NONPARAMETRIC COMBINATION METHODOLOGY

- A Better Way to Handle Composite Endpoints?

Supervisors:
Johan Bring, Statisticon AB
Måns Thulin, Uppsala University

Author:
Yvette Baurne
Abstract

Composite endpoints are widely used in clinical trials. The outcome of a clinical trial can affect many individuals and it is therefore of importance that the methods used are as effective and correct as possible. Improvements of the standard method of testing composite endpoints have been proposed, and in this thesis, the alternative method using nonparametric combination methodology is compared to the standard method. Performing a simulation study, the power of three combining functions (Fisher, Tippett and the Logistic) are compared to the power of the standard method. The performances of the four methods are evaluated for different compositions of treatment effects, as well as for independent and dependent components. The results show that using the nonparametric combination methodology leads to higher power in both dependent and independent cases. The combining functions are suitable for different compositions of treatment effects, the Fisher combining function being the most versatile. The thesis is written with support from Statisticon AB.

Keywords: composite endpoints, multivariate randomization tests, combining p-values, Fisher combining function, Tippett combining function, Logistic combining function.
# Table of Contents

1 Introduction ........................................................................................................... 1
   1.1 Research Questions ....................................................................................... 2
   1.2 Restrictions .................................................................................................... 2
   1.3 Outline of the Paper ...................................................................................... 2

2 Background ............................................................................................................. 3
   2.1 Composite Endpoints .................................................................................... 3
      2.1.1 The Belimumab Case – an Example of the Use of Composite Endpoints .. 4
   2.2 Combining P-values ....................................................................................... 5

3 Theory ...................................................................................................................... 8
   3.1 Chi-square Tests ............................................................................................ 8
   3.2 Permutation and Randomization Tests .......................................................... 9
   3.3 Nonparametric Combination Methodology .................................................... 10
      3.3.1 Combining Functions .............................................................................. 11
   3.4 Statistical Power ............................................................................................ 11

4 Method .................................................................................................................... 13
   4.1 Outline for the Simulation .............................................................................. 13
   4.2 Simulation Design ......................................................................................... 14
      4.2.1 Generating Data ...................................................................................... 14
      4.2.2 Testing Procedure ................................................................................... 14

5 Results ..................................................................................................................... 16
   5.1 Same Treatment Effect on All Components .................................................. 16
      5.1.1 Situation 1 .............................................................................................. 16
   5.2 Different Treatment Effects on the Components ............................................. 18
      5.2.1 Situation 2 .............................................................................................. 18
      5.2.2 Situation 3 .............................................................................................. 21
   5.3 A Remark Regarding Significance Levels ..................................................... 23

6 Discussion ............................................................................................................... 24
   6.1 Conclusions .................................................................................................... 24
   6.2 Methodological Discussion .......................................................................... 26
   6.3 Suggestions for Further Research................................................................. 28

Appendix A – Values for Generating Components ..................................................... 31
Appendix B - Code Used for Simulations ................................................................. 32
Appendix C - Results ............................................................................................... 37
1 Introduction

A composite endpoint is a combination of variables (or endpoints) that are relevant to include when evaluating the effect of a treatment in a clinical trial. For example; for measuring serious cardiac morbidity the composite endpoint can include all-cause mortality, stroke and myocardial infarction, and a composite for allergic rhinitis can include itchy nose/mouth/throat/eyes, sneezing and rhinorrhea.

The use of composite endpoints in clinical trials is motivated by several advantages. Composite endpoints, if designed properly, enhances the statistical precision and is ethically and economically justified as it reduces the needed number of participants to reach adequate power to detect a potential treatment effect (Sankoh, et al., 2014). As the results of the clinical trials underlie the potential approval or disapproval of a drug, it is crucial that the proper methods are used. The standard method for testing a composite endpoint is to combine the components into a single outcome and then conduct the group comparison based on this (Mascha & Sessler, 2011). As an attempt of improving the method, alternative multivariate methods have been proposed where the components are individually tested and then combined to estimate the treatment effect (Pesarin, 2001; Mascha & Sessler, 2011; Pesarin & Salmaso, 2012).

The multivariate methods offer advantages including the possibility to handle different types of variables, varying frequency in the components and the possibility to adjust for correlation among the components (Mascha & Sessler, 2011). Research implies that statistical power differs within different conditions regarding treatment effect size, number of components, correlation between components etc. (Pesarin, 2001; Mascha & Sessler, 2011; Corain & Salmaso, 2015). For data that have favourable characteristics, such as a multivariate normal distribution, the multivariate testing procedure is well established. If data do not follow a multivariate normal distribution, for example if some of the component are categorical, other methods are available. Pesarin (2001) suggests using nonparametric combination (NPC) methodology to deal with situations like this. Different combination methods have different critical regions, and thus different power. The question at hand is how different conditions in clinical trials affect the power for tests based on these combination methods. Are the multivariate NPC methods preferable to use instead of the standard methods of composite endpoints?
1.1 Research Questions

The purpose of this thesis is to evaluate if the nonparametric combination (NPC) methodology offers a better way to test composite endpoints, in terms of power, than the standard method. This will be done by a simulation study, where the following research questions will be answered:

- Which of the two methods – the standard method or the NPC method – have the highest power?
  - Does the power differ depending on whether there are dependence relations between components or not?
  - Does the power differ depending on the composition of different treatment effects on the components?

1.2 Restrictions

The nonparametric combination methodology can be used in a variety of situations and since it is not possible to test them all at once, this thesis is restricted to a case with three binary components. A case with only binary components will not illustrate one of the major strengths with NPC – that variables with different properties can be used since they are tested one by one – but has the advantage that it can be seen as a simple entry point to the nonparametric combination methodology, which later can be extended to include all sorts of variables.

Further, there exists many different combination methods to apply within the nonparametric combination methodology. In this thesis, three of the most common combining functions will be used; Fisher (1950), Logistic (Mudholkar & George, 1979) and Tippett (1931).

1.3 Outline of the Paper

The paper is outlined as follows; Section 2 gives a background with a description of the concept of composite endpoints, as well as a review of the research on binary composite endpoints and nonparametric combination methodology. The theory of multivariate permutation tests and nonparametric combining functions is explained in Section 3. The methodology of the simulation procedure is described in Section 4. In Section 5, the results of the simulations are presented. The results and the method are discussed, along with implications for further research, in Section 6.
2 Background

2.1 Composite Endpoints

A composite endpoint can consist of any number of components, but should include at least two. There are different ways to combine the components. In the binary case, one way is to use a collapsed composite, also called any event versus none, where the patient is considered a responder if he/she shows a response in at least one of the components. Another way is to use counts, where the composite endpoint gets one point for each component that shows a response (Mascha & Sessler, 2011). In the continuous case an index is often used, and the composite endpoint then consists of an, either weighted or unweighted, combination of counts or scores. The composite can also be based on failure rates for the treatment (Huque, et al., 2011).

It is important to design the composite endpoint carefully. First of all, the components have to be clinically important for the disease of interest. If components that are not affected by the disease are included in the composite, the power decreases (Mascha & Sessler, 2011). Further it is important to be careful when using components that are heterogeneous in the composite (Huque, et al., 2011; Mascha & Sessler, 2011; Sankoh, et al., 2014). In this context, heterogeneous components imply that the treatment effects are trending in different directions. If heterogeneity is present, the effects of the components could cancel each other out. In connection to this, it is good to aim at including components with similar frequencies, severity and treatment effects (Mascha & Sessler, 2011; Sankoh, et al., 2014). If a clinically less important component occurs more frequently, this component can affect the significance in a larger extent than a more important but less frequently occurring component, and thereby distort the results. In reality it can be difficult to ensure that the components have approximately the same frequency and treatment effects, and because of that it is essential to aim at including equally important components.

Mascha & Sessler (2011) investigates factors of importance regarding the power of tests, using five different methods of combining binary components in composite endpoints. For the methods that they compare, the power increases as the correlation between components decreases. They explain this by the fact that when responses on the components are less like each other, more information exists in the data and would thus naturally increase the power. Further, they state that the power will be lower for the overall test if only one or a few components are affected by the treatment, in comparison to if all components are affected. They have also found that the power is affected by the control group incidence (the treatment effect...
in the control group). Mascha & Sessler examines other methods than the ones used in this thesis, but their results are of a general character and it is probable that they also apply for other methods.

2.1.1 The Belimumab Case – an Example of the Use of Composite Endpoints

A clinical trial testing a drug called Belimumab will be used to more thoroughly explain how composite endpoints can be used. Belimumab (Human Genome Sciences, Inc, 2010) is a drug that is used to treat Systemic Lupus Erythematosus (SLE). The SLE disease is a chronic autoimmune disorder that can affect multiple organs and is, easily explained, caused by an overactive immune system.

In the trials, the composite endpoint used is called the SLE Responder Index (SRI). The SRI is constructed by three components: the SELENA SLEDAI index, the BILAG Classic Index, and the Physician’s Global Assessment (PGA). To be considered a SRI responder, a patient has to be considered a responder in all three components.

The SELENA SLEDAI index is a SLE disease activity index, which measures the presence of 24 manifestations of the disease. The patient gets a score for each manifestation and depending on how severe the manifestation is, the scores are weighted from 1 to 8 points. The patient gets a baseline value at the beginning of the trial, and are then evaluated again after the trial. If the score is reduced with minimum 4 points, the patient is considered a responder in the SELENA SLEDAI component.

The BILAG classic index measures the increase in disease activity in 8 organ domains for the prior month. For each organ domain, the patient is given a score from A to E, where A is the highest increase, and an E score means that the organ domain has not been involved in the disease at all. The BILAG classic index is included as a component in the composite endpoint to measure the worsening of the disease. To be considered a responder on the BILAG component, the patient cannot have any new A scores, or 2 new B scores.

The Physician’s Global Assessment (PGA) is used to measure the patients’ general health status, in order to control that an improvement in the SELENA SLEDAI does not come with the patients general health getting worse. The PGA is measured by a physician, who evaluates the overall condition of the patient on a visual scale from 0 (no lupus disease activity) to 3 (severe lupus disease activity). If the PGA score is increased by 10% (0.3 points) or more, the
general health status is considered to have worsened. Thus, to be considered a responder in the PGA component, the PGA score cannot have been increased by 0.3 points or more.

The composite endpoint in the Belimumab trial is thus a binary endpoint, consisting of three binary components which are covering three important aspects of the SLE disease, relevant to include when testing a new treatment. The Belimumab case is a good example of some of the advantages of composite endpoints. By including all three components, the researcher will get a better understanding of how the treatment works than if they only would have included one component, for example the SELENA SLEDAI index. The method for combining the components; response has to be shown in all three components to be considered a response in the composite, is a common method and will further on be referred to as the “standard method”. The simulated trial in this thesis will mimic the Belimumab trials in their use of three binary components, and the standard method used to combine them.

2.2 Combining P-values

The basic idea in nonparametric combination methodology is to combine p-values. Several functions for combining the p-values have been proposed. The combining functions can be divided into two types: quantile combination methods and order statistic methods. The quantile combination methods are based on cumulative distribution functions. The p-values are transformed and the sampling distribution of the combining function is calculated by using the rules for probabilities involving sums of independent, identically distributed, random variables\(^1\) (Loughin, 2004). P-values are order statistics, which is the foundation for the order statistic methods (Loughin, 2004).

Loughin (2004) has compared six different methods for combining p-values from independent tests; four quantile combination methods (uniform, chi-square, normal and logistic) and two order statistic methods (minimum and maximum), which he calls by the name of their probability distributions. There have been earlier comparisons, but Loughin considers a wider range of circumstances to be able to make general recommendations about the use of the combining functions. His results show that different functions are suitable in different situations, depending on the composition of evidence against the null in the partial tests. If there is equal evidence in each test, for \(k\) tests, where \(k\) is any number, a normal or a logistic combining function is recommended. If there is evidence in a majority of the tests, a normal or logistic combining function is recommended. If there only exists evidence against the null

\(^1\) See Theorem 6.2 (Wackerly, et al., 2008)
hypothesis in one of the tests, the minimum function is recommended if there is strong total evidence. If there is moderate total evidence, a chi-square function is recommended, and for weak total evidence, a normal or logistic function is recommended.

One of the special cases that Loughin (2004) examines is when \( k=3 \), which is the case this thesis covers. The results slightly deviate from the general recommendations for any \( k \). For the cases with \( k=3 \), he concludes that the minimum function is one of the best methods for when only one out of the three tests conclude a false null, and only when there is quite strong evidence. When evidence is evenly spread, the uniform performs the best. The normal is slightly better than logistic and chi-square, but both normal, logistic and chi-square gets to maximum power when the evidence is strong and evenly spread. The maximum and the uniform combining functions are no good at all when the evidence is unevenly spread across the three tests. For the normal, chi-square and the logistic, the performance varies based on the strength of evidence when the evidence is unevenly spread across the tests. For weak evidence, the normal is to prefer. For stronger evidence, the chi-square is to prefer.

Loughin (2004) further notes that the normal function has the disadvantage of possibly cancelling out large p-values with the small p-values. This can be a problem when there are only some of the hypotheses that are true. Using the chi-square, this problem is avoided since it focuses more on the small p-values. But, the chi-square then loses power when the evidence is evenly spread in the hypotheses. He suggests using the logistic function as a compromise, since it is quite robust for different compositions of evidence in the hypotheses.

Westberg (1985) has compared the power for the Fisher (chi-square \((2k)\) distributed) and the Tippett (minimum) combining functions when combining three p-values \((k=3)\), originating from normal or chi-square distributed tests. Her conclusion is, similarly to Loughins conclusions, that it cannot generally be recommended which one of the Fisher or Tippett functions that should be used, how the evidence is spread across the components has to be taken into account. But, on the contrary to Loughins results, she concludes that when there is evidence in one out of the three tests, the Fisher function is to prefer before Tippett. She also concludes that when a small number of partial tests are combined, the difference in power between Fisher and Tippett is smaller than if \( k \) is larger.

In conclusion, previous research indicates that there should be a difference in power between the three combination methods used in this thesis; Fisher, Logistic and Tippett, depending on the composition of treatment effects in the three components. There is also indications that there
should be a difference in power depending on whether the components are dependent or not. No previous research has been found that compares the standard method with the nonparametric combining methods.
3 Theory

When several hypotheses are to be tested at the same time, a multivariate method can be used. There exist several multivariate methods, which are suitable for different conditions depending on their assumptions and the nature of the data. For example, the Hotelling $T^2$ is a multivariate t-test that can be used to test for differences between groups on multiple variables. One assumption for Hotelling $T^2$ is that the data should be multivariate normally distributed and continuous (Hair, et al., 2014). The Hotelling $T^2$ test is robust to deviations of the multivariate normality assumption, but data still has to be continuous. In the situation that this thesis regards, where data is binary, the Hotelling $T^2$ test would not be suitable. By combining the components into a single composite endpoint it is possible to work around this problem and use a univariate test instead, which is what is done in the standard method. Another way to work around the possible problem of multivariate testing is to use the nonparametric combination methodology and test each component one by one, and thus also be able to use univariate tests.

3.1 Chi-square Tests

As the components of interest are binary, chi-square tests of independence are used to test for differences between the treatment groups. The chi-square test is a nonparametric test that is used to analyse simple frequency tables. The hypotheses tested are $H_0$: there is no difference between the treatment groups, versus $H_1$: there is a difference between the treatment groups. What the chi-square test actually does is that it tests the differences between the observed frequencies and the expected frequencies (Körner & Wahlgren, 2006). The chi-square test uses the following formula:

$$
\chi^2_{df} = \sum \left[ \frac{O_{i,j} - E_{i,j}}{E_{i,j}} \right]^2
$$

$O$ is the observed frequencies, $E$ is the expected frequencies, $i$ is the row number and $j$ is the column number. The degrees of freedom, $df$, indicates which chi-square distribution the test statistic follows, and are obtained through:

$$
df = (\text{number of rows} - 1)(\text{number of columns} - 1)
$$

Assumptions for the chi-square test are that at least 80% of the expected frequencies are $\geq 5$, none of them are $< 1$ and that the groups are independent (McHugh, 2012).
3.2 Permutation and Randomization Tests

Usually, p-values or critical regions are used to decide whether to reject a hypothesis or not. If the distribution of the test performed is unknown, and thereby also the rejection region is unknown, permutation or randomization tests can be used instead. The theory of permutation and randomization tests used in this thesis follows the outline by Pesarin (2001). Permutation tests builds on the concept of calculating all possible permutations of the data, and using these to estimate the distribution of the test statistic under the null hypothesis. Using this simulated null distribution, critical regions can be calculated. If dealing with a large sample, it can be difficult to make exact calculations of the null distribution using permutations, as the null distribution depends on the specific data set and the number of permutations quickly becomes massive if the data set is large. This problem can be solved by using a randomization test instead. Using the randomization test, the group belongings are randomly reassigned to the observed data. These randomly reassigned group belongings illustrate the conditions in the null hypothesis – there is no difference between the groups, data is completely random. The random reassigning of group belongings is done repeatedly, and is used to simulate the null distribution. If enough independent replications are conducted, the estimation of the distribution can be very exact. By only reassigning the group belongings and keep the other data for each individual together, the dependence relations between the components are kept. In this thesis, the sample will be large, so the randomization procedure will be used.

Since the null distribution is estimated using the observed data, permutation and randomization tests are inferential procedures that are conditioned on the observed data. This conditioning together with an important property called exchangeability, are the two important properties of the permutation procedure (and consequently, in extension, also for the randomization procedure). These two important properties are explained by Pesarin (2001). In hypothesis testing of the kind we are interested in in this thesis, data is obtained in some way, for example an experiment that is performed \( n \) times on a population variable \( X \) and results in the observed data set \( x = \{x_1, ..., x_n\} \). The data set is divided into groups based on the experiments treatment levels, and takes on values in the sample space \( \mathcal{X}_n \). The factorization theorem gives that under \( H_0 \), the observed data set \( x \) is a set of sufficient statistics for the underlying distribution. By likelihood, sufficiency, and the conditionality principles of inference, \( x \) and \( x^* \) from the same sample space \( \mathcal{X}_n \) are equivalent for making inferences if, given a sample point \( x, x^* \in \mathcal{X}_n \) is such that the likelihood ratio \( f_p^{(n)}(x)/f_p^{(n)}(x^*) = \rho(x, x^*) \) is not dependent on \( f_p \) for whatever \( P \in \mathcal{P} \), where \( P \) is the parent population, \( f_p \) denotes the density of \( P \) and \( \mathcal{P} \) is a nonparametric
family of distributions. The points that are equivalent to \( x \) can together be called the orbit associated with \( x \), denoted \( \mathcal{X}^n_{/x} \), so that \( \mathcal{X}^n_{/x} = \{ x^* : \rho(x, x^*) \text{ is } f_P - \text{independent} \} \). If exchangeability is assumed for the observed data, \( f_P(n)(x_1, \ldots, x_n) = f_P(n)(x_{u_1}, \ldots, x_{u_n}) \), where \((u_1^*, \ldots, u_n^*)\) is any permutation of \((1, \ldots, n)\). Thus, the observed data is exchangeable with respect to groups.

When exchangeability is assumed in the null hypothesis and \( x \) is known, the orbit associated with \( x \) is still entirely determined by \( x \). Due to the conditioning and the exchangeability, a permutation test is not dependent of the likelihood model related to \( P \) and this enables \( P \) to be unknown. Since we condition on a set of sufficient statistics, the permutation tests obtain good general properties.

### 3.3 Nonparametric Combination Methodology

In the nonparametric combination methodology each variable is tested one by one and the obtained p-values are combined. Then the permutation or randomization testing principle is used to estimate the distribution for the test statistic under the null hypothesis, and this will make it possible to determine the overall effect even if the variables are tested one by one with univariate tests.

Pesarin (2001) states the following assumptions for nonparametric combination methodology:

1. The null hypothesis must take the form \( H_0 : \{ P_1 = \ldots = P_C \} = \{ X_1 = \ldots = X_C \} \), which implies equality of multivariate distributions of responses on \( C \) groups, that is, the exchangeability property.

2. The null hypothesis can be broken down into a finite set of sub-hypotheses \( H_{0i}, i = 1, \ldots, k \), that each are appropriate for the partial component. Thus, the overall null hypothesis can be written \( \bigcap_{i=1}^k H_{0i} \).

3. The alternative hypothesis can be broken down into a finite set of sub-hypotheses \( H_{1i}, i = 1, \ldots, k \). If at least one of the alternative sub-hypotheses are true, \( H_1 \) is true. The overall alternative hypothesis can thus be written \( \bigcup_{i=1}^k H_{1i} \).

4. A set of suitable partial tests has to exist.

5. All \( k \) partial tests are jointly analysed.
3.3.1 Combining Functions

To combine the p-values, a test statistic called combining function is used. Three of the most common combining functions are presented below. The p-values are noted $\lambda_i$ in the formulae, $i=1,\ldots,k$.

The Fisher (1950) omnibus combining function:

$$T_F'' = -2 \cdot \sum_i \log(\lambda_i)$$

The Logistic (Mudholkar & George, 1979) combining function (which is a version of the Liptak (1958) combining function):

$$T_L'' = \sum_i \log[(1 - \lambda_i)/\lambda_i]$$

The Tippett (1931) combining function:

$$T_T'' = \max_{1 \leq i \leq k} (1 - \lambda_i)$$

When the $k$ number of partial test statistics are independent, the Fisher function follows a chi-square ($2k$) distribution under the null hypothesis (Pesarin, 2001). The logistic function approximately follows a t ($5k+4$) distribution (Loughin, 2004). Both the Fisher function and the logistic combining function are quantile combination methods. The Tippett function is an order statistic method, which behaves according to the smallest p-value.

When choosing between these combining functions, Pesarin (2001) states the following guidelines: When one or some sub-alternatives are expected to be true, but not all of them, the Tippett combining function is recommended. If all the sub-alternatives are expected to be jointly true, the logistic combining function is suggested. If there are no particular sub-alternatives expected, the Fisher combining function is suggested. The combining functions have different rejection regions, and therefore also different statistical power.

3.4 Statistical Power

The concept of power is used to evaluate the performance of a test. The power of a hypothesis test is the probability of rejecting the null hypothesis. When $H_0$ is true, the power is equal to the $\alpha$-level. The formal definition of power is stated by Wackerly et al. (2008): If a test with the test statistic $W$ and the rejection region $RR$, which tests a hypothesis regarding the parameter $\theta$
is used, the power is the probability of that test leading to the rejection of $H_0$ when the true parameter value is $\theta$. Also written:

$$\text{power}(\theta) = P(W \text{ in } RR \text{ when the parameter value is } \theta)$$

Since the power depends on the values of the parameter in the alternative hypothesis, a statistical test can have many different powers. Typically, it can be said that the power increases when the difference between the assumed parameter value and the actual value increases (Triola & Uppsala Universitet, 2015).

In this thesis, since the parameter value in the population is simulated and therefore known, the power can be estimated by calculating the proportion of rejected null hypotheses.
4 Method

In this section an outline for the simulation is given, followed by a description of the simulation design.

4.1 Outline for the Simulation

Like in the Belimumab trial, the simulated clinical trial consists of two treatment groups; A and B. Treatment group A obtains the standard treatment for the disease. Treatment group B obtains the standard treatment and the new treatment. The composite endpoint Z consists of three components: $Y_1$, $Y_2$, and $Y_3$. Each of the three components are a binary event, where 1 indicates that the patient showed enough improvement to be considered a successful outcome, and 0 otherwise.

The composite endpoint Z is a binary event. A patient is considered a responder ($Z=1$) if success (1) is obtained in all three components; $Y_1$, $Y_2$, and $Y_3$. Otherwise, Z takes the value 0.

The simulations illustrate three situations regarding the composition of different treatment effects on the components. Situation 1: the treatment has the same effect on all three components. Situation 2: the treatment affects only one of the components. Situation 3: the treatment affects two of the components in the same extent. The last component is not affected by the treatment.

The above stated situations are simulated for both the case where the components are independent of each other, and where all or some of the components are dependent. In Situation 1, all components have the same effect and therefore it is possible to simulate all of them dependent of each other. In Situations 2 and 3, the two components with the same effect are simulated as dependent. That is, in Situation 2, the components that are not affected by the treatment are dependent of each other, and in Situation 3, the two components that are affected by the treatment are dependent of each other. It is important to have the same treatment effect for both the independent and the dependent cases so that the potential difference between the independent and the dependent cases only originates from the fact that the components depend on each other.

The treatment effects that are simulated ranges from no treatment effect (treatment effect on all components in both group A and B = 0.5) to a treatment effect (for one, two or all components in group B) of 0.85 in Situation 1 and 3, and 0.95 in Situation 2. These maximum treatment effects are chosen in order to be able to maintain the same dependency between components in
the dependent cases. The treatment effect in the control group A is always 0.5 for all components.

The simulation is performed using the software R which is an open-source software, suitable for all kinds of data analysis. Throughout the simulations, a significance level of $\alpha=0.05$ will be used.

4.2 Simulation Design

4.2.1 Generating Data

$N_a = 100$ observations in group A for each of the components are generated from a binomial distribution $\text{Bin}(1,p_a)$ with treatment effect $p_a$. $N_b = 100$ observations in group B for each of the components are generated from a binomial distribution $\text{Bin}(1,p_b)$ with treatment effect $p_b$. Thus, the total sample size in each simulation is $N=200$. In the dependent cases, observations for treatment group A are generated as follows. Component $Y_1$ is generated as stated above. $N_a$ observations for component $Y_2$ are generated by conditioning on component $Y_1$. If $Y_1$ equals 1, then $Y_2$ is generated from a binomial distribution $\text{Bin}(1,p_{a1})$ with treatment effect $p_{a1}$. If $Y_1$ equals 0, $Y_2$ is generated from a binomial distribution $\text{Bin}(1,p_{a0})$ with treatment effect $p_{a0}$. $N_a$ observations for component $Y_3$ are generated in the same way as $Y_2$ in the dependent situations, and in the same way as $Y_1$ in the situations where $Y_3$ is independent of $Y_1$ and $Y_2$. $N_b$ observations for treatment group B in the dependent situations are generated with the same procedure as the observations for treatment group A, but using binomial distributions $\text{Bin}(1,p_{b1})$ and $\text{Bin}(1,p_{b0})$. See appendix A for a presentation of all values of $p$.

4.2.2 Testing Procedure

The difference between treatment groups A and B are tested for each component; $Y_1$, $Y_2$, $Y_3$, and the composite endpoint $Z$, with a chi-square test of independence. From these tests, $p$-values for each of the components are obtained, as well as for the composite endpoint $Z$: $p_1$, $p_2$, $p_3$, and $p_z$.

Using the obtained $p$-values $p_1$, $p_2$, $p_3$, the combining functions of Fisher, Logistic and Tippett are calculated in accordance to the formulae stated in Section 3.3.1. This provides the observed values of $T_F''$, $T_L''$, and $T_T''$.

To obtain a realisation of the distributions of the combining functions $T_F''$, $T_L''$, and $T_T''$ under the null hypothesis, the treatment groups are randomly reassigned to the patients. The differences between treatment group A and B for each of the components are tested using the
same procedure as stated above – chi-square tests of independence – and new p-values are obtained and $T_F''$, $T_L''$, and $T_T''$ are calculated. This is repeated B=200 times.

The whole procedure, from generating observations to estimating the distributions of the combining functions, is repeated B=200 times. For the combining functions, the percentages of the observed $T_F''$, $T_L''$, and $T_T''$ that are bigger than the 95th percentile of the distribution under the null hypotheses are calculated. For the test of the composite endpoint Z, the percentage of the observed $p_z$ that are smaller than 0.05 is calculated. The procedure described above is performed for each of the situations described in section 4.1, for both the independent and the dependent cases.
5 Results

5.1 Same Treatment Effect on All Components

5.1.1 Situation 1

The following results were obtained in the situation where all components have the same treatment effect.

![Independent components](image1.png)

![Dependent components](image2.png)

**Figure 1.** Power as a function of treatment effect when components are independent, and when components are dependent of each other. Same treatment effect on all three components. Treatment effect in group A=0.5.

When the components are independent of each other and have the same treatment effect, the Fisher and the Logistic combining functions are almost equal in terms of power and performs the best of the four evaluated test statistics. The Tippett combining function performs slightly less good, followed closely by the standard method. In Figure 1, it is visible that small increases in treatment effect considerably improves power. All four test statistics perform in a similar way when all three components are independent and have the same treatment effect, but the Fisher or the Logistic combining function is to prefer since their power is consistently somewhat higher than the Tippett function and the standard method.

When the components are dependent and have the same treatment effect, the Fisher and the Logistic function still perform almost identical in terms of power. The Tippett function starts with the same power as Fisher and Logistic at a treatment effect of 0.55 on all components. After that, it has slightly lower power than Fisher and Logistic until a treatment effect of 0.7, where it has higher power than them; almost 1 and Fisher and Logistic only has a power of around 0.9. However, the difference is small and due to the relatively small number of
replications, it cannot be stated definitely. The standard method performs less good than the three combining functions just like in the independent case, but the difference is greater when the components are dependent of each other.

![Graphs showing power as a function of treatment effect on all three components for both independent and dependent cases, for each method. Treatment effect in group A=0.5.](image)

**Figure 2.** Power as a function of treatment effect on all three components for both independent and dependent cases, for each method. Treatment effect in group A=0.5.

In Figure 2, the power curves for each method in both the independent and the dependent cases are displayed together, in order to compare the cases. When the treatment effect is equal for all components, all four methods have higher power when the components are independent of each other. For the three combining functions, the difference is small but visible. For the standard
method, the difference in power between the independent and the dependent cases is greater than for the combining functions.

5.2 Different Treatment Effects on the Components

5.2.1 Situation 2

The following results were obtained in the situation where the treatment affects only one of the components.

**Figure 3.** Power as a function of treatment effect on component Y₃. Treatment effect on components Y₁ and Y₂ in group B = 0.5. Treatment effect in group A=0.5. In the right graph, Y₁ and Y₂ are dependent.

When there is a treatment effect on only one of the components out of the three, in this case Y₃, and the components are independent of each other, the Fisher and the Tippett combining functions perform the best in terms of statistical power. Both the Tippett and the Fisher combining functions reaches a power of 1 around when the treatment effect of Y₃ reaches 0.8. The Logistic combining function performs similar to Tippett and Fisher when the treatment effect is small, but as the treatment effect increases, it cannot keep up with their performance. The Logistic combining function peaks at a treatment effect of 0.85 where the power is slightly above 0.7. The power should not be able to peak and then decrease again when the treatment effect is increasing. Despite that, the power curve for the Logistic combining function is decreasing. This is most certainly due to the relatively small number of replications that are performed, which makes the estimates less accurate than if more replications had been performed. Thus, no importance should be attributed to this, merely the trend should be considered. The standard method performs poorly in this setting, at most reaching a power just below 0.5.
In the dependent case, the two components that have no treatment effect \((Y_1 \text{ and } Y_2)\) are dependent of each other. In this setting, the Tippett function has the highest power. It is followed by the Fisher function that has only slightly lower power than the Tippett function. The Logistic function shows a similar behaviour as in the independent case; for small treatment effects, it has the same power as the Fisher function. When the treatment effect increases, it cannot keep up with the Fisher function and deviates, increasing slower and evens out completely at a treatment effect of 0.8. The standard method is yet again the worst out of the four methods, the power slowly increasing and at most reaching a power of 0.8.
Figure 4. Power as a function of treatment effect on component $Y_3$ for both independent and dependent cases ($Y_1$ and $Y_2$ dependent, $Y_3$ independent), for each method. Treatment effect on components $Y_1$ and $Y_2$ in group $B = 0.5$. Treatment effect in group $A = 0.5$.

When the power curves for the four methods are compared between the independent and the dependent cases (Figure 4) the trend of the power being higher in the independent cases from Situation 1 (Figure 2) can no longer be seen in Situation 2. For the Fisher function, the power in the independent case is slightly higher than in the dependent case. For the Logistic function, the power for the independent and the dependent cases intersects multiple times when the treatment effect reaches over 0.75 and therefore it cannot be stated in which situation it performs best. The Tippett function performs equally well in both cases. The standard method has markedly higher power in the dependent case. This is intuitive, since the standard method requires a response in all three components to obtain a response in the composite endpoint.
When the two components without treatment effect are correlated, the probability of getting a response in the composite increases and thus the power is higher when the components are correlated than when they are not.

5.2.2 Situation 3

The following results were obtained in the situation where two of the components are affected by the treatment.

**Figure 5.** Power as a function of treatment effect on components $Y_1$ and $Y_2$. Treatment effect on component $Y_3$ in group B = 0.5. Treatment effect in group A=0.5. In the right graph, $Y_1$ and $Y_2$ are dependent.

The power behaviour for the four tested methods when the treatment affects two out of three components are shown in Figure 5. When the components are independent of each other, the Fisher combining function performs best out of the four, but is closely followed by the Tippett function. The Logistic combining function follows the Fisher function until the treatment effect reaches 0.65 where it deviates and evens out, at best obtaining a power of 0.85. The standard method performs worst out of the four until it reaches a treatment effect of 0.75, where it is equally good as the Logistic function and then outperforms it, but is still not as good as the Fisher and the Tippett functions until a treatment effect of above 0.85.

When the two components affected by the treatment are dependent, the same pattern as in the independent case are shown for the combining functions. The Fisher function has the highest power, closely followed by the Tippett function. The Logistic function keeps up with Fisher and Tippett until a treatment effect of 0.65, where it evens out at a maximum power of 0.8. The standard method performs poorly, with a slowly increasing power at most reaching 0.6.
Figure 6. Power as a function of treatment effect on components $Y_1$ and $Y_2$ for both independent and dependent cases ($Y_1$ and $Y_2$ dependent, $Y_3$ independent), for each method. Treatment effect on component $Y_3$ in group B = 0.5. Treatment effect in group A=0.5.

Comparing the power of the three combining functions between the independent and the dependent cases, there are no big differences between them. There is a tendency that the power is slightly higher when the components are independent of each other for both the Fisher, Logistic and Tippet functions but the difference is very small, and the power curves intersects on several occasions, making it hard to state a definite difference. For small treatment effects on the two components affected by the treatment, for the standard method the power is almost the same in both the independent and the dependent case. However, the difference gets bigger as the treatment effect increases, with the power being higher in the independent case.
5.3 A Remark Regarding Significance Levels

In the cases where the components are independent of each other, for all three situations it can be noted that the power is lower for the standard method than the nonparametric methods when the treatment effects are the same in both group A and group B (i.e. no difference between the groups). Since a significance level of 0.05 is used, the power for each method should be close to this value when there is no difference between the groups. However, the power of the standard method is far below 0.05, at 0.005. This is an extreme value, most likely due to chance, and if the procedure is repeated the power would be expected be closer to the chosen significance level of 0.05. Comparing the power of the standard method in the independent cases for all three situations, the power is the same: 0.005. This is explained by that the scenario, where both treatment groups have the same treatment effect on all components, exists as a starting point in all three situations. In order to avoid multiplicity problems, this scenario was only tested once and this estimation of the power is used in all three situations. Hence the unnaturally low power appear in all three situations, misleadingly indicating that the true significance level for the standard method is lower than the significance level for the nonparametric methods.
6 Discussion
6.1 Conclusions

Based on the results of this thesis, the nonparametric combining methodology offers a better way to test composite endpoints than the standard method in terms of power. The power differs depending on whether the components are dependent or not; in general, independent components lead to higher power than if the components are dependent of each other. The power of the four tested methods differs depending on the composition of treatment effects.

Comparing the standard method and the NPC methods, the simulations indicate that the NPC methods have higher power than the standard method, both in cases where the components are independent and cases where the components are dependent of each other. The fact that the standard method in general has the lowest power of all four methods is reasonable due to its constitution of requiring a response in all three components simultaneously to be considered a response in the composite endpoint.

When there is equal treatment effect on all components, the Fisher and the Logistic function have the highest power both in the independent and the dependent cases. This is in line with both theory provided by Pesarin (2001) and previous research by Loughin (2004). However, the Tippett function has almost as high power as the other two NPC methods in Situation 1. Since the Tippett function uses only one of the p-values, it is reasonable to think that it would perform similarly to the other two NPC methods if the treatment effects are the same for all components, since that would lead to three similar p-values.

For Situations 2 and 3 where there is not equal treatment effect on all components, Loughins (2004) research indicates that the Logistic function should work well when evidence is weak. It further indicates that the Fisher function is to prefer when evidence is moderately strong, and Tippett when evidence is strong. This is exactly what can be seen in Figures 3 and 5, the Logistic function works well until a treatment effect around 0.6-0.7, but cannot keep up with the performance of the Fisher or the Tippett functions above that treatment effect. The Fisher and the Tippett functions intersects in Situation 3, which is in line with the previous research by Westberg (1985). However, the results of Fisher and Tippett in Situation 2, where Tippett has slightly higher power than Fisher in the dependent situation, goes against Westbergs results for the corresponding situation. In the independent case, Fisher and Tippett has almost the same power. In the dependent case, the power of Tippett is the same as in the independent case, but
the power of the Fisher function is lower. This suggests that the Fisher function is somehow “punished” by the dependence.

The Fisher combining function has the highest power in all situations and cases, except for the dependent case in Situation 2 where the Tippett function has slightly higher power. This is in accordance with the theory provided by Pesarин (2001) and previous research by Loughin (2004) which states that the Fisher function is a good choice when there is no specific knowledge about the compositions of treatment effects.

Overall, if there is no knowledge about the compositions of evidence in the hypotheses, Loughin states that the Logistic combining function is a good choice since it is robust. The results in this thesis is not supportive of this. The results show good performance by the Logistic function in Situation 1, but in Situations 2 and 3, both the Fisher and the Tippett functions are much better choices. Loughin further states that the Logistic function should work slightly better than the Fisher function when the evidence is evenly spread (Situation 1). The results in this thesis does not support that, instead the Logistic and the Fisher functions has almost the same power. This could be due to the limited number of replications, or that there are no big differences between the two functions when $k=3$.

Further, the results support Maschas & Sesslers (2011) research that concludes that the power decreases if the correlation between components are increased. The results show that the power of all four methods is lower in the dependent cases than in the independent cases in all situations with one exception; the standard method in situation 2. Here, the standard method shows the opposite and is a reasonable result due to the nature of the composite endpoint, as mentioned in Section 5.2.1. However, this result would be expected to be seen for the standard method in situations 1 and 3 as well. The standard method should have higher power in the dependent cases than in the independent cases, as the dependency should increase the probability of getting a response in the composite endpoint. A possible explanation for this is that the trend of the power decreasing if the correlation between components is strong, is stronger than the effect of the composition of the composite endpoint in situation 1 and 3.

The final conclusions that are made from the results of this thesis, are that it is preferable to use a NPC method instead of the standard method when testing a composite endpoint in situations similar to the ones covered in this thesis. If completely certain that there is evidence in only one out of three components, use the Tippett combining function. Otherwise, use the Fisher
combining function. Be aware that strong dependency between components decreases the power.

Despite the fact that the nonparametric methods are to prefer in terms of power, other aspects of the advantages and disadvantages of the method has to be taken into account when deciding which method to use. It can be argued that the standard method is to prefer on the basis of interpretational aspects. From a medical point of view, the type of composite endpoint used in this thesis, where response has to be shown in all three components, offers a clean interpretation of what a significant result indicates – the treatment successfully affects all three components. The nonparametric methods can be somewhat more unclear from this point of view. A significant result using a NPC method indicates an overall effect, but as shown in this thesis, not all components have to show a treatment effect in order to obtain an overall significance. This makes the interpretation from a medical point of view a little ambiguous, as a significant result does not necessarily mean that the treatment was successful for all the components. Thus, it is important to consider more than just which method that has the highest power when deciding which method to use.

6.2 Methodological Discussion

The assumptions for the nonparametric combination method stated in Section 3.3 are mild and easily fulfilled. The exchangeability property in the null hypothesis is fulfilled since the randomization procedure is used. The hypotheses can be broken down into the required sub-hypotheses, suitable tests (chi-square tests in this case) exists, and all three tests are jointly analysed.

Regarding error sources, as always when simulating data there are possible error sources in the outline of the simulation, writing of the code, and counting errors. The outline, the code and the calculations have been thoroughly revised to minimize the risk of errors.

Due to time restrictions for the thesis, the number of replications performed was forced to the relatively low amount of 200 replications for the testing procedures and 200 replications for the resampling procedure. A higher number of replications would have been to prefer in order to get more accurate estimates of the power. As can be seen in the figures in Section 5, especially for the Logistic function in Situations 2 and 3, the estimates are not consistent. In theory, it is not possible for the power to decrease when the treatment effect is increased and everything else is held constant. The fact that it still does in some places, show the
shortcomings of a small number of replications. However, the number of replications is enough to be able to show trends in the power behaviour for the methods.

To investigate the research questions at hand, it would have been possible to analyse already existing data instead of performing a simulation study. An application of that method can be seen in Mascha and Sessler (2011). Using already existing data could have the benefit of being more representative of real life situations than simulated data, if the simulations are not done properly. Worth discussing is the method of generating dependence structures between components in this thesis. It is difficult to model and generate dependence structures between variables when dealing with multivariate problems, especially when dealing with binary variables (Pesarin, 2001; Oman & Zucker, 2001). Because of this, the dependence relations between the components in the simulation are made up without reference to real empirical data. Despite it being difficult to model the dependence relations between the components, methods to generate correlated binary variables have been suggested. Multivariate Bernoulli distributions could have been used, or some of the methods proposed by Lunn & Davies (1998) or Oman & Zucker (2001). The simulation could possibly have been improved by being more representative of real life situations by applying one of these methods.

What is lost in connection to empirical dependence relations can be said to be gained by the possibility to generate a lot of different situations in the simulation study. To find empirical data that possesses the desirable dependence structures that are modelled in this study could take time, and thus well performed simulations are a good substitute for real empirical data.

Further, an advantage of simulating data is that one avoids many of the troubles that follows with using already existing empirical data. Empirical data can suffer from drop-outs and typing-as well as measurement errors. When using empirical data, one also has to keep in mind that there could be other factors than the treatment of interest influencing the data. There are several methods to control for these potential problems, but by choosing to perform a simulation instead of using existing data, the work that has to be spent on cleaning up the data set is minimized.

Another advantage of using simulated data is related to the above discussion about using already existing data. In real life, one cannot perform experiments on human beings to the extent that would be preferable, where all conditions are exactly the same for each patient during the course of the trial, due to ethical aspects. When data is simulated, the “experiment” can be conducted exactly as the researcher wishes and thus disturbing factors can be avoided. The researcher can choose her own setting for the simulation.
Worth mentioning is also the potential problem of multiplicity. As long as the results of the tests of each component are not individually interpreted, there will be no problem with multiplicity. If the researcher is interested in treatment effects on the individual components, suitable correction techniques have to be used, for example Bonferroni correction. The problem of multiplicity is therefore present for both the standard methods and the NPC methods, but not when only the overall effect is of interest as it is in this thesis.

6.3 Suggestions for Further Research

The main result is that the nonparametric combining methods have higher power than the standard method when testing a composite endpoint, and is thus preferable to use. As the results for the standard method when comparing between the dependent and the independent situations did not follow expectations, further research on the effect of dependent variables is suggested.

The standard method used in this thesis is only one of many standard methods, and it would therefore be of relevance to further test standard methods such as a collapsed composite versus the nonparametric combination methods. Moreover, it would be appropriate to validate the findings in this thesis with a larger number of replications. Further research could also include extending the subject of this thesis to include more partial tests, other types of variables such as ordinal and continuous variables, and varying control group incidences, since all these factors possibly could affect the performance of the different methods.
References


Human Genome Sciences, Inc, 2010. *Arthritis Advisory Committee Meeting Briefing Document; Belimumab*.


Appendix A – Values for Generating Components

Components are generated in the following ways in the dependent cases.

In group A, if \( Y_1 \) equals 1, then \( Y_2 \) is generated from a binomial distribution \( Bin(1, p_{a1}) \) with treatment effect \( p_{a1} = 0.9 \). If \( Y_1 \) equals 0, \( Y_2 \) is generated from a binomial distribution \( Bin(1, p_{a0}) \) with treatment effect \( p_{a0} = 0.1 \). N observations for component \( Y_3 \) are generated in the same way as \( Y_2 \) in the dependent situations.

N observations for treatment group B in the dependent situations are generated with the same procedure as the observations for treatment group A, but using binomial distributions \( Bin(1, p_{b1}) \) and \( Bin(1, p_{b0}) \). \( p_{b1} \) is always 0.9. In order to maintain the decided treatment effect, \( p_{b0} \) has to be altered. Table 1 display the values of \( p_{b0} \) that are used.

**Table 1.** Values of \( p_{b0} \) used in the dependent cases of Situations 1, 2 and 3.

<table>
<thead>
<tr>
<th>Treatment effect of ( Y_1 )</th>
<th>Treatment effect of ( Y_2 )</th>
<th>( p_{b1} )</th>
<th>( p_{b0} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>0.50</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>0.55</td>
<td>0.55</td>
<td>0.9</td>
<td>0.122</td>
</tr>
<tr>
<td>0.60</td>
<td>0.60</td>
<td>0.9</td>
<td>0.15</td>
</tr>
<tr>
<td>0.65</td>
<td>0.65</td>
<td>0.9</td>
<td>0.186</td>
</tr>
<tr>
<td>0.70</td>
<td>0.70</td>
<td>0.9</td>
<td>0.233</td>
</tr>
<tr>
<td>0.75</td>
<td>0.75</td>
<td>0.9</td>
<td>0.3</td>
</tr>
<tr>
<td>0.80</td>
<td>0.80</td>
<td>0.9</td>
<td>0.4</td>
</tr>
<tr>
<td>0.85</td>
<td>0.85</td>
<td>0.9</td>
<td>0.567</td>
</tr>
</tbody>
</table>
Appendix B - Code Used for Simulations

# Probability for response in each group
pa1<-0.5
pa2<-0.5
pa3<-0.5
pb1<-0.x
pb2<-0.x
pb3<-0.x

# Number of patients in each treatment group
n<-100

# Creating group coding
group<-c(rep('A',n),rep('B',n))

# Creating vectors
pv1<-numeric()
pv2<-numeric()
pv3<-numeric()
Tfres<-numeric()
Tlres<-numeric()
Ttres<-numeric()
Zres<-numeric()

pvZ<-numeric()

kombTf<-numeric()
kombTl<-numeric()
kombTt<-numeric()
pZ<-numeric()
# Performing B number of simulations
for(i in 1:B){

  **Independent cases:**

  # Generating n responders per group and component
  a1<-rbinom(n,1,pa1)
a2<-rbinom(n,1,pa2)
a3<-rbinom(n,1,pa3)

  b1<-rbinom(n,1,pb1)
b2<-rbinom(n,1,pb2)
b3<-rbinom(n,1,pb3)

**Dependent case, Situation 1:**

# Generating n responders per group and component
  a1<-rbinom(n,1,pa1)
  for(k in 1:n){
    a2[k]<-ifelse(a1[k]==1,rbinom(1,1,0.9),rbinom(1,1,0.1))
    a3[k]<-ifelse(a1[k]==1,rbinom(1,1,0.3),rbinom(1,1,0.1))
  }

  b1<-rbinom(n,1,pb1)
  for(l in 1:n){
    b2[l]<-ifelse(b1[l]==1,rbinom(1,1,0.9),rbinom(1,1,0.1))
    b3[l]<-ifelse(b1[l]==1,rbinom(1,1,0.9),rbinom(1,1,0.1))
  }

**Dependent cases, Situation 2 and 3:**

# Generating n responders per group and component
  a1<-rbinom(n,1,pa1)
a3<-rbinom(n,1,pa3)

  for(k in 1:n){

33
a2[k]<-ifelse(a1[k]==1,rbinom(1,1,0.9),rbinom(1,1,0.1))

b1<-rbinom(n,1,pb1)
b3<-rbinom(n,1,pb3)
for(l in 1:n){
  b2[l]<-ifelse(b1[l]==1,rbinom(1,1,0.9),rbinom(1,1,0.1))
}

Proceeding with the same code for all situations and cases:

# Creating the composite endpoint
za<-ifelse((a1+a2+a3)==3,1,0)
zb<-ifelse((b1+b2+b3)==3,1,0)

# Creating result vector for the standard method
resZ<-c(za,zb)

# Creating result vectors for each component
res1<-c(a1,b1)
res2<-c(a2,b2)
res3<-c(a3,b3)

# Saving the observed p-values
pv1<-chisq.test(res1,group)$p.value
pv2<-chisq.test(res2,group)$p.value
pv3<-chisq.test(res3,group)$p.value
pvZ<-chisq.test(resZ,group)$p.value

# Test statistics
# Fisher combining function
Tf<-(-2)*(log(pv1)+log(pv2)+log(pv3)))
# Liptak combining function
Tl<-\( \log\left(\frac{1-pv1}{pv1}\right)+\log\left(\frac{1-pv2}{pv2}\right)+\log\left(\frac{1-pv3}{pv3}\right) \)

# Tippett combining function
Tt<-\( \max\left(1-pv1, 1-pv2, 1-pv3\right) \)

# Loops B times to obtain the null distribution for each test statistic
for\( (j in 1:B) \) {
  # Randomizing new treatment group belonging
  gr2<-sample(group,replace=F)

  # Calculating p-values for each test statistic under the null hypothesis
  pv1<-chisq.test(res1,gr2)$p.value
  pv2<-chisq.test(res2,gr2)$p.value
  pv3<-chisq.test(res3,gr2)$p.value

  # Calculating the test statistics
  # Fisher combining function
  kombTf[j]<-\((-2)\ast(\log(pv1)+\log(pv2)+\log(pv3))\)\)

  # Liptak combining function
  kombTl[j]<-\(\log\left(\frac{1-pv1}{pv1}\right)+\log\left(\frac{1-pv2}{pv2}\right)+\log\left(\frac{1-pv3}{pv3}\right) \)

  # Tippett combining function
  kombTt[j]<-\( \max\left(1-pv1, 1-pv2, 1-pv3\right) \)

  # calculating p-values for the standard method
pZ[j]<-chisq.test(resZ,gr2)$p.value

# are the values from the combining functions among the 5% highest in the null distribution?
Tfres[i]<-ifelse(Tf>quantile(kombTf,0.95),1,0)
Tlres[i]<-ifelse(Tl>quantile(kombTl,0.95),1,0)
Ttres[i]<-ifelse(Tt>quantile(kombTt,0.95),1,0)

# How often is the null hypothesis rejected with the standard method?
Zres[i]<-ifelse(pvZ<0.05,1,0)

# Calculating power for each test statistic
mean(Tfres)
mean(Tlres)
mean(Ttres)
mean(Zres)
### Appendix C - Results

**Table 2.** Situation 1, same treatment effect on all components.

<table>
<thead>
<tr>
<th>Treatment effect for group B (A=0.5)</th>
<th>0.5</th>
<th>0.55</th>
<th>0.6</th>
<th>0.65</th>
<th>0.7</th>
<th>0.75</th>
<th>0.8</th>
<th>0.85</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher</td>
<td>0.06</td>
<td>0.2</td>
<td>0.545</td>
<td>0.91</td>
<td>0.99</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Liptak</td>
<td>0.06</td>
<td>0.155</td>
<td>0.525</td>
<td>0.9</td>
<td>0.985</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Tippett</td>
<td>0.045</td>
<td>0.16</td>
<td>0.41</td>
<td>0.77</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Standard</td>
<td>0.005</td>
<td>0.07</td>
<td>0.34</td>
<td>0.755</td>
<td>0.935</td>
<td>0.99</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 3.** Situation 1, same treatment effect on all components. All components are dependent.

<table>
<thead>
<tr>
<th>Treatment effect for group B (A=0.5)</th>
<th>0.5</th>
<th>0.55</th>
<th>0.6</th>
<th>0.65</th>
<th>0.7</th>
<th>0.75</th>
<th>0.8</th>
<th>0.85</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher</td>
<td>0.085</td>
<td>0.14</td>
<td>0.375</td>
<td>0.695</td>
<td>0.89</td>
<td>0.99</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Liptak</td>
<td>0.085</td>
<td>0.145</td>
<td>0.385</td>
<td>0.695</td>
<td>0.89</td>
<td>0.99</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Tippett</td>
<td>0.065</td>
<td>0.12</td>
<td>0.325</td>
<td>0.61</td>
<td>0.97</td>
<td>0.975</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Standard</td>
<td>0.045</td>
<td>0.07</td>
<td>0.175</td>
<td>0.355</td>
<td>0.65</td>
<td>0.75</td>
<td>0.94</td>
<td>0.975</td>
</tr>
</tbody>
</table>

**Table 4.** Situation 2, treatment effect on one component.

<table>
<thead>
<tr>
<th>Treatment effect for Y3 in group B (A=0.5, Y1 och Y2 in group B=0.5)</th>
<th>0.5</th>
<th>0.55</th>
<th>0.6</th>
<th>0.65</th>
<th>0.7</th>
<th>0.75</th>
<th>0.8</th>
<th>0.85</th>
<th>0.9</th>
<th>0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher</td>
<td>0.06</td>
<td>0.095</td>
<td>0.23</td>
<td>0.415</td>
<td>0.715</td>
<td>0.855</td>
<td>0.985</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Liptak</td>
<td>0.06</td>
<td>0.09</td>
<td>0.185</td>
<td>0.31</td>
<td>0.525</td>
<td>0.56</td>
<td>0.59</td>
<td>0.72</td>
<td>0.635</td>
<td>0.63</td>
</tr>
<tr>
<td>Tippett</td>
<td>0.045</td>
<td>0.14</td>
<td>0.205</td>
<td>0.405</td>
<td>0.75</td>
<td>0.88</td>
<td>0.985</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Standard</td>
<td>0.005</td>
<td>0.015</td>
<td>0.065</td>
<td>0.085</td>
<td>0.12</td>
<td>0.14</td>
<td>0.215</td>
<td>0.36</td>
<td>0.36</td>
<td>0.47</td>
</tr>
</tbody>
</table>

**Table 5.** Situation 2, treatment effect on one component. The two non-affected components are dependent.

<table>
<thead>
<tr>
<th>Treatment effect for Y3 in group B (A=0.5, Y1,Y2 in group B=0.5)</th>
<th>0.5</th>
<th>0.55</th>
<th>0.6</th>
<th>0.65</th>
<th>0.7</th>
<th>0.75</th>
<th>0.8</th>
<th>0.85</th>
<th>0.9</th>
<th>0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher</td>
<td>0.03</td>
<td>0.07</td>
<td>0.145</td>
<td>0.305</td>
<td>0.58</td>
<td>0.795</td>
<td>0.935</td>
<td>0.995</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Liptak</td>
<td>0.04</td>
<td>0.05</td>
<td>0.135</td>
<td>0.235</td>
<td>0.42</td>
<td>0.575</td>
<td>0.7</td>
<td>0.675</td>
<td>0.705</td>
<td>0.655</td>
</tr>
<tr>
<td>Tippett</td>
<td>0.035</td>
<td>0.095</td>
<td>0.2</td>
<td>0.41</td>
<td>0.715</td>
<td>0.89</td>
<td>0.97</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Standard</td>
<td>0.02</td>
<td>0.045</td>
<td>0.1</td>
<td>0.125</td>
<td>0.255</td>
<td>0.365</td>
<td>0.505</td>
<td>0.645</td>
<td>0.76</td>
<td>0.82</td>
</tr>
</tbody>
</table>
### Table 6. Situation 3, treatment effect on two components.

<table>
<thead>
<tr>
<th></th>
<th>0.5</th>
<th>0.55</th>
<th>0.6</th>
<th>0.65</th>
<th>0.7</th>
<th>0.75</th>
<th>0.8</th>
<th>0.85</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher</td>
<td>0.06</td>
<td>0.095</td>
<td>0.31</td>
<td>0.735</td>
<td>0.95</td>
<td>0.99</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Liptak</td>
<td>0.06</td>
<td>0.115</td>
<td>0.315</td>
<td>0.64</td>
<td>0.74</td>
<td>0.795</td>
<td>0.83</td>
<td>0.86</td>
</tr>
<tr>
<td>Tippett</td>
<td>0.045</td>
<td>0.1</td>
<td>0.21</td>
<td>0.64</td>
<td>0.88</td>
<td>0.985</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Standard</td>
<td>0.005</td>
<td>0.045</td>
<td>0.115</td>
<td>0.315</td>
<td>0.55</td>
<td>0.765</td>
<td>0.89</td>
<td>0.965</td>
</tr>
</tbody>
</table>

### Table 7. Situation 3, treatment effect on two dependent components.

<table>
<thead>
<tr>
<th></th>
<th>0.5</th>
<th>0.55</th>
<th>0.6</th>
<th>0.65</th>
<th>0.7</th>
<th>0.75</th>
<th>0.8</th>
<th>0.85</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher</td>
<td>0.03</td>
<td>0.115</td>
<td>0.265</td>
<td>0.61</td>
<td>0.88</td>
<td>0.96</td>
<td>0.995</td>
<td>1</td>
</tr>
<tr>
<td>Liptak</td>
<td>0.04</td>
<td>0.105</td>
<td>0.195</td>
<td>0.535</td>
<td>0.755</td>
<td>0.81</td>
<td>0.82</td>
<td>0.815</td>
</tr>
<tr>
<td>Tippett</td>
<td>0.035</td>
<td>0.11</td>
<td>0.23</td>
<td>0.525</td>
<td>0.815</td>
<td>0.955</td>
<td>0.98</td>
<td>1</td>
</tr>
<tr>
<td>Standard</td>
<td>0.02</td>
<td>0.035</td>
<td>0.075</td>
<td>0.2</td>
<td>0.265</td>
<td>0.34</td>
<td>0.495</td>
<td>0.615</td>
</tr>
</tbody>
</table>