MASTER'S THESIS IN MATHEMATICAL STATISTICS

MULTIDIMENSIONAL DECISION MAKING IN EARLY PHASE CLINICAL TRIALS

Author: David Abraham Deniz
Supervisor: Ziad Taib
Examiner: Petter Mostad
University: University of Gothenburg
Department: Department of Mathematical Sciences
Division: Division of applied Mathematics and Statistics

UNIVERSITY OF GOTHENBURG

Master's thesis 2018-2019

Multidimensional decision-making in early phase clinical trials

David Abraham Deniz

UNIVERSITY OF GOTHENBURG

Department of Mathematical Sciences
Division of Mathematical Statistics
University of Gothenburg
Gothenburg, Sweden 2018-2019

Multidimensional decision making in early phase clinical trials

Supervisor: Ziad Taib, Department of Mathematical Sciences
Examiner: Petter Mostad, Department of Mathematical Sciences

Master's Thesis 2018-2019
Department of Mathematical Sciences
Division of Mathematical Statistics
University of Gothenburg
SE-412 96 Gothenburg

Abstract

Despite increasing development costs, only a few of all drug candidates are being launched in the market. This is due to the fact that the majority of drugs undergoing the drug development process fails to meet the high demands set on new drugs. With this background, decision making in the pharmaceutical industry is very important. At the end of each clinical trial, one must decide whether one should continue with the upcoming clinical trials or just end the drug development process for this specific drug. This is termed a GO/NO-GO decision, where GO means to continue and NO-GO means to stop, taking one factor into account.

It is common that the GO/NO-GO decision depends on more than one factor, making the decision framework multidimensional. In this thesis, we focus on different methods for multidimensional decision-making, and how the overall risks can be compared to the individual risks of each factor. The analysis is based on generalising existing one-dimensional decision-making into a multidimensional framework.

The one-dimensional decision-making is based on the normal distribution with unknown mean and standard deviation. The standard deviation is estimated using historical data. The decision is made by replacing the unknown mean with the sample mean since the sample mean is our best guess of the value of the unknown mean. Since the sample mean is approximately normally distributed according to the central limit theorem, one may calculate the risks before observing the data.

Contents

List of Figures

List of Tables

Acknowledgment

I would like to express my gratitude to everyone who has supported me throughout the work on this thesis. I would like to acknowledge my supervisor Ziad Taib for all the help, support and inspiration. I am truly grateful for his assistance, which went beyond all my expectations for a supervisor. I would also like to express my gratitude to Astra Zeneca for giving me this incredible opportunity to work on their data and to explore this interesting research topic.

I would also like to acknowledge my co-workers at Astra Zeneca for getting me installed and keeping me company during my time at Astra Zeneca. I would like to acknowledge my co-worker Szilárd Nemes, who is working with the case study that is being used as an example in this thesis.

Abbreviations and symbols with explanation

In table 1, abbreviations and symbols used throughout this report are presented.

Abbreviation	**Explanation**
LRV	Lower Reference Value, definition 2.1.
TV	Target Value, definition 2.2.
CLT	Central Limit Theorem, theorem 3.1.
CDF	Cummulative Distribution Function.
IID	Independent and Identically Distributed.
GNG	GO/NO-GO.
$P(A)$	Probability that the event A occur.
$P(A \cap B)$	Probability that both events A and B occur.
$P(A\|\Delta)$	Posterior probability that the event A occur, given Δ.
FGR	False Go Risk, Explanation is given here.
FSR	False Stop Risk, Explanation is given here.
$N(\mu, \sigma^2)$	Univariate Normal distribution with mean μ and variance σ^2.
$X \overset{a}{\sim} D$	X is approximately distributed as D.
$X \sim \mathrm{D}$	X is distributed as D.

Table 1: Abbreviations and symbols used throughout the report.

1 Introduction

1.1 Background

The drug development process is the process starting by discovering a drug candidate and all the way until the drug is launched in the market. The entire process of finding a potential drug until it is marketed involves various types of experiments, both pre-clinical and Clinical Trials, resulting in huge expenses.

Definition 1.1: A Clinical Trial is an experiment conducted on human subjects to evaluate some hypothesis related to a new treatment.

The drug development process consists of two parts: pre-clinical and clinical development. The latter can be divided into four sub processes, called phases, where each phase involves designing, conducting and analyzing several Clinical Trials. Here is an illustration of the clinical development process:

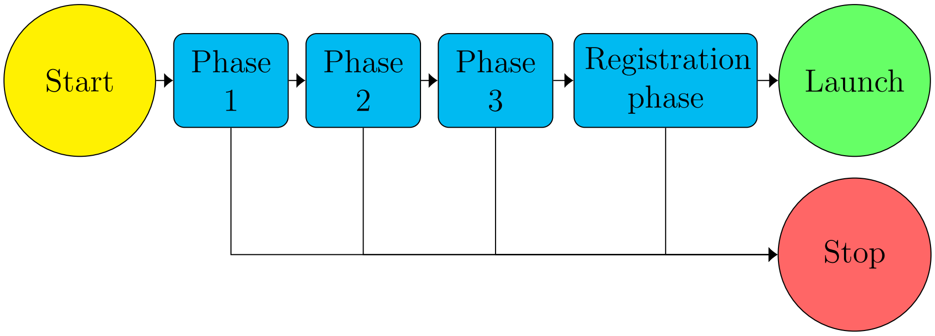

Figure 1: A figure over the four phases in a clinical development process. At the end of each phase (and each clinical trial in a phase), a GNG (GO/NO-GO) decision is made.

The first phase focuses on safety and tolerability of the drug in healthy subjects to assess if the drug has the effect that it is assumed to have. The second phase focuses on safety and tolerability of the drug on patients and on assessing the optimal dose for those patients. The third phase focuses on confirming the effect of the drug in the target population for the drug.

Most drug projects under development fail to meet the high demands set on new drugs and are thus terminated. This subsequently results in a financial loss for the sponsor, in which case human subjects have been exposed to unnecessary risk. With this background, it is very important that inefficient drug projects are being terminated as soon as possible to save time and resources and that efficacious drug projects are not being terminated. A Clinical Trial is a study that requires a lot of resources and planning. Before a Clinical Trial start, one must describe its intentions and write this in a document, called the clinical study protocol. This protocol is the most important document and serves as the blueprint of the Clinical Trial. After the Clinical Trial has been conducted, collected data from all subjects that participated in the trial are organised into a data base for further analysis. Clinical Trial data can be of any form, e.g. continuous (such as change of cholesterol level) or categorical (e.g. binary, such as dead/alive at some time t after receiving the treatment). Usually, data consists of a large number of variables describing different aspects of the subjects (age, gender, etc) as well as clinical endpoints.

Definition 1.2: A clinical endpoint is a Clinical Trial outcome value that is measured for each subject that participates in the Clinical Trial and that describes some aspect of the efficacy or safety of the drug under development.

The most common type of clinical endpoints are efficacy endpoints, but there are other types of clinical endpoints as well. It is common that a Clinical Trial has more than one endpoint. When a Clinical Trial has been conducted and analyzed, one must decide whether one should continue the project or whether one should terminate it. This is called a GO/NO-GO (GNG) decision, where GO means continuing the drug development process with the next step and NO-GO means terminating the project for this specific treatment. If one should terminate the project and not continue then that would mean a loss for the sponsor in both time and resources. However, if the new treatment that is being tested does not appear to have a clinically relevant effect then, ideally, one would like to terminate that project as soon as possible to save time and resources.

One could plan to make one (or more) analyses while the Clinical Trial is being conducted, to see early if the new drug does appear promising or not. These analyses are termed interim analyses. However, since the Clinical Trial is not complete at the interim analysis, this results in a lower sample size and more uncertainty. That could lead to termination, even if the new treatment does have a clinically relevant effect. The fact that early phase Clinical Trials usually have more than one endpoint complicates the GNG decision at the end of the trial.

1.2 Aim

The aim of this project is to investigate the following:

1. How does one make an overall GNG decision in a Clinical Trial when there are multiple endpoints to consider?

2. Should we use the information in all endpoints or should we ignore some endpoints? If we ignore some endpoints, which ones should we ignore and why?

3. Can we ignore the correlation between the different endpoints or do we need to take the correlation into account when making the overall GNG decision?

Due to the fact that the problem becomes too complicated when the number of endpoints increases, we have limited ourselves to only consider the cases of two endpoints and three endpoints in this thesis.

The thesis is organised as follows: In chapter 2, the GNG criteria for one endpoint is presented and it is also shown how the probability to make each decision depends on the underlying assumptions of the trial. In chapter 3, the GNG criteria is generalised to two and three endpoints, and different methods on how to make an overall GNG decision are presented. In chapter 4, the GNG methods presented in chapter 2 and 3 are applied to an example of how it could look like in a real-life clinical trial. In chapter 5, a case study that has been used in this thesis is presented, and the results from previous chapters are applied to data from the case study. Conclusions of the thesis is given in chapter 6.

1.3 Related Research

The topic of decision making in clinical drug development has been an ongoing research topic ever since the beginning of clinical development. Paul Frewer et al. investigated how one-dimensional decision making (i.e. decision making with only one clinical endpoint) is made using an approach proposed by Lalonde [3], in [1]. The approach, presented in this article, is based on using the 90th percentile and 20th percentile of the treatment effect to categorize the strength of observed results. The 20th percentile is compared to the lowest clinically relevant treatment effect and the 90th percentile is compared to the desired treatment effect to establish that this treatment works better than the alternative treatment. This thesis is based on generalising this approach to a multi-dimensional framework. Research has also been done in the topic of multiple endpoints in clinical trials [9].

Another approach to multicriteria decision making is based on the benefit-risk relationship when including multiple endpoints. Methods like Multiple-Criteria Decision Analysis (MCDA), Stochastic Multi-criteria Acceptance Analysis (SMAA) and Dirichlet SMAA have been developed by Gaëlle Saint-Hilary for the sake of decision-making [6]. These methods have been used by Filip Mussen et al. [7] and Ed Waddingham et al. [8]. In MCDA, one summarizes the benefit-risk relationship in a unified utility score, based on some utility score function. The utility score is calculated by putting weights on each endpoint to reflect the importance of each endpoint. In SMAA, one assumes that the weights in MCDA are random so that the utility score also becomes random. In Dirichlet SMAA, one assumes that the weights are Dirichlet distributed.

2 Theory

2.1 Decision making with one endpoint

Decision making is a crucial part of the clinical development process. In all phases of the clinical development process, especially in early phases, the need of making evidence-based decisions is essential. If a drug under development fails in phase 3 then patients who participated in the clinical trials at phase 2 has been exposed to potential side-effects of the drug for no reason. Assuming that we have a new drug, we want to study its effect μ_d (unknown to us). If it already exists another drug in the market against the disease in question for this new drug, then one would like to prove that our new drug has a better effect than the existing drug. Also, if other sponsors are developing another drug that has a greater effect than the existing drug, then one would like to prove that our new drug has a better effect than the other sponsors' drug. Otherwise, our new drug is not considered clinically relevant.

In this thesis, we focus on comparing the effect μ_d with the effect μ_p of matching placebo. We assume that we have access to data from older clinical trials (placebo data) that we can use to calculate summary statistics, such as the mean μ_p and standard deviation σ. The Target Product Profile (TPP) of a new drug defines the desired efficacy and safety criteria for the drug [4]. What defines the TPP is the concept of a Lower Reference Value and Target Value.

Definition 2.1: The Lower Reference Value (**LRV**) represents the lowest level of efficacy for a new treatment to be considered clinically relevant (a decency line for a new treatment) [4].

Definition 2.2: The Target Value (**TV**) represents the desired level of efficacy for a new treatment to establish that the new treatment works better than any other existing treatment [4].

Assuming that the new drug fulfills the TV, the goal is to make the decision GO after we have observed the data at the end of the trial. However, if the new treatment fulfills the LRV or below then we would like to make the decision NO-GO instead. Whenever one makes a GNG decision, one needs to be prepared to take some risks, similar to the two risks involved in statistical hypothesis testing.

Definition 2.3: The False Stop Risk (**FSR**) is the risk that a stop (i.e. NO-GO) decision has been made, given that the true effect is equal to the **TV** (this is similar to the risk of type II error in statistical hypothesis testing) [4].

Definition 2.4: The False Go Risk (**FGR**) is the risk that a GO decision has been made, given that the true effect is equal to the **LRV** (this is similar to the risk of type I error in statistical hypothesis testing, i.e. the significance level) [4].

For the remainder of this chapter, denote the **FSR** by β and the **FGR** by α. In clinical development, one usually uses $\beta = 10\%$ and $\alpha = 20\%$ but one could use other values. Once the TV, LRV and these two risks have been set, one can then derive a GNG criterion for a given total sample size, where the total sample size refers to the total number of human subjects enrolling in the Clinical Trial. The total sample size is denoted by N and is assumed to be a positive even number.

The decision made at the end of the trial depends on what type of endpoint the trial has. In this thesis, we will focus mainly on normally distributed efficacy endpoints since they are the most common endpoints in early phase clinical trials. Other cases are survival analysis, where one looks at the time until some event of interest is observed. Another special case is when the endpoint is binary so that μ_d and μ_p are proportions. To that end, define $\Delta = \mu_d - \mu_p$ as the difference of the drug-effect and the placebo-effect. Since we want to compare the placebo-effect μ_p with the effect μ_d of the new drug, $n = \frac{N}{2}$ subjects are assigned to test the new drug (drug-arm) and the remaining n subjects are assigned to test the placebo (placebo-arm). Assuming that no patients drop out of the trial, we will observe the data $\mathbf{X} = \{X_1, \dots, X_n\}$ from the drug-arm and the data $\mathbf{Y} = \{Y_1, \dots, Y_n\}$ from the placebo-arm at the end of the trial, where X_i,Y_i are observations of the endpoint of the trial.

Since Δ was defined as a difference of means, a natural estimate of Δ is the difference of the sample means, i.e.

$$\hat{\Delta} = \bar{X} - \bar{Y} = \frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} Y_i}{n}.$$

Note that Δ also could have been defined as a ratio of means, so that $\hat{\Delta}$ would be defined as a ratio of the sample means. We will see later in this chapter why the difference in means is more frequently used instead of the ratio of the means. By letting $P(A)$ denote the probability that an event A occurs, the decision framework proposed by Lalonde [3] looks like this:

1. GO if $P(\hat{\Delta} > \Delta_{\text{LRV}}) > 1 - \alpha$ is fulfilled,
2. NO-GO if $P(\hat{\Delta} > \Delta_{\text{TV}}) < \beta$ is fulfilled,
3. Indeterminate otherwise,

where Δ_{TV} and Δ_{LRV} are the values of Δ associated with the TV and LRV. Note that Lalonde specifically used $\alpha = 20\%$ and $\beta = 10\%$ and that we consider arbitrary values α and β in the interval (0,1) in this thesis.

To derive a more compact version of the GNG criterion, we need the distribution of the random variable $\hat{\Delta}$. The mean of $\hat{\Delta}$ is

$$\mathbb{E}[\hat{\Delta}] = \mathbb{E}[\bar{X} - \bar{Y}] = \mathbb{E}[\bar{X}] - \mathbb{E}[\bar{Y}] = \mu_d - \mu_p = \Delta$$

i.e. $\hat{\Delta}$ is an unbiased estimator of Δ. The variance of $\hat{\Delta}$ equals

$$\mathrm{Var}[\hat{\Delta}] = \mathrm{Var}[\bar{X} - \bar{Y}] = \mathrm{Var}[\bar{X}] + \mathrm{Var}[\bar{Y}] - 2 \cdot \mathrm{Cov}[\bar{X}, \bar{Y}].$$

Due to the fact that the measurements $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are taken from different subjects enrolling in the trial w.r.t. the same endpoint, we may assume WLOG that both samples $\mathbf{X}$ and $\mathbf{Y}$ are IID samples and independent of each other. Thus $\mathrm{Cov}[\bar{X}, \bar{Y}] = 0$. We also assume that the variance of the drug-arm and the placebo-arm are equal, i.e. $\mathrm{Var}(X_i) = \mathrm{Var}(Y_i) = \sigma^2$ for all $i \in \{1, \ldots, n\}$. Thus, the variance of $\hat{\Delta}$ equals

$$\mathrm{Var}[\hat{\Delta}] = \mathrm{Var}[\bar{X}] + \mathrm{Var}[\bar{Y}] = \frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \frac{2\sigma^2}{n} := \tau^2.$$

Note that if $\hat{\Delta}$ would have been defined as the ratio of sample means then it would become harder to calculate a theoretical variance of $\hat{\Delta}$. For the remainder of this chapter, $\tau = \sqrt{\tau^2}$ will be referred to as the standard error of $\hat{\Delta}$. Note that τ is calculated (estimated) by using the placebo-data from older clinical trials to estimate σ. Now that we know the mean of $\hat{\Delta}$ and variance of $\hat{\Delta}$, we can derive the distribution of $\hat{\Delta}$. If we were to assume that $X_i \sim N(\mu_d, \sigma^2)$ and $Y_i \sim N(\mu_p, \sigma^2)$ for all $i \in \{1, \ldots, n\}$ then the exact distribution of $\hat{\Delta}$ would be normal, due to the independence assumption and the fact that a sum, and difference, of independent normal random variables is normally distributed. But, it will not always hold that $X_i \sim N(\mu_d, \sigma^2)$ and $Y_i \sim N(\mu_p, \sigma^2)$, so the exact distribution of $\hat{\Delta}$ will not always be normal. To overcome this issue, the one-dimensional Central Limit Theorem (CLT) is used:

Theorem 3.1 Let $X_1, X_2, \ldots$ be a sequence of IID random variables with (common) mean μ and variance σ^2. Let $\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}$ and $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$. Then, as $n \to \infty$, Z_n converges in distribution to the standard normal distribution.

According to the one-dimensional CLT, $\hat{\Delta} \overset{a}{\sim} N(\Delta, \tau^2)$ for large values of the sample size n with the approximation becoming better for larger values of n. Thus, we do not need the assumption that $X_i \sim N(\mu_d, \sigma^2)$ and $Y_i \sim N(\mu_p, \sigma^2)$ for all $i \in \{1, \ldots, n\}$, assuming that n is considerably large. Note that the ratio of two normal distributed random variables is not normally distributed, another reason for not using the ratio of means. Now that we have derived the (approximate) distribution for the random variable $\hat{\Delta}$, we can specify the GNG criterion in terms of Δ.

The GO criterion is formulated as

$$P(\hat{\Delta} > \Delta_{\mathrm{LRV}}) > 1 - \alpha.$$

This is rewritten as:

$$P(\hat{\Delta} > \Delta_{\mathrm{LRV}}) > 1 - \alpha \iff$$
$$P(\hat{\Delta} \leq \Delta_{\mathrm{LRV}}) < \alpha.$$

Since $\hat{\Delta} \overset{a}{\sim} N(\Delta, \tau^2)$, we have that $\frac{\hat{\Delta}-\Delta}{\tau} \overset{a}{\sim} N(0,1)$. By letting Φ denote the CDF (Cummulative Distribution Function) for the standard normal distribution, we arrive at GO if

$$\Phi\left(\frac{\Delta_{\text{LRV}} - \Delta}{\tau}\right) < \alpha \iff$$

$$\iff \frac{\Delta_{\text{LRV}} - \Delta}{\tau} < \Phi^{-1}(\alpha) \iff$$

$$\iff \Delta > \Delta_{\text{LRV}} - \Phi^{-1}(\alpha) \cdot \tau := a, \tag{1}$$

where Φ^{-1} is the quantile function (inverse CDF) of the standard normal distribution. The NO-GO criterion is formulated as

$$P(\hat{\Delta} > \Delta_{\text{TV}}) < \beta.$$

This is rewritten as:

$$P(\hat{\Delta} > \Delta_{\text{TV}}) < \beta \iff$$

$$P(\hat{\Delta} \leq \Delta_{\text{TV}}) > 1 - \beta$$

Using a similar approach, one arrives at NO-GO if

$$P(\hat{\Delta} \leq \Delta_{\text{TV}}) > 1 - \beta \iff$$

$$\iff \Delta < \Delta_{\text{TV}} - \Phi^{-1}(1-\beta) \cdot \tau := b. \tag{2}$$

Using the numbers a and b that are given at the right-hand-side of equations 1 and 2, the decision framework looks like this:

- GO if $\Delta > a$,
- Indeterminate if $\Delta \in [b, a]$,
- NO-GO if $\Delta < b$.

However, there is a problem with this approach. In reality, we do not know the true value of Δ and thus can not use it to form the decision framework. Instead, we use the observed $\hat{\Delta}$ that is being calculated when we have obtained the data $\mathbf{X}$ and $\mathbf{Y}$ at the end of the trial, due to it being an unbiased estimate of Δ and our best guess of the value of Δ.

Since $\hat{\Delta}$ is a random variable that is approximately normally distributed with unknown mean Δ and estimated variance τ^2, we can (before starting the trial) calculate the probabilities of each of the three decisions if we assume that the unknown mean Δ is either Δ_{TV} or Δ_{LRV}. These probabilities are termed the operating characteristics of the trial [1]. If this approach is used to make a decision then we want the following to hold:

1. The probability of making an Indeterminate decision should be low, preferably less than 30%. If P(Indeterminate)> 30% then the operating characteristics is not acceptable.

2. The probability of making a GO decision should be large when the true unknown effect is the TV. This is analogous to the power in statistical hypothesis testing.

3. The probability of making a NO-GO decision should be large when the true unknown effect is the LRV.

The false stop risk, β, is the probability of making a NO-GO decision when $\Delta = \Delta_{\text{TV}}$ and the false go risk, α, is the probability of making a GO decision when $\Delta = \Delta_{\text{LRV}}$. Moreover, the probability of making a GO decision when the mean is the TV is

$$P(\hat{\Delta} > a|\Delta = \Delta_{\text{TV}}) = 1 - \Phi\left(\frac{\Delta_{\text{LRV}} - \Delta_{\text{TV}}}{\tau} - \Phi^{-1}(\alpha)\right), \tag{3}$$

the probability of making a NO-GO decision when the mean is the LRV is

$$P(\hat{\Delta} < b|\Delta = \Delta_{\text{LRV}}) = \Phi\left(\frac{\Delta_{\text{TV}} - \Delta_{\text{LRV}}}{\tau} - \Phi^{-1}(1-\beta)\right), \tag{4}$$

the probability of making an Indeterminate decision when the mean is the LRV is

$$\begin{gathered} P(\hat{\Delta} \in [b, a]|\Delta = \Delta_{\text{LRV}}) = \\ = P(\hat{\Delta} \leq a|\Delta = \Delta_{\text{LRV}}) - P(\hat{\Delta} \leq b|\Delta = \Delta_{\text{LRV}}) = \\ 1 - \alpha - \Phi\left(\frac{\Delta_{\text{TV}} - \Delta_{\text{LRV}}}{\tau} - \Phi^{-1}(1-\beta)\right) \end{gathered} \tag{5}$$

and the probability of making an Indeterminate decision when the mean is the TV is

$$\begin{gathered} P(\hat{\Delta} \in [b, a]|\Delta = \Delta_{\text{TV}}) = \\ = P(\hat{\Delta} < a|\Delta = \Delta_{\text{TV}}) - P(\hat{\Delta} < b|\Delta = \Delta_{\text{TV}}) = \\ \Phi\left(\frac{\Delta_{\text{LRV}} - \Delta_{\text{TV}}}{\tau} - \Phi^{-1}(\alpha)\right) - \beta. \end{gathered} \tag{6}$$

One interesting thing to note is that these probabilities only depends on the TV and the LRV through their difference.

2.2 Bayesian Approach for decision making

One could employ a Bayesian approach instead and assume a prior distribution for the true effect Δ. After obtaining the data from the trial, one would use the data to calculate the posterior distribution of Δ, given the observed $\hat{\Delta}$. Then, instead, the decision framework looks like this:

1. GO if $P(\Delta > \Delta_{\text{LRV}}|\hat{\Delta}) > 1 - \alpha$ is fulfilled,

2. NO-GO if $P(\Delta > \Delta_{\text{TV}}|\hat{\Delta}) < 1 - \beta$ is fulfilled,

3. Indeterminate otherwise,

where $P(A|\hat{\Delta})$ is the posterior probability that the event A occur, given the observed value of $\hat{\Delta}$ from the data [4].

An equivalent way to make a decision is to calculate an $100 \cdot (1-\alpha-\beta)\%$ credibility interval, say $(\hat{\Delta}_{\text{Lower}}, \hat{\Delta}_{\text{Upper}})$ with the posterior distribution of Δ and base your decision on whether the credibility interval contains Δ_{LRV} and/or Δ_{TV}:

1. GO if $\hat{\Delta}_{\text{Upper}} > \Delta_{\text{TV}}$ and $\hat{\Delta}_{\text{Lower}} > \Delta_{\text{LRV}}$.
2. NO-GO if $\hat{\Delta}_{\text{Upper}} < \Delta_{\text{TV}}$.
3. Indeterminate if $\hat{\Delta}_{\text{Upper}} > \Delta_{\text{TV}}$ and $\hat{\Delta}_{\text{Lower}} < \Delta_{\text{LRV}}$.

The problem with this is that we can only use the placebo data from the older clinical trials in order to choose a prior distribution for Δ. If it looks like we can assume that the prior distribution of Δ is normal with known parameters then we can use the normal-normal conjugacy to conclude that the posterior distribution is also normal with some updated parameters. However, if we can not assume that the prior distribution is normal then the posterior distribution will not be easy to derive and thus would require simulations. Due to this, the bayesian approach will not be considered in this thesis and is left for further research. But, it is worth mentioning that the bayesian approach has been used in a single-arm study [10].

3 Methodology

In this chapter, we assume that we have a drug that we want to test whether it has a clinically relevant effect w.r.t. more than one endpoint. We assign $n = \frac{N}{2}$ randomly chosen subjects to test the drug (drug-arm) and we assign the remaining n subjects to test the placebo (placebo-arm), just like in the previous chapter.

3.1 Two endpoints

When we have two endpoints instead of one, the settings for the decision framework needs to be generalised. We have three possible decisions for each of the two endpoints, giving us in total $3^2 = 9$ different scenarios. From now on, denote GO by G_i, NO-GO by R_i and Indeterminate by A_i for the ith endpoint (G=Green, A=Amber, R=Red). The nine scenarios in the two-dimensional decision framework can be illustrated with a table:

endpoint 2 \ endpoint 1	G_1	A_1	R_1
G_2	$\mathrm{G}_1\mathrm{G}_2$	$\mathrm{A}_1\mathrm{G}_2$	$\mathrm{R}_1\mathrm{G}_2$
A_2	$\mathrm{G}_1\mathrm{A}_2$	$\mathrm{A}_1\mathrm{A}_2$	$\mathrm{R}_1\mathrm{A}_2$
R_2	$\mathrm{G}_1\mathrm{R}_2$	$\mathrm{A}_1\mathrm{R}_2$	$\mathrm{R}_1\mathrm{R}_2$

Table 2: Table for two-dimensional decision framework.

Assuming that we have two endpoints in the trial and that no dropouts occur, at the end of the trial, we have collected the data

$$\mathbf{X}_1 = \{X_{11}, \ldots, X_{1n}\},$$

$$\mathbf{X}_2 = \{X_{21}, \ldots, X_{2n}\},$$

$$\mathbf{Y}_1 = \{Y_{11}, \ldots, Y_{1n}\},$$

$$\mathbf{Y}_2 = \{Y_{21}, \ldots, Y_{2n}\},$$

where $\mathbf{X}_i$ is the data from the drug-arm w.r.t. the ith endpoint and $\mathbf{Y}_i$ is the data from the placebo-arm w.r.t. the ith endpoint.

Now that we have two endpoints, the random variable $\hat{\Delta}$ instead becomes a random vector with two components. More specifically, $\hat{\Delta} = (\hat{\Delta}_1, \hat{\Delta}_2)$, where

$$\hat{\Delta}_i = \bar{X}_i - \bar{Y}_i = \frac{\sum_{j=1}^{n} X_{ij}}{n} - \frac{\sum_{j=1}^{n} Y_{ij}}{n}.$$

We still assume that the standard deviation is the same w.r.t. group, but the standard deviation is not the same w.r.t. the two different endpoints. More specifically, $\sigma_1 = \sigma_{1d} = \sigma_{1p}$ and $\sigma_2 = \sigma_{2d} = \sigma_{2p}$, where σ_i is the standard deviation w.r.t. the ith endpoint, d=drug and p=placebo.

The idea to make a decision in a two-dimensional setting is to make a decision w.r.t. each of the two endpoints and end up in one of the nine regions presented in table 2. Then, one has to decide the overall decision for all nine regions and then make your decision based on which of the nine regions you end up in when observing the data at the end of the trial. The problem with this approach is that there is no unique way of determining which region should correspond to which overall decision. The only thing that is unique is that the region G_1G_2 always corresponds to an overall G decision and that the region R_1R_2 always corresponds to an overall R decision.

First, let us introduce some notations:

- $\Delta_{\text{TV}} = (\Delta_{\text{TV}_1}, \Delta_{\text{TV}_2})$ is the vector of TV's and $\Delta_{\text{LRV}} = (\Delta_{\text{LRV}_1}, \Delta_{\text{LRV}_2})$ is the vector of LRV's.
- $\alpha' = (\alpha_1, \alpha_2)$ and $\beta' = (\beta_1, \beta_2)$ are the vectors of the individual False Stop Risks and False GO Risks for each endpoint.
- $s = (s_1, s_2)$, where $s_i = \sqrt{\frac{2 \cdot \sigma_i^2}{n}}$, is the vector of standard errors for each endpoint.
- $a = (a_1, a_2)$ and $b = (b_1, b_2)$, where $a_i = \Delta_{\text{LRV}_i} - \Phi^{-1}(\alpha_i) \cdot s_i$ and $b_i = \Delta_{\text{TV}_i} - \Phi^{-1}(1 - \beta_i) \cdot s_i$, are the vectors of lower and upper bound for the G and R zone of the individual endpoints.

3.1.1 Stepwise GNG criteria

Sometimes, one of the two endpoints will be more important than the other, i.e. one endpoint is primary and the other is secondary. In that case, we do not want to make a G decision if the primary endpoint says R, even if the secondary endpoint says G. In stepwise GNG, one first looks at the endpoint that is considered as primary and makes a decision according to that endpoint. If the decision of the primary endpoint is A then the overall decision is determined by the decision of the secondary endpoint. If the decision of the primary endpoint isn't A then the overall decision is determined by the primary endpoint. This GNG criteria corresponds to translating the two-dimensional decisions in table 2 into the following one-dimensional table:

secondary endpoint \ primary endpoint	G_1	A_1	R_1
G_2	G	G	R
A_2	G	A	R
R_2	G	R	R

Table 3: Table for two-dimensional decision framework transformed into a one-dimensional decision framework using the stepwise GNG criteria.

Note that, the reason it is called stepwise GNG is because we are stepwise looking at the next endpoint if we end up in the A zone w.r.t. the current endpoint that we are looking at. If we would have had more than two endpoints then we would continue to look at the next endpoint if we end up in the A zone according to the primary and secondary endpoint.

Now that we have translated table 2 into table 3, we are now able to calculate the probability of each decision by summing up the probabilities of the regions corresponding to that decision. Using that the nine regions in table 2 are disjoint regions, the probability to end up in the G zone is

$$P(\mathrm{G}) = P(\mathrm{G}_1\mathrm{G}_2 \cup \mathrm{G}_1\mathrm{A}_2 \cup \mathrm{G}_1\mathrm{R}_2 \cup \mathrm{A}_1\mathrm{G}_2) = P(\mathrm{G}_1\mathrm{G}_2) + P(\mathrm{G}_1\mathrm{A}_2) + P(\mathrm{G}_1\mathrm{R}_2) + P(\mathrm{A}_1\mathrm{G}_2), \tag{7}$$

the probability to end up in the R zone is

$$P(\mathrm{R}) = P(\mathrm{R}_1\mathrm{G}_2 \cup \mathrm{R}_1\mathrm{A}_2 \cup \mathrm{R}_1\mathrm{R}_2 \cup \mathrm{A}_1\mathrm{R}_2) = P(\mathrm{R}_1\mathrm{G}_2) + P(\mathrm{R}_1\mathrm{A}_2) + P(\mathrm{R}_1\mathrm{R}_2) + P(\mathrm{A}_1\mathrm{R}_2) \tag{8}$$

and the probability to end up in the A zone is

$$P(\mathrm{A}) = P(\mathrm{A}_1\mathrm{A}_2). \tag{9}$$

Now, all we need to calculate the probabilities is to derive the distribution of the random vector $\hat{\Delta} = (\hat{\Delta}_1, \hat{\Delta}_2)$. Recall that we used the one-dimensional CLT to argue that, in the case of one endpoint, $\hat{\Delta}$ was approximately normally distributed with the approximation becoming better for larger values of n. Now that we have two endpoints, causing $\hat{\Delta}$ to be a two-dimensional random vector instead of an univariate random variable, one uses the two-dimensional CLT [2]:

Theorem 3.2 Let $\mathbf{X}_i = \begin{bmatrix} X_{i(1)} \\ X_{i(2)} \end{bmatrix}$ for $i = 1, \ldots, n$ with mean vector $\mu = \mathbb{E}[\mathbf{X}_i] = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$. Also let

$$\bar{\mathbf{X}}_n = \frac{1}{n} \cdot \begin{bmatrix} \sum_{i=1}^n X_{i(1)} \\ \sum_{i=1}^n X_{i(2)} \end{bmatrix} = \frac{1}{n} \cdot \sum_{i=1}^n \mathbf{X}_i$$

denote the sample average. Then, as $n \to \infty$, $\frac{1}{\sqrt{n}} \cdot \sum_{i=1}^n (\mathbf{X}_i - \mu) = \sqrt{n} \cdot (\bar{\mathbf{X}}_n - \mu)$ converges to a bivariate normal distribution with mean vector $\mathbf{0} = (0, 0)$ and covariance matrix

$$\Sigma = \begin{bmatrix} \mathrm{Var}(X_{1(1)}) & \mathrm{Cov}(X_{1(1)}, X_{1(2)}) \\ \mathrm{Cov}(X_{1(1)}, X_{i(2)}) & \mathrm{Var}(X_{1(2)}) \end{bmatrix}.$$

Let ρ denote the correlation between the two endpoints on a single measurement, which is estimated using the placebo-data from the older clinical trials. Then, it can be shown that the correlation between $\hat{\Delta}_1$ and $\hat{\Delta}_2$ is also equal to ρ. According to Theorem 3.2, the random vector $\hat{\Delta} = (\hat{\Delta}_1, \hat{\Delta}_2)$ is approximately bivariate normal distributed with mean vector $\Delta = (\Delta_1, \Delta_2)$ and covariance matrix

$$\Sigma_{\hat{\Delta}} = \begin{pmatrix} s_1^2 & \rho \cdot s_1 \cdot s_2 \\ \rho \cdot s_1 \cdot s_2 & s_2^2 \end{pmatrix}$$

for large values of n. The bivariate normal distribution with mean vector $\mu = \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}$ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho \cdot \sigma_1 \cdot \sigma_2 \\ \rho \cdot \sigma_1 \cdot \sigma_2 & \sigma_2^2 \end{pmatrix}$$

has probability density function

$$f(x, y|\mu, \Sigma) = \frac{\exp\left(- \frac{1}{2 \cdot (1-\rho^2)} \left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-\mu_2}{\sigma_2}\right)^2 - 2\rho \cdot \left(\frac{x-\mu_1}{\sigma_1}\right) \cdot \left(\frac{y-\mu_2}{\sigma_2}\right) \right] \right)}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}}.$$

To calculate the probabilities, one integrates the probability density function over the appropriate subspace of $\mathbb{R}^2$, e.g.

$$P(\mathrm{G}_1\mathrm{G}_2) = P(\hat{\Delta}_1 > a_1 \cap \hat{\Delta}_2 > a_2) = \int_{a_1}^{\infty} \int_{a_2}^{\infty} f(x, y|\mu = \Delta, \Sigma = \Sigma_{\hat{\Delta}}) dxdy.$$

This integral can not be written in a closed form, forcing us to rely on numerical methods to approximate it. The probabilities for the decisions for the secondary endpoint, conditioning that we ended up in the A zone w.r.t. the primary endpoint, are given by

$$P(\mathrm{G}_2|\mathrm{A}_1) = \frac{P(\mathrm{A}_1\mathrm{G}_2)}{P(\mathrm{A}_1)},$$

$$P(\mathrm{A}_2|\mathrm{A}_1) = \frac{P(\mathrm{A}_1\mathrm{A}_2)}{P(\mathrm{A}_1)}$$

and

$$P(\mathrm{R}_2|\mathrm{A}_1) = \frac{P(\mathrm{A}_1\mathrm{R}_2)}{P(\mathrm{A}_1)}.$$

Now that we know how to calculate the probability of each decision, what is the overall false go risk and overall false stop risk when we use the stepwise GNG criteria? The overall false go risk, denoted by α, is the probability of ending up in the G zone assuming that the drug fulfills the LRV:

$$\alpha = P(\mathrm{G}|\Delta = \Delta_{\mathrm{LRV}}).$$

According to equation 7,

$$\alpha = P(\mathrm{G}_1\mathrm{G}_2|\Delta = \Delta_{\mathrm{LRV}}) + P(\mathrm{G}_1\mathrm{R}_2|\Delta = \Delta_{\mathrm{LRV}}) + P(\mathrm{G}_1\mathrm{A}_2|\Delta = \Delta_{\mathrm{LRV}}) + P(\mathrm{A}_1\mathrm{G}_2|\Delta = \Delta_{\mathrm{LRV}}).$$

In the case of no correlation between the two endpoints ($\rho = 0$), the two endpoints will be independent because uncorrelated implies independent in the case of the multivariate normal distribution. If $\rho = 0$, we can rewrite the overall false go risk in a closed form:

$$\alpha = P(\mathrm{G}_1|\Delta_1 = \Delta_{\mathrm{LRV}_1}) \cdot P(\mathrm{G}_2|\Delta_2 = \Delta_{\mathrm{LRV}_2}) + P(\mathrm{G}_1|\Delta_1 = \Delta_{\mathrm{LRV}_1}) \cdot P(\mathrm{A}_2|\Delta_2 = \Delta_{\mathrm{LRV}_2}) +$$

$$+P(\mathrm{G}_1|\Delta_1 = \Delta_{\mathrm{LRV}_1}) \cdot P(\mathrm{R}_2|\Delta_2 = \Delta_{\mathrm{LRV}_2}) + P(\mathrm{A}_1|\Delta_1 = \Delta_{\mathrm{LRV}_1}) \cdot P(\mathrm{G}_2|\Delta_2 = \Delta_{\mathrm{LRV}_2}) =$$

$$= \alpha_1 + \alpha_2 \cdot \left(1 - \alpha_1 - \Phi\left(\frac{\Delta_{\mathrm{TV}_1} - \Delta_{\mathrm{LRV}_1}}{s_1} - \Phi^{-1}(1-\beta_1)\right)\right) = w_1 \cdot \alpha_1 + w_2(\alpha_1, \beta_1) \cdot \alpha_2.$$

For $\rho = 0$, α can be seen as a linear combination of the individual false go risks with coefficients

$$w_1 = 1,$$
$$w_2(\alpha_1, \beta_1) = 1 - \alpha_1 - \Phi\left(\frac{\Delta_{\mathrm{TV}_1} - \Delta_{\mathrm{LRV}_1}}{s_1} - \Phi^{-1}(1-\beta_1)\right).$$

The overall false stop risk, denoted by β, is the probability of ending up in the R zone assuming that the drug fulfills the TV:

$$\beta = P(\mathrm{R}|\Delta = \Delta_{\mathrm{TV}}).$$

According to equation 8,

$$\beta = P(\mathrm{R}_1\mathrm{G}_2|\Delta = \Delta_{\mathrm{TV}}) + P(\mathrm{R}_1\mathrm{A}_2|\Delta = \Delta_{\mathrm{TV}}) + P(\mathrm{R}_1\mathrm{R}_2|\Delta = \Delta_{\mathrm{TV}}) + P(\mathrm{A}_1\mathrm{R}_2|\Delta = \Delta_{\mathrm{TV}}).$$

In the case of no correlation ($\rho = 0$),

$$\beta = P(\mathrm{R}_1|\Delta_1 = \Delta_{\mathrm{TV}_1}) \cdot P(\mathrm{R}_2|\Delta_2 = \Delta_{\mathrm{TV}_2}) + P(\mathrm{R}_1|\Delta_1 = \Delta_{\mathrm{TV}_1}) \cdot P(\mathrm{G}_2|\Delta_2 = \Delta_{\mathrm{TV}_2}) +$$

$$+P(\mathrm{R}_1|\Delta_1 = \Delta_{\mathrm{TV}_1}) \cdot P(\mathrm{A}_2|\Delta_2 = \Delta_{\mathrm{TV}_2}) + P(\mathrm{A}_1|\Delta_1 = \Delta_{\mathrm{TV}_1}) \cdot P(\mathrm{R}_2|\Delta_2 = \Delta_{\mathrm{TV}_2}) =$$

$$= \beta_1 + \beta_2 \cdot \left(\Phi\left(\frac{\Delta_{\mathrm{LRV}_1} - \Delta_{\mathrm{TV}_1}}{s_1} - \Phi^{-1}(\alpha_1)\right) - \beta_1\right) = w_3 \cdot \beta_1 + w_4(\alpha_1, \beta_1) \cdot \beta_2,$$

which can also be seen as a linear combination of the individual false stop risks with coefficients

$$w_3 = 1,$$
$$w_4(\alpha_1, \beta_1) = \Phi\left(\frac{\Delta_{\mathrm{LRV}_1} - \Delta_{\mathrm{TV}_1}}{s_1} - \Phi^{-1}(\alpha_1)\right) - \beta_1.$$

3.1.2 1-of-2 GNG criteria

In some trials, both endpoints will be equally important to determine whether one should continue or not continue with the next step in the drug development process. If that is the case then it might not be optimal to look at the endpoints stepwise, mostly due to the fact that there is no easy way to determine which endpoint to look at first since they are equally important. Instead, one can choose to make a decision according to this:

1. G zone: at least one of the two endpoints ends up in the G zone and the other endpoint does not end up in the R zone.

2. R zone: if at least one of the two endpoints ends up in the R zone and the other endpoint does not end up in the G zone.

3. A zone otherwise.

This corresponds to translating table 2 into the following one-dimensional table:

endpoint 2 \ endpoint 1	G_1	A_1	R_1
G_2	G	G	A
A_2	G	A	R
R_2	A	R	R

Table 4: Table for two-dimensional decision framework transformed into a one-dimensional decision framework using the 1-of-2 GNG criteria.

This table differs from table 3 in the sense that $\mathrm{G}_1\mathrm{R}_2$ (and $\mathrm{R}_1\mathrm{G}_2$) corresponds to the overall decision A instead of G (and R). As a consequence, the probability to end up in the G zone is now given by

$$P(\mathrm{G}) = P(\mathrm{G}_1\mathrm{G}_2 \cup \mathrm{G}_1\mathrm{A}_2 \cup \mathrm{A}_1\mathrm{G}_2) = P(\mathrm{G}_1\mathrm{G}_2) + P(\mathrm{G}_1\mathrm{A}_2) + P(\mathrm{A}_1\mathrm{G}_2), \tag{10}$$

the probability to end up in the R zone is

$$P(\mathrm{R}) = P(\mathrm{R}_1\mathrm{A}_2 \cup \mathrm{R}_1\mathrm{R}_2 \cup \mathrm{A}_1\mathrm{R}_2) = P(\mathrm{R}_1\mathrm{A}_2) + P(\mathrm{R}_1\mathrm{R}_2) + P(\mathrm{A}_1\mathrm{R}_2) \tag{11}$$

and the probability to end up in the A zone is

$$P(\mathrm{A}) = P(\mathrm{A}_1\mathrm{A}_2 \cup \mathrm{G}_1\mathrm{R}_2 \cup \mathrm{R}_1\mathrm{G}_2) = P(\mathrm{A}_1\mathrm{A}_2) + P(\mathrm{G}_1\mathrm{R}_2) + P(\mathrm{R}_1\mathrm{G}_2). \tag{12}$$

We can see that the probability to end up in the A zone has increased compared to when using the stepwise GNG criteria. According to equation 10, the overall False Go Risk becomes

$$\alpha = P(\mathrm{G}_1\mathrm{G}_2|\Delta = \Delta_{\mathrm{LRV}}) + P(\mathrm{G}_1\mathrm{A}_2|\Delta = \Delta_{\mathrm{LRV}}) + P(\mathrm{A}_1\mathrm{G}_2|\Delta = \Delta_{\mathrm{LRV}}).$$

When $\rho = 0$,

$$\begin{aligned}
\alpha = P(\mathrm{G}_1|\Delta_1 = \Delta_{\mathrm{LRV}_1}) \cdot P(\mathrm{G}_2|\Delta_2 = \Delta_{\mathrm{LRV}_2}) + P(\mathrm{G}_1|\Delta_1 = \Delta_{\mathrm{LRV}_1}) \cdot P(\mathrm{A}_2|\Delta_2 = \Delta_{\mathrm{LRV}_2}) + \\
+ P(\mathrm{A}_1|\Delta_1 = \Delta_{\mathrm{LRV}_1}) \cdot P(\mathrm{G}_2|\Delta_2 = \Delta_{\mathrm{LRV}_2}) = \\
= \alpha_1 \cdot \left(1 - \Phi\left(\frac{\Delta_{\mathrm{TV}_2} - \Delta_{\mathrm{LRV}_2}}{s_2} - \Phi^{-1}(1-\beta_2)\right)\right) + \alpha_2 \cdot \left(1 - \alpha_1 - \Phi\left(\frac{\Delta_{\mathrm{TV}_1} - \Delta_{\mathrm{LRV}_1}}{s_1} - \Phi^{-1}(1-\beta_1)\right)\right) = \\
= w_1(\beta_2) \cdot \alpha_1 + w_2(\beta_1, \alpha_1) \cdot \alpha_2,
\end{aligned}$$

which is also a linear combination of the individual False Go Risks.

According to equation 11, the overall False Stop Risk becomes

$$\beta = P(\mathrm{R}_1\mathrm{A}_2|\Delta = \Delta_{\mathrm{TV}}) + P(\mathrm{R}_1\mathrm{R}_2|\Delta = \Delta_{\mathrm{TV}}) + P(\mathrm{A}_1\mathrm{R}_2|\Delta = \Delta_{\mathrm{TV}})$$

and when $\rho = 0$,

$$\beta = \beta_1 \cdot \left(\Phi\left(\frac{\Delta_{\mathrm{LRV}_2} - \Delta_{\mathrm{TV}_2}}{s_2} - \Phi^{-1}(\alpha_2)\right)\right) + \beta_2 \cdot \left(\Phi\left(\frac{\Delta_{\mathrm{LRV}_1} - \Delta_{\mathrm{TV}_1}}{s_1} - \Phi^{-1}(\alpha_1)\right) - \beta_1\right).$$

3.1.3 2-of-2 GNG criteria

Sometimes, when the two endpoints are equally important, it may not be ethical to continue even if one of the endpoints say G and the other endpoint say A. In that case, one makes a G decision if and only if both endpoints say G and one makes a R decision if and only if both endpoints say R. This corresponds to translating table 2 into the following one-dimensional table:

endpoint 2 \ endpoint 1	G_1	A_1	R_1
G_2	G	A	A
A_2	A	A	A
R_2	A	A	R

Table 5: Table for two-dimensional decision framework transformed into a one-dimensional decision framework using 2-of-2 GNG criteria.

As can be seen in table 5, the probability to end up in the A zone has further increased. The overall False Go Risk is given by

$$\alpha = P(G_1G_2|\Delta = \Delta_{\text{LRV}}),$$

and with $\rho = 0$,

$$\alpha = \alpha_1 \cdot \alpha_2.$$

The overall False Stop Risk is given by

$$\beta = P(R_1R_2|\Delta = \Delta_{\text{TV}}),$$

and with $\rho = 0$,

$$\beta = \beta_1 \cdot \beta_2.$$

Since both endpoints are considered equally important, both endpoints should have the same False Go Risk and the same False Stop Risk. Thus, the overall False Go Risk and False Stop Risk are equal to $\beta = \beta_1^2$ and $\alpha = \alpha_1^2$ (when $\rho = 0$). If $\alpha_1 = \alpha_2 = 20\%$ then $\alpha = 4\%$, and if $\alpha_1 = \alpha_2 = 45\%$ then $\alpha = 20.25\%$. If $\beta_1 = \beta_2 = 10\%$ then $\beta = 1\%$, and if $\beta_1 = \beta_2 = 30\%$ then $\beta = 9\%$.

3.2 Three Endpoints

When we have three endpoints, the problem becomes even more complicated. Due to having three possible decisions for each of the three endpoints, there are $3^3 = 27$ different scenarios to consider in this setting:

$G_1G_2G_3$, $G_1G_2A_3$, $G_1G_2R_3$, $G_1A_2G_3$, $G_1R_2G_3$, $A_1G_2G_3$, $R_1G_2G_3$, $G_1A_2A_3$, $G_1R_2R_3$, $G_1A_2R_3$, $G_1R_2A_3$

$A_1G_2R_3$, $R_1G_2A_3$, $A_1G_2A_3$, $R_1G_2R_3$, $A_1A_2G_3$, $R_1R_2G_3$, $A_1R_2G_3$, $R_1A_2G_3$, $A_1A_2A_3$, $A_1A_2R_3$, $A_1R_2A_3$

$$\mathrm{A}_1\mathrm{R}_2\mathrm{R}_3, \mathrm{R}_1\mathrm{A}_2\mathrm{R}_3, \mathrm{R}_1\mathrm{A}_2\mathrm{A}_3, \mathrm{R}_1\mathrm{R}_2\mathrm{A}_3, \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3$$

Assuming that we have three endpoints in the trial and that no dropouts occur, at the end of the trial, we have collected the data

$$\mathbf{X}_1 = \{X_{11}, \ldots, X_{1n}\},$$

$$\mathbf{X}_2 = \{X_{21}, \ldots, X_{2n}\},$$

$$\mathbf{X}_3 = \{X_{31}, \ldots, X_{3n}\},$$

$$\mathbf{Y}_1 = \{Y_{11}, \ldots, Y_{1n}\},$$

$$\mathbf{Y}_2 = \{Y_{21}, \ldots, Y_{2n}\},$$

$$\mathbf{Y}_3 = \{Y_{31}, \ldots, Y_{3n}\},$$

where $\mathbf{X}_i$ is the data from the drug-arm w.r.t. the ith endpoint and $\mathbf{Y}_i$ is the data from the placebo-arm w.r.t. the ith endpoint.

The case of three endpoints is similar to the case of two endpoints. The only difference is that we increase the dimension, so that the random vector $\hat{\Delta} = (\hat{\Delta}_1, \hat{\Delta}_2, \hat{\Delta}_3)$ instead has dimension three. The same goes for all other vectors introduced in the section Two Endpoints, i.e. the dimension is increased from two to three and the components are calculated in the same way.

The idea to make a decision is the same idea as when we had two endpoints: make a decision w.r.t. each endpoint and end up in one of the 27 three-dimensional regions, and then make your decision based on which of the 27 regions you end up in when observing the data at the end of the trial. We still have the same issue as when we had only two endpoints: there is no unique way to translate the 27 three-dimensional regions into 27 one-dimensional regions.

3.2.1 Stepwise GNG criteria

Now that we have three endpoints instead of two, the stepwise GNG criteria becomes more complicated. The idea is still the same as when we had only two endpoints, but we look at the third endpoint if and only if we have made the decision A w.r.t. the first and second endpoint. This GNG criteria is used whenever the first endpoint is considered more important than the second and third endpoint, and when the second endpoint is considered more important than the third endpoint. Using this GNG criteria, the probability to end up in the A zone is

$$P(\mathrm{A}) = P(\mathrm{A}_1\mathrm{A}_2\mathrm{A}_3), \tag{13}$$

the probability to end up in the G zone is

$$P(\mathrm{G}) = P(\mathrm{G}_1\mathrm{G}_2\mathrm{G}_3)+P(\mathrm{G}_1\mathrm{G}_2\mathrm{A}_3)+P(\mathrm{G}_1\mathrm{G}_2\mathrm{R}_3)+P(\mathrm{G}_1\mathrm{A}_2\mathrm{G}_3)+P(\mathrm{G}_1\mathrm{R}_2\mathrm{G}_3)+P(\mathrm{A}_1\mathrm{G}_2\mathrm{G}_3)+P(\mathrm{G}_1\mathrm{A}_2\mathrm{A}_3)+$$

$$+ P(\mathrm{G}_1\mathrm{R}_2\mathrm{R}_3) + P(\mathrm{G}_1\mathrm{A}_2\mathrm{R}_3) + P(\mathrm{G}_1\mathrm{R}_2\mathrm{A}_3) + P(\mathrm{A}_1\mathrm{G}_2\mathrm{R}_3) + P(\mathrm{A}_1\mathrm{G}_2\mathrm{A}_3) + P(\mathrm{A}_1\mathrm{A}_2\mathrm{G}_3) \tag{14}$$

and the probability to end up in the R zone is

$$P(\mathrm{R}) = P(\mathrm{R}_1\mathrm{R}_2\mathrm{R}_3)+P(\mathrm{R}_1\mathrm{R}_2\mathrm{A}_3)+P(\mathrm{R}_1\mathrm{R}_2\mathrm{G}_3)+P(\mathrm{R}_1\mathrm{G}_2\mathrm{R}_2)+P(\mathrm{R}_1\mathrm{A}_2\mathrm{R}_3)+P(\mathrm{A}_1\mathrm{R}_2\mathrm{R}_3)+P(\mathrm{A}_1\mathrm{R}_2\mathrm{G}_2)+$$

$$+P(\mathrm{R}_1\mathrm{A}_2\mathrm{G}_3)+P(\mathrm{A}_1\mathrm{A}_2\mathrm{R}_3)+P(\mathrm{A}_1\mathrm{R}_2\mathrm{A}_3)+P(\mathrm{R}_1\mathrm{G}_2\mathrm{G}_3)+P(\mathrm{R}_1\mathrm{G}_2\mathrm{A}_3)+P(\mathrm{A}_1\mathrm{R}_2\mathrm{R}_3). \quad (15)$$

Using that AAA⊂AA, the probability to end up in the A zone has decreased when using stepwise GNG criteria with three endpoints instead of two.

Since $\hat{\Delta}$ is a random vector of dimension three, we use the three-dimensional CLT [2]:

Theorem 3.3: Let $\mathbf{X}_i = \begin{bmatrix} X_{i(1)} \\ X_{i(2)} \\ X_{i(3)} \end{bmatrix}$ for $i = 1, \ldots, n$ with mean vector $\mu = \mathbb{E}[\mathbf{X}_i] = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$. Also let

$$\bar{\mathbf{X}}_n = \frac{1}{n} \cdot \begin{bmatrix} \sum_{i=1}^n X_{i(1)} \\ \sum_{i=1}^n X_{i(2)} \\ \sum_{i=1}^n X_{i(3)} \end{bmatrix} = \frac{1}{n} \cdot \sum_{i=1}^n \mathbf{X}_i$$

denote the sample average. Then, as $n \to \infty$, $\frac{1}{\sqrt{n}} \cdot \sum_{i=1}^n (\mathbf{X}_i - \mu) = \sqrt{n} \cdot (\bar{\mathbf{X}}_n - \mu)$ converges in distribution to a multivariate normal distribution of dimension three with mean vector $\mathbf{0} = (0,0,0)$ and covariance matrix

$$\Sigma = \begin{bmatrix} \mathrm{Var}(X_{1(1)}) & \mathrm{Cov}(X_{1(1)}, X_{1(2)}) & \mathrm{Cov}(X_{1(1)}, X_{1(3)}) \\ \mathrm{Cov}(X_{1(1)}, X_{1(2)}) & \mathrm{Var}(X_{1(2)}) & \mathrm{Cov}(X_{1(3)}, X_{1(2)}) \\ \mathrm{Cov}(X_{1(1)}, X_{1(2)}) & \mathrm{Cov}(X_{1(3)}, X_{1(2)}) & \mathrm{Var}(X_{1(3)}) \end{bmatrix}.$$

According to the three-dimensional CLT, the random vector $\hat{\Delta}$ follows (approximately) a multivariate normal distribution of dimension three with mean vector $\Delta = (\Delta_1, \Delta_2, \Delta_3)$ and covariance matrix

$$\Sigma_{\hat{\Delta}} = \begin{pmatrix} s_1^2 & \rho_{12} \cdot s_1 \cdot s_2 & \rho_{13} \cdot s_1 \cdot s_3 \\ \rho_{12} \cdot s_1 \cdot s_2 & s_2^2 & \rho_{23} \cdot s_3 \cdot s_2 \\ \rho_{13} \cdot s_1 \cdot s_3 & \rho_{23} \cdot s_3 \cdot s_2 & s_3^2 \end{pmatrix}$$

for large values of n, where ρ_{ij} is the correlation between endpoint i and endpoint j for a single measurement. The probabilities are calculated by integrating the probability density function

$$f(x|\mu, \Sigma) = f(x_1, x_2, x_3|\mu, \Sigma) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu)\right)}{\sqrt{(2\pi)^3|\Sigma|}}$$

over the appropriate subspace of $\mathbb{R}^3$, e.g.

$$P(\mathrm{G}_1\mathrm{G}_2\mathrm{G}_3) = P(\hat{\Delta}_1 > a_1 \cap \hat{\Delta}_2 > a_2 \cap \hat{\Delta}_3 > a_3) = \int_{a_1}^{\infty}\int_{a_2}^{\infty}\int_{a_3}^{\infty} f(x_1, x_2, x_3|\mu = \Delta, \Sigma = \Sigma_{\hat{\Delta}})dx_1dx_2dx_3.$$

Once again, this integral is not available in a closed form, forcing us to rely on numerical methods to approximate it. The probabilities for the decisions of the third endpoint, conditioning that we ended up in the A zone w.r.t. the first and second endpoint, are given by

$$P(\mathrm{G}_3|\mathrm{A}_1\mathrm{A}_2) = \frac{P(\mathrm{A}_1\mathrm{A}_2\mathrm{G}_3)}{P(\mathrm{A}_1\mathrm{A}_2)},$$

$$P(\mathrm{A}_3|\mathrm{A}_1\mathrm{A}_2) = \frac{P(\mathrm{A}_1\mathrm{A}_2\mathrm{A}_3)}{P(\mathrm{A}_1\mathrm{A}_2)}$$

and

$$P(\mathrm{R}_3|\mathrm{A}_1\mathrm{A}_2) = \frac{P(\mathrm{A}_1\mathrm{A}_2\mathrm{R}_3)}{P(\mathrm{A}_1\mathrm{A}_2)}.$$

3.2.2 2-of-3 GNG criteria

Sometimes, it will not be the case that the first endpoint is more important than the second and that the second will be more important than the third. In that case, one could use the 2-of-3 GNG criteria instead of the stepwise GNG criteria. This GNG criteria can be seen as some sort of majority GNG criteria:

1. G zone: If at least two endpoints end up in the G zone.
2. R zone: If at least two endpoints end up in the R zone.
3. If two endpoints end up in the A zone then the decision is based on the decision of the endpoint that did not end up in the A zone.
4. A zone: If one G, one A and one R, or if all endpoints end up in the A zone.

If this GNG criteria is used then the probability to end up in the G zone is

$$P(\mathrm{G}) = P(\mathrm{G}_1\mathrm{G}_2\mathrm{G}_3)+P(\mathrm{G}_1\mathrm{G}_2\mathrm{A}_3)+P(\mathrm{G}_1\mathrm{G}_2\mathrm{R}_3)+P(\mathrm{G}_1\mathrm{R}_2\mathrm{G}_3)+P(\mathrm{G}_1\mathrm{A}_2\mathrm{G}_3)+P(\mathrm{A}_1\mathrm{A}_2\mathrm{G}_3)+P(\mathrm{A}_1\mathrm{G}_2\mathrm{A}_3)+ \\ + P(\mathrm{G}_1\mathrm{A}_2\mathrm{A}_3) + P(\mathrm{A}_1\mathrm{G}_2\mathrm{G}_3) + P(\mathrm{R}_1\mathrm{G}_2\mathrm{G}_3), \tag{16}$$

the probability to end up in the A zone is

$$P(\mathrm{A}) = P(\mathrm{A}_1\mathrm{G}_2\mathrm{R}_3)+P(\mathrm{A}_1\mathrm{R}_2\mathrm{G}_3)+P(\mathrm{A}_1\mathrm{A}_2\mathrm{A}_3)+P(\mathrm{G}_1\mathrm{A}_2\mathrm{R}_3)+P(\mathrm{R}_1\mathrm{A}_2\mathrm{G}_3)+P(\mathrm{G}_1\mathrm{R}_2\mathrm{A}_3)+P(\mathrm{R}_1\mathrm{G}_2\mathrm{A}_3) \tag{17}$$

and the probability to end up in the R zone is

$$P(\mathrm{R}) = P(\mathrm{R}_1\mathrm{R}_2\mathrm{R}_3)+P(\mathrm{R}_1\mathrm{R}_2\mathrm{G}_3)+P(\mathrm{R}_1\mathrm{R}_2\mathrm{A}_3)+P(\mathrm{R}_1\mathrm{A}_2\mathrm{R}_3)+P(\mathrm{R}_1\mathrm{G}_2\mathrm{R}_3)+P(\mathrm{A}_1\mathrm{R}_2\mathrm{R}_3)+P(\mathrm{G}_1\mathrm{R}_2\mathrm{R}_3)+ \\ + P(\mathrm{A}_1\mathrm{A}_2\mathrm{R}_3) + P(\mathrm{A}_1\mathrm{R}_2\mathrm{A}_3) + P(\mathrm{R}_1\mathrm{A}_2\mathrm{A}_3). \tag{18}$$

3.2.3 Stepwise 1-of-2 GNG criteria

If the first endpoint is considered primary and the other two endpoints are both considered secondary then one could combine the stepwise GNG criteria with the 1-of-2 GNG criteria into a new GNG criteria:

1. Make a decision w.r.t. the first endpoint. If the decision is G then the overall decision is G. If the decision is R then the overall decision is R. If the decision is A then continue to step 2.

2. Base the overall decision on the 1-of-2 GNG criteria w.r.t. the two secondary endpoints.

If this combined GNG criteria is used then the probability to end up in the G zone is

$$\begin{aligned} P(\mathrm{G}) = {} & P(\mathrm{G}_1\mathrm{G}_2\mathrm{G}_3)+P(\mathrm{G}_1\mathrm{G}_2\mathrm{A}_3)+P(\mathrm{G}_1\mathrm{G}_2\mathrm{R}_3)+P(\mathrm{G}_1\mathrm{R}_2\mathrm{G}_3)+P(\mathrm{G}_1\mathrm{A}_2\mathrm{G}_3)+P(\mathrm{A}_1\mathrm{A}_2\mathrm{G}_3)+P(\mathrm{A}_1\mathrm{G}_2\mathrm{A}_3)+ \\ & + P(\mathrm{G}_1\mathrm{A}_2\mathrm{A}_3) + P(\mathrm{A}_1\mathrm{G}_2\mathrm{G}_3) + P(\mathrm{G}_1\mathrm{R}_2\mathrm{R}_3) + P(\mathrm{G}_1\mathrm{R}_2\mathrm{A}_3) + P(\mathrm{G}_1\mathrm{A}_2\mathrm{R}_3), \end{aligned} \tag{19}$$

the probability to end up in the A zone is

$$P(\mathrm{A}) = P(\mathrm{A}_1\mathrm{A}_2\mathrm{A}_3) + P(\mathrm{A}_1\mathrm{G}_2\mathrm{R}_3) + P(\mathrm{A}_1\mathrm{R}_2\mathrm{G}_3) \tag{20}$$

and the probability to end up in the R zone is

$$\begin{aligned} P(\mathrm{R}) = {} & P(\mathrm{R}_1\mathrm{R}_2\mathrm{R}_3)+P(\mathrm{R}_1\mathrm{R}_2\mathrm{A}_3)+P(\mathrm{R}_1\mathrm{A}_2\mathrm{R}_3)+P(\mathrm{R}_1\mathrm{R}_2\mathrm{G}_3)+P(\mathrm{R}_1\mathrm{G}_2\mathrm{R}_3)+P(\mathrm{A}_1\mathrm{A}_2\mathrm{R}_3)+P(\mathrm{A}_1\mathrm{R}_2\mathrm{A}_3)+ \\ & + P(\mathrm{R}_1\mathrm{A}_2\mathrm{A}_3) + P(\mathrm{A}_1\mathrm{R}_2\mathrm{R}_3) + P(\mathrm{R}_1\mathrm{G}_2\mathrm{G}_3) + P(\mathrm{R}_1\mathrm{G}_2\mathrm{A}_3) + P(\mathrm{R}_1\mathrm{A}_2\mathrm{G}_3). \end{aligned} \tag{21}$$

4 Application of methods

In this chapter, the results from the analysis of the probabilities and risks will be applied to a situation where the parameters are specified.

4.1 Two endpoints

Consider a clinical trial with two endpoints designed to compare the effect of a new drug against placebo using the difference in means. In this trial, let

$$\Delta_{\text{TV}} = (10, 15),$$

$$\Delta_{\text{LRV}} = (5, 10),$$
$$\sigma = (\sigma_1, \sigma_2) = (15, 20),$$
$$\beta' = (0.1, 0.1),$$
$$\alpha' = (0.2, 0.2)$$

and

$$n = 50,$$

corresponding to a total of $N = 100$ subjects. Then

$$s = (3, 4),$$

$$a = (7.525, 13.366)$$

and

$$b = (6.155, 9.874).$$

If we would assume that $\rho = 0$ then we can list how the overall risks α and β are affected by the individual risks α_1, α_2, β_1 and β_2, for the 1-of-2 GNG criteria and stepwise GNG criteria:

α	β	α_1	α_2	β_1	β_2
13.25%	4.46%	20%	20%	10%	10%
16.07%	3.94%	20%	20%	10%	5%
16.08%	3.25%	20%	20%	5%	10%
11.75%	6.17%	20%	10%	10%	10%
10.13%	5.92%	10%	20%	10%	10%
10.45%	4.06%	10%	10%	5%	5%
18.9%	2.48%	20%	20%	5%	5%
7.63%	7.63%	10%	10%	10%	10%

Table 6: Overall risks for the 1-of-2 GNG criteria assuming no correlation between the two endpoints, as a function of the individual risks.

α	β	α_1	α_2	β_1	β_2
23%	11%	20%	20%	10%	10%
23%	10.5%	20%	20%	10%	5%
25.8%	6.5%	20%	20%	5%	10%
21.5%	11%	20%	10%	10%	10%
15%	12.5%	10%	20%	10%	10%
13.91%	6.5%	10%	10%	5%	5%
25.8%	5.77%	20%	20%	5%	5%
12.5%	12.5%	10%	10%	10%	10%

Table 7: Overall risks for the stepwise GNG criteria, assuming no correlation between the two endpoints, as a function of the individual risks.

To see how the correlation ρ between the two endpoints affects the probabilities of the three decisions when the drug fulfills the TV or the LRV, the following six plots have been produced:

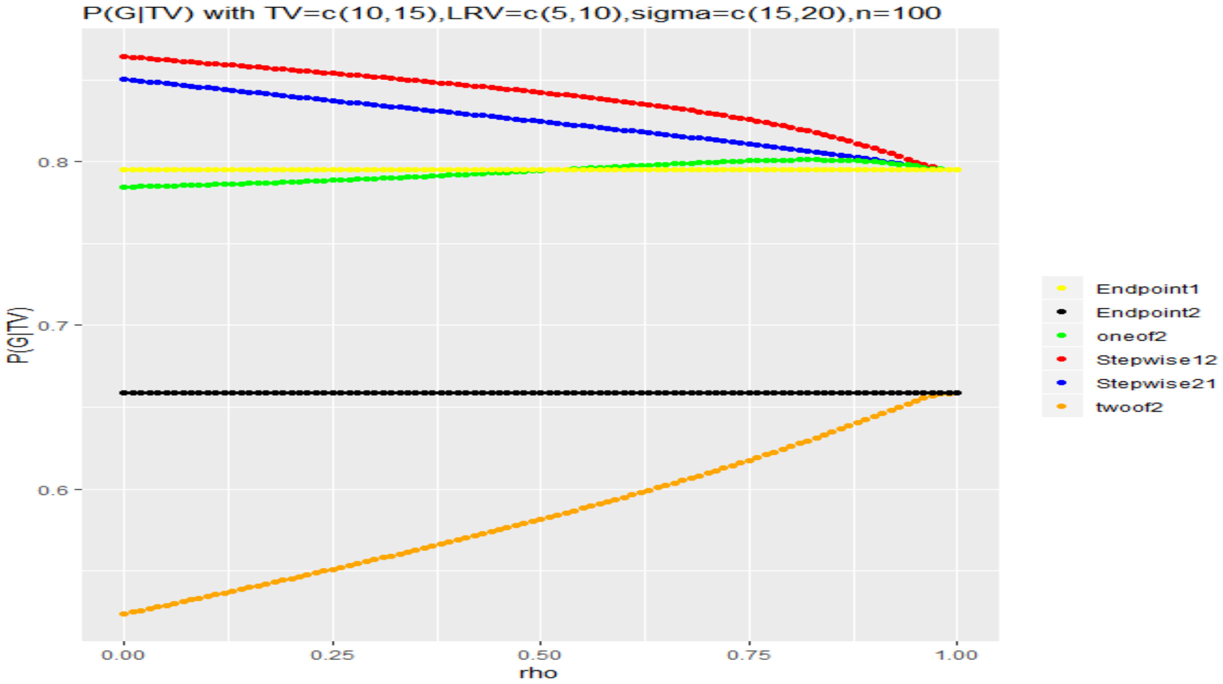

Figure 2: The probability to end up in the G zone assuming that the drug fulfills the TV, as a function of correlation and method.

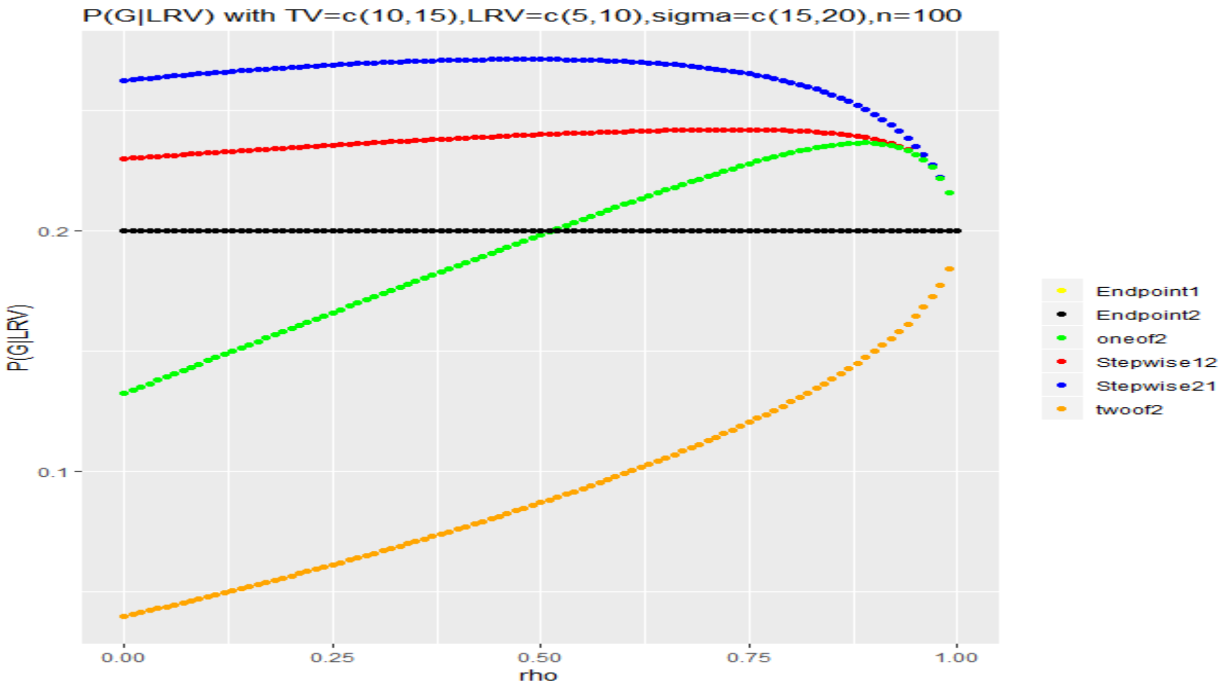

Figure 3: The probability to end up in the G zone assuming that the drug fulfills the LRV, as a function of correlation and method. In other words, this is the overall False Go Risk.

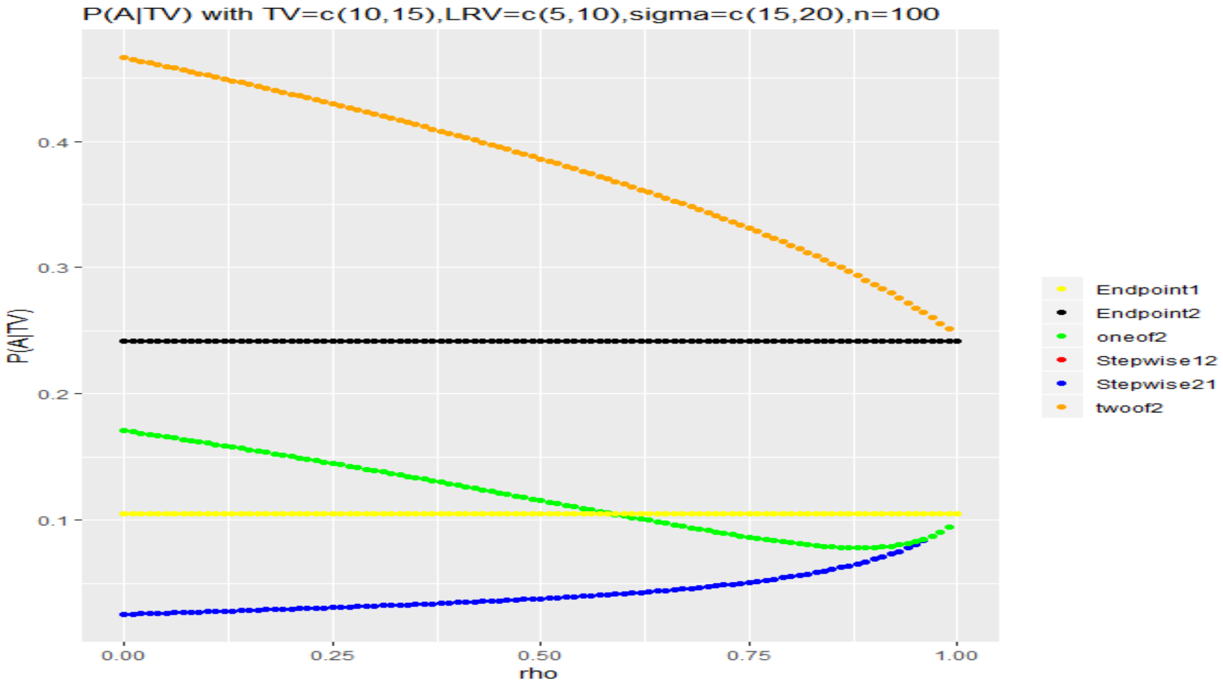

Figure 4: The probability to end up in the A zone assuming that the drug fulfills the TV, as a function of correlation and method.

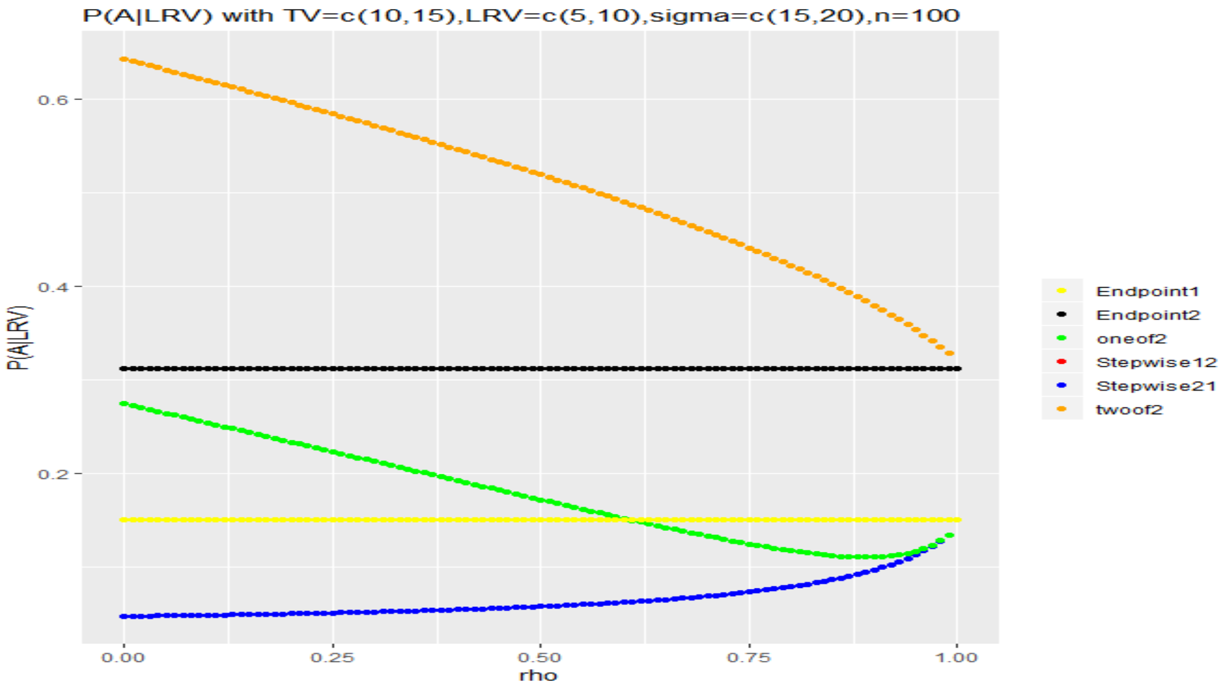

Figure 5: The probability to end up in the A zone assuming that the drug fulfills the LRV, as a function of correlation and method.

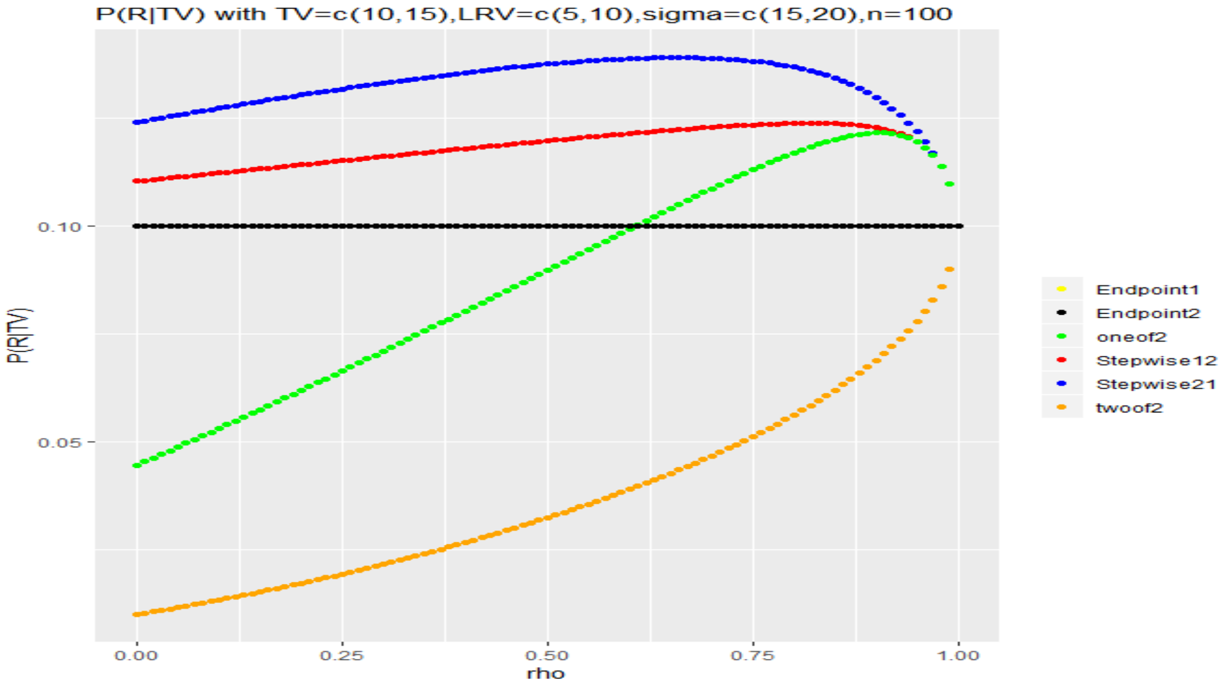

Figure 6: The probability to end up in the R zone assuming that the drug fulfills the TV, as a function of correlation and method. In other words, this is the overall False Stop Risk.

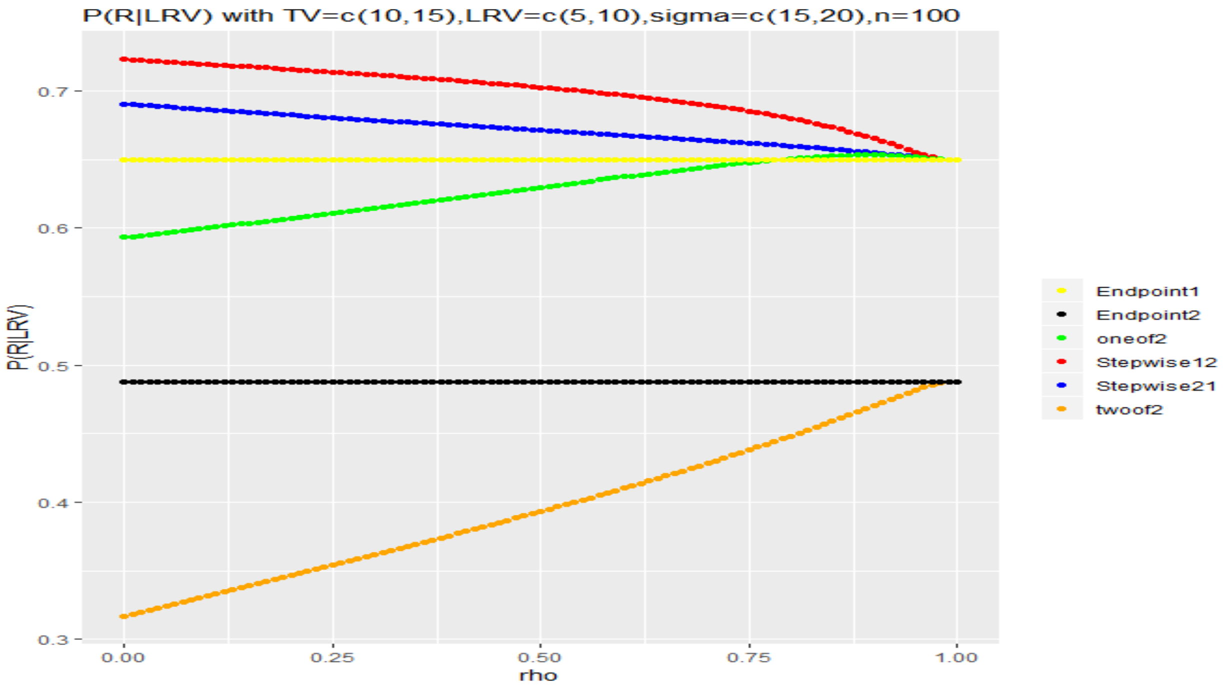

Figure 7: The probability to end up in the R zone assuming that the drug fulfills the LRV, as a function of correlation and method.

Figure 2 shows how the probability to make an overall G decision, with each method, is affected by the magnitude of the correlation ρ between the two endpoints when the drug fulfills the TV. Figure 3 is the same plot as figure 2, but when the drug fulfills the LRV instead of the TV. Figures 4 and 5 are the corresponding plots for the probability of A, and figures 6 and 7 are the corresponding plots for the probability of R.

When we have two endpoints then we can either base the GNG decision on both endpoints or choose to ignore one of the endpoints and base the GNG decision on the other endpoint. In figure 2, we can see that $P(\mathrm{G}|\mathrm{TV})$ reaches either $P(\mathrm{G}_1|\mathrm{TV}_1)$ or $P(\mathrm{G}_2|\mathrm{TV}_2)$ when $\rho \to 1$. That makes sense, since the more correlated the two endpoints are, the higher probability that they give the same decision. Thus, ignoring one of them will not matter that much when the correlation is close to 1. The closer to zero the correlation is, the more information we obtain by using both endpoints.

We can also see, in figure 2, that the stepwise GNG criteria is dominating all other GNG criteria, no matter if the correlation is weak or strong. This is expected, considering equations 7 and 10. However, as a consequence of equations 7 and 10, the overall false Go Risk, that is shown in figure 3, is also dominated by the stepwise GNG criteria. Also, the overall false Stop Risk, that is shown in figure 6, is also dominated by the stepwise GNG criteria. Thus, there is a trade-off between the probability to end up in the A zone and the overall risks.

If the overall risks are low then the probability to end up in the A zone is high, but if the overall risks are high then the probability to end up in the A zone is low. In general, we can not have a low probability to end up in the A zone and low overall risks. The only parameter that we can change to control the overall risks are the risks for the individual endpoints, i.e. α_1, α_2, β_1 and β_2. If we choose to ignore the correlation (i.e. assume $\rho = 0$) then we can calculate the overall risks as a function of the individual risks, as seen in chapter 3.

Moreover, the 2-of-2 GNG criteria seems to be problematic, considering the fact that $P(\text{A}|\text{LRV}) > 30\%$ for any $\rho \in [0, 1]$ according to figure 5. One way to resolve this is to have larger values of α_1, α_2, β_1 and β_2 than 20% and 10% so that the overall risks become larger.

5 Case Study

This chapter gives a short description of a clinical trial where the GNG decision is based on the difference in means w.r.t. three endpoints. The result of this case study, using the methods described in chapter 3 and chapter 2, is presented in this chapter after the short description.

This study is an allergen inhalation challenge study conducted on healthy human subjects for a new drug that is supposed to help against asthma. The aim of the study is to study the effect of the drug compared to placebo w.r.t. the following primary endpoint: LAR (Late Asthmatic Response). Two other endpoints that will be used as secondary endpoints are Sputum and PC20.

An allergen inhalation challenge is a clinical trial model commonly used as an experimental tool to better understand the pathophysiology of allergic asthma [5]. The subjects are exposed to an allergen and then receive the drug (or placebo) that is supposed to help the subjects withstand the allergen. The primary endpoint can be the reduction in either LAR (as in this study) or EAR (Early Asthmatic Response).

In this study, we have obtained the following information from the historical placebo-data: Placebo-effect vector

$$\mu_p = (\mu_p^{(\text{LAR})}, \mu_p^{(\text{Sputum})}, \mu_p^{(\text{PC20})}) = (-24.014, 19.338, -3.42),$$

standard deviation vector

$$\sigma = (\sigma^{(\text{LAR})}, \sigma^{(\text{Sputum})}, \sigma^{(\text{PC20})}) = (12.653, 17.394, 7.711)$$

and correlation between the three endpoints:

	LAR	Sputum	PC20
LAR	1	-0.644	-0.214
Sputum	-0.644	1	-0.024
PC20	-0.214	-0.024	1

Table 8: Correlation between endpoints.

The TV is to have a 50% reduction in LAR, 50% increase in Sputum and 50% reduction in PC20. The LRV is to have a 25% reduction in LAR, 25% increase in Sputum and 25% reduction in PC20. Thus, we obtain

$$\Delta_{\text{TV}} = (12.007, 9.669, 1.71)$$

and

$$\Delta_{\text{LRV}} = (6.003, 4.835, 0.855).$$

This, in turn, gives us the following decision boundaries when using a false go risk of 20% and false stop risk of 10% for each endpoint:

$$a = (a^{(\text{LAR})}, a^{(\text{Sputum})}, a^{(\text{PC20})}) = (9.553, 9.714, 3.018),$$

$$b = (b^{(\text{LAR})}, b^{(\text{Sputum})}, b^{(\text{PC20})}) = (6.602, 2.239, -1.584).$$

The number of subjects per arm is $n = 18$. When it comes to the overall GNG criteria, we can choose to either use only LAR (since it is the primary endpoint), or LAR combined with either Sputum or PC20 (since both Sputum and PC20 are secondary endpoints), or we can choose to use all three endpoints. The following two tables shows the probabilities for each of the three decisions, assuming that the drug fulfills either the TV or the LRV:

GNG criteria	P(G|TV)	P(A|TV)	P(R|TV)
Stepwise LAR, Sputum	85.82%	4.03%	10.15%
Stepwise LAR, PC20	78.68%	10.13%	11.19%
Stepwise LAR, Sputum, PC20	87.72%	1.98%	10.3%
Stepwise LAR, PC20, Sputum	86.77%	1.98%	11.25%
1-of-2 LAR/Sputum	69.5%	17.23%	13.27%
1-of-2 LAR/PC20	70.33%	22.93%	6.74%
2-of-3	48.58%	49.97%	1.45%
Only LAR	71.96%	18.04%	10%
Stepwise 1-of-2	86.69%	1.97%	11.34%

Table 9: The probabilities for each of the three decisions, assuming that the TV holds.

GNG criteria	P(G|LRV)	P(A|LRV)	P(R|LRV)
Stepwise LAR, Sputum	21.7%	12.27%	66.03%
Stepwise LAR, PC20	24.08%	15.58%	60.34%
Stepwise LAR, Sputum, PC20	23.4%	7.88%	68.72%
Stepwise LAR, PC20, Sputum	25.1%	7.88%	67.02%
1-of-2 LAR/Sputum	25.95%	21.94%	52.11%
1-of-2 LAR/PC20	19%	34.12%	46.88%
2-of-3	6.55%	67.27%	26.18%
Only LAR	20%	24.36%	55.64%
Stepwise 1-of-2	22.86%	7.88%	69.26%

Table 10: The probabilities for each of the three decisions, assuming that the LRV holds.

We can clearly see that the 2-of-3 GNG criteria, and the 1-of-2 GNG criteria with LAR and PC20, are not acceptable decision methods in this case study, since the probability to end up in the A zone is too high (approximately 67% and 34% when the LRV holds, which is larger than 30%), according to table 8. We can also see that the 1-of-2 GNG criteria with

LAR and Sputum is not an optimal GNG criteria compared with the stepwise GNG criteria, even if it is still deemed acceptable.

Moreover, the stepwise GNG criteria with all three endpoints seems to be the best GNG criteria in this case study. Ignoring one of the secondary endpoints seems to be a bad idea, since the probability to end up in the A zone increases too much in comparison with how the overall false go risk (i.e. P(G|LRV)) and overall false stop risk (i.e. P(R|TV)) decreases. Considering that one of the endpoints is primary (LAR) and the other two endpoints are secondary (Sputum and PC20), it makes more sense to use the stepwise 1-of-2 GNG criteria, even if it is not the optimal GNG criteria according to table 7 and 8.

If one is unlucky then one would end up in the A zone w.r.t. LAR. Then one would be forced to look at the secondary endpoints to make the overall GNG decision. The probabilities of the three decisions w.r.t. the secondary endpoints, conditioning that LAR ended up in the A zone, are given in the following table:

Endpoint	P(G|TV)	P(A|TV)	P(R|TV)	P(G|LRV)	P(A|LRV)	P(R|LRV)
Sputum	76.83%	22.35%	0.82%	6.98%	50.37%	42.65%
PC20	37.26%	56.14%	6.60%	16.76%	63.95%	19.29%

Table 11: The probabilities for each of the three decisions w.r.t. the secondary endpoints, conditioning that we ended up in the A zone w.r.t. LAR, with $n = 18$ subjects per arm. The three leftmost probabilities holds when the TV holds and the remaining three probabilities holds when the LRV holds.

If one were to look at the endpoints stepwise then one should first look at LAR and secondly Sputum. If we would ignore the correlation between LAR and the secondary endpoints (i.e. assume that $\rho_{12} = \rho_{13} = 0$) then the conditional probabilities would become unconditional probabilities, since uncorrelated implies independent in the multivariate normal distribution case. The corresponding unconditional probabilities are given in the following table:

Endpoint	P(G|TV)	P(A|TV)	P(R|TV)	P(G|LRV)	P(A|LRV)	P(R|LRV)
Sputum	49.69%	40.31%	10%	20%	47.28%	32.72%
PC20	30.54%	59.46%	10%	20%	62.87%	17.13%

Table 12: The probabilities for each of the three decisions w.r.t. the secondary endpoints, with $n = 18$ subjects per arm. The three leftmost probabilities holds when the TV holds and the remaining three probabilities holds when the LRV holds.

In table 10, we can see that if we end up in the A zone w.r.t. LAR, then we should look at Sputum and not PC20. If we are unlucky again and end up in the A zone w.r.t. Sputum, then the conditional probabilities of the three decisions w.r.t. PC20 are given in this table:

	P(G|TV)	P(A|TV)	P(R|TV)	P(G|LRV)	P(A|LRV)	P(R|LRV)
PC20	47.39%	49.02%	3.59%	13.83%	64.25%	21.92%

Table 13: The probabilities for each of the three decisions w.r.t. PC20, conditioning that we ended up in the A zone w.r.t. LAR and Sputum, with $n = 18$ subjects per arm. The three leftmost probabilities holds when the TV holds and the remaining three probabilities holds when the LRV holds.

If one is unlucky to end up in the A zone at both LAR and Sputum, should we use PC20 to make our overall decision or should we ignore it? If we decide to use PC20 and end up in the A zone, then we do not have any other endpoint to look at. In the same time, if we end up in the G zone then the risk of making the wrong decision is low, since the False go risk (i.e. P(G|LRV)) is low ($\approx 14\%$).

Moreover, if we end up in the R zone then the risk of making the wrong decision is also low, since the False stop risk (i.e. P(R|TV)) is low ($\approx 4\%$). Thus, ending up in the A zone is the most dangerous risk one takes if one decides to include PC20 in the decision process.

But, we still need to take into account that these probabilities are calculated using the multivariate normal distribution. Considering that the number of subjects per arm is $n = 18$, the approximation of the multivariate normal distribution (according to theorem 3.2 and theorem 3.3) might not be so accurate.

Thus, we can not ignore the fact that these approximations of the probabilities might be poor and could deviate a lot from the true probabilities. A way to obtain more accurate probabilities would be to simulate data according to the LRV and TV repeatedly and calculate how many times we make the decisions G/A/R according to the different decision methods. But, that would be computationally heavy and would require us to simulate at least 10000 times to obtain probabilities with high accuracy. Plus, we do not know the exact distribution of $\hat{\Delta}$ under the LRV and TV.

Moreover, we also need to take into account that the correlation between the three endpoints are sample correlations, calculated using the placebo-data from older clinical trials. Considering that the correlation between Sputum and PC20 is approximately -0.02, the true underlying correlation could be 0, i.e. Sputum and PC20 could be uncorrelated. The sample correlation is an unbiased estimator of the population correlation ρ if $\rho = 0$, $\rho = 1$ or $\rho = -1$. The sample correlation also becomes unbiased when $n \to \infty$, where n is the sample size. The sample size of the historical placebo-data is $n = 48$, which is considered low compared to ∞.

6 Conclusions and further research

In this chapter, the conclusions drawn from the discussions in the end of chapter 4 and 5 will be presented. Also, possible further research that can be made in this research topic will be discussed.

6.1 One endpoint or two endpoints?

To conclude the discussion in chapter 4, use two endpoints if the correlation between the two endpoints is low, so that one obtains much more information by using the other endpoint. If the probability to end up in the A zone is too high, consider to increase the individual risks of the endpoints, or consider to use another GNG criteria. Each of the presented GNG criteria in chapter 3 has its own disadvantages, but in my opinion, the stepwise GNG criteria seems to be the best. In real life situations, I would strongly recommend to use the stepwise GNG criteria with a lower false go risk than 20%, e.g. 10%, and lower false stop risk than 10%, e.g. 5%, in order to balance out the probability of the A zone and the overall risks.

6.2 Case study

To conclude the discussion of the Case Study, I would use the probabilities with caution. I would not ignore the correlation between the endpoints, except in the case of the correlation between Sputum and PC20. The fact that the correlation between PC20 and the other two endpoints is pretty low indicates that we should not ignore PC20 and thus use it in the decision process. However, considering that the probability to end up in the A zone w.r.t. PC20 is high when conditioning that we ended up in the A zone w.r.t. LAR and Sputum, it does not seem to help very much at all to include PC20 in the decision process. Thus, I would not recommend to use PC20 in the decision process. However, if one should decide to use PC20, one should be aware of the risks involved.

6.3 Further Research

As mentioned in chapter 2, one could employ a bayesian approach to construct the GNG criteria. This approach was not used in this thesis and could form the base for further research. If one would employ this approach then the first step would be to use the normal-normal conjugacy. The next step would be to try to derive the posterior distribution for other prior distributions than the normal distribution. Sampling methods like MCMC (Markov Chain Monte Carlo) and rejection sampling could be useful to obtain a sample of the posterior distribution when the posterior distribution is not available in a closed form.

Moreover, this thesis was restricted to two and three endpoints. Further research could be to investigate how the results would differ if more endpoints are included. Also, the results presented are based on the difference in means to construct the decision framework.

Even if the ratio of means has disadvantages, it would be interesting to see if the ratio of means would give similar results as the difference in means.

7 References

[1] Frewer P, Mitchell P, Watkins C and Matcham J. Decision-making in early clinical drug development. *Pharmaceutical Statistics* **2016**, 15 255-263. Published March 17, 2016 in Wiley online library.

[2] Van der Vaart A.W. *Asymptotic Statistics* (page 16), 1998. New York: Cambridge University Press.

[3] Lalonde R.L., Kowalski K.G., Utmacher M.M., Ewy W, Nichols D.J., Milligan P.A., Corrigan B.W., Lockwood P.A., Marshall S.A., Benincosa L.J., Tensfeldt T.G., Parivar K, Amantea M, Glue P, Koide H. Miller R. Model-based Drug Development. *Clinical Pharmacology & Therapeutics* 2007; **82**:21-32.

[4] Dmitrienko A, Pulkstenis E. *Clinical Trial Optimization Using R* (page 253-257), 2017. Chapman & Hall/CRC Biostatistics Series.

[5] Gauvreau G.M., O'Byrne P.M., Boulet L.P., Wang Y, Cockcraft D, Bigler J, FitzGerald J.M., Boedigheimer M, Davis B.E., Dias C, Gorski K.S., Smith L, Bautista E, Comeau M.R., Leigh R, Parnes J.R. *Effects of an Anti-TSLP Antibody on Allergen-Induced Asthmatic Responses. State of the art asthma. New England Journal of Medicine* 370;22. May 29, 2014.

[6] **Saint-Hilary G**, Cadour S, Robert V, and Gasparini M. *A simple way to unify multicriteria decision analysis (MCDA) and stochastic multicriteria acceptability analysis (SMAA) using a Dirichlet distribution in benefit-risk assessment.* Biometrical Journal, 59(3):567-578, 2017.

[7] Mussen F, Salek S, Walker S. *A quantitative approach to benefit-risk assessment of medicines - part 1: The development of a new model using Multi-Criteria Decision Analysis (MCDA).* Pharmacoepidemiology and drug safety 2017; **16**: S2-S15.

[8] Waddingham E, Mt-Isa S, Nixon R, Ashby D. *A Bayesian approach to probabilistic sensitivity analysis in structured benefit-risk assessment.* Biometrical Journal **58** (2016) 1, 28-42.

[9] Food and Drug Administration (FDA). *Multiple Endpoints in Clinical Trials Guidance for Industry.* Draft Guidance. January 2017.
(https://www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials-guidance-industry)

[10] Mitchell P.D. A Bayesian single-arm design using predictive probability monitoring. *Biometrics & Biostatistics International Journal.* Published July 26, 2018.