2011
2012

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

The application of Gatekeeping strategies in dose-response clinical trials with multiple endpoints.

Promotor :
Prof. dr. Tomasz BURZYKOWSKI

Promotor :
Ms. AGNES BALOGH

Guohui Hu

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

De transnationale Universiteit Limburg is een uniek samenwerkingsverband van twee universiteiten in twee landen: de Universiteit Hasselt en Maastricht University

universiteit hasselt
KNOWLEDGE IN ACTION

Maastricht University

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

Maastricht University

universiteit hasselt
KNOWLEDGE IN ACTION

2011
2012

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

The application of Gatekeeping strategies in dose-response clinical trials with multiple endpoints.

Promotor :
Prof. dr. Tomasz BURZYKOWSKI

Promotor :
Ms. AGNES BALOGH

Guohui Hu

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
KNOWLEDGE IN ACTION

The application of Gatekeeping strategies

in dose-response clinical trials with multiple endpoints

Guohui Hu

SUPERVISORS: Ms. Agnes Balogh[1] and Prof. dr. Tomasz Burzykowski[2]

[1] Bristol-Myers Squibb, Braine-l'Alleud, Belgium

[2] Hasselt University, Diepenbeek, Belgium

September 1, 2012

Contents

Abstract

Gatekeeping strategies are methods which deal with the analysis of the hierarchically ordered multiple objectives. The methods control the study-wise error rate in the strong sense. The objective of this report is to apply the Dunnett-Bonferroni-based parallel and other gatekeeping procedures in a clinical trial with ordered endpoints and multiple doses for Type II diabetes patients. The original hierarchical testing strategy does not take into account multiplicity in the family of the secondary endpoint. The efficacy of the experimental drug will be evaluated using the primary endpoint: the mean change from baseline in glycohaemoglobin A1C (HbA1C) and the first secondary endpoint: fasting plasma glucose (FPG). Three doses of the active treatment will be compared versus placebo. The results show an increase in p-values compared to the p-values presented in the final clinical study report (CSR). The comparison of the estimated power is performed between Dunnett-Bonferroni-based parallel gatekeeping procedure and the study design of CSR based on simulation. In addition, the changes in the adjusted p-values are analyzed in details to illustrate the working mechanism of different gatekeeping procedures.

Keywords: Type II diabetes, Dunnett test, Dunnett-Bonferroni-based parallel gatekeeping procedure

List of Abbreviations

ANCOVA	Analysis of covariance
CSR	Clinical study report
DB	Dunnett-Bonferroni-based
FPG	Fasting plasma glucose
HbA1C	Glycohaemoglobin
LOCF	Last observation carried forward
MCP	Multiplicity control procedure
mITT	Modified intent-to-treat
SE	Standard error

1 Introduction

The term diabetes mellitus describes a metabolic disorder of carbohydrate, fat and protein metabolism resulting from defects in insulin secretion, insulin action, or both [1]. The three main types of diabetes are: type I diabetes results from the body's failure to produce insulin, type II diabetes results from insulin resistance and type III gestational diabetes. Among them, type II diabetes makes up about 90% of cases of diabetes. Obesity is thought to be the primary cause of type II diabetes in people who are genetically predisposed to the disease. Long-term complications from high blood sugar lead to heart disease, strokes, kidney failure, blindness and lower limb amputations. Management in diabetes can usually be accomplished with diet, exercise, and use of appropriate anti-diabetic medications.

The number of diabetes patients is expected to approximately double between 2000 and 2030, due to population growth, aging, urbanization and increasing prevalence of obesity and physical inactivity. The greatest relative increases will occur in the developing countries and most of the expected population growth will be concentrated in the urban areas of the world [2].

The measurement of glycohaemoglobin A1C (HbA1C) in blood has become the gold standard for the long-term control of the glycaemic state of diabetic patients. [3] In addition, fasting plasma glucose (FPG) level which is defined as the concentration of glucose in the plasma after the patient has not eaten for at least 8 hours is also measured to diagnose diabetes. HbA1C is better than FPG for determining risks of cardiovascular disease and death from any cause [4].

The clinical trial selected to be investigated in this report is designed to compare three doses of a new anti-diabetes medicine versus placebo. The primary endpoint is the mean change in HbA1C (%) from baseline while the first secondary endpoint is the mean change in FPG (mg/dL) [5]. Significant decrease in both endpoints reflects a beneficial effect of treatment.

In clinical trials it is often necessary to answer more than one question about the efficacy or safety of one or more treatments in a specific disease, because the success of a drug development program may depend on a positive answer to more than a single question. From a regulatory point of view, the usual concern with multiple testing is that, if it is not properly handled, unsubstantiated claims for the effectiveness of a drug may be made as a consequence of an inflated rate of false positive conclusions. Multiplicity can have a substantial influence on the rate of false positive conclusions which may affect approval and labeling of an investigational drug whenever there is an opportunity to choose the most favorable result from two or more analyses. Control of the study-wise rate of type I error at an acceptable alpha level is an important principle and is often of great value in the assessment of the results of confirmatory clinical trials [6].

Once there is hierarchy between the primary and secondary endpoints in the sense that without success on the primary endpoint, the results on the secondary endpoints have limited relevance. Nevertheless, it is widely recognized that secondary endpoints play an important role in characterizing the efficacy profile of an investigational drug and contribute to a better understanding of its properties. For this reason, findings with respect to secondary endpoints are often included in the product label [7].

In order to handle these hierarchically ordered endpoints and multiple doses while controlling the study-wise error rate in the strong sense, gatekeeping strategies seem ideal. In general, in the context of gatekeeping strategies, the primary endpoints of a trial could be considered as the first family of null hypotheses, while the secondary endpoints as the secondary family, etc. The families of hypotheses are tested in a sequential manner in the sense that the acceptance or rejection of hypotheses in a particular family depends on the preceding families [8]. In other words, the hypotheses family of the primary endpoints serves as gatekeeper for those of the secondary endpoints. When controlling type I error rate in the strong sense is considered critical, the gatekeeping approach helps maximize the power with respect to more important endpoints and, at the meanwhile enables clinical researchers to test less important endpoints if the primary analysis was significant [9].

1.1 Objectives

- Writing SAS program to obtain critical values and adjusted p-values for Dunnett-Bonferroni-based (DB) parallel gatekeeping procedure for a clinical trial testing a primary and a secondary endpoint for 3 dose levels of an active drug versus placebo.

- Applying the DB parallel gatekeeping procedure and other 3 gatekeeping methods for the selected clinical trial; comparing and discussing the results.

2 Methodology

2.1 Study design of anti-diabetes trial (summary of final clinical study report)

The anti-diabetes study is a multicenter, randomized, double-blind, placebo-controlled phase 3 trial to evaluate the efficacy of an experiment drug in subjects with type I diabetes having inadequate glycemic control. The research hypothesis supposed a greater mean reduction from baseline in HbA1C achieved with the treatment groups compared to the placebo group. Subjects with screening HbA1C≥7.0% and ≤10.0% entered a 2-week placebo lead-in period. The lead-in period was followed by being randomized to the low, medium high doses or placebo group, and by entering a 24-week double-blind short term treatment period.

The primary efficacy endpoint is the mean change from baseline to week 24 in glycohaemoglobin A1C (HbA1C). There are three secondary efficacy endpoints in this study. However in this report, only the first secondary endpoint, the mean change from baseline to week 24 in fasting plasma glucose (FPG) level will be included. The main reason for not taking the other two secondary endpoints is that the second secondary endpoint is a proportion, thus it does not satisfy the normality criterion [10]. Also, including one more family in the structure (i.e. testing 9 instead of 6 original hypotheses) would require the construction of 2^9-1=511 intersection hypotheses with the corresponding decision rules and p-values, which seemed computationally too tedious for the scope of the report with limited extra benefit.

Both the mean changes in HbA1C and in FPG were compared between each of the three treatment groups and the placebo group. In the primary family the alpha level of each comparison is set at 0.019 according to the Dunnett adjustment so that the overall type I error rate within the first family is controlled at 0.05. The anti-diabetes study interpreted the tests of primary and secondary endpoints in a sequential way. P-value of the test in secondary endpoint FPG will be reported only if test of the primary endpoint is significant [5].

The sample size of this study is determined to achieve at least 90% power of detecting a mean difference of 0.7% in HbA1C in any of the three doses of experiment medicine versus placebo. The population of the primary and secondary

endpoints is based on the modified intent-to-treat (mITT). All randomized subjects who took at least one dose of double-blind treatment and who had baseline and at least one post-baseline measurements were analyzed in the treatment group to which they were randomized even if the treatment they received was different. If no measurement was available at the end of the study in the week 24, last observation carried forward (LOCF) was applied, which was the standard method to address missingness at the time of the study.

2.2 Statistical analysis

2.2.1 Analysis of covariance

Analysis of covariance (ANCOVA) is a technique that combines features of analysis of variance and regression. The basic idea is to augment the analysis of variance model containing the factor effects with one or more additional quantitative variables that are related to the response variable. This augmentation aims to reduce the variance of the error terms in the model and thus to make the analysis more precise. [11].

In the clinical trial selected for this report, the analysis of the primary endpoint is performed using ANCOVA with treatment group as a factor effect and the baseline value as the covariate.

$$Y_{ij} = \mu. + \tau_i + \gamma X_{ij} + \varepsilon_{ij}; \quad i = 1,2,3,4; \; j = 1...n_i \qquad (1)$$

Where:

$\mu.$ is the overall mean.

τ_i are the fixed treatment effects subject to the restriction $\sum \tau_i = 0$.

γ is the regression coefficient for the relation between the measurement of primary endpoint at the end of the study (Y_{ij}) and baseline measurement(X_{ij}).

Within the framework of the ANCOVA model, point estimates of the mean changes within the low, medium, high doses and placebo group are calculated. In addition, the difference in mean changes between each dose versus placebo is obtained together

with the p-values for each comparison. A similar ANCOVA model is applied when analyzing the secondary endpoint FPG.

The p-values obtained from ANCOVA models will be referred as p-values of the final clinical study report (CSR) throughout the report, and will be the basis for calculating the adjusted p-values for different gatekeeping methods.

2.3 Multiple testing issues in clinical trials

Different methods have been developed to control the rate of the false positive findings. One of the most common though conservative multiple testing control procedures is Bonferroni adjustment, where the significance level of the individual tests is derived by dividing the pre-specified overall alpha level by the number of comparisons.

When interests are in comparing J-1 active treatment groups with a single control group, the multiplicity control procedure (MCP) credited to Dunnett could be used, as the procedure exploits the correlation between the test statistics and thus gives higher statistical power than the Bonferroni method [12]. With equal sample sizes and the control mean as the J^{th} mean, the decision rule for the two sided Dunnett MCP is to reject H_0 for a comparison:

If $\left|t_{\hat{\psi}}\right| \geq D^{\alpha}_{J,df_W}$. (2)

where:

α is the pre-specified family wise significant level,

$t_{\hat{\psi}}$ is T statistics for comparing the mean of the active treatment group μ_j to the placebo group μ_J,

D is the critical value of Dunnett test,

df_W is the total degree of freedom.

In practice, SAS function PROBMC can be used to apply the Dunnett procedure when having equal or unequal sample size.

In the clinical trial selected for this report, the comparisons in primary family are between each of the three active treatment groups and the placebo group. The alpha level of each comparison is set at 0.019 according to the Dunnett adjustment so that

the overall type I error rate is controlled at 0.05. It's important to point out that the Dunnett test only protects the Type I error rate within the primary family. In other words, the nature of ordered hypotheses among primary endpoint and subsequent endpoints cannot be taken into account via the Dunnett test.

2.3.1 Dunnett-Bonferroni-based (DB) and Dunnett-based parallel gatekeeping procedure

In order to handle the hierarchical structure between families of hypotheses in a proper way, a gatekeeping approach groups the null hypotheses into families with pre-specified order. Each family becomes a gatekeeper for subsequent families in the sense that the less important goals are not evaluated before the more important goals have been reached [13].

Dmitrienko et al. [7] proposed the ***Dunnett-based Parallel Gatekeeping procedure*** which was designed for managing multiple comparisons among different doses versus control and several endpoints. This method depends on two assumptions namely (a) the multivariate-t distribution of test statistics from different endpoints and (b) that the true correlation between multiple endpoints needs to be well estimated from the observed data. However these assumptions are difficult to justify and hence could be challenged by the drug regulatory agencies [13]. Furthermore, the Dunnett-based parallel gatekeeping procedure involves the Genz and Bretz method [14] requiring the same estimated standard errors across multiple endpoints. This requirement is questionable since endpoints from different families are based on very different parts of the dataset.

In this report, the so called ***Dunnett-Bonferroni-based (DB) parallel gatekeeping procedure*** [13] is applied as the main approach. The DB parallel gatekeeping procedures is suitable for a clinical trial with two hierarchically ordered efficacy endpoints to be tested and with three active treatment dose groups to be compared with the placebo for each endpoint, i.e. in a study design where the hypotheses to be tested are as follows:

$$H_1^P : \mu_1^P - \mu_4^P \neq 0 \;\; H_2^P : \mu_2^P - \mu_4^P \neq 0, \;\; H_3^P : \mu_3^P - \mu_4^P \neq 0 \qquad (3)$$

$$H_1^S : \mu_1^S - \mu_4^S \neq 0, \;\; H_2^S : \mu_2^S - \mu_4^S \neq 0, \;\; H_3^S : \mu_3^S - \mu_4^S \neq 0$$

where $\mu_1^P, \mu_2^P, \mu_3^P$ and μ_4^P represent the mean change of HbA1C of low, medium, high doses and placebo groups from their baseline level of HbA1C, while $\mu_1^S, \mu_2^S, \mu_3^S$ and μ_4^S are the mean change of FPG for the corresponding treatment groups.

The major motivation of using the DB parallel gatekeeping procedure is not only that the preceding assumptions required by Dunnett-based parallel gatekeeping procedure could be relaxed with the modified procedure, but also the computation becomes easier [13]. In addition the power loss comparing to the Dunnett-based parallel gatekeeping procedure is expected to be very limited [13].

In general the gatekeeping methods are formulated based on the closed testing procedure: an individual null hypothesis can be rejected only if all the intersection hypotheses containing it are rejected. For the hypotheses structure described in (3), in total 63(=2^6-1) intersection hypotheses can be derived (Table 1).

For example, when testing the null hypotheses of high and medium doses from primary family and low dose from secondary family simultaneously, the corresponding intersection is defined as $H_{011100} = H_2^P \cap H_3^P \cap H_1^S$. The closed testing procedure then constructs a decision rule for each of these intersection hypotheses. The decision rules (Table 1) are defined by the following three rules:

- If H includes all primary hypotheses, the decision rule for H should not include the hypotheses from secondary family: this will ensure that a secondary hypothesis cannot be rejected unless at least one primary hypothesis has been rejected.
- The same critical value should be used for testing all primary hypotheses. In this way, the inference made in the primary family is not affected by the secondary test statistics.

- If H includes a primary hypothesis and a matching secondary hypothesis, the decision rule for H should not depend on the test statistic for the secondary hypothesis. This guarantees that hypotheses from secondary family cannot be rejected unless those from primary family were rejected. [7]

The computation of the critical values for DB parallel gatekeeping procedure can be classified into three groups according to the corresponding decision rules.

- the critical value c_1 (Table 1) contains any hypotheses from the primary family, it can be obtained from a Dunnett-t distribution so that: $\Pr\{\max(T_1^P, T_2^P, T_3^P) > c_1\} = \alpha$.
- for a decision rule containing intersection hypotheses from both primary and secondary families (line 8 to 19), for example H_{110001}, α' is first calculated for primary family as $\alpha' = \Pr\{T_1^P > c_1 \text{ or } T_2^P > c_1\}$, where (T_1^P, T_2^P) follows a Dunnett-t distribution . c_2 can be obtained so that $\Pr\{T_3^S > c_2\} = \alpha - \alpha'$, where T_3^S follows an univariate-t distribution. Thus the significance level is split between the two families indicating the major modification from the DB parallel gatekeeping approach over the Dunnet gatekeeping approach.
- For decision rules that only involve hypotheses from the secondary family (line20 to 26), the critical values (c_5, c_6, c_7) can be obtained based on the univariate t or Dunnett-t distribution.

The computation of the adjusted p-values for each intersection hypothesis is based on the same principles. Once the adjusted p-value for each intersection hypothesis is calculated, according to the principle of closed testing procedure, the adjusted p-value of each of the six hypotheses equals the maximum over the p-values associated with the intersection hypotheses containing it.

Table 1: Decision rules for 63 intersection hypotheses

	Inter section Hypotheses	Decision rule
1	H_{111111},H_{111000},$H_{111001,}H_{111010}$, H_{111001},$H_{111110,}H_{111011}$,H_{111101}	$T_1^P>c_1$ or $T_2^P>c_1$ or $T_3^P>c_1$
2	H_{110110},H_{110000},$H_{110001,}H_{110010}$	$T_1^P>c_1$ or $T_2^P>c_1$
3	H_{101101},H_{101000},$H_{101001,}H_{101100}$	$T_1^P>c_1$ or $T_3^P>c_1$
4	H_{011011},H_{011000},$H_{011001,}H_{011010}$	$T_2^P>c_1$ or $T_3^P>c_1$
5	H_{100100},H_{100000}	$T_1^P>c_1$
6	H_{010010},H_{010000}	$T_2^P>c_1$
7	H_{001001},H_{001000}	$T_3^P>c_1$
8	$H_{110001,}H_{110101}$,H_{110011},H_{110111}	$T_1^P>c_1$ or $T_2^P>c_1$ or $T_3^S>c_2$
9	H_{101111},H_{101110},$H_{101011,}H_{101010}$	$T_1^P>c_1$ or $T_3^P>c_1$ or $T_2^S>c_2$
10	H_{011111},H_{011110},$H_{011101,}H_{011100}$	$T_2^P>c_1$ or $T_3^P>c_1$ or $T_1^S>c_2$
11	H_{100111},H_{100011}	$T_1^P>c_1$ or $T_2^S>c_3$ or $T_3^S>c_3$
12	H_{010111},H_{010101}	$T_2^P>c_1$ or $T_1^S>c_3$ or $T_3^S>c_3$
13	H_{001111},H_{001110}	$T_3^P>c_1$ or $T_1^S>c_3$ or $T_2^S>c_3$
14	H_{100110},H_{100010}	$T_1^P>c_1$ or $T_2^S>c_4$
15	H_{100101},H_{100001}	$T_1^P>c_1$ or $T_3^S>c_4$
16	H_{010110},H_{010100}	$T_2^P>c_1$ or $T_1^S>c_4$
17	H_{010011},H_{010001}	$T_2^P>c_1$ or $T_3^S>c_4$
18	H_{010101},H_{010100}	$T_3^P>c_1$ or $T_1^S>c_4$
19	H_{010011},H_{010010}	$T_3^P>c_1$ or $T_2^S>c_4$
20	H_{000111}	$T_1^S>c_5$ or $T_2^S>c_5$ or $T_3^S>c_5$
21	H_{000110}	$T_1^S>c_6$ or $T_2^S>c_6$
22	H_{000101}	$T_1^S>c_6$ or $T_3^S>c_6$
23	H_{000011}	$T_2^S>c_6$ or $T_3^S>c_6$
24	H_{000100}	$T_1^S>c_7$
25	H_{000010}	$T_2^S>c_7$
26	H_{000001}	$T_3^S>c_7$

2.3.2 Other gatekeeping procedures

Different types of gatekeeping procedures with ordered objectives for clinical trials have been proposed in the literature [15]. ***Serial Gatekeeping Method***, which was suggested by Westfall and Krischen, Maurer et al. and Bauer et al., considered a multistage testing procedure in the following form. The hypotheses in the first family are tested using any multiple tests that control the study-wise error rate at a pre-specified alpha level. Testing can only pass through first gate if all hypotheses of the first family are rejected. Otherwise the testing is stopped at first failure of rejection. In the secondary family, the overall significance used is alpha [10].

Dmitrienko et al. discussed an alternative scenario, namely the ***Parallel Gatekeeping Method***. Instead of all hypotheses, at least one of them in a primary gatekeeper family should be declared significant in order to proceed with testing in subsequent lower level families. The structure of the parallel gatekeeping for two families is established. In the primary family, test the null hypotheses using Bonferroni test at the overall alpha level. In the secondary family, test the null hypotheses using the weighted Holm test at an overall significance level. This overall significance level of the secondary family is equivalent to alpha in case where all hypotheses in the primary family are rejected. In case not all primary hypotheses are rejected the overall significance level in the secondary family is less than alpha [10].

As trial designs becoming more complex, clinical researchers encounter situation beyond the simple serial or parallel gatekeeping frameworks where some null hypotheses are tested serially and the others in a parallel fashion. A general framework for setting up hybrid multistage testing procedures, namely the ***Tree Gatekeeping approach*** was proposed by Dimitrienko et al. [15]. This type of flexibility was achieved via defining for each hypothesis a serial rejection set and a parallel rejection set. Both rejection sets include selected hypotheses from the preceding families, and the given hypothesis can be tested if and only if all hypotheses from the serial set and at least one from the parallel set are rejected [10].

3 Results

3.1 Analysis of covariance (summary of final clinical study report)

A total of 400 subjects (with 100 per group) were planned to be included in the trial. According to the modified intent-to-treat principle, 390 subjects were included in the analysis of the primary endpoint (change from baseline in HbA1C measurement). The distribution of the subjects over the treatment groups is slightly unbalanced (100:103:95:92).The alpha level of each comparison is set at 0.019 according to the Dunnett adjustment so that the overall type I error rate is controlled at 0.05. All p-values are less than 0.019, indicating a significant mean reduction in A1C from baseline level for all three active treatment groups in comparison to the placebo group.

As the secondary endpoint change from baseline in FPG level provides additional information of treatment efficacy. There are in total 395 subjects included in the secondary statistical analysis. Similarly to the previous primary endpoint analysis, the sample size for the secondary analysis is also slightly unbalanced (101:105:97:92). As planned in the final clinical study report, the tests of the primary and secondary endpoints are interpreted in a sequential way. P-values of secondary hypotheses family are reported only if the primary endpoint test is significant. As shown in table 2, all p-values of the secondary endpoint are less 0.05, indicating a significant reduction in FPG in all active treatment groups compared to the placebo group.

Table 2: Primary and secondary efficacy results

Efficacy parameter		Dose			
		Low	Medium	High	Placebo
A1C (%)	Sample size	100	103	95	92
	Baseline mean	7.91	7.98	7.85	7.88
	(SE)	(0.09)	(0.11)	(0.09)	(0.10)
	Adjusted mean change	-0.43	-0.46	-0.54	0.19
	from baseline (SE)	(0.10)	(0.10)	(0.10)	(0.10)
	Difference from	-0.62	-0.64	-0.73	
	placebo (SE)	(0.14)	(0.14)	(0.15)	
	p-value (*)	<0.0001*	<0.0001*	<0.0001*	
FPG(mg/dL)	Sample size	101	105	97	92
	Baseline mean	177.72	171.31	176.51	171.85
	(SE)	(4.12)	(4.09)	(4.43)	(4.80)
	Adjusted mean change	-14.53	-8.67	-16.75	6.06
	from baseline (SE)	(3.82)	(3.74)	(3.89)	(4.00)
	Difference from	-20.60	-14.73	-22.81	
	placebo (SE)	(5.53)	(5.48)	(5.58)	
	p-value (*)	0.0002*	0.0075*	<0.0001*	

* Between group comparisons significant at alpha = 0.019 (applying Dunnett adjustment) for A1C and at alpha=0.05 for FPG.

3.2 Dunnett-Bonferroni-based parallel gatekeeping procedure

This section presents the computed critical values (Section 3.2.1), the adjusted p-values based on Dunnett-Bonferroni-based (DB) parallel gatekeeping procedure (Section 3.2.2) and the power comparison between the study design according to CSR and DB parallel gatekeeping procedure (Section 3.2.3).

3.2.1 Computation of critical values

As described in Section 2.1.4, the Dunnett-Bonferroni gatekeeping procedure is constructed using the principle of closed testing. There are in total 63 intersections of all possible combinations of the 6 null hypotheses. For each of these intersections, decision rules were set up (Table 1) on 7 critical values. These critical values are calculated both for the planned and actual sample sizes.

3.2.1.1 Planned balanced study

The study was planned to recruit 400 subjects in total with 100 in each of 4 groups. The critical values under this balanced design are presented in the second column of Table 3. The degrees of freedom for calculating each critical value is 386.

3.2.1.2 Unbalanced dataset

The actual sample size in the study is slightly less than planned and unbalanced among treatment groups. Instead of a unique value, three different critical values of c_2, c_3 and c_6 are obtained. This is because the computation of the Dunnett-t distribution now involves an extra parameter in the PROBMC function which handles unequal sample sizes and depends on the ratio of sample sizes for each comparison.

Table 3: The critical values of the Dunnett-Bonferroni-based parallel gatekeeping procedure

Critical values	Balanced design	Unbalanced design
c_1	2.3580	2.3547
c_2	2.4504	2.4527
		2.4553
		2.4544
c_3	2.4060	2.4055
		2.4060
		2.4065
c_4	2.1630	2.1653
c_5	2.3580	2.3537
c_6	2.2200	2.2168
		2.2174
		2.2180
c_7	1.9660	1.9661

3.2.2 Computation of adjusted p-values

The adjusted p-values of the Dunnett-Bonferroni-based parallel gatekeeping procedure are presented in Table 4. Comparing to the original design, the adjusted p-

values are all larger except for H_2^S. This is due to the fact that the Dunnett-Bonferroni-based parallel gatekeeping procedure is stricter in the sense of controlling the type I error. Moreover, the adjusted p-values from the secondary family are larger than those from the primary family which is in agreement with the logical restriction of gatekeeping procedure, declaring that the hypotheses from subsequent families can be only rejected unless the corresponding primary hypotheses can be declared significant.

Table 4: p-values of original design and Dunnett-Bonferroni-based parallel gatekeeping procedure (adjusted)

Hypothesis	p-value(CSR)	Adjusted p-value
H_1^P	0.00001979	0.00004457
H_2^P	0.00000768	0.00001736
H_3^P	0.00000052	0.00000117
H_1^S	0.00022580	0.00044478
H_2^S	0.00745100	0.00745100
H_3^S	0.00005271	0.00015537

3.2.3 Power comparison through simulation

A series of simulations are performed to compare the power of DB gatekeeping procedure and the MCPs according to CSR. For the reason of simplicity, the balanced sample size is used (100 patients per group) and the correlation between the primary and secondary endpoints is set to zero. Four scenarios are examined and the means (SE) of the simulation are displayed in Table 5.

The results (Table 6) show that within the primary endpoint, the estimated powers are almost the same or equal for both methods throughout all the scenarios and treatment groups. With respect to the secondary endpoint, it is worthwhile to point out that a huge difference for the estimated power is found under scenario 3 for the high dose group. While for scenarios 1, 2 and 4, the difference is not obvious when applying DB gatekeeping procedure.

Table 5: The means and standard errors of simulated data per scenarios

Scenario	Adjusted mean change from baseline(SE)							
	HbA1C				FPG			
	low	medium	high	placebo	Low	medium	High	placebo
1	-0.43 (0.10)	-0.46 (0.10)	-0.54 (0.10)	0.19 (0.10)	-14.54 (3.82)	-8.67 (3.74)	-16.75 (3.89)	6.06 (4.00)
2	0.15 (0.10)	-0.46 (0.10)	-0.54 (0.10)	0.19 (0.10)	5 (3.82)	-8.67 (3.74)	-16.75 (3.89)	6.06 (4.00)
3	0.15 (0.10)	0.15 (0.10)	-0.54 (0.10)	0.19 (0.10)	5 (3.82)	5 (3.82)	-16.75 (3.89)	6.06 (4.00)
4	0.15 (0.10)	0.15 (0.10)	0.15 (0.10)	0.19 (0.10)	5 (3.82)	5 (3.82)	5 (3.82)	6.06 (4.00)

Table 6: Power estimates based on 10000 simulations

scenario		HbA1C			FPG		
		H_1^P	H_2^P	H_3^P	H_1^S	H_2^S	H_3^S
1	CSR	0.9791	0.9883	0.9970	0.9410	0.7560	0.9941
	DB	0.9791	0.9882	0.9970	0.9517	0.7644	0.9943
2	CSR	0.0245	0.9883	0.9970	0.0017	0.7560	0.9941
	DB	0.0243	0.9882	0.9970	0.0243	0.6856	0.9887
3	CSR	0.0245	0.0239	0.9970	0.0017	0.0009	**0.9941**
	DB	0.0243	0.0236	0.9970	0.0243	0.0236	**0.0441**
4	CSR	0.0245	0.0239	0.0261	0.0017	0.0009	0.0042
	DB	0.0243	0.0236	0.0258	0.0243	0.0236	0.0258

3.3 Other gatekeeping procedures

Parallel, serial and tree gatekeeping approaches were also applied to the dataset. The adjusted p-values are inflated comparing to the p-values reported in the CSR in both endpoints. In addition the results are very close to those obtained from DB parallel gatekeeping procedure in section 3.2.2. When performing hypotheses testing in secondary family, the parallel gatekeeping method requires rejection of at least one hypothesis while the serial gatekeeping method needs the rejection of all hypotheses of the primary family. Finally tree gatekeeping procedure requires the rejection of the hypothesis in the primary family before the hypothesis testing for the secondary family within the same treatment groups can be engaged. Since all hypotheses for the primary endpoint were rejected in this study, these three gatekeeping procedures should provide logically identical adjusted p-values (Table 7).

Table 7: Adjust p-values of parallel, serial and tree gatekeeping procedure

Hypothesis	Parallel gatekeeping	Serial gatekeeping	Tree gatekeeping
H_1^P	0.00004512	0.00004512	0.00004512
H_2^P	0.00001752	0.00001752	0.00001752
H_3^P	0.00000118	0.00000118	0.00000118
H_1^S	0.00045160	0.00045160	0.00045160
H_2^S	0.00745100	0.00745100	0.00745100
H_3^S	0.00015810	0.00015810	0.00015810

4 Discussions and conclusions

4.1 Discussions

The control of the study-wise type I error rate at an acceptable level of alpha is an important principle in clinical trials, and is also a crucial element of the regulatory requirement towards study designs. Clinical trials involving hierarchically ordered objectives and multiple treatment groups give rise to more sophisticated analysis methods to deal with multiple comparisons. This report describes the application of the recently developed gatekeeping procedures in an anti-diabetes clinical trial. This trial investigated an active treatment of 3 different doses versus placebo having multiple endpoints. The Dunnett-Bonferroni-based parallel gatekeeping procedure, which controls the type I error in the strong sense, is expected to deal with multiplicity in a stricter way than the reported testing strategy of the CSR, which accounted for the multiplicity within the primary endpoint and between the primary and the first secondary endpoint. However it does not address the multiplicity within the secondary endpoint.

The Dunnett-Bonferroni-based parallel gatekeeping procedure is first applied to the dataset. Both the critical values and the adjusted p-values are calculated. The program is built based on different articles [10, 13]. One of them contained an addendum to the article introducing the Dunnett-based parallel gatekeeping approach with further details of the calculations, and was uploaded on the BioPharmNet's web site. Our SAS program was validated by reproducing the results from Xu et al. [13]. While working on this program, the corresponding author was contacted to clarify persistent inconsistencies. One of them turned out to be an error in his paper.

The power of the DB parallel gatekeeping procedure is further compared to the power of the study design in CSR. Four different scenarios are selected and under each 10000 simulated data are generated. The results show that within the primary endpoint, the difference in terms of power is very small when applying both methods. However with respect to the secondary endpoint, different conclusion was drawn. Huge difference of estimated power is found for the high dose group under scenario 3.

The adjusted p-value of parallel and serial gatekeeping procedures are obtained by applying the SAS tree gatekeeping macro written by Dimitrienko et al. [8]. In order to proceed with testing the hypotheses in the secondary family, the parallel gatekeeping procedure requires rejection of at least one hypothesis from the primary family whilst the serial gatekeeping procedure needs rejection of all hypotheses. The tree gatekeeping as an extension to these two above allows processing the hypothesis from the secondary family only when the corresponding hypothesis in the primary family have already has been rejected. However in the anti-diabetes study, the adjusted p-values for the three types of the gatekeeping methods are found to be identical since the decreases of mean HbA1C from baseline were significant for all doses. In addition the Dunnett-Bonferroni-based parallel gatekeeping procedure has slightly lower p-values than the other gatekeeping procedures due to the Dunnett adjustment within families.

No confidence intervals were presented in the references of gatekeeping procedures [7, 8, 9, 10,13 and 15] which illustrate a serious limitation of these methods. This is because step-wise procedure cannot be used to compute confidence interval [12].

4.2 Conclusions

- All four gatekeeping procedures lead to the same conclusion as the final clinical study report (CSR): there is significant reduction from baseline for both HbA1C and FPG in all 3 active treatment doses groups versus the placebo group.
- The slight differences in the adjusted p-values across the different methods are consistent with their working mechanism.
- Power comparisons between the DB parallel gatekeeping procedure and the study design according to CSR show very little difference in the primary endpoint. However important difference of estimated power is found under one of the simulation scenarios when testing the hypothesis of the secondary endpoint.

5 Acknowledgements

First I would like to express my gratitude to my supervisors Prof. dr. Tomasz Burzykowski and Ms. Agnes Balogh for their advice, encouragement and guidance. This work would not have been possible without their great ideals, mentoring, and patience towards my constant questions.

I would like to acknowledge Bristol-Myers Squibb Pharmaceutical in Braine-l'Alleud, for kindly providing the anti-diabetes data. Special thanks go to Dr. Harry Goyvaets, for his feedbacks and suggestions to an earlier version of this report. Moreover, I really appreciate the short yet pleasant summer internship in company of all colleagues.

I extend my gratitude to all the professors and staff of I-BioStat at Hasselt University for their great inspiration and support over the passing two years of my master degree. Finally, I will thank my family for the infallible support.

6 References

[1] "Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications" (http://whqlibdoc.who.int.hq/1999/WHO_NCD_NCS_99.2.pdf). World Health Organization. 1999.

[2] Wild, S., Sicree, R., Roglic, G., King, H., and Green, A. (2004). Global Prevalence of Diabetes. *Diabetes Care*, **27**: 1047-1053.

[3] Jeppsson, J., Kobold, U., Barr, J. et al. (2002). Approved IFCC Reference Method for the Measurement of HbA1C in Human Blood. *Clin Chem Lab Med* 2002, **40** (1): 78-89.

[4] Selvin, E., Steffes, MW, Zhu, H. et al. (2010). Glycated hemoglobin, diabetes, and cardiovascular risk in nondiabetic adults. *N Engl J Med*, **362** (9): 800-11.

[5] *Final Clinical Study Report.* Bristol-Myers Squibb.

[6] Committee for Proprietary Medicinal Products: Points to consider on multiplicity issues in clinical trials. The European Agency for the Evaluation of Medicinal Products Evaluation of Medicines for Human Use. 2002.

[7] Dmitrienko, A., Offen, W., Wang, O., and Xiao, D. (2006). Gatekeeping procedures in dose-response clinical trials based on the Dunnett test. *Pharmaceutical Statistics*, **5**: 19-28.

[8] Dmitrienko, A., Wiens, BL., Tamhane, AC., and Wang, X. (2007). Tree-structured gatekeeping tests in clincal trials with hierarchically ordered multiple objectives. *Statistics in medicine*, **26**: 2465-2478.

[9] Dmitrienko, A., Molenberghs, G., Chuang-Stein,C., and Offen,W. (2005). Analysis of clinical trials using SAS: a practical guide. Cary, NC: SAS Press Biometrics.

[10] Dmitrienko, A., Tamhane, AC. (2007). Gatekeeping procedures with clinical trial applications. Pharmaceutical Statistics, **6**: 171-180.

[11] Kutner, M H., Nachtsheim, C J., Neter, J., and Li, W. (2005). Applied Linear Statistical Models. NY: McGraw-Hill.

[12] Toothaker, L E. (1993). Multiple Comparison Procedures. *Sage University Paper series on Quantitative Applications in the Social Sciences*, 07-089. Newbury Park, CA: Sage.

[13] Xu, H., Nuamah, I., Liu, J., Lim, P., and Sampson, A. (2009). A Dunnett-Bonferroni-based parallel gatekeeping procedures for dose-response clinical trials with multiple endpoints. *Pharmaceutical Statistics*, **8**: 301-316.

[14] Gentz, A., and Bretz, F.(2002). Methods for the computation of multivariate t-probabilities. *Journal of Computational and Graphical Statistics*, **11**: 950-971.

[15] Dmitrienko, A., Wiens, BL., Tamhane, AC., and Wang, X. (2007). Tree-structured gatekeeping tests in clincal trials with hierarchically ordered multiple objectives. *Statistics in medicine*, **26**: 2465-2478.

7 Appendix

SAS codes:

Computation of critical values

```
data C_unequal;
n1=100;n2=103;n3=95;n4=92;
n5=101;n6=105;n7=97;n8=92;
df1=n1+n2+n3+n4-4;
df2=n5+n6+n7+n8-4;
lambda1=sqrt(n1/(n1+n4));
lambda2=sqrt(n2/(n2+n4));
lambda3=sqrt(n3/(n3+n4));
lambda4=sqrt(n5/(n5+n8));
lambda5=sqrt(n6/(n6+n8));
lambda6=sqrt(n7/(n7+n8));
c1=probmc("dunnett1",.,0.975,df1,3
,of lambda1-lambda3);
p1a=1-
probmc("dunnett1",c1,.,df1,2,of
lambda1,lambda2);
p1b=1-
probmc("dunnett1",c1,.,df1,2,of
lambda1,lambda3);
p1c=1-
probmc("dunnett1",c1,.,df1,2,of
lambda2,lambda3);
C21=tinv(1-(0.025-p1a),df2);
C22=tinv(1-(0.025-p1b),df2);
C23=tinv(1-(0.025-p1c),df2);
p1=1-cdf("T",c1,df1);
p2=1-cdf("T",c1,df1);
p3=1-cdf("T",c1,df1);
C31=probmc("dunnett1",.,1-(0.025-
p1),df2,2,of lambda4-lambda5);
C32=probmc("dunnett1",.,1-(0.025-
p2),df2,2,of lambda5-lambda6);
C33=probmc("dunnett1",.,1-(0.025-
p3),df2,2,of lambda4,lambda6);
p4=1-cdf("T",c1,df2);
C4=tinv(1-(0.025-p4),df2);
C5=probmc("dunnett1",.,0.975,df2,3
,of lambda4-lambda6);
C61=probmc("dunnett1",.,0.975,df2,
2,of lambda4,lambda5);
C62=probmc("dunnett1",.,0.975,df2,
2,of lambda5,lambda6);
C63=probmc("dunnett1",.,0.975,df2,
2,of lambda4,lambda6);
C7=tinv(0.975,df2);
run;
```

Computation of adjusted p-values

```
data P_value;
n1=100;n2=103;n3=95;n4=92;
n5=101;n6=105;n7=98;n8=95;
df1=n1+n2+n3+n4-4;
df2=n5+n6+n7+n8-4;
lambda1=sqrt(n1/(n1+n4));
lambda2=sqrt(n2/(n2+n4));
lambda3=sqrt(n3/(n3+n4));
lambda4=sqrt(n5/(n5+n8));
lambda5=sqrt(n6/(n6+n8));
lambda6=sqrt(n7/(n7+n8));
t1=4.38406204335659;
t2=4.59619407771256;t3=5.161879502
6618;
t4=3.72262742117815;
t5=2.68987372204188;
t6=4.08809591698787;
c1=probmc("dunnett1",.,0.975,df1,3
,of lambda1-lambda3);
/*P1*/
p1=1-
probmc("dunnett1",max(t1,t2,t3),.,
df1,3,of lambda1-lambda3);
/*p2-p4*/
p2=1-
probmc("dunnett1",max(t1,t2),.,df1
,3,of lambda1-lambda3);
p3=1-
probmc("dunnett1",max(t1,t3),.,df1
,3,of lambda1-lambda3);
p4=1-
probmc("dunnett1",max(t2,t3),.,df1
,3,of lambda1-lambda3);
/*p5-p7*/
p5=1-
probmc("dunnett1",t1,.,df1,3,of
lambda1-lambda3);
p6=1-
probmc("dunnett1",t2,.,df1,3,of
lambda1-lambda3);
p7=1-
probmc("dunnett1",t3,.,df1,3,of
lambda1-lambda3);
/*p8-p10*/
p1a=1-
probmc("dunnett1",max(t1,t2),.,df1
,2,of lambda1,lambda2);
p1b=1-
probmc("dunnett1",max(t1,t3),.,df1
,2,of lambda1,lambda3);
p1c=1-
probmc("dunnett1",max(t2,t3),.,df1
,2,of lambda2,lambda3);
alpha1=1-
probmc("dunnett1",c1,.,df1,2,of
lambda1,lambda2);
alpha2=1-
probmc("dunnett1",c1,.,df1,2,of
lambda1,lambda3);
```

```
alpha3=1-
probmc("dunnett1",c1,.,df1,2,of
lambda2,lambda3);
p8a=alpha1+1-
probmc("dunnett1",t6,.,df2,1);
p9b=alpha2+1-
probmc("dunnett1",t5,.,df2,1);
p10c=alpha3+1-
probmc("dunnett1",t4,.,df2,1);
p8=min(p1a,p8a);
p9=min(p1b,p9b);
p10=min(p1c,p10c);
/*pname="p8a";p=1-
probmc("dunnett1",t6,.,df2,1);outp
ut;
pname="p8b";p_value=p+p1a;output;*
/
/*p11-p13*/
p1d=1-
probmc("dunnett1",t1,.,df1,1);
p2e=1-
probmc("dunnett1",t2,.,df1,1);
p3f=1-
probmc("dunnett1",t3,.,df1,1);

alpha11=1-
probmc("dunnett1",c1,.,df1,1);
alpha12=1-
probmc("dunnett1",c1,.,df1,1);
alpha13=1-
probmc("dunnett1",c1,.,df1,1);
p11a=alpha11+1-
probmc("dunnett1",max(t5,t6),.,df2
,2,of lambda5,lambda6);
p12b=alpha12+1-
probmc("dunnett1",max(t4,t6),.,df2
,2,of lambda4,lambda6);
p13c=alpha13+1-
probmc("dunnett1",max(t4,t5),.,df2
,2,of lambda4,lambda5);
p11=min(p1d,p11a);
p12=min(p2e,p12b);
p13=min(p3f,p13c);
/*p14-p19*/
p1g=1-
probmc("dunnett1",t1,.,df1,1);
p2h=1-
probmc("dunnett1",t2,.,df1,1);
p3i=1-
probmc("dunnett1",t3,.,df1,1);
alpha14=1-
probmc("dunnett1",c1,.,df2,1);
p14a=alpha14+1-
probmc("dunnett1",t5,.,df2,1);
p15b=alpha14+1-
probmc("dunnett1",t6,.,df2,1);
p16c=alpha14+1-
probmc("dunnett1",t4,.,df2,1);
p17d=alpha14+1-
probmc("dunnett1",t6,.,df2,1);
p18e=alpha14+1-
probmc("dunnett1",t4,.,df2,1);
p19f=alpha14+1-
probmc("dunnett1",t5,.,df2,1);
p14=min(p1g,p14a);
p15=min(p1g,p15b);
p16=min(p2h,p16c);
p17=min(p2h,p17d);
p18=min(p3i,p18e);
p19=min(p3i,p19f);
/*p20*/
p20=1-
probmc("dunnett1",max(t4,t5,t6),.,
df2,3,of lambda4-lambda6);
/*p21-23*/
p21=1-
probmc("dunnett1",max(t4,t5),.,df2
,2,of lambda4,lambda5);
p22=1-
probmc("dunnett1",max(t4,t6),.,df2
,2,of lambda4,lambda6);
p23=1-
probmc("dunnett1",max(t5,t6),.,df2
,2,of lambda5,lambda6);
/*p24-p26*/;
p24=1-
probmc("dunnett1",t4,.,df2,1);
p25=1-
probmc("dunnett1",t5,.,df2,1);
p26=1-
probmc("dunnett1",t6,.,df2,1);

ph1=2*max(p1,p2,p3,p5,p8,p9,p11,p1
4,p15);
ph2=2*max(p1,p2,p4,p6,p8,p10,p12,p
16,p17);
ph3=2*max(p1,p3,p4,p7,p9,p10,p13,p
18,p19);
ps1=2*max(p1,p2,p3,p5,p8,p9,p10,p1
1,p12,p13,p14,p15,p16,p18,p20,p21,
p22,p24);
ps2=2*max(p1,p2,p4,p6,p8,p9,p10,p1
1,p12,p13,p14,p16,p17,p19,p20,p21,
p23,p25);
ps3=2*max(p1,p3,p4,p7,p8,p9,p10,p1
1,p12,p13,p15,p17,p18,p19,p20,p22,
p23,p26);
run;
```

Tree gatekeeping macro

```
data diabetes;
    input hyp $ family weight
raw_p parallel $ serial $;
    datalines;
H11 1 0.3333 0.000015039 000000
000000
    H12 1 0.3333 0.000005838
000000 000000
    H13 1 0.3334 0.000000392
000000 000000
    H21 2 0.3333 0.000225798
000000 111000
    H22 2 0.3333 0.007451001
000000 111000
    H23 2 0.3334 0.000052712
000000 111000
    ;
      %Treegatekeeper(diabetes,adj
      p);

data diabetes;
    input hyp $ family weight
raw_p parallel $ serial $;
    datalines;
H11 1 0.3333 0.000015039 000000
000000
    H12 1 0.3333 0.000005838
000000 000000
    H13 1 0.3334 0.000000392
000000 000000
    H21 2 0.3333 0.000225798
111000 000000
    H22 2 0.3333 0.007451001
111000 000000
    H23 2 0.3334 0.000052712
111000 000000
    ;
%Treegatekeeper(diabetes,adjp)

data diabetes;
    input hyp $ family weight
raw_p parallel $ serial $;
    datalines;
    H11 1 0.3333 0.000015039
000000 000000
    H12 1 0.3333 0.000005838
000000 000000
    H13 1 0.3334 0.000000392
000000 000000
    H21 2 0.3333 0.000225798
000000 100000
    H22 2 0.3333 0.007451001
000000 010000
    H23 2 0.3334 0.000052712
000000 001000
;
      %Treegatekeeper(diabetes,adj
      p);
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
The application of Gatekeeping strategies in dose-response clinical trials with multiple endpoints.

Richting: **Master of Statistics-Biostatistics**
Jaar: **2012**

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,

Hu, Guohui

Datum: **14/09/2012**