using the normal approximation test for rates, the study had a power of 0.9 to detect a decrease in the composite endpoint from 0.45 in the standard treatment group to 0.1 in the intervention group with a one-sided significance level of 0.025. The actual one-sided type I error rate was 0.0235, the simulated power was 0.9143, which shows that the application of the chi-square approximation is appropriate in this case. Although the event rates for the components are smaller, the simulation results of Lydersen [10] show, that for event rates around 0.2 the actual type I error is just slightly anti-conservative. Under the null hypotheses that all rates in the intervention group equal the corresponding rates in the control group and under the additional assumption that all event rates are given as specified above, the correlations are given as:

\[ \text{Corr}(T_{CE}, T_{EP₁}) = 0.638, \]
\[ \text{Corr}(T_{CE}, T_{EP₂}) = 0.553, \]
\[ \text{Corr}(T_{EP₁}, T_{EP₂}) = -0.289. \]

Instead of testing exclusively the composite at the full one-sided level of 0.025, an alternative would be to test the composite at the only slightly reduced level of 0.02. The power to test exclusively the composite at the reduced level 0.02 is given by 0.88, which corresponds to a very small power loss when compared to the original test procedure that applies the full level 0.025. The other way round, a total of 68 patients would be required to reach a power of 0.9 when the test is performed at the reduced level 0.02.

The components are tested at equal local levels. Solving \(\text{Equation 6}\) yields

\[ \alpha_{EP₁ \text{ local}} = \alpha_{EP₂ \text{ local}} = 0.0039, \]

which is far above the Bonferroni adjusted level of 0.01/2 = 0.0025. The robustness of the type I error rate when using these adjustments under small parameter deviations was investigated by simulations. We simulated the maximum over all type I errors for the following parameter constellations

\[ p_{CE} \in [0.4; 0.5], \]
\[ p_{EP₁} \in [0.4 \cdot p_{CE}^C; 0.6 \cdot p_{CE}^C], \]
\[ p_{EP₂} = p_{CE}^C - p_{EP₁}. \]

The parameter ranges were chosen such that the normal approximation is appropriate for the composite and the components. The simulated maximal type I error was 0.027 which supports the robustness of our method under small parameter deviations. Thus, even for small sample sizes, our approach might be applicable as long as the parameters only slightly deviate from the assumptions made.

In another randomized, controlled study in pregnant women, prenatal repair of myelomeningocele was compared to repair after delivery which corresponds to the standard treatment [26]. The primary outcome was defined as a composite of fetal or neonatal death (EP₁) or the need for a cerebrospinal fluid shunt placement among children who survived (EP₂). The study was planned to detect a decrease in the composite endpoint from 0.85 in the standard treatment group to 0.57 in the intervention group at a one-sided significance level of 0.025 with a power of 0.99. This results in a required total number of 200 patients. The composite rate of 0.85 in the standard treatment group was based on the assumption of a 0.8 shunt rate and a 0.05 death rate. Under the assumption that all null hypotheses are true and the event rates are given as specified above, the correlations correspond to

\[ \text{Corr}(T_{CE}, T_{EP₁}) = 0.096, \]
\[ \text{Corr}(T_{CE}, T_{EP₂}) = 0.84, \]
\[ \text{Corr}(T_{EP₁}, T_{EP₂}) = -0.459. \]

If again the composite is tested at the reduced level of 0.02, the power to test exclusively the composite at the reduced level 0.02 still amounts to 0.99, so the power loss is extremely small and does not affect sample size.

The local levels for the components can be chosen as

\[ \alpha_{EP₁ \text{ local}} = \alpha_{EP₂ \text{ local}} = 0.0045. \]

However, in this trial the event rates for the individual components are of very different magnitude. An unexpectedly large effect is therefore much more likely to be observed in the shunt endpoint. Therefore, it might be more appropriate to choose the local levels for the components differently. An alternative choice of the local levels fulfilling \(\text{Equation 6}\) is given by

\[ \alpha_{EP₁ \text{ local}} = 0.001, \alpha_{EP₂ \text{ local}} = 0.1. \]

With this \(\alpha\) allocation a large effect in the shunt components is more likely to be detected whereas an overwhelming effect in
the death component seems unrealistic. Note that the above setting of the local significance levels highly profits from the underlying correlation structure, as the sum of the local levels considerably exceeds the one-sided family wise level of 0.025 by

\[ \alpha_{CE}^{local} + \alpha_{EP}^{local} + \alpha_{PCE}^{local} = 0.02 + 0.01 + 0.001 = 0.031. \]

Again, we simulated the robustness of the type I error rate using the above local levels under parameter deviations. We simulated the maximum over all type I error rates for the following parameter constellations

\[ p_{CE}^C \in [0.75; 0.9], \]
\[ p_{EP}^C \in [0.05 \cdot p_{CE}^C ; 0.1 \cdot p_{CE}^C], \]
\[ p_{PCE}^C = p_{CE}^C - p_{EP}^C. \]

The maximal type I error rate for these parameter constellations was 0.0287. Thus, our method is reasonably robust for this clinical study example.

5. Discussion

In the statistical literature and in current guidelines [2, 5, 8, 9], it is recommended that clinical trials using a composite binary endpoint as primary outcome measure should additionally assess the effects observed for the constituting components in order to guarantee a valid interpretation of the outcome for the composite. Usually, this is done by including the components as secondary endpoints which are analyzed descriptively. However, it would be desirable to include at least the most important components in the confirmatory analysis. This results in a multiple test problem which requires adjustment of the local significance levels. In the literature, compare e.g. Bretz et al. [27], there exist two main approaches a) ad-hoc methods like the standard Bonferroni-corrrection assign each hypothesis to be tested a predefined local significance level; b) sequentially rejective procedures like the well-known Bonferroni-Holm method allow to shift the local significance level once a hypothesis has been rejected. Although the latter methods guarantee more satisfactory use of the global significance level, post-hoc methods can be more easier integrated in the planning stage of the clinical trial as the sample size calculation is performed at fixed known levels. The standard Bonferroni method is very conservative. However, a composite endpoint and its components are related by a special correlation structure which can be used to obtain less conservative local significance levels. We demonstrated how to calculate the adjusted local levels based on the underlying correlation structure. Thereby, the composite is tested at a level which is only slightly reduced as compared to the global significance level and the remaining level is spent to the components. The power loss for the test of the composite null hypothesis is limited, which was demonstrated by several examples.

A main concern about the general idea of incorporating the correlation between test statistics for the adjustment of the local levels is that the global type I error rate might not be strictly preserved in case of deviations from the parameter assumptions. We provided simulation results which demonstrated that the maximal type I error rate only slightly exceeds the global significance level if the parameters deviate within a reasonable range from the assumptions made in the planning stage and if the underlying sample size is large enough.

6. Conclusions

Our proposed multiplicity adjustment method requires assumptions on the event rates for the composite and the components under the corresponding null hypotheses of no treatment effect.

If reasonable parameter assumption are available from the literature, our proposed method provides a valid and less conservative alternative to the standard Bonferroni adjustment, which is robust under misspecification of the parameter assumptions.

Acknowledgments

We would like to thank the referees for their careful reviews and helpful comments which lead to a substantially improved version of the manuscript.

References


