2014•2015
FACULTY OF SCIENCES
Master of Statistics

Master's thesis

Multiple Testing Procedures: The Application of Graphical Approaches in the Design of Clinical Trials

Promotor :
dr. Tatsiana KHAMIAKOVA

Promotor :
Ms. AGNES BALOGH

Maria Cristina Mingala
Thesis presented in fulfillment of the requirements for the degree of Master of Statistics

universiteit hasselt | Maastricht University

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

universiteit hasselt
KNOWLEDGE IN ACTION

Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek

2014•2015
FACULTY OF SCIENCES
Master of Statistics

Master's thesis

Multiple Testing Procedures: The Application of Graphical Approaches in the Design of Clinical Trials<u></u>

Promotor :
dr. Tatsiana KHAMIAKOVA

Promotor :
Ms. AGNES BALOGH

Maria Cristina Mingala
Thesis presented in fulfillment of the requirements for the degree of Master of Statistics

universiteit hasselt | Maastricht University

universiteit hasselt
KNOWLEDGE IN ACTION

MASTER THESIS PROJECT

MULTIPLE TESTING PROCEDURES: THE APPLICATION OF GRAPHICAL APPROACHES IN THE DESIGN OF CLINICAL TRIALS

Author:
Maria Cristina Mingala

Supervisors:
Dr. Tatsiana KHAMIAKOVA
(Univeristeit Hasselt)
Mevrouw AgnesBALOGH
(Bristol-Myers Squibb)

January, 2015

Contents

List of Tables

List of Figures

Acknowledgement

I would like to express my deep gratitude to all the people who helped me with this paper, most especially to my dedicated supervisors Ms. Agnes Balogh and Dr. Tatsiana Khamiakova, to Bristol-Myers Squibb especially to Mr. Harry Goyvaerts, to my friends here in Belgium, in the Philippines whose names are not mentioned but are all remembered in my heart. I would like to thank also Lazaro Mwakesi who pushed me to my limits in order to make this paper better. I would also like to thank the VLIR-UOS who funded for my studies here in Belgium. Most of all, I would like to thank our Creator, who is always guiding and protecting me.

1 Introduction

Myriad of applications of statistics involve simultaneous testing of many hypotheses. In clinical trials, **multiplicity** may easily arise in situations such as multiple endpoints, multiple primary variables, multiple comparisons of treatments, repeated measurements over time, and interim analyses [7].

Such multiplicity issues have to be addressed accordingly in order to avoid false findings and misleading conclusions. Some occasionally preferable possible solutions to lessen or avoid multiplicity have been developed such as identification of the key primary variable (multiple variables), choosing a critical treatment contrast (multiple comparisons), and use of a summary measure like area under the curve (repeated measures). According to the ICH Harmonised Tripartite Guideline Statistical Principles for Clinical Trials (ICH E9), in **confirmatory analyses**, any aspects of multiplicity which remain after the above procedure(s) should be addressed and should be determined in the protocol, while the details of any adjustment procedure or an explanation of why the adjustment is not necessary should be incorporated in the analysis plan. In hypothesis testing, interpreting presumptive statistically significant findings when there is no multiplicity adjustment should be dealt with care [7].

According to ICH E9, if hypothesis tests are used, committing the **Type II error** (failure to reject a false null hypothesis) is also of concern: in a clinical trial, committing a Type II error means failure to demonstrate that the treatment works when in fact it does [7]. The choice of Type II error rate is done by the trial sponsor.

In a case of testing a single hypothesis, a statistical test in the absence of a treatment effect can result in a conclusion in favor of a treatment effect by chance. Such error, also called as **Type I error**, can become excessive when many hypotheses are tested and proper multiplicity adjustment is not imposed. This condition inflates the **familywise error rate (FWER)** discussed in Section 3 and can lead to significant results for ineffective treatments. Thus, it is imperative to control this error probability at a prespecified level. This can be done through appropriate design and well-planned analyses strategies [6]. Controlling the Type I error is more of interest for, and is imposed by, regulatory authorities and for drug approval because it controls the probability of putting an inefficient drug on the market and/or putting a wrong claim in the product label. Controlling the familywise Type I error in strong sense (defined in Section 3) is a prerequisite of the regulatory agencies for Phase III confirmatory claims. Such trials are necessary to provide clear and firm evidence of efficacy, safety or evidence in support of claims about the treatment [13].

The conceptual definitions of Type I and Type II error rates can be illustrated in a statistical inference decision table as shown in Table 1. In this table, the unknown truth is crossed with the statistical decision. The null hypothesis is presumed to be true until statistical evidence in a form of a hypothesis test indicates otherwise. Not rejecting the null hypothesis implies two things: either the null hypothesis is actually true, implying a correct decision in lower left cell of Table 1, or that the null hypothesis is actually false but there was no enough evidence observed to reject it hence committing a Type II error. Similarly, rejecting a null hypothesis implies that either the null hypothesis is actually false, reflecting a correct decision in the upper right cell, or that the null hypothesis is actually true but we concluded otherwise hence committing a Type I error [12].

Table 1: Statistical Inference Decision Table

Inferential Decision	**True Null Hypothesis**	**False Null Hypothesis**
Reject Null Hypothesis	Type I error Probability = a	Correct Decision Probability = 1-b
Do not Reject Null Hypothesis	Correct Decision Probability = 1-a	Type II error Probability = b

a = Type I error rate
b = Type II error rate ; power = 1-b
Source:[12]

Various classes methods of adjustments for multiplicity has been developed throughout the years from the simplest Bonferroni adjustment to the methods that can handle complex designs. According to the amount of distributional information utilized, these methods can be broadly classified into three classes: nonparametric procedures, that do not make any assumptions about the joint distribution of the hypothesis test statistics; semiparametric procedures, that require additional distributional assumptions to establish FWER control but do not explicitly depend on the joint distribution of the hypothesis test statistics; and the parametric procedures, that require explicit assumptions about the joint distribution of the hypothesis test statistics [1] [5].

In the recent years, graphical approaches have been proposed to facilitate the visualization and communication of **Bonferroni-based closed test procedures** for multiple test problems similar to the clinical trial considered in this paper, such as comparing several treatments with control and assessing the benefit of a new drug for more than one endpoint. This enabled us to first derive the suitable weighting strategies that reflect the given study objectives and then apply the appropriate test procedures. We consider two families of test procedures based on **weighted Bonferroni tests** and **weighted Simes tests** [3].

1.1 Statements of objectives

Based on a completed clinical trial, investigate the effect of different multiple testing procedures within the framework of graphical approaches with respect to the power of the results.

In Section 2, we show the testing strategy used in the CSR. We then introduce the graphical approach in multiple testing in Section 3 and explain how it will be used in the paper. The weighted Bonferroni test and weighted Simes test will also be discussed. Moreover, the simulation settings used to investigate the power of the different testing strategies considered will be described in the same section. The rest of the paper discusses the results and offers some conclusions and recommendations.

2 Data description and CSR testing strategy

The clinical trial utilized in this paper is from a completed study investigating a currently marketed drug. Due to confidentiality reasons, all the information that will identify the study drug, its therapeutic area, dosage and endpoints of the clinical trial study used in this paper were masked. Hence, this paper will focus more on the methodological aspect of multiple testing procedures and its effect on the power of the test. Furthermore, the working data set is not the full CSR data set: for the purposes of this thesis, 5 observations per treatment arm were randomly deleted for each endpoint, and all the results presented in the paper are based on this data set. Nonetheless, the conclusions based in this truncated data set are entirely valid because the deletion of observations did not generate any substantial difference from the CSR results that would affect interpretation.

The trial was a confirmatory Phase III, multicenter, randomized, four-arm, parallel group, double-blind, placebo-controlled trial. The patients were randomized in a 1:1 ratio to one of the following treatment arms: 3 active dose levels of study drug (low, medium, and high doses) or placebo. All endpoints were then evaluated for the 3 dose levels versus placebo. Since only the testing of efficacy endpoints are subject to the control for Type I error, the safety endpoints were not discussed in this paper.

The trial efficacy endpoints are as follows:

Primary efficacy endpoint

- **Primary**: Change from baseline at Week T of the primary efficacy variable
  (continuous variable, negative value indicates improvement)

Secondary efficacy endpoints

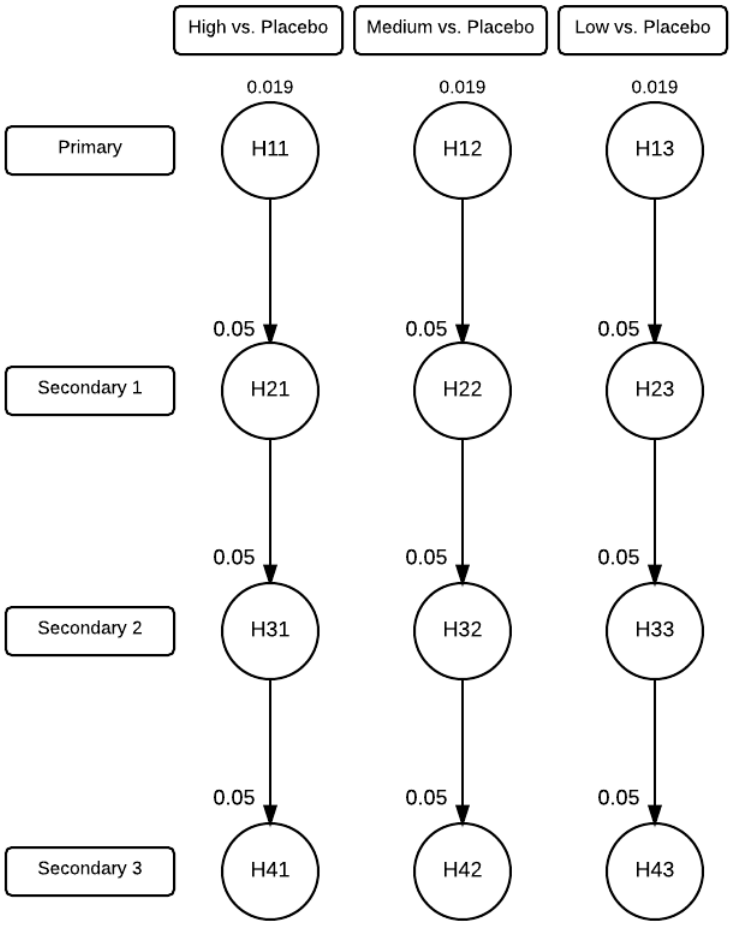
- **First Secondary (S1)**: Change from baseline at Week T in first secondary efficacy variable
  (continuous variable, negative value indicates improvement)
- **Second Secondary (S2)**: Proportion of responders based on primary efficacy variable. A subject was considered as responder if its corresponding primary efficacy endpoint value at Week T is below a certain threshold.
  (proportion; if a patient is responder, he/she is coded as 1)
- **Third Secondary (S3)**: Change from baseline at Week T in third secondary efficacy variable
  (continuous variable, negative value indicates improvement)

For all the endpoints, in case that there is no available measurement at Week T, the Last Observation Carried Forward (LOCF) method was used to approximate the measurement. All endpoints included subjects from the Randomized Subjects Data Set, which consisted all randomized subjects who took at least one dose of double-blind treatment.

Figure 1 presents the testing strategy used in the clinical trial considered in this paper. There were 12 elementary hypotheses of different importance to be tested. These hypotheses are denoted as H_{mj}, where m corresponds to the m^{th} efficacy endpoint following the sequential testing methodology and j is the j^{th} pairwise comparison ($m = 1,2,3,4$ denote the primary, S1, S2, and S3 endpoints; $j = 1,2,3$ denote the high vs. placebo, medium vs. placebo, and low vs. placebo comparisons). The diagram also shows the sequential testing methodology. In order

to control the Type I error rate within each treatment group at the 0.05 level, the statistical testing of all the 4 study endpoints proceeded sequentially. More specifically, the sequential testing procedure is reflected by the arrows linking the hypotheses within the treatment comparisons and the values at each hypothesis are the respective levels of significance per pairwise comparison. The significance of the treatment comparisons for the primary efficacy endpoint was a requirement in order to proceed in testing the secondary efficacy endpoints. That is, for the change from baseline to Week T in secondary efficacy endpoint S1, only those active drug treatment groups significantly superior to placebo under the primary efficacy endpoint were tested against placebo. Similarly, testing for the secondary efficacy endpoints proceeded sequentially. At each step in the testing sequence, only the active drug treatment groups significantly superior to placebo were tested at the subsequent step. For primary efficacy endpoint, individual comparison of an active drug treatment group against the placebo group was performed at $\alpha = 0.019$ level. The (pairwise) Type I error rate for comparing each active drug treatment group to placebo for each secondary efficacy endpoint was 0.05.

Figure 1: CSR Testing Strategy diagram



The primary efficacy analysis in this study was based on analysis of covariance (**ANCOVA**)-adjusted mean change from baseline to Week T, for 3 dose levels of active drug (low, medium, high) and the placebo group. Specifically, ANCOVA was used to compare the change from baseline values across treatment groups, adjusting for the baseline weight and is based on the following model [10]:

$$\mathbf{Y_{kl}} = \mu + \alpha_{\mathbf{k}} + \beta(\mathbf{X_{kl}} - \bar{\mathbf{X}}_{..}) + \varepsilon_{\mathbf{kl}}$$

where

$\mathbf{Y_{kl}}$ is the change from baseline to Week T primary efficacy variable value for the l^{th} subject belonging to a treatment group k with baseline value X_{kl}

$\bar{\mathbf{X}}_{..}$ is the mean of the baseline values X_{kl}

μ is overall mean

$\alpha_{\mathbf{k}}$ is the effect of the k^{th} treatment group

β is the regression coefficient parameter that represents the degree of linear relationship between Y_{kl} and X_{kl}

ε_{kl} independent $\sim N(0, \sigma^2)$

$k = 1,2,3,4$; $l = 1,...,n_k$

Under the structure of ANCOVA model, point estimates for mean changes for each active drug treatment group and the placebo group were obtained. A mentioned earlier, individual comparison of an active drug treatment group against the placebo group was also performed at $\alpha = 0.019$ level. The CSR testing strategy controlled the overall Type I error rate at the 0.05 significance level for the primary endpoint because in testing the 3 primary endpoint hypotheses, each of the 3 significance levels were adjusted to $\alpha = 0.019$ according to Dunnett's multiple comparison procedure for comparing several treatments with a control. This procedure performs the pairwise comparisons only for the placebo group against with each of the treatment group, hence reduces the multiple comparisons price to pay, while accounting for the correlation of the comparisons since they all use the same placebo group as control [14].

The secondary efficacy analysis was also subjected to the sequential testing methodology and was done in two different ways based on the nature of the endpoint. The continuous secondary efficacy endpoints, S1 and S3, were analyzed similarly as the primary endpoint (ANCOVA). The ANCOVA model used for the primary analysis described earlier was used, but the involved variables are changed according to S1 (and to S3). Within the structure of ANCOVA model, point estimates for the mean change from baseline within each treatment group as well as for the difference in the mean change from baseline between each active drug treatment group and the placebo group was calculated. The binary secondary efficacy endpoint S2 was also subjected to sequential testing. Using Fisher's exact test, the percentage of responders at Week T between each of the active drug treatment groups and the placebo group were compared. The frequency and percentage of responders at Week T were presented. Point estimates of the difference in response rates between each of the active drug treatment groups and the placebo group were obtained. The pairwise Type I error rate for comparing each active drug treatment group versus placebo group for each secondary efficacy endpoint hypothesis was kept at 0.05. This implies that for each secondary endpoint, the 3 hypotheses simultaneously tested were not adjusted for multiplicity.

Table 2 presents the p-values obtained from the CSR. It can be seen that all, except for the low versus placebo comparison for secondary efficacy endpoint S2, were lower than their respective predetermined significance levels. Thus, following the sequential testing method, the said

comparison cannot be tested further for S3. Hence, the p-value obtained there (value=0.0073) cannot be interpreted, even if it is lower than 0.05.

Table 2: P-values from ANCOVA (Primary, S1 and S3) and Fisher's exact p-values (S2)

	Treatment group comparison		
Endpoint	**Low vs. Placebo**	**Medium vs. Placebo**	**High vs. Placebo**
Primary	<.0001	<.0001	<.0001
S1	0.0001	0.0085	0.0001
S2	0.0983^{ns}	0.0245	0.0048
S3	(0.0073)	0.0096	0.0032

ns = not significantly lower than 0.05

Table 9 in Appendix A.2 shows the summary statistics at baseline, at Week T and change from baseline per arm for primary efficacy endpoint and secondary efficacy endpoints S1 and S3. Pairwise correlations at baseline and at Week T were presented in Table 11. The proportion of responders based on primary efficacy variable per arm is shown in Table 10.

3 Methodology

3.1 The Graphical Approach in Multiple Testing Adjustment

The **closure principle** defined in Appendix A.1 was used by Bretz et al. (2011b) to construct powerful multiple test procedures. This procedure controls the **familywise error rate (FWER)** in a **strong sense** at a level α (where α lies between 0 and 1) [3]. FWER is the probability of incorrectly rejecting at least one true null hypothesis. Strong control of FWER refers to controlling FWER under all possible combinations of true and false null hypotheses. This type of control allows us to control the maximum probability of incorrectly rejecting at least one true null hypothesis. Testing procedures control the FWER in **weak sense** when the FWER control is secured only when all the null hypotheses are true [5]. Since in clinical trials, the researchers aim for strong control of FWER, testing procedures which control only in a weak sense were not the focus of this paper.

The main idea about graphical approaches is to express the resulting multiple test procedures by directed, weighted graphs, where each node corresponds to an elementary hypothesis, together with a simple algorithm to generate such graphs while sequentially testing the individual hypotheses. By using graphical approaches, the statisticians are able to explore different testing strategies together with the clinical team and therefore tailor the multiple test procedure to the given study objectives [3]. In this paper, the graphical approaches are applied using weighted Bonferroni testing procedures and weighted Simes test.

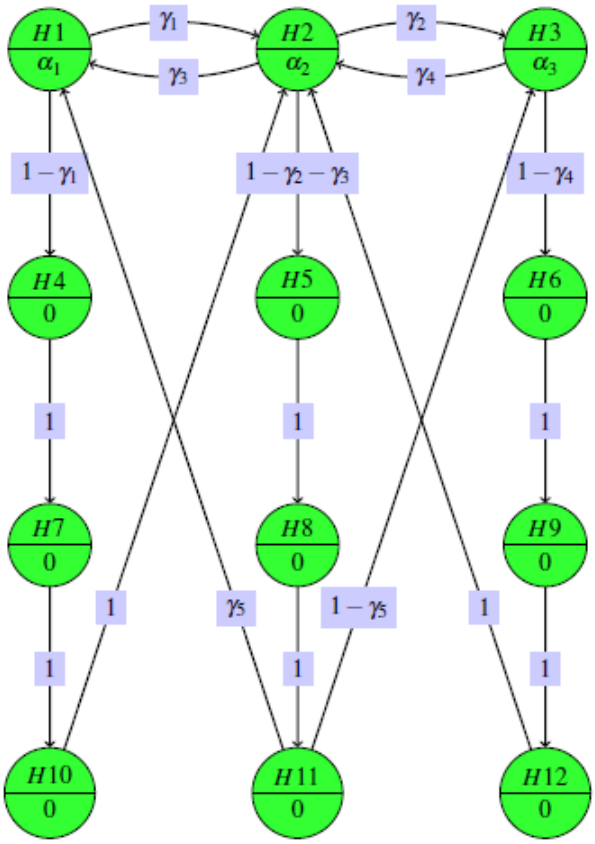
The notation of the hypotheses introduced in the previous section will be adopted. For ease in writing, the hypotheses will be recoded such that H_{mj} becomes H_i, where $i = 1, ..., 12$, so that $H_{11} = H_1$, $H_{12} = H_2$, $H_{13} = H_3$, $H_{21} = H_4$, $H_{22} = H_5$, and so on. Let $I = 1, ..., 12$ be the corresponding index set. Weighting strategies are formally defined through the **weights (or local significance levels)** $\mathbf{W} = (\alpha_1, ..., \alpha_{12})$ for the global null hypothesis H_I, and through the 12x12 **transition matrix** $\mathbf{G} = (g_{ij})$, where $0 \leq g_{ij} \leq 1$ and $\sum_{j=1}^{12} g_{ij} \leq 1$ for all $i, j = 1, ..., 12$. The local significance levels α_i are pre-determined such that they sum up to α. The transition matrix fully determines the edges, i.e., the weight g_{ij} controls the fraction of the local level α_i that is allocated to H_j in case H_i was rejected. Specifically, for a given index set $J \subseteq I$, let $J^c = I \backslash J$ be the set of indices that are excluded in J. The graph is updated using the algorithm described in Table 12 in Appendix A.3 as follows: Test the null hypothesis H_i having an assigned local significance α_i larger than zero. If the hypothesis is rejected, update the graph by splitting and reallocating its α_i to the succeeding null hypothesis(es) as determined by the pre-specified rules represented by a weighted graph. The edges g_{ij} will also be updated. All the edges coming from the rejected hypothesis will be removed and all the edges going towards it will be deflected to the subsequent hypothesis(es) with the updated local levels. With the reduced graph and updated local significance levels α_i, repeat the step for the remaining 11 non-rejected hypotheses. This process is repeated until no further hypothesis can be rejected. The algorithm of the above procedure is detailed in Section 2 of Bretz et al. (2011b) [3] [2].

The procedure is illustrated in Figure 5 of Appendix A.5. The initial α weights for the 3 primary endpoint hypotheses are $\frac{0.05}{3} = 0.0167$ and zero for all the secondary endpoint hypotheses. The transition matrix used is given is Table 15 in Appendix A.5. When a hypothesis, say H_1, is rejected, its local level is divided and propagated to the subsequent hypotheses (H_2 and H_4) according to the weights determined in the edges ($\frac{1}{2}$, $\frac{1}{2}$). The edges coming from H_1 to H_2 and to H_4 will then be removed. The other edge coming from H_{11} to H_1 will be detached

and attach to the H_2 and H_4. Since there are 2 subsequent hypotheses following H_1, the edge will split into two and its weight will be divided equally (from 1, two $\frac{1}{2}$'s will be produced).

The chosen weighting scheme of our improved CSR testing strategy is visualized in Figure 2. The hypotheses $H_1,...,H_{12}$ are represented by vertices with associated weights (local significance levels) $\alpha_1,...,\alpha_{12}$. Any two vertices H_i and H_j are linked together through directed edges (H_j is the tail), in which the associated weights g_{ij} determine the part of the significance level α_i to be propagated to H_j if H_i has been rejected. In a case when no propagation of the significance level is anticipated, the edge will be given a weight $g_{ij} = 0$ and is not shown in the graph [3]. For simplicity purposes, the edges of Figure 2 are renamed such that $\gamma_1 = g_{12}$, $\gamma_2 = g_{23}$, $\gamma_3 = g_{21}$, $\gamma_4 = g_{32}$, and $\gamma_5 = g_{11,1}$.

Figure 2: General Scheme of Improved CSR Testing Strategy Using Graphical Approach



The CSR testing procedure described in Section 2 is enhanced by establishing a suitable graph designed for the study objectives. This is done by allocating to α_1, α_2, and α_3 of the 3 primary hypotheses H_1, H_2, and H_3 some values greater than zero, while setting all the other initial local significance levels of the secondary hypotheses to zero. The edge pointing toward the subsequent S1 indicates that the primary hypotheses needs to be rejected first before the associated S1 will be tested. The same concept applies for the associated S2 and S3 hypotheses. In addition, only the primary hypotheses are tested in the initial step because they are given more importance. As mentioned earlier, if a primary hypothesis is rejected, its local significance level will be divided in a way that some will be propagated to the remaining primary hypotheses to be tested and some are propagated to the associated S1. This reflects the desire to further test also the other primary hypotheses rather than testing only the S1 associated with an already rejected primary hypothesis at higher significance level. In a case when S3 is rejected, its local

significance level will be propagated back to the remaining primary hypothesis(es) to be tested. This increases the chance for significant results in another active dose. Furthermore, the three primary hypotheses are treated equally, as reflected by the edges connecting them. It is also worth noting that there are no edges that connect any two non-adjacent dose groups (high to low or low to high edges), hence the two adjacent doses becoming significant is preferred [2].

Based on the **raw (unadjusted) p-values** denoted as $\mathbf{p} = (p_1, ..., p_{12})$ obtained from the primary and secondary efficacy analyses described in Section 2, a test procedure such as weighted Bonferroni or weighted Simes were then performed. In order to describe the outcome of the multiple testing strategy, the **adjusted p-values** defined by the graph (and hence, on the imposed testing strategy) was obtained. Any elementary null hypothesis is rejected if its corresponding adjusted p-value is below the predetermined significance level 0.05 [3].

3.2 Weighted Bonferroni procedure

The most basic multiplicity adjustment is the Bonferroni-based adjustment, which is simply splitting the significance level into the number of hypotheses to be tested [4]. In the weighted Bonferroni procedure, the weighted p-values were incorporated. The hypotheses to be tested $H_1, ..., H_{12}$ will be given nonnegative weights $w_1, ..., w_{12}$, where $\sum_{i=1}^{12} w_i = 1$. For hypothesis H_i, when $w_i > 0$, reject H_i if $pvalue_i \leq w_i\alpha$, and fail to reject H_i otherwise. The weights are pre-specified using available prior information [8].

The weighted Bonferroni procedure controls FWER at level α. This procedure is known to be conservative and lacks power. [4] [8]

3.3 Weighted Simes procedure

Simes test is a modification of the Bonferroni procedure which is based on the ordered p-values of the individual tests. This procedure is still simple to apply and is less conservative than the Bonferroni procedure. It is more advantageous over the latter when many highly correlated tests are involved [11].

The weighted Simes test rejects the null hypothesis if for some i $p_(i) \leq \frac{i}{12}\alpha$, $i = 1, ..., 12$ where the $p_(i)$ are the ordered values of $\frac{p_i}{w_i}$, $\sum_{i=1}^{12} = 12$ and $max(w_i) \leq \frac{1}{\alpha}$ [15].

[11].

3.4 Calculations

3.4.1 Exploratory data analysis on observed data

Summary statistics and and plots were used to obtain insight about the data. The four variance-covariance matrices for the 3 active drug treatment groups and the placebo group were calculated for the simulations.

3.4.2 Simulation

In order to examine the underlying power and FWER of the CSR testing strategy and to investigate the multiple testing procedures in the framework of graphical approaches, simulation was performed as follows: Three independent treatment groups of different dose levels of active study drug and a placebo group were considered. The 4 groups were treated as independent since the observations obtained from the different patients were independent from the other treatment groups. For each group, 4 correlated study endpoints were considered and measured. These endpoints were treated as dependent with each other because for each patient, the measurement of one endpoint is related to the other. Following the CSR, the secondary endpoint S2 was derived from the simulated primary endpoint. In order to be consistent with the assumptions in the CSR sample size calculation, increasing dose groups were assumed not to be ordinal in efficacy.

For each of these simulated samples (10,000 for weighted Bonferroni, 2,000 for weighted Simes), raw p-values were calculated as specified in the CSR (using ANCOVA and Fisher exact test). These raw p-values were then fed in the gMCP package to obtain the adjusted p-values.

Once obtaining the adjusted p-values, the local power for the 12 hypotheses were calculated. By definition, power is the probability of rejecting the null hypothesis when the alternative is true. Bretz et al. (2011a) proposed a power calculation by simulating the power under different realistic scenarios in order to understand the operating characteristics of a given multiple test procedure. The power is calculated for the non-null scenarios (described in Table 3), in which we assume that the alternative hypotheses are true. The power is obtained by calculating the percentage of correctly rejecting the null hypothesis for all the simulated date sets, giving the approximate power of the specific scenario considered. The FWER is calculated in the same manner as the power calculation, but to the null scenarios (described in Table 4). In the null scenarios, we assume that the null hypotheses are true. [2].

The Improved Testing Strategy based on Weighted Bonferroni tests

In an attempt to improve the CSR testing strategy discussed in Section 2, a weighted Bonferroni was applied to each of the intersection hypotheses (H_i, where $i = 1, ..., 12$) under the framework of graphical approaches. This method ensures that the strong control of FWER is preserved and allows recycling of the significance level α_j corresponding to the rejected hypothesis H_j.

Within the framework of graphical weighting strategy considered in this paper, the application of Bonferroni tests leads to shortcut procedures given that monotonicity condition (see Appendix A.1 for definition) is satisfied. The algorithm that is based on the weighted Bonferroni tests is presented in Table 13 [1][3].

Bonferroni-based test procedures are simple and often easier to communicate with other researchers as compared to other potentially more powerful tests available. However, one distinguished general disadvantage of using Bonferroni-based approaches is power loss, hence, another testing strategy called weighted Simes test was also utilized [3].

The Improved Testing Strategy Based on Weighted Simes tests

The Simes test is a known test when certain restriction in the correlations between the test

statistics are imposed, although these correlations are unknown. It assumes that the test statistics follow a joint multivariate normal distribution with positive correlations. The weighted Simes test rejects H_I for some $j \in I$ $p_{(j)} \leq \sum_{i=1}^{j} \alpha_{(i)}$, where $\alpha_{(i)} = w_{(i)}\alpha$ where $w_{(i)}$ is the weight associated with $p_{(i)}$.

The same weights vectors ($W1, W2, W3$), transition matrices ($M1, M2, M3$), and diagrams used for the weighted Bonferroni were utilized to apply the weighted Simes test on the framework of graphical approach.

Scenarios

It is essential to know how much significance level can be passed from one hypothesis to the other (for instance, from H_1 to H_2). As mentioned earlier, this is determined by the weights vector **W** and **G**. In order to investigate the effects of assigning different weights, 3 sets of initial alpha weights (denoted by $W1$, $W2$, and $W3$) and 3 transition matrices (denoted by $M1$, $M2$ and $M3$), hence nine combinations of **W** and **G** are considered. Sample diagrams [combinations: (a) M1W1, (b) M2W2, and (c) M3W3] of the weighting schemes and an illustration of the alpha propagation are presented in Figure 4 in Appendix A.4.

The values of the weights in $W1$ are predetermined in a way that there is a balanced allocation of the weights for the 3 primary hypotheses H_1, H_2, and H_3 ($\alpha_1 = \alpha_2 = \alpha_3 = \frac{0.05}{3}$) and no allocation for the weights of the secondary hypotheses. The values of initial alpha weights in $W2$ are assigned so that half of weights correspond to the high dose vs. placebo pairwise comparison of the primary hypothesis ($\alpha_1 = \frac{0.05}{2}$, $\alpha_2 = \alpha_3 = \frac{0.05}{4}$). The weights in $W3$ reflect the extreme situation wherein most of the weights are allocated to high dose vs. placebo pairwise comparison of the primary hypothesis, thus giving the highest significance level to it ($\alpha_1 = \frac{0.05*9}{10}$, $\alpha_2 = \alpha_3 = \frac{0.05}{20}$). Similarly, the values of g_{ij}'s in transition matrices $M1$, $M2$, and $M3$ were designed in a way that $M1$ propagates the local level α in a balanced manner, $M2$ propagates half of the local level α_j of the rejected hypothesis H_j to the high dose vs. placebo pairwise comparison of the primary hypothesis, and $M3$ is the extreme case of $M2$ wherein most of the local level of the rejected hypothesis is propagated to the high dose vs. placebo pairwise comparison of the primary hypothesis. The values assigned for weights vectors and transition matrices are presented in Tables 15 to 18 Appendix A.5.

Tables 3 summarizes the weighting schemes where the observed means of the baseline and Week T measurements of the subjects for the primary and first and third secondary endpoints were used as starting values for the simulation of different scenarios. It also presents the non-null scenarios which will be used to calculate the power. Corresponding null scenarios will be calculated to obtain the FWER.

Table 3: Description of Non-null Scenarios

	Description	
Scenario	**Simulated data**	**Testing strategy**
1. Without graphical approach		
1.1 CSROrig	CSR-based means, variances and correlations	per CSR
1.3 CSRIndep	Independent correlations	per CSR
*2. Weighted Bonferroni-based improved testing strategy using graphical approach**		
2.1a WBOrigM1W1	CSR-based means, variances and correlations	Balanced α propagation, equal initial α weights for 3 primary endpoint hypotheses
2.1b WBOrigM1W2	CSR-based means, variances and correlations	Balanced α propagation, More initial weights allocated to high dose vs. placebo comparison of primary hypothesis
2.1c WBOrigM1W3	CSR-based means, variances and correlations	Balanced α propagation, Most initial weights allocated and adding to high dose vs. placebo comparison of primary hypothesis
2.1d WBOrigM2W1	CSR-based means, variances and correlations	More α weights adding to high dose vs. placebo comparison of primary hypothesis, equal initial α weights for 3 primary endpoint hypotheses
2.1e WBOrigM2W2	CSR-based means, variances and correlations	More weights allocated to high dose vs. placebo comparison of primary hypothesis
2.1f WBOrigM2W3	CSR-based means, variances and correlations	More α weights adding to high dose vs. placebo comparison of primary hypothesis, most initial α weights allocated to high dose vs. placebo comparison of primary hypothesis
2.1g WBOrigM3W1	CSR-based means, variances and correlations	Most α weights adding to high dose vs. placebo comparison of primary hypothesis, equal initial α weights for 3 primary endpoint hypotheses
2.1h WBOrigM3W2	CSR-based means, variances and correlations	Most α weights adding to high dose vs. placebo comparison of primary hypothesis, More initial weights allocated to high dose vs. placebo comparison of primary hypothesis
2.1i WBOrigM3W3	CSR-based means, variances and correlations	Most weights allocated and adding to high dose vs. placebo comparison of primary hypothesis
3. Weighted Simes improved testing strategy using graphical approach		
*Same description apply for respective scenarios in weighted Simes test		

In order to establish that the CSR testing strategy does not control the FWER in a strong sense, different CSR null scenarios were also performed.

Table 4: Description of CSR Null Scenarios

Scenario	Description	
	Simulated data	**Testing strategy**
1. Without graphical approach		
1.2a CSRNulla	All means, variances and correlations set to equal to those of the placebo group	per CSR
1.2b CSRNullb	Equal means, except primary endpoint hypotheses CSR-based variances and correlations	per CSR
1.2c CSRNullc	Equal means and variances, independent correlations	per CSR
1.2d CSRNulld	Equal means, variance and correlations for medium dose, low dose, and placebo groups	per CSR

3.5 Software used

Exploratory analysis for all endpoints was done in SAS 9.3 and R version 3.1.2. Preliminary simulation analyses, covariance matrices, means, ANCOVA, and t-tests for the primary and secondary endpoints were obtained using SAS 9.3. Data simulation, ANCOVA, Fisher's exact test, multiplicity adjustment techniques, and graphical approaches using gMCP package were performed using R version 3.1.2 [9].

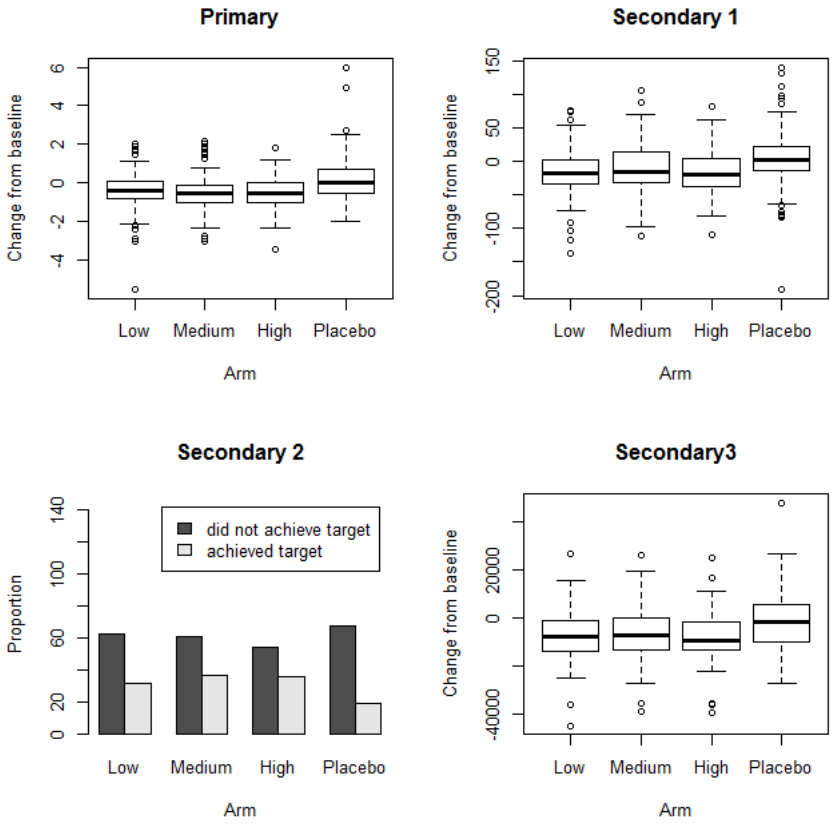
4 Results

4.1 Exploratory data analysis

Summary statistics of the 3 continuous endpoint variables P, S1, and S3 for each dose group were presented in Appendix A.2. Boxplots of the mean change from baseline values presented in Table 9 are shown in Figure 3. It can be seen that for P, S1, and S3, the placebo group is always the closest to zero, indicating that the placebo group has the least treatment effect as compared to the active treatment group. There is an ordinality of efficiency for low-medium-high dose groups for Primary and for S2. It was observed that for S1 and S3, the medium dose seems to perform slightly worse than the low dose. Furthermore, normality seems to be met for the continuous variables.

Moderate to strong positive correlations were observed. Stronger positive correlations were observed for the same variables measured at different times (for example, Baseline P vs. Week T P correlation = 0.65), while weaker correlations are observed different variables measured at different times (for example, Baseline P vs. Week T S3 correlation = 0.34). In general, the correlations seem to be comparable across the treatment groups.

Figure 3: Summary Plots of the Change for Baseline values of the Endpoint variables



4.2 Simulation

4.2.1 CSR Testing Strategy

Table 5 presents the proportion of rejections or the approximate power of the original CSR testing strategy. These two scenarios were simulated using the CSR-observed means. As expected based on the hierarchical approach, for each group (high, medium, low), there is a decreasing trend in the power as the testing is done sequentially from testing the primary variable P to testing the third secondary variable S3. Ignoring the underlying correlations between the variables, the power goes down slightly for each individual hypothesis. This was due to ignoring the positive correlations, hence the generated samples are expected to have larger variability than those with correlation. In addition, the powers obtained for P and S1 are much higher than for S2 and S3. Although the original CSR testing strategy gives reasonable power, this strategy does not control the FWER in strong sense, as shown in Table 6, where the FWER in scenario 1.2b is larger than 0.05.

In addition, the p-values in Table 5 shows ordinality for P and S2. There is also a big drop in the p-values for S2, may be due to losing information by categorizing the continuous variable P to binary. The p-values of S2 and S3 were also close to each other, indicating that once S2 is significant, S3 will be significant as well.

Table 5: Proportion of Rejections per Individual Hypotheses for CSR-based Testing Strategy under Non-Null Scenarios

	P			S1			S2			S3		
Scenario	**High**	**Med**	**Low**	**High**	**Med**	**Low**	**High**	**Med**	**Low**	**High**	**Med**	**Low**
1.1 CSROrig	0.9988	0.9887	0.9782	0.9833	0.8262	0.9547	0.4219	0.2980	0.1913	0.4218	0.2973	0.1912
1.3 CSRIndep	0.9703	0.8129	0.8232	0.8313	0.5667	0.6176	0.3616	0.2180	0.1435	0.3616	0.2012	0.1376

Table 6: Familywise Error Rate for CSR-based Testing Strategy under Null Scenarios

Scenario	**FWER**
1.2a CSRNulla	0.0245
1.2b CSRNullb	0.1523
1.2c CSRNullc	0.0026
1.2d CSRNulld	0.0314

4.2.2 Graphical Approaches-based Testing Strategies

The powers for the individual hypothesis obtained from generating different scenarios of graphical approaches using weighted Bonferroni and weighted Simes under the structure of graphical approaches are presented in Table 7.

As compared to 1.1, there is a monotone decrease of power from P to S3 and the power for each individual hypothesis is smaller in scenario 2.1. This is due to smaller initial alpha values used in 2.1 , that is 0.0167 is smaller than 0.019. There is also a big drop of power from P to S1 due to splitting the propagated α to the other primary endpoints. For P, S1 and S3, ordinality on the powers was also observed.

The above observations are the same for the other scenarios in Table 7. This indicates that any of the transition matrix-initial α weight combination does not affect these trend. Nevertheless, some of these combinations lead to a slightly higher power.

Similar to the results in scenarios 1.1 and 1.2, there is a drastic drop of power from S1 to S2. Furthermore, the powers for S2 and S3 within each dose level are very close to each other, indicating that once the adjusted p-value was rejected for S2 the adjusted p-value of the following hypothesis corresponding to S3 tend to be rejected as well.

Based on Table 7, the weighted Simes give very similar results as in weighted Bonferroni in terms of power, but not larger. Both multiple testing procedures also control the FWER is strong sense, as shown in Table 8.

Table 7: Proportion of Rejections per Individual Hypotheses for Graphical Approaches-based Testing Strategies under Non-Null Scenarios

	P			**S1**			**S2**			**S3**		
Scenario	**High**	**Med**	**Low**	**High**	**Med**	**Low**	**High**	**Med**	**Low**	**High**	**Med**	**Low**
2.1a WBOrigM1W1	0.9824	0.9412	0.8871	0.7849	0.3755	0.6888	0.0503	0.0261	0.0146	0.0502	0.0257	0.0144
2.1b WBOrigM1W2	0.9854	0.9393	0.8790	0.7995	0.3711	0.6736	0.0565	0.0258	0.0134	0.0564	0.0255	0.0132
2.1c WBOrigM1W3	0.9905	0.9364	0.8443	0.8275	0.3596	0.6211	0.0672	0.0245	0.0103	0.0671	0.0242	0.0101
2.1d WBOrigM2W1	0.9830	0.9367	0.8815	0.8176	0.3345	0.6716	0.0629	0.0209	0.0127	0.0628	0.0206	0.0125
2.1e WBOrigM2W2	0.9857	0.9327	0.8687	0.8289	0.3242	0.6504	0.0691	0.0206	0.0119	0.0690	0.0203	0.0118
2.1f WBOrigM2W3	0.9905	0.9179	0.8135	0.8501	0.2892	0.5645	0.0855	0.0197	0.0096	0.0855	0.0195	0.0094
2.1g WBOrigM3W1	0.9838	0.9346	0.8708	0.8395	0.2173	0.6428	0.0804	0.0141	0.0117	0.0804	0.0140	0.0117
2.1h WBOrigM3W2	0.9862	0.9263	0.8569	0.8462	0.2100	0.6162	0.0857	0.0153	0.0104	0.0857	0.0151	0.0103
2.1i WBOrigM3W3	0.9905	0.8946	0.7769	0.8596	0.1721	0.4972	0.0912	0.0155	0.0089	0.0912	0.0154	0.0088
3.1a SiOrigM1W1	0.9820	0.9465	0.8890	0.7915	0.3835	0.6860	0.0410	0.0265	0.0150	0.0410	0.0265	0.0150
3.1b SiOrigM1W2	0.9840	0.9450	0.8805	0.8050	0.3770	0.6720	0.0480	0.0265	0.0145	0.0480	0.0265	0.0145
3.1c SiOrigM1W3	0.9900	0.9385	0.8455	0.8300	0.3640	0.6180	0.0560	0.0255	0.0120	0.0560	0.0255	0.0120
3.1d SiOrigM2W1	0.9835	0.9420	0.8835	0.8215	0.3415	0.6700	0.0525	0.0205	0.0135	0.0525	0.0205	0.0135
3.1e SiOrigM2W2	0.9845	0.9365	0.8720	0.8320	0.3295	0.6485	0.0585	0.0190	0.0125	0.0585	0.0190	0.0125
3.1f SiOrigM2W3	0.9900	0.9215	0.8080	0.8545	0.2885	0.5575	0.0740	0.0170	0.0100	0.0740	0.0170	0.0100
3.1g SiOrigM3W1	0.9845	0.9390	0.8750	0.8455	0.2165	0.6395	0.0690	0.0130	0.0125	0.0690	0.0130	0.0125
3.1h SiOrigM3W2	0.9860	0.9335	0.8585	0.8520	0.2110	0.6160	0.0750	0.0130	0.0105	0.0750	0.0130	0.0105
3.1i SiOrigM3W3	0.9900	0.8965	0.7725	0.8635	0.1655	0.4940	0.0795	0.0130	0.0090	0.0795	0.0130	0.0090

Table 8: Familywise Error Rate for Graphical Approaches-based Testing Strategies under Null Scenarios

Scenario	FWER
2.2a WBNullM1W1	< 0.001
2.2b WBNullM1W2	< 0.001
2.2c WBNullM1W3	< 0.001
2.2d WBNullM2W1	< 0.001
2.2e WBNullM2W2	< 0.001
2.2f WBNullM2W3	< 0.001
2.2g WBNullM3W1	< 0.001
2.2h WBNullM3W2	< 0.001
2.2i WBNullM3W3	< 0.001
3.2a SiNullM1W1	< 0.001
3.2b SiNullM1W2	< 0.001
3.2c SiNullM1W3	< 0.001
3.2d SiNullM2W1	< 0.001
3.2e SiNullM2W2	< 0.001
3.2f SiNullM2W3	< 0.001
3.2g SiNullM3W1	< 0.001
3.2h SiNullM3W2	< 0.001
3.2i SiNullM3W3	< 0.001

5 Discussion and conclusion

Multiplicity issues are common in clinical trials and may easily arise in many situations. If ignored, multiplicity can lead to false findings and misleading conclusions. Through the years, many multiplicity adjustment procedures have been developed to handle specific multiplicity problems. Recently, Bretz et al. (2009) proposed a method called graphical approaches that offers advantages such as flexible to cater complex designs, allows for recycling of significance level, and ensures strong FWER control for closed test procedures. It also enables the researchers to show the flow of statistical thinking in trial design and can coordinate different requirements from the clinical trial team [1][2]. Since graphical approaches are newly introduced in clinical trials, there is still a lot of room for research.

This paper is investigating the ways to improve the testing strategy of an already completed Phase III confirmatory clinical trial by finding a better multiplicity adjustment technique under the framework of graphical approach. The said CSR testing strategy was adjusted for multiplicity only for the primary endpoint hypotheses; the secondary hypotheses, however, are just adjusted within each dose group. The results of the trials are fully valid, and the strategy is the best that could have been designed with the multiplicity techniques of that time - however, with the most recent developments in statistical literature, this strategy leaves place for improvement. As a result, the FWER is only weakly controlled.

Under the framework of graphical approaches, two testing strategies were applied to improve the testing strategy: weighted Bonferroni test and weighted Simes test using the gMCP package in R. In order to compare the multiple testing procedures, simulations of different scenarios were performed and the local power for each hypotheses as calculated. Bretz et al (2011b) showed that the graphical approach provides a strong control of FWER. Indeed, we illustrated that when improving the original CSR testing strategy using the graphical approach, FWER strongly controlled. In addition, the power of the testing strategy depends on the predetermined weighting schemes.

In conclusion, the use the graphical approach offers improvement to the original testing strategy by ensuring strong control of FWER while flexibly allowing the recycling of significance level. However, the individual power decreases.

6 Limitations and Recommendations

For further studies, it is recommended to investigate the improvement of CSR testing strategy using other powerful tests that take into account the correlation between the test statistics such as parametric tests or Dunnett-based tests under the framework of graphical approaches.

References

[1] Bretz, F., Maurer, W., Brannath, W., and Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*, 28(4):586–604.

[2] Bretz, F., Maurer, W., and Hommel, G. (2011a). Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. *Statistics in Medicine*, 30(13):1489–1501.

[3] Bretz, F., Posch, M., Glimm, E., Klinglmueller, F., Maurer, W., and Rohmeyer, K. (2011b). Graphical approaches for multiple comparison procedures using weighted bonferroni, simes, or parametric tests. *Biometrical Journal*, 53(6):894–913.

[4] Dmitrienko, A. and D'Agostino, R. (2013). Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine*, 32(29):5172–5218.

[5] Dmitrienko, A., D'Agostino, R. B., and Huque, M. F. (2012). Key multiplicity issues in clinical drug development. *Statistics in Medicine*, 32(7):1079–1111.

[6] Dmitrienko, A., Tamhane, A., and Bretz, F. (2010). *Multiple Testing Problems in Pharmaceutical Statistics*. Taylor & Francis Group, LLC, Boca Raton, Florida.

[7] International Conference of Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (1998). Statistical Principles for Clinical Trials (E9). `http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf`. Accessed on 15-October-2014.

[8] Kang, G., Ye, K., Liu, N., Allison, D., and Gao, G. (2009). Weighted multiple hypothesis testing procedures. *Statistical Aapplications in Genetics and Molecular Biology*, 8(1):Article 23.

[9] Rohmeyer, K. and Klinglmueller, F. (2014). Graph based multiple comparison procedures. `http://cran.r-project.org/web/packages/gMCP/gMCP.pdf`. Accessed on 27-December-2014.

[10] Rutherford, A. (2001). *Introducing ANOVA and ANCOVA*. SAGE Publications Ltd, London.

[11] Simes, R. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–4.

[12] Singh, A. K., Kelley, K., and Agarwal, R. (2008). Interpreting results of clinical trials: A conceptual framework. *Clinical Journal of the American Society of Nephrology*, 3(5):1246–1252.

[13] The European Agency for the Evaluation of Medicinal Products (2002). Points to Consider on Multiplicity Issues in Clinical Trials. `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003640.pdf`. Accessed on 15-October-2014.

[14] Westfall, G., Tobias, R., and Wolfinger, R. (2011). *Multiple Comparisons and Multiple Tests Using SAS*. SAS Institute Inc., Cary, North Carolina, 2nd edition.

[15] Yoav, B. and Hochberg, Y. (1997). Multiple hypothesis testing with weights. *Board of the Foundation of the Scandinavian Journal of Statistics*, 24:407–418.

A Appendix

A.1 Glossary

Term	Definition
Familywise error rate (FWER)	probability to erroneously reject at least one true null hypothesis, beyond the pre-specified significance level $\alpha \in (0,1)$
Transition matrix G	is a mxm matrix containing the elements g_{ij} in which each element is the fraction of the level of H_i that is allocated to H_j. [1]
α **allocation rule**	is used in nonparametric and parametric chain procedures. This rule defines the initial distribution of the overall error rate among the null hypotheses. [6]
α **propagation rule**	is used in nonparametric and parametric chain procedures. This rule defines the process of redistributing the available error rate among the nonrejected null hypotheses after each rejection according to the pre-specified logical relationships among the null hypotheses. [6]
Closure principle	Given a set of m hypotheses to test simultaneously with pre-specified α-level test for each H_J. If *all* intersection hypotheses H_J, J is a subset of I, are rejected by their corresponding α-level tests, then the resulting closed test procedure rejects H_i, i is an element of I. The closure principle states that an FWER-controlling testing procedure can be constructed by testing each hypothesis in the closed family using a suitable local α-level test. This procedure rejects a hypothesis if all intersection hypotheses containing this hypothesis are rejected by the associated local tests. [6]

A.2 Summary statistics

Table 9: Summary statistics of Primary, S1 and S3 at Baseline, at Week T, and Change from Baseline per Group

		Low			Medium		
		P	**S1**	**S3**	**P**	**S1**	**S3**
Baseline	**N**	95	96	69	98	100	74
	Mean	7.945	178.100	45910.000	8.004	171.500	46197.000
	SD	0.955	40.659	11095.050	1.095	42.657	10860.830
	Min	6.100	85.000	22800.000	6.100	109.000	21680.000
	Max	11.200	263.000	71500.000	10.600	303.000	74970.000
Week T	**N**	95	96	72	98	100	76
	Mean	7.508	161.800	38790.000	7.508	162.800	39180.000
	SD	1.109	45.969	10539.800	1.307	50.102	11925.470
	Min	5.400	94.000	20320.000	4.100	94.000	17680.000
	Max	10.400	305.000	71840.000	11.900	353.000	78260.000
Change from baseline	**N**	95	96	69	98	100	74
	Mean	**-0.437**	**-16.320**	**-7254.000**	**-0.496**	**-8.700**	**-6965.000**
	SD	1.048	37.714	11602.750	1.003	39.849	11263.110
	Min	-5.500	-137.000	-44980.000	-3.000	-112.000	-39150.000
	Max	2.000	76.000	26910.000	2.200	107.000	26120.000
		High			**Placebo**		
		P	**S1**	**S3**	**P**	**S1**	**S3**
Baseline	**N**	90	92	68	87	87	61
	Mean	7.878	176.000	44680.000	7.868	172.900	46339.000
	SD	0.887	43.721	11607.570	0.915	46.808	11635.630
	Min	6.300	106.000	22980.000	6.300	92.000	17300.000
	Max	10.300	315.000	75030.000	10.100	315.000	81680.000
Week T	**N**	90	92	70	87	87	64
	Mean	7.331	159.300	36430.000	8.114	180.100	44821.000
	SD	0.983	38.529	10152.790	1.610	55.756	12567.100
	Min	5.500	99.000	13560.000	5.600	91.000	16780.000
	Max	10.700	265.000	62520.000	13.600	339.000	80700.000
Change from baseline	**N**	90	92	68	87	87	61
	Mean	**-0.547**	**-16.660**	**-7837.000**	**0.246**	**7.241**	**-1608.000**
	SD	0.784	32.873	11043.620	1.245	52.299	13677.400
	Min	-3.400	-110.000	-39600.000	-2.000	-191.000	-27500.000
	Max	1.800	83.000	24840.000	6.000	141.000	47970.000

Table 10: Proportion of responders based on Primary per group

	Low	**Medium**	**High**	**Placebo**
N	95	98	90	87
***Responders (Proportion %)**	32(33.68)	37 (37.76)	36(40.00)	19 (21.84)

*Proportion of responders based on primary efficacy variable

Table 11: Pairwise correlations of Primary, S1 and S3 at Baseline and at Week T per group

			Baseline			Week T		
			P	**S1**	**S3**	**P**	**S1**	**S3**
High	**Baseline**	**P**	1	0.6057	0.5952	0.6528	0.5030	0.3436
		S1	0.6057	1	0.6168	0.5171	0.6872	0.4014
		S3	0.5952	0.6168	1	0.3738	0.4197	0.4821
	Week T	**P**	0.6528	0.5171	0.3738	1	0.7443	0.6608
		S1	0.5030	0.6872	0.4197	0.7443	1	0.5721
		S3	0.3436	0.4014	0.4821	0.6608	0.5721	1
Medium	**Baseline**	**P**	1	0.6600	0.6668	0.6638	0.4790	0.4725
		S1	0.6600	1	0.6336	0.4901	0.6415	0.6287
		S3	0.6668	0.6336	1	0.4281	0.5066	0.5218
	Week T	**P**	0.6638	0.4901	0.4281	1	0.7252	0.7137
		S1	0.4790	0.6415	0.5066	0.7252	1	0.8302
		S3	0.4725	0.6287	0.5218	0.7137	0.8302	1
Low	**Baseline**	**P**	1	0.5649	0.5380	0.4926	0.3828	0.2920
		S1	0.5649	1	0.6374	0.5041	0.6270	0.4106
		S3	0.5380	0.6374	1	0.3996	0.4151	0.4338
	Week T	**P**	0.4926	0.5041	0.3996	1	0.8095	0.7895
		S1	0.3828	0.6270	0.4151	0.8095	1	0.7041
		S3	0.2920	0.4106	0.4338	0.7895	0.7041	1
Placebo	**Baseline**	**P**	1	0.7308	0.5771	0.6378	0.5299	0.5037
		S1	0.7308	1	0.5682	0.5418	0.4913	0.4399
		S3	0.5770	0.5682	1	0.2187	0.2866	0.3792
	Week T	**P**	0.6378	0.5418	0.2187	1	0.8565	0.7032
		S1	0.5299	0.4913	0.2866	0.8565	1	0.7135
		S3	0.5037	0.4399	0.3792	0.7032	0.7135	1

A.3 Algorithms for Generation of Weighted Graphs

Table 12: Algorithm for Weighting Strategy

Step 1	Select $j \in J^c$ and remove H_j	
Step 2	Update the graph:	
	old value	*new value*
	I	$I \backslash \{j\}$
	J^c	$J^c \backslash \{j\}$
	$w_l(I)$	$w_l(I) + w_j(I) g_{jl}$, if $l \in I$; 0, otherwise
	g_{lk}	$\frac{g_{lk}+g_{lj}g_{jk}}{1-g_{lj}g_{jl}}$, if $l,k \in I$, $l \neq k$,$g_{lj}g_{jl} < 1$; 0, otherwise
Step 3	If $\lvert J^c \rvert \geq 1$, go to step 1; otherwise $w_l(J) = w_l(I)$,$l \in J$, and stop.	

Source: [3]

Table 13: Algorithm for Weighted Bonferroni Test

Step 1	Select $j \in I$ such that $p_j \leq w_j(I)\alpha$ and remove H_j, otherwise stop.	
Step 2	Update the graph:	
	old value	*new value*
	I	$I \backslash \{j\}$
	$w_l(I)$	$w_l(I) + w_j(I) g_{jl}$, if $l \in I$; 0, otherwise
	g_{lk}	$\frac{g_{lk}+g_{lj}g_{jk}}{1-g_{lj}g_{jl}}$, if $l,k \in I$, $l \neq k$,$g_{lj}g_{jl} < 1$; 0, otherwise
Step 3	If $\lvert I \rvert \geq 1$, go to step 1; otherwise stop.	

Source: [3]

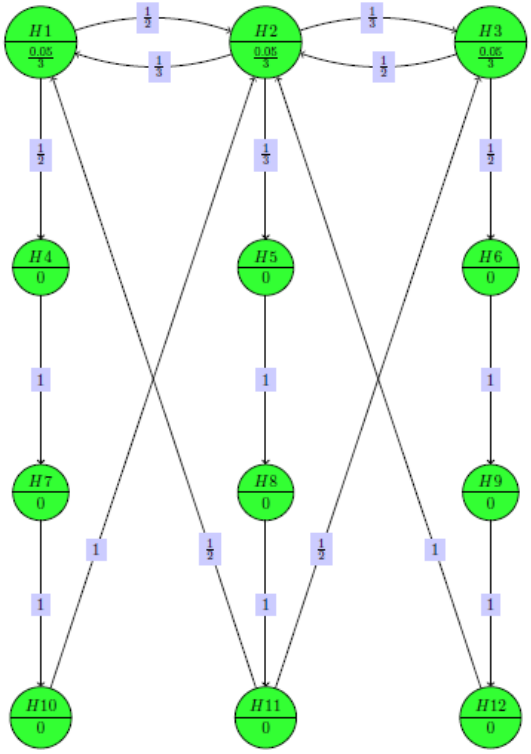
Table 14: Algorithm for Weighted Simes Test

Step 1	If $p_i > \alpha$ for all $i \in I$, stop and retain all 12 hypotheses.
Step 2	If $p_i \leq \alpha$ for all $i \in I$, stop and reject all hypotheses.
Step 3	Perform the Bonferroni-based graphical test procedure in the previous subsection. If the $\lvert I_r^c \rvert < 3$, stop and retain the remaining hypotheses.
Step 4	If $I_r^c \geq 3$ use the weights $w_i(I_r^c)$, $i \in I_r^c$ and transition matrix $\mathbf{G}$ defined on I_r^c as the new initial graph for the remaining hypotheses. Using the algorithm in Table 12, calculate the weights $w_k(J)$ for all $J \subseteq I_r^c$.
Step 5	If for each $J \subseteq I_r^c$ with $i \in J$, there exists an index such that $p_j \leq \alpha \sum_{k \in J_j} w_k(J)$, then reject H_i

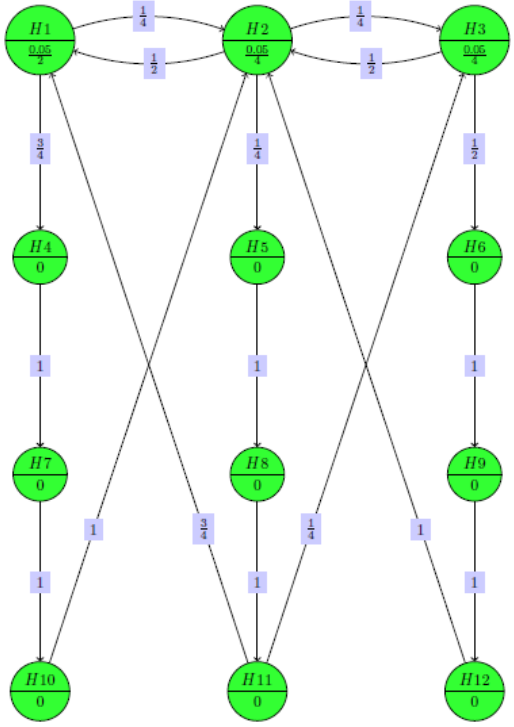
I_r = index set of rejected hypotheses
$\lvert I_r^c \rvert$ = complement of I_r in I
Source: [3]

A.4 Weighted Bonferroni-based Improved Testing Strategy diagram

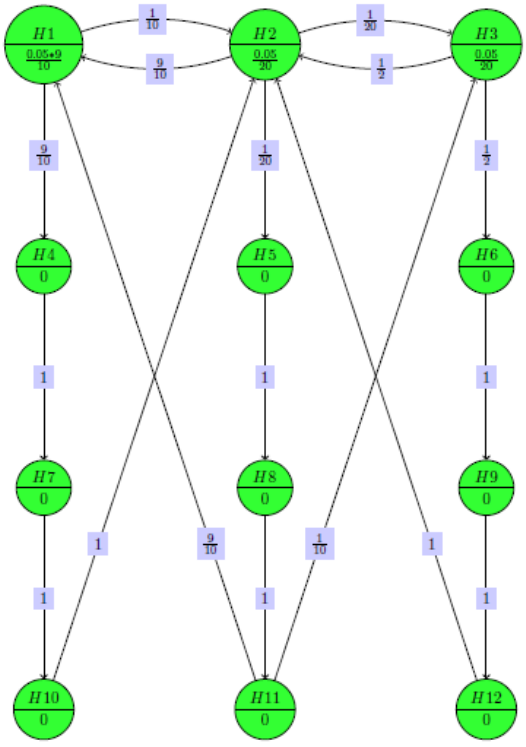
Figure 4: Weighted Bonferroni-based Improved Testing Strategy diagram



(a) Balanced alpha propagation,
Equal initial weights primary hypotheses

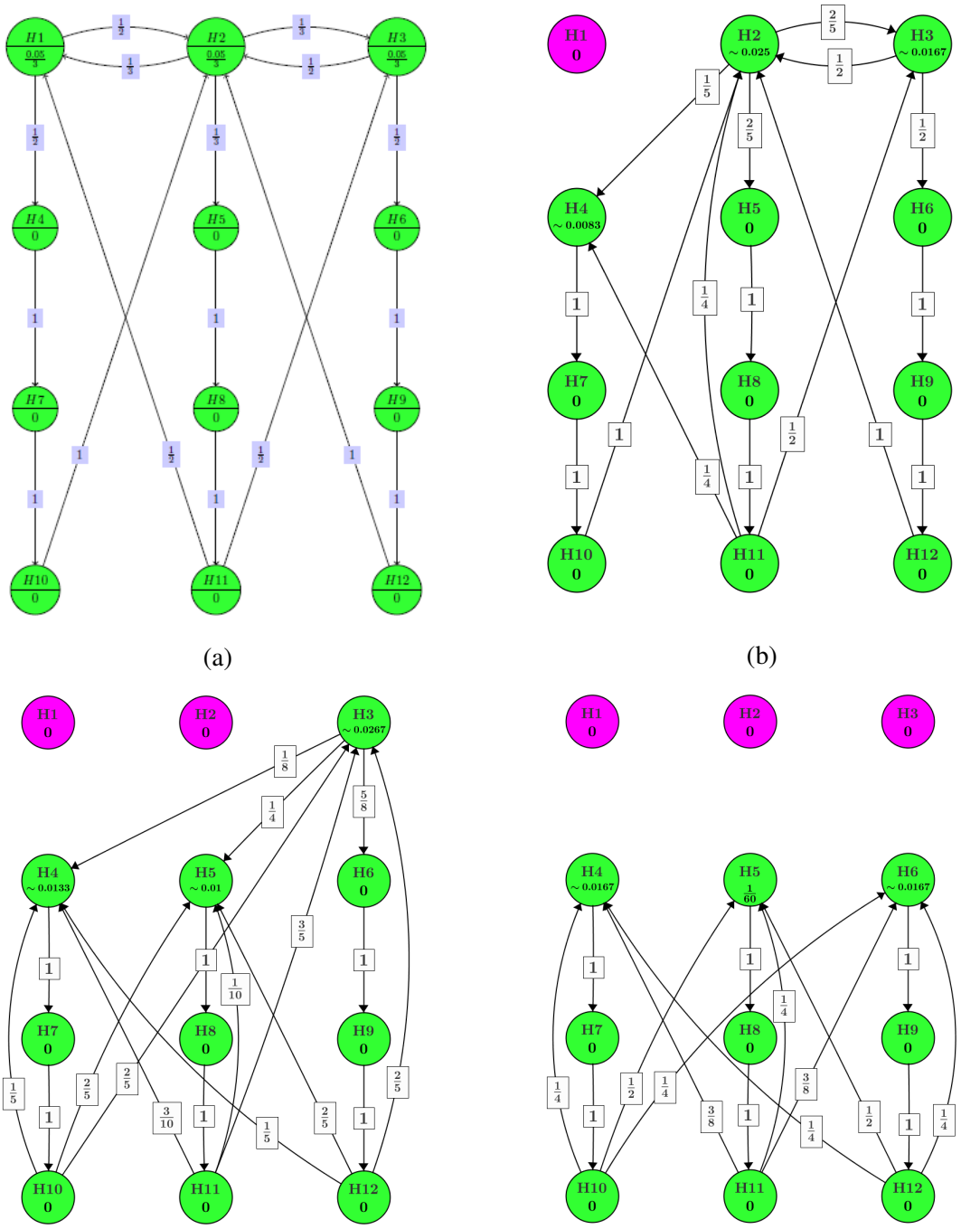


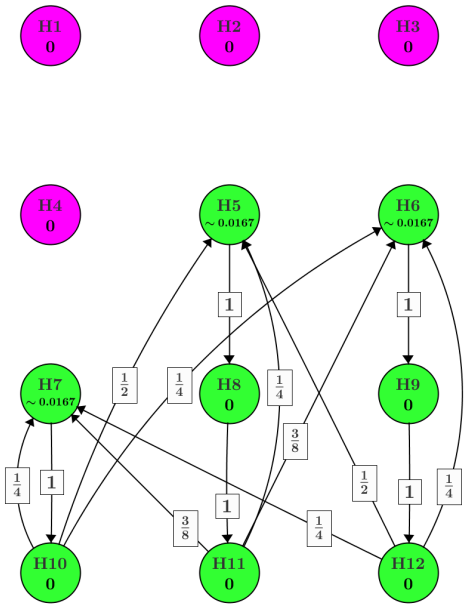
(b) More weights adding to
high dose primary hypothesis



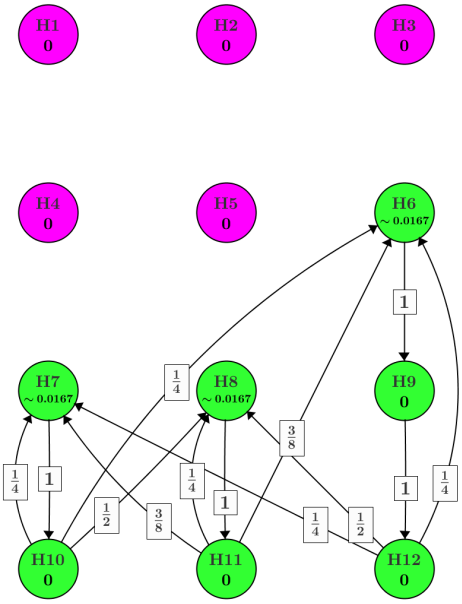
(c) Most weights adding to
high dose primary hypothesis

Figure 5: Improved testing strategy using Bonferroni-based sequentially rejective multiple test procedure for proposed weighting strategy M1W1

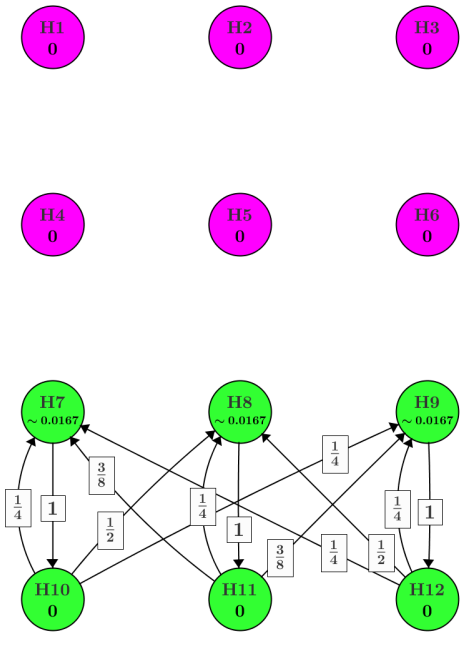



H1 0
H2 0
H3 0
H4 0
H5 ~ 0.0167
H6 ~ 0.0167
H7 ~ 0.0167
H8 0
H9 0
H10 0
H11 0
H12 0

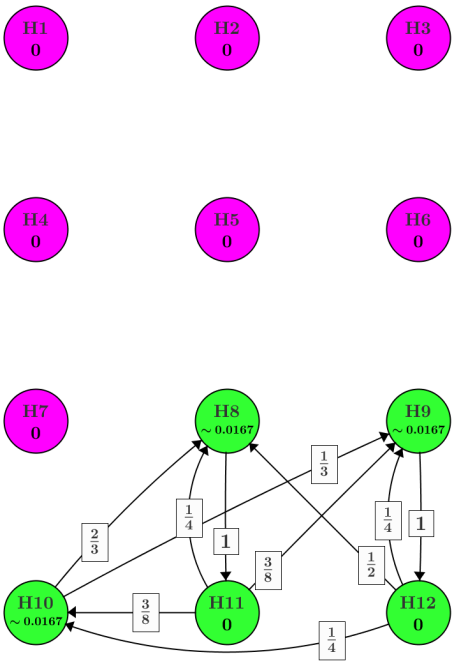
(e)


H1 0
H2 0
H3 0
H4 0
H5 0
H6 ~ 0.0167
H7 ~ 0.0167
H8 ~ 0.0167
H9 0
H10 0
H11 0
H12 0

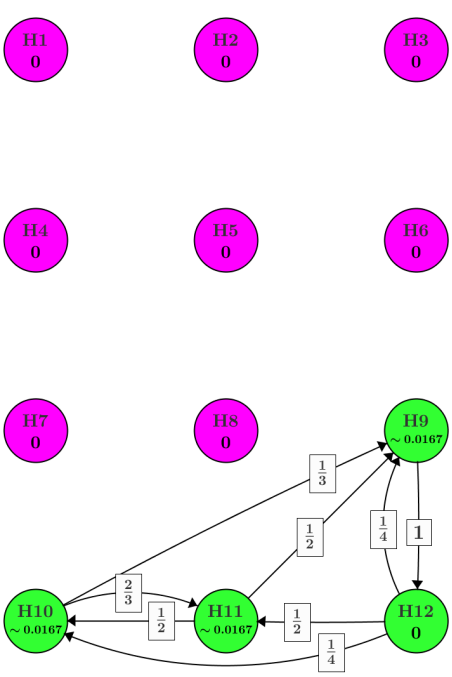
(f)


H1 0
H2 0
H3 0
H4 0
H5 0
H6 0
H7 ~ 0.0167
H8 ~ 0.0167
H9 ~ 0.0167
H10 0
H11 0
H12 0

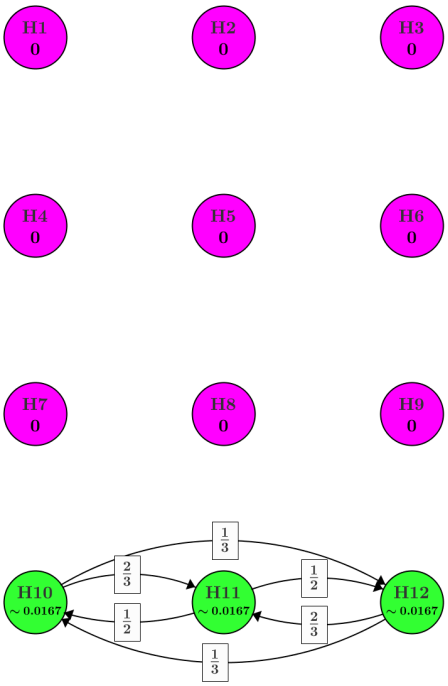
(g)


H1 0
H2 0
H3 0
H4 0
H5 0
H6 0
H7 0
H8 ~ 0.0167
H9 ~ 0.0167
H10 ~ 0.0167
H11 0
H12 0

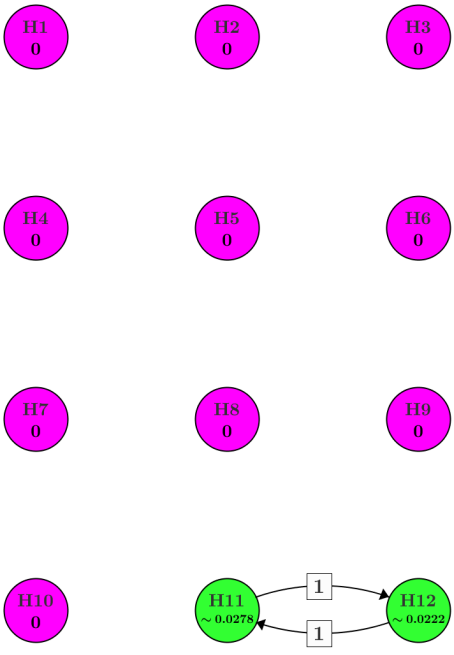
(h)


H1 0
H2 0
H3 0
H4 0
H5 0
H6 0
H7 0
H8 0
H9 ~ 0.0167
H10 ~ 0.0167
H11 ~ 0.0167
H12 0
1/3
1/2
1/4
1
2/3
1/2
1/2
1/4

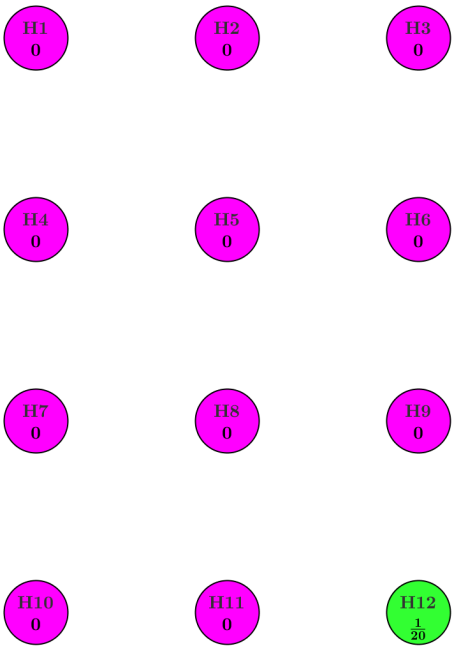
(i)


H1 0
H2 0
H3 0
H4 0
H5 0
H6 0
H7 0
H8 0
H9 0
H10 ~ 0.0167
H11 ~ 0.0167
H12 ~ 0.0167
1/3
2/3
1/2
1/2
2/3
1/3

(j)


H1 0
H2 0
H3 0
H4 0
H5 0
H6 0
H7 0
H8 0
H9 0
H10 0
H11 ~ 0.0278
H12 ~ 0.0222
1
1

k


H1 0
H2 0
H3 0
H4 0
H5 0
H6 0
H7 0
H8 0
H9 0
H10 0
H11 0
H12 1/20

l

A.5 Transition matrices and initial alpha weights

Table 15: M1

	1	2	3	4	5	6	7	8	9	10	11	12
H1	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H2	0.33	0.00	0.33	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H3	0.00	0.50	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
H4	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
H5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
H6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
H7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
H8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
H9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
H10	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H11	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H12	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 16: M2

	1	2	3	4	5	6	7	8	9	10	11	12
H1	0.00	0.25	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H2	0.50	0.00	0.25	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H3	0.00	0.50	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
H4	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
H5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
H6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
H7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
H8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
H9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
H10	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H11	0.75	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H12	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 17: M3

	1	2	3	4	5	6	7	8	9	10	11	12
H1	0.00	0.10	0.00	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H2	0.90	0.00	0.05	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H3	0.00	0.50	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
H4	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
H5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
H6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
H7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
H8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
H9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
H10	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H11	0.90	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H12	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 18: Initial α weights

	H1	**H2**	**H3**	**H4**	**H5**	**H6**	**H7**	**H8**	**H9**	**H10**	**H11**	**H12**
W1	0.0167	0.0167	0.0167	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
W2	0.0250	0.0125	0.0125	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
W3	0.0444	0.0028	0.0028	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Multiple Testing Procedures: The Application of Graphical Approaches in the Design of Clinical Trials<u></u>

Richting: **Master of Statistics-Biostatistics**
Jaar: **2015**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Mingala, Maria Cristina

Datum: **23/01/2015**