Thesis for the Degree of
Master of Science in Mathematical Statistics

Decision Policies for Early Stage Clinical Trials with Multiple Endpoints

Víctor López Juan

UNIVERSITY OF
GOTHENBURG

Department of Mathematical Sciences
University of Gothenburg
Gothenburg, Sweden 2022

Decision Policies for Early Stage Clinical Trials with Multiple Endpoints
Víctor López Juan

Department of Mathematical Sciences
University of Gothenburg
SE-412 96 Gothenburg
Sweden
Telephone +46 (0)31-772 10 00

Printed in Gothenburg, Sweden

Decision Policies for Early Stage Clinical Trials with Multiple Endpoints

Víctor López Juan
Department of Mathematical Sciences
University of Gothenburg

Abstract:
Before a drug can be prescribed to patients, it must be shown to be safe and effective for a certain indication in a controlled clinical trial (known as Phase III). Such studies are costly to run and expose patients to potential risks. Therefore, after initial studies in human subjects show the drug's safety (Phase I), studies with a small number of patients are run to assess the prospects of the drug (Phase II). If the number of patients in a Phase II study is not be sufficient to detect differences in the variable of interest (e.g. number of hospitalizations due to heart failure); a surrogate variable which is predictive of the variable of interest is used instead. A decision framework originally proposed by Lalonde (2007) is used in industry to determine, based on a single surrogate endpoint, whether to "Go" ahead with a Phase III study, or to "Stop" development of the drug. In some therapeutic areas, a single endpoint is not sufficient to predict the Phase III variable of interest; several related endpoints are used instead. Endpoints which are considered clinically related may be grouped into domains. How to best combine several disease markers across different domains to achieve the desired probabilities of correct/incorrect decisions is an open question.

This report presents an extension to multiple endpoints of the decision framework proposed by Lalonde. In this extension, decision policies are formulated in two levels. First, a Go or Stop decision is made for each domain, for example by individually comparing each of the relevant endpoints to certain thresholds. Performing multiple comparisons heightens the risk of an incorrect Go decision. This risk can be controlled effectively by using the Simes procedure (1986), which is a special case of the Benjamini-Hochberg (1995) method. Domain-level decisions are then combined into policies fulfilling a monotonicity property. This property enables the calculation of upper bounds for the probability of an incorrect decision, and lower bounds for the probability of a correct decision. These calculations are performed both for purely synthetic endpoints and for a case study involving endpoints related to heart failure. The resulting bounds are analogues to the statistical notions of Type I error and power, respectively. Heuristics are derived to help practitioners decide which endpoints to include, depending on the statistical power of these endpoints and on which combinations of true effects are of clinical interest.

Overall, the framework proposed in this report can represent many of the policies used by practitioners when designing Phase II studies with multiple endpoints. The outcome of the simulations presented in this theses can guide the selection of endpoints in order to achieve the desired bounds on the probabilities of correct and incorrect decisions.

Keywords: clinical trials, early stage, multivariate, Lalonde

Acknowledgements

I wish to thank my supervisors, Karin Nelander and Marcus Millegård, for their dedication to making this report comprehensive, relevant and applicable; and the examiner, José Sánchez, for getting this project off the ground and for his support throughout. I am grateful to both my supervisors and my examiner for the copious amount of helpful feedback and suggestions they have provided on the many iterations of this manuscript.

This thesis was written in collaboration with AstraZeneca. I want to thank my colleagues for making me feel welcome from the first day, and for all the stimulating discussions and enjoyable coffee breaks during my time here. These six months have gone by so fast.

I extend my thanks to Stefan and Franz, for being such pleasant companions in this statistical journey.

And to Henrik, for encouraging me to work on what I love.

Contents

List of Definitions

List of Figures

List of Tables

Chapter 1

Introduction

Before a new drug or treatment can be marketed to patients, sufficient evidence must be collected regarding its safety and effectiveness. The clinical studies required to collect this evidence (i.e. Phase III studies) entail a sizable commitment of resources, manpower and time to be carried out, and expose large numbers of patients to potential side effects. To manage these risks, shorter studies with fewer patients (i.e. Phase II) are run to assess the viability of the drug and decide whether to pursue a larger study. However, in some therapeutic areas, Phase II studies enrol too few patients over too short time spans to yield sufficient information on the Phase III endpoint. In these cases, a surrogate endpoint which can be estimated with lower variance is then used for the Phase II studies. In some therapy areas, there might not a single surrogate endpoint which can predict the Phase III endpoint with high accuracy. Instead, multiple endpoints reflecting different aspects of the disease are observed in order to make a decision.

A framework originally proposed by Lalonde [LKH$^+$07], is used in industry for making decisions regarding the viability of a drug candidate based on a single endpoint. In this report the framework is extended to support decision making for multiple endpoints. The statistical properties of the decision policies built within this framework are evaluated under different scenarios of interest.

The background section of report is written from the perspective of clinical studies performed to gain approval by FDA (Food and Drug Administration), as the US is typically one the first countries in which marketing permission for a new treatment is sought. However, the issues at hand are shared by clinical trials performed within other regulatory frameworks.

1.1 Background

Since the early 1920s, new drugs brought to market in the US must demonstrate safety and efficacy in a randomized clinical trial before being marketed and prescribed to patients. This means that the expected benefit for a patient taking the drug is commensurate with the risk of potential side effects. In a typical setup to demonstrate efficacy, the treatment must be shown to be better than the existing standard of care (or, in the absence of that, a placebo)

Table 1.1: Phases of clinical trials up until commercial approval. Before Phase I, the drug must have been suitably tested in cell cultures and non-human animals to allay concerns of potential serious side effects. After approval, further studies (so called Phase IV and V) may be conducted both to ensure safety and to gain approval for more indications and patient populations. Among other information, the table shows the "Success rate" for each phase, which is the overall proportion of treatments that progress from said phase to the next. For instance, 60% of drugs that enter a Phase III trial result in a New Drug Application (NDA). After Phase III, if the NDA is successful (which occurs at an overall rate of 85%), the drug receives a marketing authorization for the indications shown to be safe and effective in Phase III. Overall, 10% of drugs that enter Phase I trials result in an approval for at least one indication.

<table>
<tr><td></td><td>Phase I</td><td>▶</td><td colspan="2">Phase II</td><td>▶</td><td colspan="2">Phase III</td><td>▶</td></tr>
<tr><td></td><td></td><td></td><td>IIa</td><td>IIb</td><td></td><td>III</td><td>NDA</td><td></td></tr>
<tr><td>Main goals</td><td>Safety, Dosage, PK/PD</td><td></td><td colspan="2">Safety and Effect (Phase IIa), Dosage (Phase IIb)</td><td></td><td colspan="2">Risk/Benefit relationship, Indications</td><td></td></tr>
<tr><td rowspan="2">Participants[a]</td><td>Healthy volunteers[b]</td><td></td><td colspan="4">Afflicted patients</td><td></td></tr>
<tr><td>10s</td><td></td><td>10s-100s</td><td>100s</td><td></td><td colspan="2">100s-1000s</td><td></td></tr>
<tr><td>Duration[c]</td><td>Months</td><td></td><td colspan="2">Months – 2 years</td><td></td><td>1–4 years</td><td>1 year</td><td></td></tr>
<tr><td>Cost[d] (MUSD)</td><td>1–6</td><td></td><td colspan="2">7–20</td><td></td><td>12–50</td><td>2</td><td></td></tr>
<tr><td rowspan="2">Success rate[e]</td><td>64%</td><td></td><td colspan="2">32%</td><td></td><td>60%</td><td>85%</td><td></td></tr>
<tr><td colspan="7">10%</td></tr>
</table>

[a] Population from which the trial participants are drawn and order of magnitude of their number [CL13].

[b] For drugs with particularly strong side effects (e.g. chemotherapy), the Phase I involves patients from the target population instead.

[c] Typical duration of a clinical study in this phase [FDA18]. For the NDA (New Drug Application), the time to review the application is given.

[d] Average cost for clinical trials across a broad range of therapeutic areas [SBBE14]. For the NDA, the FDA's application fee is given.

[e] Proportion of drugs in this phase that reach the following phase (top). In the case of the NDA, percentage of applications that result in a marketing authorization. Below, the overall percentage of drugs that enter clinical trials (Phase I) which result in a marketing authorization [HTC+14].

by a clinically-meaningful margin, and this difference must be shown to be statistically significant.

For certain conditions, achieving a sufficiently high probability of the study detecting the relevant effect (i.e. high-power) requires recruiting large numbers of patients and following them over long periods of time. Launching such a large-scale clinical trial for every possible drug candidate would be prohibitive and expose study participants to unwarranted risks.

Instead, investigational drugs are tested in phases. In the pre-clinical stage, the drugs are tested on cell cultures and animals. The collected data is submitted to the regulator as an Investigational New Drug (IND) application. If the evidence shows that the drug is safe and effective in animal models, the regulator authorizes the drug for clinical studies in humans.

Phase I trials are the first occasion in which the drug is tested in humans. The goals of these studies are usually to (i) assess if and at which doses the drug is safe to use in humans, and (ii) measure the pharmacological properties of the drug; e.g. how fast the drug is metabolized, how long it remains in the body and how it is excreted. The participants in a Phase I trial are typically a small number of healthy males; although, if the mentioned goals cannot be realized by observing only healthy volunteers, or in the case of certain inherently toxic treatments where using healthy volunteers who cannot gain any benefit from the treatment would be unethical (e.g. certain cancer drugs), patients suffering the target disease may be enrolled [CL13, §6.2].

Provided that the drug is deemed sufficiently safe in a Phase I trial, Phase II trials are performed to assess effectiveness in the target population. This is usually the first time that the drug is tested on sick patients. Phase II trials are often divided into two phases. In Phase IIa the emphasis is on assessing whether the drug has an effect. Usually, a single dose which is deemed high enough to be likely to cause a measurable effect, but still within safety margins is used. If the results of the Phase IIa are promising, a Phase IIb study is performed. The Phase IIb will usually have more patients than the Phase IIa and follow them during a longer time period. One or two different doses are tested, from which the dose used in an eventual Phase III study is selected.

If the Phase II studies show that the drug is safe and that it may have a clinically-relevant effect, then a larger, Phase III study may be performed. In Phase III the indications for which the benefits of the drug to the patients are commensurate with its risks are ascertained. If sufficient evidence is obtained for one or more indications (typically statistical significance at the $\alpha = 0.05$ level), an NDA (New Drug Application) is submitted to the regulator. The NDA includes the results of the Phase III study together with the evidence collected in previous phases. While the NDA is under review, further trials may be conducted to gain additional evidence of the drug's effectiveness. These are sometimes known as Phase IIIb trials.

If the evidence included with the NDA is sufficient to show effectiveness for one or more indications, the drug receives a marketing authorization. An authorized drug can be legally marketed for the corresponding indications and prescribed to patients.

After the approval of the drug, additional studies may be performed. Some of these may be observational studies involving patients which have been prescribed the drug by their doctors, both to assess real-world efficacy and to

detect relatively rare side effects that may have been missed in Phase III. Additional studies may also be performed to obtain marketing authorization for new indications, or for new patient subpopulations (e.g. children).

1.2 Motivation

In 2004, the FDA published a report regarding the diminishing returns of investment in clinical trials. That is, the proportion of Phase III trials that succeed is becoming smaller, despite the advancements in biomedical research. This has increased the cost of bringing new medicines to market and reduced the number of new treatments that reach patients [FDA04].

During the previous decades biomedical research had been greatly successful at finding molecules that target the specific mechanisms behind many diseases. For the remaining ailments, either their cause is not completely understood in terms of a single therapeutic target, or a molecule that can bind to this target effectively has not yet been found. This lack of understanding makes it difficult to predict, at the early stages of the clinical development, which treatments are likely to be safe and effective for a given indication. This means that many Phase III trials will often not read; that is, they will not produce the a statistically significant result that leads to approval. There are many possible reasons for a trial not reading, including errors when conducting the study that invalidate the results, toxicity issues that the sample sizes in earlier phases were not sufficient to discover, or the inherent random variation in measurements. More importantly, the failure to read may be due to incorrect assumptions about the effect of the drug, or the drug not having a clinically significant effect at all.

In order to try and predict the success of an eventual Phase III study, a smaller, Phase II study is performed beforehand. However, in the case of medicines that aim to minimize the occurrence of events that are already relatively infrequent (e.g. myocardial infarction), thousands of patients followed for a long period of time are required to achieve sufficient statistical power. Early-stage clinical trials are comparatively short in duration and enrol too few patients to reliably estimate such an endpoint.

Among the remedies to this situation, the FDA's report [FDA04] proposes the adoption of new disease makers for early stage trials that can act as a proxy for the endpoint that will be measured in Phase III. For example, in the case of drugs targeting heart function, this may involve biomarkers associated the disease in question; image-based diagnostics, questionnaires for self-evaluation of well-being; associated physiological conditions such as kidney function, physical fitness measurements such as the distance that a patient can run or walk in a fixed time period; or, more recently, the use of wearable activity trackers to monitor the patients activity. Treatments that improve life expectancy, by e.g. reducing the frequency of major adverse cardiovascular events (MACE) can be presumed to also produce clinically-relevant improvements in one or more of the aforementioned endpoints. As another example, in the case of cancer drugs, a decrease in the size of the tumour can be a surrogate for the Phase III endpoint, which is usually progression-free survival.

In 2012 AstraZeneca adopted a methodology for early stage clinical trials

based on the decision-making framework proposed by Lalonde [LKH+07]. In this frequentist approach, the healthcare practitioners behind the study first choose an endpoint and determine two reference values: a lower reference value (LRV), which is the smallest effect for the endpoint that is clinically relevant; and a target value (TV), which is a desired effect for a drug to be marketable. Using these values and the results from a study, the framework produces a decision to either go ahead with a Phase III study or to stop clinical development of the drug. If the sample is too small or the variation in measurements too high compared to the size of the true effect, the result may be inconclusive. The user of the framework can establish an upper bound on the probability of stopping a drug that reaches the target value (False Stop), and of going ahead with a drug that does not reach the lower reference value (False Go). The lower these bounds are, the lower the probability of producing an erroneous decision, but also the higher the probability of the study being inconclusive.

As observed by Frewer and others [FMWM16], this framework has successfully become a common language for decision making across the whole company. However, in many studies there are multiple disease markers that may be of interest, and not all of them may indicate a Go decision when taken into account individually. How to best put this information together to assist in deciding whether to carry on with a Phase III trial is an open question. The authors of the paper [FMWM16] suggest different approaches for combining decisions, such as basing the decision on the most important of the endpoints, and using the other endpoints when the first endpoint is inconclusive. However, the probability distribution of a combined decision is not as straightforward as in the single-endpoint case, due to a combination of factors:

(i) The distribution of statistics in the multivariate case is more complicated than in the single variable case.

(ii) The potential decision thresholds and associated regions have more complex shapes compared with the single variable case, where the thresholds are always scalar values and the regions delimited by these thresholds are intervals.

(iii) The potential increase in the number of parameters for adjusting a policy compared to the one variable case.

Evaluating the statistical properties of the combination of multiple Go/Stop decisions may be done either analytically or by simulations. Deniz [Den19] follows the former approach, addressing in detail how to combine up to two or three Go/Stop decisions, and obtaining formulas that bound the resulting False Go and False Stop risks. These calculations do however not deal with policies involving a larger number of endpoints and/or spanning multiple domains. They also rely of having an exact knowledge of the correlations between endpoints, which in practice may need to be estimated from data. The limitations of a purely analytic approach encourage the use of simulations to provide estimates of the probability of success of a given study.

1.3 Contributions

This report focuses on the design of Phase IIb studies in which multiple endpoints are use to make a decision on whether to proceed with an eventual Phase III study. This report presents a framework to formulate and evaluate policies for a given study and estimate the probability a correct decision. The following contributions are made:

(i) An extension of the single-endpoint decision-making framework to yield decisions based on multiple endpoints, together with a number of building blocks which, when combined according to specified rules, produce policies with a monotonicity property. For a given a monotone policy, lower bounds on the probability of a correct decision (respectively upper bounds on the probability of an incorrect decision) can be obtained by simulation.

(ii) A demonstration of how correcting for multiple comparisons can be used to enforce upper bounds on the probabilities of an incorrect decision, and the effect of this correction on the probability of a correct decision.

(iii) A quantitative analysis on how the choice of policy, the number of domains and endpoints in each domain, the power of the endpoints included in a policy, the correlation between these endpoints, and the addition of safety conditions affect the overall probability of correct and incorrect decisions.

(iv) A hypothetical case study in which the endpoints under consideration by a policy are chosen, and the number of patients needed to achieve a sufficiently high probability of a correct decision is obtained.

The proposed framework is implemented as an R package [`gonogo`] with which the analyses outlined above are performed.

1.4 Structure

In Chapter 2 the theoretical underpinnings of the different approaches for combining multiple endpoints into a single decision are introduced. The result is a framework for defining, combining and evaluating decision policies in a multivariate setting.

In Chapter 3 the relative merits of different policies under several scenarios of interest are compared, using and developing the tools introduced in Chapter 2.

In Chapter 4, an early-stage study for a hypothetical heart failure drug is designed. Simulations are used to calculate the number of patients required to achieve the desired probability of producing the right decision under different scenarios of interest. The insights from Chapter 3 are used to interpret the results and modify the policies so as to reduce the number of patients required.

In Chapter 5, the main takeaways from this report are summarized, together with a short exploration of the directions in which the framework can be further extended and developed.

Chapter 2

Background

In this chapter the decision framework proposed by Lalonde [LKH+07] for a study with one endpoint is introduced. The study definition and the policy are then generalized to multiple endpoints. Metrics are defined for assessing the statistical properties and the fitness for purpose of the resulting policies.

2.1 The univariate decision framework

Consider a clinical trial with a single endpoint and two arms: a control arm ($x = 0$) and a treatment arm ($x = 1$). There are N patients in each of the two arms, giving $2N$ patients in total.

For each patient, a single measurement is taken. The measurements in the control arm are normally distributed independent random variables $Y_n^0 \sim N(\mu^0, \sigma^2)$ $n = 1, ..., N$, with μ^0 as the expected measurement. For the active arm, $Y_n^1 \sim N(\mu^1, \sigma^2)$ $n = 1, ..., N$ i.r.v. For simplicity, the measurements are assumed to have a common variance $\sigma^2 \in \mathbb{R}_{>0}$, which is a known constant. The true effect $\mu := \mu^1 - \mu^0 \in \mathbb{R}$ is the endpoint, which can be estimated by $\hat{\mu} = \bar{Y}^1 - \bar{Y}^0$. By construction, $\hat{\mu} \sim N(\mu, \sigma_\mu^2)$, where $\sigma_\mu^2 := \frac{2}{N}\sigma^2$.

The clinical team determines values $\mathrm{TV}, \mathrm{LRV} \in \mathbb{R}$, where:

- TV is the *Target Value* for the variable, which is the desired effect for the drug to be marketable.
- LRV is the *Lower Reference Value* for the variable, which is the smallest clinically-relevant effect.

Without loss of generality, it is assumed that $\mathrm{TV} > \mathrm{LRV}$.

For a given TV and LRV, a univariate decision policy computes a Stop threshold ($\mathrm{thr}^{\mathrm{stop}}$) and a Go threshold ($\mathrm{thr}^{\mathrm{go}}$). The decision from the study is Stop if $\hat{\mu} \leq \mathrm{thr}^{\mathrm{stop}}$ (i.e. $\hat{\mu} \in (-\infty, \mathrm{thr}^{\mathrm{stop}}]$). The decision from the study is Go if $\hat{\mu} \geq \mathrm{thr}^{\mathrm{go}}$ and the decision of the trial is not Stop (i.e. $\hat{\mu} \in [\mathrm{thr}^{\mathrm{go}}, +\infty) \backslash (-\infty, \mathrm{thr}^{\mathrm{stop}}]$). This way, avoiding a False Go decision has precedence over avoiding a False Stop decision. Finally, the outcome where the study leads to neither a Go nor a Stop decision (i.e. $\mathrm{thr}^{\mathrm{stop}} < \hat{\mu} < \mathrm{thr}^{\mathrm{go}}$) is denoted "Discuss"; also known as "amber". These thresholds are illustrated for a synthetic endpoint in Figure 2.1.

The value of thr^{stop} is chosen as the highest threshold such that the risk of a Stop when $\mu \geq \text{TV}$ (i.e. $\mu \in [\text{TV}, +\infty)$) is ≤ 0.1. By continuity, this is equivalent to choosing thr^{stop} so that when $\mu = \text{TV}$ the risk of a Stop is exactly 0.1. Given $\alpha \in (0,1)$, define $z_\alpha = \Phi^{-1}(\alpha)$. That is, for any $Z \sim N(0,1)$, $\mathbb{P}(Z \leq x) = \Phi(x)$; thus $\mathbb{P}(Z \leq z_\alpha) = \alpha$. Note that when $\mu = \text{TV}$, we have $\frac{\hat{\mu} - \text{TV}}{\sigma_\mu} \sim N(0,1)$. Therefore:

$$\begin{gathered}
\mathbb{P}(\hat{\mu} \leq \text{thr}^{\text{stop}} \mid \mu = \text{TV}) = 0.1 \iff \\
\mathbb{P}\left(\frac{\hat{\mu} - \text{TV}}{\sigma_\mu} \leq \frac{\text{thr}^{\text{stop}} - \text{TV}}{\sigma_\mu} \,\middle|\, \mu = \text{TV}\right) = 0.1 \iff \\
\Phi\left(\frac{\text{thr}^{\text{stop}} - \text{TV}}{\sigma_\mu}\right) = 0.1 \iff \\
\frac{\text{thr}^{\text{stop}} - \text{TV}}{\sigma_\mu} = \Phi^{-1}(0.1) = z_{0.1} \iff \\
\text{thr}^{\text{stop}} = \text{TV} + \sigma_\mu z_{0.1}
\end{gathered} \tag{2.1}$$

The threshold thr^{go} is chosen as the lowest threshold so that the risk that $\hat{\mu} \in [\text{thr}^{\text{go}}, +\infty)$ when $\mu \leq \text{LRV}$ (i.e. $\mu \in (-\infty, \text{LRV}]$) is ≤ 0.2. By continuity, this is equivalent to choosing thr^{go} so that when $\mu = \text{LRV}$ the risk that $\hat{\mu} \in [\text{thr}^{\text{go}}, +\infty)$ is exactly 0.2. When $\mu = \text{LRV}$, $\frac{\hat{\mu} - \text{LRV}}{\sigma_\mu} \sim N(0,1)$.

Analogously:

$$\begin{gathered}
\mathbb{P}(\hat{\mu} \geq \text{thr}^{\text{go}} \mid \mu = \text{LRV}) = 0.2 \iff \\
\mathbb{P}\left(\frac{\hat{\mu} - \text{LRV}}{\sigma_\mu} \geq \frac{\text{thr}^{\text{go}} - \text{LRV}}{\sigma_\mu} \,\middle|\, \mu = \text{LRV}\right) = 0.2 \iff \\
1 - \Phi\left(\frac{\text{thr}^{\text{go}} - \text{LRV}}{\sigma_\mu}\right) = 0.2 \iff \\
\frac{\text{thr}^{\text{go}} - \text{LRV}}{\sigma_\mu} = \Phi^{-1}(1 - 0.2) = z_{1-0.2} \iff \\
\text{thr}^{\text{go}} = \text{LRV} + \sigma_\mu z_{1-0.2}
\end{gathered} \tag{2.2}$$

A Go decision requires both $\hat{\mu} \geq \text{thr}^{\text{go}}$ and $\hat{\mu} > \text{thr}^{\text{stop}}$. Therefore, the risk of a Go decision when $\mu = \text{LRV}$ may be less than 0.2.

Estimating σ from the data: The reasoning in this section holds when the variance is a known constant σ^2. Alternatively, the variance can be estimated by pooling the sample variances from the two arms:

$$\begin{aligned}
\hat{\sigma}^2 &:= \frac{(N-1)s^2_{\mathbf{Y}^1} + (N-1)s^2_{\mathbf{Y}^0}}{2N-2} \\
\hat{\sigma}^2_\mu &:= \frac{2}{N}\hat{\sigma}^2
\end{aligned}$$

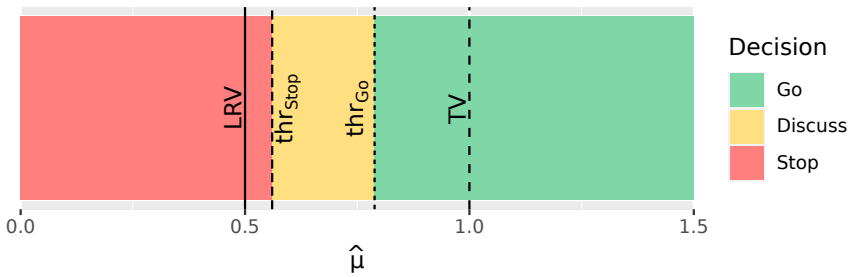


Figure 2.1: Thresholds for the decision framework, with $\text{TV} = 1$, $\text{LRV} = 0.5$, $\sigma^2 = 1$, *and* $N = 17$. *The background indicates the decision for a given value of* $\hat{\mu}$ *(x axis). The variance* σ^2 *is assumed to be a known constant, which means that* $\hat{\mu}$ *is normally distributed. By construction,* $\text{LRV} < \text{thr}^{\text{go}}$ *and* $\text{thr}^{\text{stop}} < \text{TV}$. *The ordering of vertical lines is otherwise contingent.*

If $\hat{\sigma}$ is substituted for σ in the above equations, then the reasoning holds only asymptotically (i.e. as $N \longrightarrow \infty$). Alternatively, if Student's t distribution with $2N-2$ degrees of freedom is used instead of the unit normal distribution, the reasoning holds for all N.

Choice of 0.1 **and** 0.2**:** The values 0.1 and 0.2 follow from the suggested percentiles in [LKH+07]. They represent the probabilities of two undesirable scenarios: 0.1 is the maximum probability of stopping development of a drug whose mean effect does reach the TV; while 0.2 corresponds to the probability of going ahead with a drug whose true effect is below the LRV. These probabilities may be adjusted according to the importance of these risks in the project at hand.

2.2 A multivariate trial model

In the multivariate trial, a vector of measurements is obtained from each patient, with one coordinate for each endpoint.

Notation (Vectors). Vectors are denoted using bold letters: $\mathbf{a}$, $\mathbf{b}$, $\mathbf{TV}$, and their components are denoted using small letters: a_i, b_i, TV_i.

With this notation, a clinical trial with multiple variables is modelled as follows:

- $d = 1, ..., D$ domains.
- $i = 1, ..., V_d$ variables in domain d.
- $V = \sum_{d=1}^{D} V_d$ total variables.
- $x \in \{0, 1\}$: 2 arms (0 for control group, 1 for active group)
- $\boldsymbol{\mu}^x \in \mathbb{R}^V$: the true mean response for patients in arm x.

The quantity of interest is the true mean effect of the drug, which is $\boldsymbol{\mu} := \boldsymbol{\mu}^1 - \boldsymbol{\mu}^0 \in \mathbb{R}^V$. The true effect for each of the V variables is therefore μ_i, with $i = 1, ..., V$. Each μ_i is an endpoint of the trial.

In order to gain information about $\boldsymbol{\mu}$, a study with N patients per arm ($N > V$, $2N$ patients in total) is run. For each arm ($x = 0, 1$) and each patient in that arm ($n = 1, ..., N$), the vector $\mathbf{Y}^{x,n} \in \mathbb{R}^V$ is measured. These measurements are of the form $\mathbf{Y}^{x,n} := \boldsymbol{\mu}^x + \mathbf{r}^{x,n}$, where $\mathbf{r}^{x,n} \sim N(0, \Sigma)$ are normally-distributed independent random vectors. Therefore:

$$\mathbf{Y}^{x,n} \sim N(\boldsymbol{\mu}^x, \Sigma) \text{ i.r.v.}$$

2.2.1 Estimators

The true effect is estimated by the statistic $\hat{\boldsymbol{\mu}}$, which is defined as follows:

$$\begin{aligned} \hat{\boldsymbol{\mu}}^x &:= \frac{1}{N} \sum_{n=1}^{N} \mathbf{y}^{x,n} \\ \hat{\boldsymbol{\mu}} &:= \hat{\boldsymbol{\mu}}^1 - \hat{\boldsymbol{\mu}}^0 \end{aligned}$$

By construction, $\hat{\boldsymbol{\mu}} \sim N(\boldsymbol{\mu}, \Sigma_\mu)$, where the covariance matrix is:

$$\Sigma_\mu := \frac{2}{N} \Sigma$$

When estimating the covariance matrix from data, no covariance structure is assumed. Thus, the estimators $\hat{\Sigma}$ and $\hat{\Sigma}_\mu$ are defined directly from the sample covariance matrix:

$$\begin{aligned} \hat{\mathbf{r}}^{x,n} &:= \mathbf{Y}^{x,n} - \hat{\boldsymbol{\mu}}^x \\ \hat{R}^x &= (\hat{\mathbf{r}}^{x,n})_{n=1}^N \in \mathbb{R}^{N \times V} \\ \hat{R} &= \begin{pmatrix} \hat{R}^0 \\ \hat{R}^1 \end{pmatrix} \in \mathbb{R}^{2N \times V} \\ \hat{\Sigma} &:= \frac{1}{2N-2} (\hat{R}^\top \hat{R}) \in \mathbb{R}^{V \times V} \\ \hat{\Sigma}_\mu &:= \frac{2}{N} \hat{\Sigma} \end{aligned}$$

Remark 2.3 (Unbiasedness, consistency and sufficiency). The estimators $\hat{\boldsymbol{\mu}}$, $\hat{\Sigma}$ and $\hat{\Sigma}_\mu$ are unbiased and consistent for $\boldsymbol{\mu}$, Σ and $\hat{\Sigma}_\mu$, respectively. Furthermore, $\hat{\boldsymbol{\mu}}$ together with either $\hat{\Sigma}$ or $\hat{\Sigma}_\mu$ is a sufficient statistic for the model parameters $\boldsymbol{\mu}$, Σ.

Proof. By construction:

$$\begin{aligned} \hat{\boldsymbol{\mu}} &\sim N(\boldsymbol{\mu}, \hat{\Sigma}_\mu) \\ (2N-2)\hat{\Sigma} &\sim W_V(\Sigma, 2N-2), \end{aligned}$$

where $W_V(\Sigma, 2N-2)$ is the Wishart distribution for V dimensions, $2N-2$ degrees of freedom and scale matrix Σ. The unbiasedness and consistency of

$\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ (and thus of $\hat{\Sigma}_\mu$) follows directly from their distribution. Sufficiency follows by rewriting the likelihood as a function of $\hat{\mu}$ and $\hat{\Sigma}$:

$$\mathcal{L}\left(\mu, \Sigma \mid (\mathbf{Y}^{x,1}, ..., \mathbf{Y}^{x,n})_{x=0,1}\right) = f(\mu, \Sigma \mid \hat{\mu}, \hat{\Sigma}) \quad \text{for some function } f$$

□

The estimators $\hat{\mu}$ and $\hat{\Sigma}_\mu$ are in fact independent. For our purposes, a weaker property is sufficient; namely, that any eventual dependency between the estimators is unaffected by the value of the true effect:

Lemma 2.4 (Shifted location)**.** *Let $\boldsymbol{\Delta}, \boldsymbol{\delta} \in \mathbb{R}^V$. The joint distribution of $(\hat{\boldsymbol{\mu}} + \boldsymbol{\delta}, \hat{\Sigma}_\mu)$ for $\boldsymbol{\mu} = \boldsymbol{\Delta}$ is the same as the joint distribution of $(\hat{\boldsymbol{\mu}}, \hat{\Sigma}_\mu)$ for $\boldsymbol{\mu} = \boldsymbol{\Delta} + \boldsymbol{\delta}$.*

Proof. Let $x \in \{0, 1\}$, $n = 1, ..., N$ and

$$\mathbf{r}_\mu := \frac{1}{N}\sum_{n=1}^{N} \mathbf{r}^{1,n} - \frac{1}{N}\sum_{n=1}^{N} \mathbf{r}^{0,n}$$

Which means:

$$\begin{aligned}
\hat{\boldsymbol{\mu}} &= \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0 \\
&= \frac{1}{N}\sum_{n=1}^{N} \mathbf{Y}^{1,n} - \frac{1}{N}\sum_{n=1}^{N} \mathbf{Y}^{0,n} \\
&= \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{\mu}^1 + \mathbf{r}^{1,n}) - \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{\mu}^0 + \mathbf{r}^{0,n}) \\
&= \boldsymbol{\mu} + \mathbf{r}_\mu
\end{aligned}$$

If $\boldsymbol{\mu} = \boldsymbol{\Delta}$, then $\hat{\boldsymbol{\mu}} + \boldsymbol{\delta} = (\boldsymbol{\Delta} + \mathbf{r}_\mu) + \boldsymbol{\delta}$. Alternatively, if $\boldsymbol{\mu} = \boldsymbol{\Delta} + \boldsymbol{\delta}$, then $\hat{\boldsymbol{\mu}} = (\boldsymbol{\Delta} + \boldsymbol{\delta}) + \mathbf{r}_\mu = (\boldsymbol{\Delta} + \mathbf{r}_\mu) + \boldsymbol{\delta}$.

As for $\hat{\Sigma}_\mu$, this matrix depends only on the $\hat{\mathbf{r}}^{x,n}$. As shown below, these random variables are the same regardless of the true effect $\boldsymbol{\mu}$:

$$\begin{aligned}
\hat{\mathbf{r}}^{x,n} &= \mathbf{Y}^{x,n} - \hat{\boldsymbol{\mu}}^x \\
&= (\boldsymbol{\mu}^x + \mathbf{r}^{x,n}) - (\boldsymbol{\mu}^x + \mathbf{r}^x_\mu) \\
&= \mathbf{r}^{x,n} - \mathbf{r}^x_\mu
\end{aligned}$$

Therefore, the two joint distributions under consideration are the same. □

2.2.2 Reference values

In the multivariate case, the target value and lower reference value are vectors $\mathbf{TV}, \mathbf{LRV} \in \mathbb{R}^V$. These vectors are such that, for each variable $i = 1, ..., V$, the target value of the corresponding endpoint is TV_i and the lower reference value is LRV_i, with the meaning described in §2.1.

In the single variable case it is assumed that $\text{TV} > \text{LRV} > 0$. To generalize this assumption, a partial order on the vectors in $\mathbb{R}^V$ is defined:

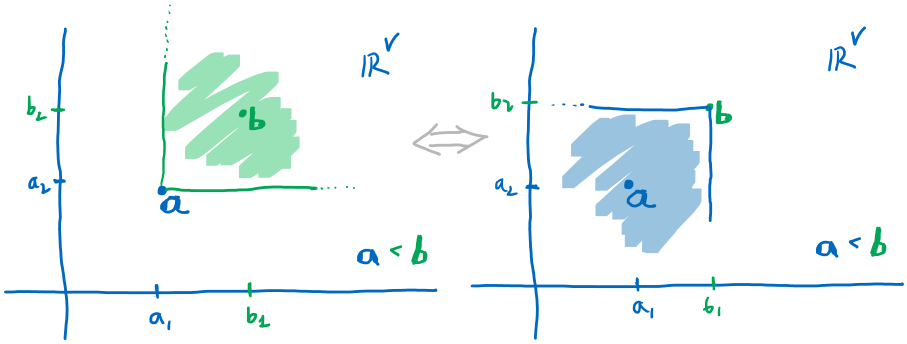


Figure 2.2: Two equivalent graphical characterizations of $\mathbf{a} < \mathbf{b}$

Definition 2.5 (Pointwise partial order on vectors)**.** The pointwise partial order $\leq$ on $\mathbb{R}^V$ is such that, for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^V$:

$$\mathbf{a} \leq \mathbf{b} \Leftrightarrow \forall i = 1, ..., V \ (a_i \leq b_i)$$

Remark. As the ordering in Definition 2.5 is partial, not all vectors can be compared to each other. For instance, if $V = 2$, it is neither the case that $(1, 0) \geq (0, 1)$ nor that $(0, 1) \geq (1, 0)$.

Remark. In the one-dimensional case ($V = 1$) the pointwise partial order reduces to the usual ordering in $\mathbb{R}$.

Notation (Pointwise comparison operators)*.* The order in Definition 2.5 motivates the following comparison operators:

$$\mathbf{a} < \mathbf{b} \Leftrightarrow (\mathbf{a} \leq \mathbf{b}) \wedge (\mathbf{a} \neq \mathbf{b})$$
$$\mathbf{a} \geq \mathbf{b} \Leftrightarrow \mathbf{b} \leq \mathbf{a}$$
$$\mathbf{a} > \mathbf{b} \Leftrightarrow \mathbf{b} < \mathbf{a}$$

For a visual definition of $\mathbf{a} < \mathbf{b}$, see Figure 2.2.

Remark. Note that $\mathbf{a} \geq \mathbf{b} \Leftrightarrow \forall i \in \{1, ..., V\} \, (a_i \geq b_i)$, which is not the same as "not $(\mathbf{a} < \mathbf{b})$". The latter is instead equivalent to $\exists i \in \{1, ..., V\} \, (a_i \geq b_i)$. As a consequence, for $\mathbf{A}, \mathbf{B}$ random vectors in $\mathbb{R}^V$, in general, $\mathbb{P}(\mathbf{A} \geq \mathbf{B}) \neq 1 - \mathbb{P}(\mathbf{A} < \mathbf{B})$.

For the multivariate model, only the case where $\mathbf{TV} > \mathbf{LRV} > \mathbf{0}$ is considered. That is: all else being equal, the more positive an endpoint is, the better.

Remark 2.6 (Negative effects)*.* In some cases, a negative value is desired for an endpoint μ_i, $i \in \{1, ..., V\}$. In other words, a Go decision is warranted when μ_i^1 is sufficiently lower than μ_i^0, and a Stop decision when the opposite is true. This situation can be modelled in our chosen framework by considering the random variables $y_i^{x,n}$ $(x = 0, 1, \ n = 1, ..., N)$ to be the negation of the measured values. The values of TV_i, LRV_i, μ_i and $\hat{\mu}_i$ must be interpreted accordingly.

Example 2.7 (Single-variable decision framework)**.** The single-variable trial in §2.1 is a special case of a multivariate trial where:

$$
\begin{aligned}
D, V, V_1 &:= 1 \\
\boldsymbol{\mu} &:= \mu \in \mathbb{R} \\
\Sigma &:= (\sigma^2) \in \mathbb{R}_{>0} \\
\hat{\Sigma} &:= (\hat{\sigma}^2) \\
\mathbf{TV} &:= \mathrm{TV} \in \mathbb{R} \\
\mathbf{LRV} &:= \mathrm{LRV} \in \mathbb{R}
\end{aligned}
$$

◀

2.3 Upwards and downwards closure of sets

The main goal of this section is to generalize the intervals $(-\infty, \mathrm{LRV}]$, $(-\infty, \mathrm{thr}^{\mathrm{stop}}]$, $[\mathrm{thr}^{\mathrm{go}}, +\infty)$, and $[\mathrm{TV}, +\infty)$ that arise in the univariate decision framework (§2.1) to suitable regions of $\mathbb{R}^V$. As explained in §2.2, the true effect $\boldsymbol{\mu}$ and the realizations of its estimator $\hat{\boldsymbol{\mu}}$ are all vectors in $\mathbb{R}^V$.

Definition 2.8 (Upwards and downwards closed sets)**.** A set $A \subseteq \mathbb{R}^V$ is upwards closed if $\forall \mathbf{a} \in A \, \forall \mathbf{b} \in \mathbb{R}^V \, (\mathbf{a} \leq \mathbf{b} \Rightarrow \mathbf{b} \in A)$. Respectively, a set $A \subseteq \mathbb{R}^V$ is downwards closed if $\forall \mathbf{b} \in A \, \forall \mathbf{a} \in \mathbb{R}^V \, (\mathbf{a} \leq \mathbf{b} \Rightarrow \mathbf{a} \in A)$.

That is, a set is upwards (resp. downwards) closed if, for any point in the set, all larger (resp. smaller) points are also in the set.

Remark 2.9. If $A = \mathbb{R}$ with the usual order $<$, the only upwards-closed sets are the intervals $(a, +\infty)$ and $[a, +\infty)$. The only downwards-closed sets are the intervals $(-\infty, a)$ and $(-\infty, a]$.

Proposition 2.10. *Both upwards and downwards closure are preserved under unions and intersections.*

Proof. Let $A, B \subseteq \mathbb{R}^V$ be upwards closed sets. Let $\mathbf{a} \in A \cup B$ and $\mathbf{b} \in \mathbb{R}^V$, with $\mathbf{a} \leq \mathbf{b}$. If $\mathbf{a} \in A \cup B$, then either $\mathbf{a} \in A$ or $\mathbf{a} \in B$. Without loss of generality, assume the former is true. Because A is upwards closed, $\mathbf{b} \in A$, which implies $\mathbf{b} \in A \cup B$. Therefore, $A \cup B$ is also upwards closed. By a similar reasoning, $A \cap B$ is upwards closed. Also, if A and B are downwards closed then, *mutatis mutandis*, $A \cup B$ and $A \cap B$ are downwards closed. □

Proposition 2.11. *The complement of an upwards closed set is a downwards closed set, and viceversa.*

Proof. Let A be an upwards closed set, and let $A^{\complement} := \mathbb{R}^V \backslash A$ be its complement. Assume $\mathbf{b} \in \mathbb{R}^V \backslash A$, and $\mathbf{a} \in \mathbb{R}^V$ such that $\mathbf{a} \leq \mathbf{b}$. Proof by contradiction; assume $\mathbf{a} \notin \mathbb{R}^V \backslash A$. Therefore $\mathbf{a} \in A$. Because A is upwards closed and $\mathbf{a} \leq \mathbf{b}$, then $\mathbf{b} \in A$. This is a contradiction, so our assumption must be false; i.e. $\mathbf{a} \in \mathbb{R}^V \backslash A$. The converse follows by symmetry. □

In $\mathbb{R}^V$ with the pointwise ordering $\leq$, the intervals in Remark 2.9 can be generalized as follows:

Definition 2.12 (Upwards and downwards cones in $\mathbb{R}^V$)**.** The upwards cone containing $\mathbf{a} \in \mathbb{R}^V$ is the set:

$$\llcorner(\mathbf{a}) := \{\boldsymbol{\Delta} \in \mathbb{R}^V \mid \mathbf{a} \leq \boldsymbol{\Delta}\}$$

The downwards cone containing $\mathbf{a} \in \mathbb{R}^V$ is the set:

$$\neg(\mathbf{a}) := \{\boldsymbol{\Delta} \in \mathbb{R}^V \mid \boldsymbol{\Delta} \leq \mathbf{a}\}$$

Remark. The upwards cone (respectively downwards cone) containing $\mathbf{a}$ is the smallest upwards (respectively downwards) closed set that contains $\mathbf{a}$.

Proposition 2.13 (Alternative characterization of upwards and downwards closure)**.** *The set A is upwards closed if and only if* $\forall \mathbf{a} \in A\ (\llcorner(\mathbf{a}) \subseteq A)$.
The set A is downwards closed if and only if $\forall \mathbf{a} \in A\ (\neg(\mathbf{a}) \subseteq A)$.

The notion of upwards and downwards cones generalizes to arbitrary sets of points:

Notation (Upwards and downwards cones for subsets of $\mathbb{R}^V$). Let $A \subseteq \mathbb{R}^V$.

$$\llcorner(A) := \bigcup_{\mathbf{a} \in A} \llcorner(\mathbf{a})$$
$$\neg(A) := \bigcup_{\mathbf{a} \in A} \neg(\mathbf{a})$$

Remark. A is upwards closed if $A = \llcorner(A)$, and A is downwards closed if $A = \neg(A)$.

Vector slices: When defining a policy, it may be desirable to focus on a specific subset of the endpoints at a time. The following notation is used for this purpose:

Notation (Vector and matrix slicing). Let $\mathbf{a} \subseteq \mathbb{R}^V$ and $I = \{i_1, ..., i_n\}$ with $1 \leq i_1 < ... < i_n \leq V$. Then:

$$\mathbf{a}_I = \begin{pmatrix} a_{i_1} & a_{i_2} & \cdots & a_{i_n} \end{pmatrix} \in \mathbb{R}^n$$

Similarly, for a square matrix $\Sigma = (\Sigma_{i,j})_{i,j=1}^V \subseteq \mathbb{R}^{V \times V}$:

$$\Sigma_{I,I} := \begin{pmatrix} a_{i_1,i_1} & a_{i_1,i_2} & \cdots & a_{i_1,i_n} \\ a_{i_2,i_1} & a_{i_2,i_2} & \cdots & a_{i_2,i_n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i_n,i_1} & a_{i_1,i_2} & \cdots & a_{i_n,i_n} \end{pmatrix} \in \mathbb{R}^{n \times n}$$

Vector slicing preserves the ordering of endpoints:

Remark. For any $\mathbf{a}$ and $\mathbf{b} \in \mathbb{R}^V$ and $I \subseteq \{1, ..., V\}$, if $\mathbf{a} \leq \mathbf{b}$, then $\mathbf{a}_I \leq \mathbf{b}_I$.

By the same token, vector slicing also preserves upwards (or downwards) closure:

Remark. For any $A \subseteq \mathbb{R}^V$, and $I \subseteq \{1, ..., V\}$, if A is upwards (respectively downwards) closed, then $A_I := \{\mathbf{a}_I \mid \mathbf{a} \in A\}$ is also upwards (respectively downwards) closed.

2.4 Policies for multivariate trials

In this section, a general framework to describe decision policies for studies with one or more endpoints is introduced.

Definition 2.14 (Study summary)**.** The space of possible outcomes of a study is denoted by $\mathcal{S}$. Each element $S \in \mathcal{S}$ is a possible summary of the results of a study. Such a summary is of the form $S = (\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c)$, where:

- $\tilde{\boldsymbol{\mu}} \in \mathbb{R}^V$ is an observed mean effect.
- $\tilde{\Sigma}_\mu \in \mathbb{R}^{V \times V}$ is a covariance matrix for the mean effect. The square roots of the elements in the diagonal of this matrix $\tilde{\Sigma}_\mu$ are the standard errors for the endpoints in the study.
- $c \in \mathcal{C}$ is a specification of how the covariance matrix for the mean effect is estimated. In this report only two possibilities are considered, with $\mathcal{C} := \{\text{Theoretical}\} \cup \{\text{Unstructured}_N \mid N \in \mathbb{N}\}$.
  - Theoretical: The covariance matrix is a constant which is assumed to be known before running the study (i.e. $\tilde{\Sigma}_\mu := \Sigma_\mu$).
  - Unstructured_N: The covariance matrix $\tilde{\Sigma}_\mu$ is estimated from the sample covariance matrix of the study data (i.e. $\tilde{\Sigma}_\mu := \hat{\Sigma}_\mu$), with N patients in each study arm.

Notation. A study summary $S = (\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c)$ may be denoted simply as $\tilde{\boldsymbol{\mu}}$, with the remaining components left implicit.

Remark 2.15 (Random study summary). The components $\tilde{\boldsymbol{\mu}}$ or $\tilde{\Sigma}$ of a study summary S are not necessarily random variables. If instantiated to estimators for $\boldsymbol{\mu}$ and Σ_μ (see §2.2.1), the result is a random element in $\mathcal{S}$. For N patients per arm, this gives either $(\hat{\boldsymbol{\mu}}, \Sigma_\mu, \text{Theoretical})$ or $(\hat{\boldsymbol{\mu}}, \hat{\Sigma}_\mu, \text{Unstructured}_N)$.

Definition 2.16 (Policy)**.** A policy is one or more rules which, based on the results of a study, produce one out of $\mathfrak{Z} := \{\text{Go}, \text{Discuss}, \text{Stop}\}$ possible decisions.

Example 2.17 (Policy)**.** "Decide Go if a univariate decision policy yields Go for any of the endpoints and no variable is negatively significant in the wrong direction. Decide Stop if a univariate decision policy yields Stop for all endpoints. Otherwise the outcome is Discuss." ◀

A policy given by such a list of rules can be interpreted as a mathematical object. In all its generality, a policy can be understood as a function which takes in the results of a study, and produces a decision. This decision may depend on the truth values of one or more predicates:

Definition 2.18 (Predicate as truth values)**.** Let $\mathfrak{2} := \{\top, \bot\}$ be the space of truth values, where $\top$ stands for "True" and $\bot$ stands for "False". A predicate P maps a study summary $S = (\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c)$ to a truth value:

$$\begin{aligned} P: \quad \mathcal{S} \quad &\to \mathfrak{2} \\ P(\tilde{\boldsymbol{\mu}}) &\mapsto \top \text{ or } \bot \end{aligned}$$

Remark (Predicates as subsets). If $\tilde{\Sigma}$, N and c are fixed, then the predicate P can be interpreted as a subset of $\mathbb{R}^V$:

$$P := \{\tilde{\boldsymbol{\mu}} \in \mathbb{R}^V \mid P(\tilde{\boldsymbol{\mu}}) = \top\} \subseteq \mathbb{R}^V$$

Remark (Interpretation of a predicate as an event). If $\tilde{\Sigma}_\mu$ and c are fixed (for $\tilde{\Sigma}_\mu$, either as a constant or as a random variable), and $\tilde{\boldsymbol{\mu}}$ is set to a random variable, (e.g. $\tilde{\boldsymbol{\mu}} := \hat{\boldsymbol{\mu}}$), then the predicate P can be interpreted as a random event.

A policy is implemented as a pair of predicates; one for the "go" decision, and one for the "stop" decision:

Definition 2.19 (Policy implementation)**.** A policy implementation is given by a pair of predicates: (P_{stop}, P_{go}). The first predicate determines whether the policy produces a "stop" decision, while the second predicate determines whether the policy produces a "go" decision.

$$\begin{aligned} G : \quad & \mathcal{S} \quad \to \mathcal{P}(\{\text{go}, \text{stop}\}) \\ & (\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) \mapsto \{\text{stop} \mid P_{\text{stop}}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c)\} \cup \{\text{go} \mid P_{\text{go}}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c)\} \end{aligned}$$

The symbol $\mathcal{P}$ denotes the powerset:

$$\mathcal{P}(\{\text{stop}, \text{go}\}) = \{\emptyset, \{\text{stop}\}, \{\text{go}\}, \{\text{stop}, \text{go}\}\}$$

Notation (Policy in case form). The policy in Definition 2.19 may be written as:

$$G(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) = \begin{cases} \text{stop} & P_{\text{stop}}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) \\ \text{go} & P_{\text{go}}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) \end{cases}$$

Notation (Decision membership). Given a policy $G = (P_{\text{stop}}, P_{\text{go}})$:

$$\begin{aligned} G(\tilde{\boldsymbol{\mu}}) \ni \text{stop} &\Leftrightarrow P_{\text{stop}}(\tilde{\boldsymbol{\mu}}) = \top \Leftrightarrow P_{\text{stop}}(\tilde{\boldsymbol{\mu}}) \\ G(\tilde{\boldsymbol{\mu}}) \ni \text{go} \quad &\Leftrightarrow \ P_{\text{go}}(\tilde{\boldsymbol{\mu}}) = \top \ \Leftrightarrow P_{\text{go}}(\tilde{\boldsymbol{\mu}}) \end{aligned}$$

Given a study summary S, a policy implementation may produce one of four possible values: {go}, {stop}, $\emptyset$ and {go, stop}. The final decision is Go if the result is {go}, Stop if the result is {stop}. If both go and stop are produced (i.e. {stop, go}), then the Stop decision prevails. If no decision is produced (i.e. the outcome is $\emptyset$), the result is Discuss, also called "amber":

Definition 2.20 (Simplification of four-valued decisions)**.** Four-valued decisions in $\mathcal{P}(\{\text{go}, \text{stop}\})$ can be interpreted as three-valued decisions in the set $\mathfrak{Z}$:

$$\begin{aligned} [\_] : \mathcal{P}(\{\text{go}, \text{stop}\}) &\to \mathfrak{Z} \\ [\{\text{go}\}] &\mapsto \text{Go} \\ [\{\text{stop}\}] &\mapsto \text{Stop} \\ [\{\text{stop}, \text{go}\}] &\mapsto \text{Stop} \\ [\emptyset] &\mapsto \text{Discuss} \end{aligned}$$

This simplification can be applied to a whole policy:

Definition 2.21 (Simplified policy implementation)**.** Let $G : \mathcal{S} \to \mathcal{P}(\{\text{go}, \text{stop}\})$ be a policy implementation. Given a study summary $S \in \mathcal{S}$, the policy G is simplified by projecting its outcome onto $\mathbb{3}$ (see Definition 2.20).

$$\begin{aligned}[G] : \quad \mathcal{S} \quad &\to \mathbb{3} \\ [G](S) &\mapsto [G(S)]\end{aligned}$$

The univariate decision policy can be implemented as an instance of Definition 2.19, assuming the simplification in Definition 2.21:

Example 2.22 (Univariate decision policy)**.** Let G^{L} be defined as follows:

$$G^{\text{L}}(\tilde{\mu}, \tilde{\sigma}^2_\mu, c) = \begin{cases} \text{stop} & \tilde{\mu} \leq \text{TV} + \tilde{\sigma}_\mu t_{\nu(c),0.1} \\ \text{go} & \tilde{\mu} \geq \text{LRV} + \tilde{\sigma}_\mu t_{\nu(c),1-0.2} \end{cases}$$

The function ν defines the degrees of freedom in the test statistic due to the way that the covariance matrix is estimated. It is defined as follows:

$$\nu(c) = \begin{cases} 2N-2 & c = \text{Unstructured}_N \\ +\infty & c = \text{Theoretical} \end{cases} \tag{2.23}$$

For N patients per arm, the estimator $\hat{\Sigma}_\mu$ has $2N-2$ degrees of freedom (Remark 2.3). Note that if $c = \text{Theoretical}$ (or, regardless of c, asymptotically as $N \longrightarrow +\infty$), $t_{\nu(c),0.1}$ reduces to $z_{0.1}$ and $t_{\nu(c),1-0.2}$ reduces to $z_{1-0.2}$.

The outcome of the univariate decision policy from §2.1 corresponds to $[G^{\text{L}}(\hat{\mu}, \sigma^2_\mu, \text{Theoretical})]$, with $[\_]$ as in Definition 2.21. ◀

Notation (Policy). In the remainder of the text, policy implementations will be referred to simply as "policies".

Notation (Abbreviated application of a policy). If c (and $\tilde{\Sigma}_\mu$) can be inferred from the context, then $G(\tilde{\boldsymbol{\mu}})$ and $G(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu)$ are abbreviations for $G(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c)$. Also, given $I \subseteq \mathbb{R}^V$, if $\tilde{\boldsymbol{\mu}}$, $\tilde{\Sigma}_\mu$ and c can be inferred from the context, then $G(I)$ is an abbreviation for $G(\tilde{\boldsymbol{\mu}}_I, (\tilde{\Sigma}_\mu)_{I,I}, c)$.

Policies based on data: There are several reasons to represent policies and predicates as a function of a study summary (i.e. $P : \mathcal{S} \to \mathcal{P}(\{\text{go}, \text{stop}\})$), instead of directly as a function of the study data (i.e. $G : \mathbb{R}^{2N \times V} \to \mathcal{P}(\{\text{go}, \text{stop}\})$):

- Sufficiency principle: The estimators $\hat{\mu}$ and $\hat{\Sigma}_\mu$ are together sufficient statistics for μ, which is the parameter about which the inferences are to be made. Under the model assumptions, any inference about μ should only depend on the underlying data through said estimators.
- Memory efficiency: Storing and processing a study summary requires fewer computational resources than dealing with the full study data. This is specially relevant when estimating the probabilities of different decisions by means of simulations.

- Legibility: By not depending on data directly, it is possible to understand the policy in terms of an observed effect $\tilde{\mu}$ and a fixed covariance matrix $\tilde{\Sigma}_\mu$. This aids in visualizing the policy and communicating it to the various stakeholders.

Therefore, the policies defined in this report do not depend on the data of study directly, but only through its summary statistics $\hat{\mu}$ and $\hat{\Sigma}$.

2.5 Monotone policies

Among the many possible policies $G : \mathcal{S} \to \mathcal{P}(\{\text{go}, \text{stop}\})$, this report is limited to those sharing a property named monotonicity. This property means that (i) if the policy yields a "go" decision for a given study outcome, then outcomes for which the observed effect is uniformly larger will (all else being equal) also lead to a "go" decision; and, similarly, (ii) if the policy yields a "stop" decision for a given study outcome, outcomes in which the estimated effect is uniformly lower will also lead to a stop decision.

Due to the monotonicity property, simulating the behaviour of the policy for a handful of scenarios (for instance $\boldsymbol{\mu} = \mathbf{TV}$ and $\boldsymbol{\mu} = \mathbf{LRV}$) is sufficient to obtain bounds on decision probabilities on swathes of the space of potential true effects (see Theorem 2.83).

In this section, the monotonicity of a policy is characterized formally.

Definition 2.24 (Upwards and downwards predicates)**.** A predicate $P^{\llcorner}$ on $\mathbb{R}^V$ is an upwards predicate if, for any $\tilde{\boldsymbol{\mu}}_\text{a}$, $\tilde{\boldsymbol{\mu}}_\text{b}$ with $\tilde{\boldsymbol{\mu}}_\text{a} \leq \tilde{\boldsymbol{\mu}}_\text{b}$, $P^{\llcorner}(\tilde{\boldsymbol{\mu}}_\text{a}) \Rightarrow P^{\llcorner}(\tilde{\boldsymbol{\mu}}_\text{b})$. Conversely, $P^{\urcorner}$ on $\mathbb{R}^V$ is a downwards predicate if, for any $\tilde{\boldsymbol{\mu}}_\text{a}$, $\tilde{\boldsymbol{\mu}}_\text{b}$ with $\tilde{\boldsymbol{\mu}}_\text{a} \leq \tilde{\boldsymbol{\mu}}_\text{b}$, $P^{\urcorner}(\tilde{\boldsymbol{\mu}}_\text{b}) \Rightarrow P^{\urcorner}(\tilde{\boldsymbol{\mu}}_\text{a})$.

Notation (Monotone predicate). A predicate which is either upwards or downwards is referred to as monotone.

The monotonicity of a policy follows from the monotonicity of its constituent predicates:

Definition 2.25 (Monotone policy)**.** A policy $G = (P^{\urcorner}_{\text{stop}}, P^{\llcorner}_{\text{go}})$ is monotone if the predicate $P^{\llcorner}_{\text{go}}$ is upwards, and the predicate $P^{\urcorner}_{\text{stop}}$ is downwards.

A trivial example of monotone predicate are the constant predicates:

Example 2.26 (Constant predicates)**.** The constant predicates $\bot$ and $\top$ are both upwards and downwards. ◀

More interesting monotone predicates can be constructed by comparing the result of a monotone function with a fixed threshold:

Proposition 2.27 (Threshold predicates)**.** *Let λ depend only on $\tilde{\Sigma}_\mu$ and c. Let $F(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c)$ be a function $F : \mathcal{S} \to \mathbb{R}$, and let P be a predicate defined as follows:*

$$P_{\lessgtr, F, \lambda}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) := F(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) \lessgtr \lambda(\tilde{\Sigma}_\mu, c)$$

If F is a non-decreasing function on each of $\tilde{\mu}_1,\ldots,\tilde{\mu}_V$, then the predicates $P_{\geq,F,\lambda}$ and $P_{>,F,\lambda}$ are upwards, and the predicates $P_{\leq,F,\lambda}$ and $P_{\leq,F,\lambda}$ are downwards. If F is instead non-increasing on each of $\tilde{\mu}_1,\ldots,\tilde{\mu}_V$, then the predicates $P_{\geq,F,\lambda}$ and $P_{>,F,\lambda}$ are downwards, and the predicates $P_{\leq,F,\lambda}$ and $P_{\leq,F,\lambda}$ are upwards.

Proof. By construction and Definition 2.24. □

An example of a monotone predicate is checking whether an endpoint is negatively significant.

Example 2.28 (Monotonicity of negative significance)**.** Given $\tilde{\boldsymbol{\mu}}$, $\tilde{\Sigma}_\mu$, $c \in \mathcal{C}$, $\alpha \in (0,1)$ and $i \in \{1,...,V\}$. Let $P^{\neg,\text{neg}}$ is defined as follows:

$$P^{\neg,\text{neg}}_{(i),\alpha}(\boldsymbol{\mu}, \tilde{\Sigma}_\mu, c) := \tilde{\mu}_i < t_{\nu(c),\alpha}\sqrt{(\tilde{\Sigma}_\mu)_{i,i}},$$

where ν is defined as in Equation 2.23. Then by Proposition 2.27, $P^{\neg,\text{neg}}_{(i),\alpha}$ is a downwards predicate. The predicate $P^{\neg,\text{neg}}_{(i),\alpha}$ can be read as "endpoint i is negatively significant at level α". ◀

The policies that we are interested in defining in practice are all monotone. For instance:

Proposition 2.29 (Monotonicity of the univariate decision policy)**.** *The policy G^{L} from Example 2.22 is monotone.*

Proof. By construction and Proposition 2.27. □

2.6 Policies with parameters

When building decision policies, it is possible to consider not only a single policy, but also for a family of policies specified by some parameters:

Notation (Parameterized policy). Let $\mathcal{H}$ be a set, where each $\mathbf{h} \in \mathcal{H}$ is a possible combination of parameters. A decision policy with parameters in $\mathcal{H}$ is denoted as $(G_{\mathbf{h}})_{\mathbf{h}\in\mathcal{H}}$.

Example 2.30 (Parameterized single-variable decision policy)**.** Consider $\mathcal{H} = [0,1]^2$, and let $\alpha^\star_{\text{FS}}, \alpha^\star_{\text{FG}} \in [0,1]$ be constants representing the allowed risk of False Stop and False Go, respectively. Let $h := (p_{\text{FS}}, p_{\text{FG}}) \in \mathcal{H}$. Then the single-variable decision policy can be parameterized in $\mathcal{H}$ as $(G^{\text{L}}_{\mathbf{h}})_{\mathbf{h}\in\mathcal{H}}$, with $G^{\text{L}}_{\mathbf{h}}$ is defined as follows:

$$G^{\text{L}}_{\mathbf{h}}(\tilde{\mu}, \tilde{\sigma}^2_\mu, c) = \begin{cases} \text{stop} & \tilde{\mu} \leq \text{thr}^{\text{stop}}_{\alpha^\star_{\text{FS}}} \\ \text{go} & \tilde{\mu} \geq \text{thr}^{\text{go}}_{\alpha^\star_{\text{FG}}} \end{cases},$$

where

$$\begin{aligned} \text{thr}^{\text{stop}}_{\alpha^\star_{\text{FS}}} &:= \text{TV} + \tilde{\sigma}_\mu t_{\nu(c),\alpha^\star_{\text{FS}}} \\ \text{thr}^{\text{go}}_{\alpha^\star_{\text{FG}}} &:= \text{LRV} + \tilde{\sigma}_\mu t_{\nu(c),1-\alpha^\star_{\text{FG}}} \end{aligned}$$

and ν is defined as in Example 2.22. The canonical policy proposed by Lalonde corresponds to $\mathbf{h} = (0.1, 0.2) \in \mathcal{H}$. ◀

Remark 2.31 (False Go and False Stop probabilities). Consider:

(i) $\tilde{\sigma}_\mu^2 := \sigma_\mu^2$ and $c = \text{Theoretical}$; or

(ii) $\tilde{\sigma}_\mu^2 := \hat{\sigma}_\mu^2$ and $c = \text{Unstructured}_N$, where N is the number of patients in each arm.

In both cases:

$$\mathbb{P}(G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\hat{\mu},\tilde{\sigma}_\mu^2,c) \ni \text{stop} \mid \mu = \text{TV}) = \alpha^\star_{\text{FS}}$$
$$\mathbb{P}(G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\hat{\mu},\tilde{\sigma}_\mu^2,c) \ni \text{go} \mid \mu = \text{LRV}) = \alpha^\star_{\text{FG}}$$

Proof. By the definition of $G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}$ and the reasoning in equations (2.1) and (2.2). □

2.7 Policies as decision tables

Monotone policies addressing only a few variables at a time (for instance, just a single variable) can be combined to form larger policies. A first approach is to apply a univariate decision policy individually to each endpoint, and combine the results using a decision table.

Definition 2.32 (Decision table)**.** A decision table for V endpoints is a mapping $T : \mathbb{3}^V \to \mathbb{3}$.

A decision table can be applied to four-valued decisions in $\mathcal{P}(\{\text{go},\text{stop}\})$ by first using Definition 2.20 to project the decisions into $\mathbb{3}$. The decisions given by the table are also elements of $\mathbb{3}$; to obtain a policy, they can be interpreted as elements of $\mathcal{P}(\{\text{go},\text{stop}\})$ through the following embedding:

Definition 2.33 (Three-valued decision as a four-valued decision)**.** A three-valued decision $\omega \in \mathbb{3}$ can be interpreted as a four-valued decision in $\mathcal{P}(\{\text{go},\text{stop}\})$ by taking $\{\text{go} \mid \omega = \text{Go}\} \cup \{\text{stop} \mid \omega = \text{Stop}\}$. This gives $\text{Go} := \{\text{go}\}$, $\text{Stop} := \{\text{stop}\}$, $\text{Discuss} := \emptyset$. In the same way, a function $H : \mathcal{S} \to \mathbb{3}$ can be directly interpreted as a policy $\bar{H} : \mathcal{S} \to \mathcal{P}(\{\text{go},\text{stop}\})$ by using the above interpretation for each of the outputs:

$$\bar{H}(S) := \begin{cases} \text{stop} & H(S) = \text{Stop} \\ \text{go} & H(S) = \text{Go} \end{cases}$$

From now on, $\bar{H}$ is denoted simply as H.

With these correspondences in mind, a decision-table policy for two variables can be defined as follows:

Example 2.34 (Piecewise policy for two variables)**.** Let $\mathcal{H} := [0,1]$, and let $(\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}) \in \mathcal{H}$. Let $V = 2$ (two endpoints) and let $G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}$ be the policy defined in Example 2.30. Consider the following decision table:

$T(\omega_1,\omega_2) :=$

$\omega_1 \setminus \omega_2$	Stop	Discuss	Go
Stop	Stop	Discuss	Go
Discuss	Discuss	Discuss	Go
Go	Go	Go	Go

with which the following policy is defined:

$$G^{\text{any}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\tilde{\boldsymbol{\mu}}) := T([G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}](\tilde{\mu}_1),[G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}](\tilde{\mu}_2))$$
$$= \begin{cases} \text{stop} & T([G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}](\tilde{\mu}_1),[G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}](\tilde{\mu}_2)) = \text{Stop} \\ \text{go} & T([G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}](\tilde{\mu}_1),[G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}](\tilde{\mu}_2)) = \text{Go} \end{cases}$$

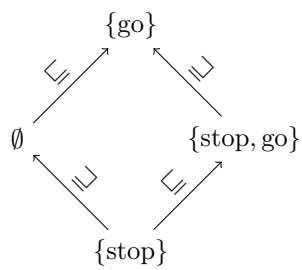
The policy G^{any} will result in a Stop decision if both of the endpoints result in a Stop decision, and will result in a Go decision if either one of the two endpoints results in a Go decision. ◀

A policy defined by a decision table is monotone if the entries in the rows and columns of the table are non-decreasing. To characterize this property formally, orderings for both three-valued and four-valued decisions are defined:

Definition 2.35 (Ordering of three-valued decisions)**.** The set $\mathbb{3}$ can be seen as an ordered set where:

$$\text{Stop} < \text{Discuss} < \text{Go}.$$

Definition 2.36 (Ordering of four-valued decisions)**.** The elements in the set $\mathcal{P}(\{\text{go},\text{stop}\})$ may be ordered as follows:



(Note that this ordering does not coincide with the usual order for sets given by "$\subseteq$").

Remark 2.37 (Order preservation: $(\mathbb{3},\leq) \subset (\mathcal{P}(\{\text{go},\text{stop}\}),\sqsubseteq)$). The ordering for three-valued decisions in $\mathbb{3}$ is consistent with the ordering of four-valued decisions in $\mathcal{P}(\{\text{go},\text{stop}\})$. That is, the simplification in Definition 2.20 and the embedding in Definition 2.33 both preserve the ordering of elements in $\mathcal{P}(\{\text{go},\text{stop}\})$ and $\mathbb{3}$ respectively, meaning that all of the following hold:

$$\{\text{stop},\text{go}\} \sqsubseteq \quad \emptyset \quad \sqsubseteq \{\text{go}\} \tag{2.38}$$
$$\text{Stop} < \text{Discuss} < \text{Go} \tag{2.39}$$
$$\{\text{stop}\} \sqsubseteq \quad \emptyset \quad \sqsubseteq \{\text{go}\} \tag{2.40}$$

where:

(i) The middle row (2.39) is obtained by simplifying either the first (2.38) or the last row (2.40) according to Definition 2.20.

(ii) The last row is obtained by interpreting the middle row as per Definition 2.33.

Definition 2.41 (Monotone decision table)**.** T is a monotone decision table if in the K-dimensional table implied by T, each row, column and higher-dimensional analogue thereof is non-decreasing according to the ordering in Definition 2.35.

Remark. The table in Example 2.34 is monotone.

Monotone tables lead to monotone policies. In order to prove this, an alternative characterization of a monotone policy is introduced:

Proposition 2.42 (Monotone if order preserving)**.** *G is a monotone policy if and only if, for any $\tilde{\boldsymbol{\mu}}_{\mathrm{a}}$ and $\tilde{\boldsymbol{\mu}}_{\mathrm{b}}$ with $\tilde{\boldsymbol{\mu}}_{\mathrm{a}} \leq \tilde{\boldsymbol{\mu}}_{\mathrm{b}}$, $G(\tilde{\boldsymbol{\mu}}_{\mathrm{a}}) \sqsubseteq G(\tilde{\boldsymbol{\mu}}_{\mathrm{b}})$.*

Proof. By Definition 2.36, $G(\tilde{\boldsymbol{\mu}}_{\mathrm{a}}) \sqsubseteq G(\tilde{\boldsymbol{\mu}}_{\mathrm{b}})$ is equivalent to the conjunction of:

(i) $G(\tilde{\boldsymbol{\mu}}_{\mathrm{a}}) \ni \mathrm{stop} \Rightarrow G(\tilde{\boldsymbol{\mu}}_{\mathrm{b}}) \ni \mathrm{stop}$

(ii) and $G(\tilde{\boldsymbol{\mu}}_{\mathrm{a}}) \ni \mathrm{go} \Rightarrow G(\tilde{\boldsymbol{\mu}}_{\mathrm{b}}) \ni \mathrm{go}$

For $G = (P^{\urcorner}_{\mathrm{stop}}, P^{\llcorner}_{\mathrm{go}})$, these conditions are equivalent to the conjunction of:

(i) $P^{\urcorner}_{\mathrm{stop}}(\tilde{\boldsymbol{\mu}}_{\mathrm{a}}) \Rightarrow P^{\urcorner}_{\mathrm{stop}}(\tilde{\boldsymbol{\mu}}_{\mathrm{b}})$

(ii) and $P^{\llcorner}_{\mathrm{go}}(\tilde{\boldsymbol{\mu}}_{\mathrm{a}}) \Rightarrow P^{\llcorner}_{\mathrm{go}}(\tilde{\boldsymbol{\mu}}_{\mathrm{b}})$

The proposition follows by Definition 2.25 and Definition 2.24. □

Corollary 2.43. *If G is monotone and $\tilde{\boldsymbol{\mu}}_{\mathrm{a}} \leq \tilde{\boldsymbol{\mu}}_{\mathrm{b}}$, then $[G](\tilde{\boldsymbol{\mu}}_{\mathrm{a}}) \leq [G](\tilde{\boldsymbol{\mu}}_{\mathrm{b}})$.*

Proof. By Remark 2.37. □

Proposition 2.44 (Monotonicity of decision table-based policies)**.** *Let $T : \mathbb{3}^K \to \mathbb{3}$ be a monotone decision table, and let G_1, ..., G_K be monotone decision policies. Let $I_1, ..., I_K \subseteq \{1, ..., V\}$, and let G^T be defined as follows:*

$$\begin{aligned} G^T : \quad \mathcal{S} \quad &\to \mathbb{3} \\ G^T(\tilde{\boldsymbol{\mu}}) &\mapsto T(G_1(\tilde{\boldsymbol{\mu}}_{I_1}), ..., G_V(\tilde{\boldsymbol{\mu}}_{I_K})) \end{aligned}$$

The policy $\bar{G}^T : \mathcal{S} \to \mathcal{P}(\{\mathrm{go}, \mathrm{stop}\})$ (where the output of the table T is interpreted as in Definition 2.33), also written simply as G^T, is monotone.

Proof. By Proposition 2.42 and Definition 2.41. □

Proposition 2.45. *The policy $G^{\mathrm{any}}_{\alpha^\star_{\mathrm{FS}}, \alpha^\star_{\mathrm{FG}}}$ in Example 2.34 is monotone for any $(\alpha^\star_{\mathrm{FS}}, \alpha^\star_{\mathrm{FG}}) \in [0, 1]^2$.*

Proof. By the monotonicity of $G^{\mathrm{L}}_{\alpha^\star_{\mathrm{FS}}, \alpha^\star_{\mathrm{FG}}}$ and Proposition 2.44. □

2.8 Policies as logical predicates

Defining a policy using the decision tables defined in §2.7 can be cumbersome if there are 3 or more decisions that need to be combined. Per Definition 2.25, a policy consists at its core of a pair of predicates $P^{\llcorner}_{\text{go}}$ and $P^{\urcorner}_{\text{stop}}$. In this section new predicates are defined either in terms of smaller predicates, or in terms of other, simpler policies.

Proposition 2.46 (Composition of monotone predicates)**.** *Let $I \subseteq \{1, ..., V\}$ be a subset of the endpoints in a study, and $S = (\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c)$ be a study summary. Predicates of the following forms are upwards:*

$$\begin{aligned} P^{\llcorner}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) ::= {} & G(\tilde{\boldsymbol{\mu}}_I, (\tilde{\Sigma}_\mu)_{I,I}, c) \ni \text{go} && \text{where } G \text{ is monotone} \\ | \quad & \text{at least } k \text{ of } (P^{\llcorner}_1, ..., P^{\llcorner}_n) && \text{where } k \leq n \\ | \quad & \text{not } P^{\urcorner} \end{aligned}$$

And predicates of the following forms are downwards:

$$\begin{aligned} P^{\urcorner}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) ::= {} & G(\tilde{\boldsymbol{\mu}}_I, (\tilde{\Sigma}_\mu)_{I,I}, c) \ni \text{stop} && \text{where } G \text{ is monotone} \\ | \quad & \text{at least } k \text{ of } (P^{\urcorner}_1, ..., P^{\urcorner}_n) && \text{where } k \leq n \\ | \quad & \text{not } P^{\llcorner} \end{aligned}$$

Proof. By construction and Definition 2.25. □

Definition 2.47 (Predicates on policy outcome)**.** Predicates defined in terms of the simplified outcome of a policy can be be realized as monotone predicates of an appropriate form:

$$G(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) \text{ is Go} \quad := [G](\tilde{\boldsymbol{\mu}}) = \text{Go} \quad \Leftrightarrow G(\tilde{\boldsymbol{\mu}}) \ni \text{go and not } G(\tilde{\boldsymbol{\mu}}) \ni \text{stop} \tag{2.48}$$

$$G(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) \text{ is Stop} := [G](\tilde{\boldsymbol{\mu}}) = \text{Stop} \Leftrightarrow G(\tilde{\boldsymbol{\mu}}) \ni \text{stop} \tag{2.49}$$

For G monotone and by Proposition 2.46, predicate (2.48) is upwards and predicate (2.49) is downwards.

Remark. For a general policy G, the predicate "$[G](\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) = \text{Discuss}$" is neither upwards nor downwards. Therefore, a predicate "$G(\tilde{\mu})$ is Discuss" is not allowed for the purposes of implementing policies. However, some predicates that would make use of " is Discuss" may be reformulated using the syntax in Proposition 2.46. For instance:

$$\text{"}G(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) \text{ is Discuss or } G(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) \text{ is Go"} \mapsto \text{not } (G(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) \text{ is Stop})$$

Notation (Shorthands for common monotone predicates). The following definitions are shorthands for already discussed upwards predicates:

$$\begin{aligned} P^{\llcorner}_1 \text{ or } P^{\llcorner}_2 &:= \text{at least 1 of } (P^{\llcorner}_1, P^{\llcorner}_2) \\ P^{\llcorner}_1 \text{ and } P^{\llcorner}_2 &:= \text{at least 2 of } (P^{\llcorner}_1, P^{\llcorner}_2) \\ \text{any of } (P^{\llcorner}_1, ..., P^{\llcorner}_n) &:= \text{at least 1 of } (P^{\llcorner}_1, ..., P^{\llcorner}_n) \\ \text{all of } (P^{\llcorner}_1, ..., P^{\llcorner}_n) &:= \text{at least } n \text{ of } (P^{\llcorner}_1, ..., P^{\llcorner}_n) \\ \text{at most } k \text{ of } (P^{\urcorner}_1, ..., P^{\urcorner}_n) &:= \text{at least } n - k \text{ of } (\text{not } P^{\urcorner}_1, ..., \text{not } P^{\urcorner}_n) \\ \text{none of } (P^{\urcorner}_1, ..., P^{\urcorner}_n) &:= \text{at least } n \text{ of } (\text{not } P^{\urcorner}_1, ..., \text{not } P^{\urcorner}_n) \end{aligned}$$

Analogous shorthands for downwards predicates are obtained by replacing all instances of an upwards predicate ($P_i^{\llcorner}$) by a downwards predicate ($P_i^{\urcorner}$) and viceversa.

Example 2.50 (Two out of three domains)**.** Consider a study with three domains, where each domain has two endpoints. A policy for this study can be defined as follows:

- A domain is Go if the decision for at least one variable in that domain is Go.
- The overall decision Go if the decision for at least 2 of the domains is Go, and no variables are negatively significant at level $\alpha = 0.05$.
- The overall decision is Stop if none of the decisions for any of the domains is Go.

This policy can be implemented as:

$$G^{\text{2of3}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(S) := \begin{cases} \text{go} & \text{at least 2 of } (G^{\text{any}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\{1,2\}) \text{ is Go}, \\ & \qquad\qquad G^{\text{any}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\{3,4\}) \text{ is Go}, \\ & \qquad\qquad G^{\text{any}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\{5,6\}) \text{ is Go}) \\ & \text{and none of } (P^{\urcorner,\text{neg}}_{(i),\alpha:=0.05}(S)) \\ \text{stop} & \text{none of } (G^{\text{any}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\{1,2\}) \text{ is Go}, \\ & \qquad\qquad G^{\text{any}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\{3,4\}) \text{ is Go}, \\ & \qquad\qquad G^{\text{any}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\{5,6\}) \text{ is Go}) \end{cases}$$

The monotonicity of the policy G^{2of3} follows from Proposition 2.46 and Example 2.28. ◀

2.9 Policies using the joint distribution of endpoints

The policy examples that we have seen so far take into account the endpoints individually, without regard to the correlation between them. In order to take into account the joint distribution of multiple endpoints, the False Stop ($\mu \in [\text{TV}, +\infty)$) and False Go ($\mu \in (-\infty, \text{LRV}]$) intervals need to be generalized to multiple dimensions (i.e. $\mathbb{R}^V$, $V > 1$).

Definition 2.51 (False Stop region)**.** The False Stop region $\Theta^{\text{FS}} \subseteq \mathbb{R}^V$ is the set of true effects for which a Stop decision must be avoided. A region Θ^{FS} is required to be upwards closed.

Definition 2.52 (False Go region)**.** The False Go region $\Theta^{\text{FG}} \subseteq \mathbb{R}^V$ is the set of true effects for which a Go decision must be avoided. A region Θ^{FG} is required to be downwards closed.

Example 2.53 (False Go and False Stop regions for the univariate decision policy)**.** For the univariate decision policy, $\Theta^{\text{FS}} = [\text{TV}, +\infty)$ and $\Theta^{\text{FG}} = (-\infty, \text{LRV}]$. ◀

A defining feature of the decision framework is that $\Theta^{\text{FS}} \cup \Theta^{\text{FG}} \subsetneq \mathbb{R}^V$; that is, there are true effects for which no decision is prescribed as incorrect.

Given Θ^{FS} and Θ^{FG}, the False Stop Risk and False Go Risk of a policy can be obtained:

Definition 2.54 (False Stop Risk)**.** Given a policy G and an associated False Stop region Θ^{FS}, the False Stop Risk (FSR) of G is:

$$\text{FSR}(G) = \sup_{\boldsymbol{\Delta} \in \Theta^{\text{FS}}} \mathbb{P}(G(\hat{\boldsymbol{\mu}}) \ni \text{stop} \mid \boldsymbol{\mu} = \boldsymbol{\Delta})$$

That is, the FSR is an upper bound on the probability that a stop decision will be produced when the true effect is in the False Stop region.

Definition 2.55 (False Go Risk)**.** Given a policy G and an associated False-Go region Θ^{FG}, the False-Go Risk (FGR) of G is:

$$\text{FGR}(G) = \sup_{\boldsymbol{\Delta} \in \Theta^{\text{FG}}} \mathbb{P}(G(\hat{\boldsymbol{\mu}}) \ni \text{go} \mid \boldsymbol{\mu} = \boldsymbol{\Delta})$$

That is, the FGR is an upper bound on the probability that a go decision will be produced when the true effect is in the False Go region.

In the univariate case, the FSR and FGR are determined by the policy parameters:

Lemma 2.56 (False Stop and False Go Risk of the univariate decision policy)**.** *Let Θ^{FS} and Θ^{FG} be defined as in Example 2.53. Then:*

$$\begin{aligned} \text{FSR}\left(G^{\text{L}}_{\alpha^\star_{\text{FS}}, \alpha^\star_{\text{FG}}}\right) &= \alpha^\star_{\text{FS}} \\ \text{FGR}\left(G^{\text{L}}_{\alpha^\star_{\text{FS}}, \alpha^\star_{\text{FG}}}\right) &= \alpha^\star_{\text{FG}} \end{aligned}$$

Proof. By construction. □

An important criterion to determine whether a policy is suitable is its ability to control the FSR and the FGR.

2.9.1 Simple hypothesis testing

In this section an attempt is made to extend the decision framework to a multivariate setting by framing it as classical inference problem. The shortcomings of this approach serve as a motivation for the approaches presented in §2.9.2 and §2.9.3.

A decision can be modelled as a pair of statistical inferences. In order to decide whether to "stop", the following setup is considered:

$$\begin{aligned} H_0^{\text{stop}} &: \boldsymbol{\mu} = \mathbf{TV} \\ H_{\text{alt}}^{\text{stop}} &: \text{not } (\boldsymbol{\mu} \geq \mathbf{TV}) \end{aligned}$$

Assume that the matrix Σ_μ is a known constant. Thus, under H_0^{stop},

$$\hat{\boldsymbol{\mu}} \sim N(\mathbf{TV}, \Sigma_\mu).$$

H_0^{stop} is rejected (thus deciding stop) when the probability of obtaining a value as low as $\hat{\boldsymbol{\mu}}$ given H_0^{stop} is low (i.e. smaller than a chosen constant $\alpha^\star_{\text{FS}} \in (0,1)$). If the rejection region Q^{stop} is defined as follows:

$$Q^{\text{stop}} := \left\{ \tilde{\boldsymbol{\mu}} \in \mathbb{R}^V \,\middle|\, \mathbb{P}\left(\boldsymbol{\Delta} \leq \tilde{\boldsymbol{\mu}} \,\middle|\, \boldsymbol{\Delta} \sim N(\mathbf{TV}, \Sigma_\mu)\right) \leq \alpha^\star_{\text{FS}} \right\}$$

then:

$$\text{Reject } H_0^{\text{stop}} \Leftrightarrow \hat{\boldsymbol{\mu}} \in Q^{\text{stop}} \Leftrightarrow F^{\text{mvn}}(\hat{\boldsymbol{\mu}}; \mathbf{TV}, \Sigma_\mu) \leq \alpha^\star_{\text{FS}}$$

where:

$$F^{\text{mvn}}(\hat{\boldsymbol{\mu}}; \mathbf{TV}, \Sigma_\mu) := \int_{\boldsymbol{\Delta} \leq \hat{\boldsymbol{\mu}}} \pi_{\text{mvn}}(\boldsymbol{\Delta}; \mathbf{TV}, \Sigma_\mu) d\boldsymbol{\Delta}.$$

and π_{mvn} is the probability density function of a multivariate normal. Analogously, in order to decide whether to "go", the following setup is considered:

$$\begin{aligned} H_0^{\text{go}} &: \boldsymbol{\mu} = \mathbf{LRV} \\ H_{\text{alt}}^{\text{go}} &: \text{not } (\boldsymbol{\mu} \leq \mathbf{LRV}) \end{aligned}$$

The null hypothesis H_0^{go} is rejected (thus deciding "go") if the probability of observing a value as high as $\hat{\boldsymbol{\mu}}$ when assuming H_0^{go} is low (i.e. smaller than a chosen constant $\alpha^\star_{\text{FG}} \in (0,1)$). If a rejection region Q^{go} is defined as follows:

$$Q^{\text{go}} := \left\{ \boldsymbol{\Delta} \in \mathbb{R}^V \,\middle|\, \mathbb{P}\left(\hat{\boldsymbol{\mu}} \geq \boldsymbol{\Delta} \,\middle|\, \hat{\boldsymbol{\mu}} \sim N(\mathbf{LRV}, \Sigma_\mu)\right) \leq \alpha^\star_{\text{FG}} \right\}$$

then:

$$\text{Reject } H_0^{\text{go}} \Leftrightarrow \hat{\boldsymbol{\mu}} \in Q^{\text{go}} \Leftrightarrow \bar{F}^{\text{mvn}}(\hat{\boldsymbol{\mu}}; \mathbf{LRV}, \Sigma_\mu) \leq \alpha^\star_{\text{FG}}$$

where:

$$\bar{F}^{\text{mvn}}(\hat{\boldsymbol{\mu}}; \mathbf{LRV}, \Sigma_\mu) := \int_{\boldsymbol{\Delta} \geq \hat{\boldsymbol{\mu}}} \pi_{\text{mvn}}(\boldsymbol{\Delta}; \mathbf{LRV}, \Sigma_\mu) d\boldsymbol{\Delta}.$$

Using these two hypothesis tests, a policy can be defined as follows:

$$G^{\text{S}}(\tilde{\boldsymbol{\mu}}) := \begin{cases} \text{stop} & F^{\text{mvn}}(\tilde{\boldsymbol{\mu}}; \mathbf{TV}, \Sigma_\mu) \leq \alpha^\star_{\text{FS}} \\ \text{go} & \bar{F}^{\text{mvn}}(\tilde{\boldsymbol{\mu}}; \mathbf{LRV}, \Sigma_\mu) \leq \alpha^\star_{\text{FG}} \end{cases} \tag{2.57}$$

When estimating the covariance matrix from the data, one may replace π_{mvn} by the density of a multivariate generalization of Student's t distribution. The null hypotheses H_0^{stop} and H_0^{go} suggest the following False Stop and False Go regions:

$$\Theta^{\text{FS}}_{\text{simple}} := \{\boldsymbol{\Delta} \in \mathbb{R}^V \mid \boldsymbol{\Delta} \geq \mathbf{TV}\} \tag{2.58}$$

$$\Theta^{\text{FG}}_{\text{simple}} := \{\boldsymbol{\Delta} \in \mathbb{R}^V \mid \boldsymbol{\Delta} \leq \mathbf{LRV}\} \tag{2.59}$$

Remark. In (2.58), the False Stop region is defined to contain those true effects where *all* endpoints are above the target value. In §2.9.2 (Example 2.66) and after, the $\Theta^{\text{FS}}_{\text{simple}}$ region is expanded to contain all true effects where *at least one* of the endpoints reaches the target value.

Note that the cumulative distribution functions F^{mvn} and $\bar{F}^{\text{mvn}}$ are monotonically increasing (respectively decreasing) on their first argument. Therefore, the FSR and the FGR thus correspond to the maximum Type I errors of the given inferences. Remembering that $\hat{\boldsymbol{\mu}} \sim N(\boldsymbol{\mu}, \Sigma_\mu)$, then:

$$\begin{aligned}
\text{FSR}(G^{\text{S}}) &= \sup_{\boldsymbol{\Delta} \in \Theta^{\text{FS}}_{\text{simple}}} \mathbb{P}(G^{\text{S}}(\hat{\boldsymbol{\mu}}) \ni \text{stop} \mid \boldsymbol{\mu} = \boldsymbol{\Delta}) \\
&= \sup_{\boldsymbol{\Delta} \in \Theta^{\text{FS}}_{\text{simple}}} \mathbb{P}(F^{\text{mvn}}(\hat{\boldsymbol{\mu}}; \mathbf{TV}, \Sigma_\mu) \leq \alpha^\star_{\text{FS}} \mid \boldsymbol{\mu} = \boldsymbol{\Delta}) \\
&= \mathbb{P}(F^{\text{mvn}}(\hat{\boldsymbol{\mu}}; \mathbf{TV}, \Sigma_\mu) \leq \alpha^\star_{\text{FS}} \mid \boldsymbol{\mu} = \mathbf{TV}) \\
&= \mathbb{P}(F^{\text{mvn}}(\hat{\boldsymbol{\mu}}; \mathbf{TV}, \Sigma_\mu) \leq \alpha^\star_{\text{FS}} \mid H_0^{\text{stop}})
\end{aligned}$$

and:

$$\begin{aligned}
\text{FGR}(G^{\text{S}}) &= \sup_{\boldsymbol{\Delta} \in \Theta^{\text{FG}}_{\text{simple}}} \mathbb{P}(G^{\text{S}}(\hat{\boldsymbol{\mu}}) \ni \text{go} \mid \boldsymbol{\mu} = \boldsymbol{\Delta}) \\
&= \sup_{\boldsymbol{\Delta} \in \Theta^{\text{FG}}_{\text{simple}}} \mathbb{P}(\bar{F}^{\text{mvn}}(\hat{\boldsymbol{\mu}}; \mathbf{LRV}, \Sigma_\mu) \leq \alpha^\star_{\text{FG}} \mid \boldsymbol{\mu} = \boldsymbol{\Delta}) \\
&= \mathbb{P}(\bar{F}^{\text{mvn}}(\hat{\boldsymbol{\mu}}; \mathbf{LRV}, \Sigma_\mu) \leq \alpha^\star_{\text{FG}} \mid \boldsymbol{\mu} = \mathbf{LRV}) \\
&= \mathbb{P}(\bar{F}^{\text{mvn}}(\hat{\boldsymbol{\mu}}; \mathbf{LRV}, \Sigma_\mu) \leq \alpha^\star_{\text{FG}} \mid H_0^{\text{go}})
\end{aligned}$$

Unless $V = 1$, the distribution of $F^{\text{mvn}}(\hat{\boldsymbol{\mu}}; \mathbf{TV}, \Sigma_\mu)$ under H_0^{Stop} and the distribution of $F^{\text{mvn}}(\hat{\boldsymbol{\mu}}; \mathbf{LRV}, \Sigma_\mu)$ under H_0^{go} are not necessarily uniform. This means that the FSR and FGR, which correspond to the Type I error rates for the given inference rules, are not directly determined by $\alpha^\star_{\text{FS}}$ and $\alpha^\star_{\text{FG}}$, but instead need to be computed on a case-by-case basis. The silver lining is that, by not being reliant on statistical inference theory to bound the Type I error, it is possible to make decision criteria that are not based on a punctual null hypothesis, but directly on regions Θ^{FS} and Θ^{FG}: this is discussed in more detail in §2.9.2. Alternatively, the hypothesis tests can be designed based on confidence sets in such a way that (non-tight) bounds on FSR and FGR can be derived formally. This approach is discussed in more detail in §2.9.3.

2.9.2 Probability measures

Consider the probability measure M defined as follows:

$$M(A; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu) = \int_A \pi_{\text{mvn}}(\boldsymbol{\Delta}; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu) d\boldsymbol{\Delta} \tag{2.60}$$

where π_{mvn} is the probability density of the multivariate normal distribution.

Example 2.61. Let Θ^{FS} and Θ^{FG} be defined as in Example 2.53 (i.e. $\Theta^{\text{FG}} := (-\infty, \text{LRV}]$ and $\Theta^{\text{FS}} := [\text{TV}, +\infty)$) The univariate decision policy can be written as:

$$G^{\text{L}}_{\alpha^\star_{\text{FS}}, \alpha^\star_{\text{FG}}}(\tilde{\mu}) := \begin{cases} \text{stop} & M(\Theta^{\text{FS}}; \tilde{\mu}, \sigma_\mu^2) \leq \alpha^\star_{\text{FS}} \\ \text{go} & M(\Theta^{\text{FG}}; \tilde{\mu}, \sigma_\mu^2) \leq \alpha^\star_{\text{FG}} \end{cases}$$

◀

Example 2.62. Let Θ^{FS} and Θ^{FG} be defined as in (2.58) and (2.59), respectively. The policy G^{S} in §2.9.1 can be written as:

$$G^{\mathrm{S}}_{\alpha^\star_{\mathrm{FS}},\alpha^\star_{\mathrm{FG}}}(\tilde{\boldsymbol{\mu}}) := \begin{cases} \text{stop} & M(\Theta^{\mathrm{FS}};\tilde{\boldsymbol{\mu}},\Sigma_\mu) \leq \alpha^\star_{\mathrm{FS}} \\ \text{go} & M(\Theta^{\mathrm{FG}};\tilde{\boldsymbol{\mu}},\Sigma_\mu) \leq \alpha^\star_{\mathrm{FG}} \end{cases}$$

◀

These policies are both monotone:

Lemma 2.63. *If Θ is an upwards (downwards) closed set, then $M(\Theta;\tilde{\boldsymbol{\mu}})$ is a non-decreasing (resp. non-increasing) function on $\boldsymbol{\mu}$.*

Proof. Let Θ be upwards closed. Consider $\tilde{\boldsymbol{\mu}}_{\mathrm{a}}$, $\tilde{\boldsymbol{\mu}}_{\mathrm{b}}$ such that $\tilde{\boldsymbol{\mu}}_{\mathrm{a}} \leq \tilde{\boldsymbol{\mu}}_{\mathrm{b}}$. Then:

$$\begin{aligned} M(\Theta;\tilde{\boldsymbol{\mu}}_{\mathrm{b}}) &= \int_\Theta \pi_{\mathrm{mvn}}(\boldsymbol{\Delta};\tilde{\boldsymbol{\mu}}_{\mathrm{b}},\tilde{\Sigma}_\mu)d\boldsymbol{\Delta} \\ &= \int_\Theta \pi_{\mathrm{mvn}}(\boldsymbol{\Delta}-(\tilde{\boldsymbol{\mu}}_{\mathrm{b}}-\tilde{\boldsymbol{\mu}}_{\mathrm{a}});\tilde{\boldsymbol{\mu}}_{\mathrm{a}},\tilde{\Sigma}_\mu)d\boldsymbol{\Delta} \\ &= \int_{\Theta-(\tilde{\boldsymbol{\mu}}_{\mathrm{b}}-\tilde{\boldsymbol{\mu}}_{\mathrm{a}})} \pi_{\mathrm{mvn}}(\boldsymbol{\Delta}\tilde{\boldsymbol{\mu}}_{\mathrm{a}},\tilde{\Sigma}_\mu)d\boldsymbol{\Delta} \\ &\geq \int_\Theta \pi_{\mathrm{mvn}}(\boldsymbol{\Delta};\tilde{\boldsymbol{\mu}}_{\mathrm{a}},\tilde{\Sigma}_\mu)d\boldsymbol{\Delta} \qquad\qquad \Theta \subseteq (\Theta-(\tilde{\boldsymbol{\mu}}_{\mathrm{b}}-\tilde{\boldsymbol{\mu}}_{\mathrm{a}})) \\ &= M(\Theta;\tilde{\boldsymbol{\mu}}_{\mathrm{a}}) \end{aligned}$$

The property for a downwards closed Θ follows analogously. □

Proposition 2.64 (Monotonicity of predicates on measures)**.** *Let $\Theta \subseteq \mathbb{R}^V$, and $p \in [0,1]$. If Θ is upwards closed, then $M(\Theta;\tilde{\boldsymbol{\mu}}) \leq p$ is a downwards predicate on $\tilde{\boldsymbol{\mu}}$. Conversely, if Θ is downwards closed, then $M(\Theta;\tilde{\boldsymbol{\mu}}) \leq p$ is an upwards predicate on $\tilde{\boldsymbol{\mu}}$.*

Proof. By Lemma 2.63 and Proposition 2.27. □

Corollary 2.65. *The policy G^{S} is monotone.*

Notation (Abbreviated measure application)*.* When applying a measure M to a set Θ (i.e. $M(\Theta;\tilde{\boldsymbol{\mu}},\tilde{\Sigma}_\mu)$), the parameters $\tilde{\boldsymbol{\mu}}$ and/or $\tilde{\Sigma}_\mu$ may be omitted if they are clear from the context.

Proposition 2.64 opens up the possibility of designing new policies by redefining the regions Θ^{FS} and Θ^{FG}:

Example 2.66 (False Stop and False Go regions for either one of two endpoints)**.**

$$\begin{aligned} \Theta^{\mathrm{FS}} &:= ([\mathrm{TV}_1,+\infty)\times\mathbb{R}) \cup (\mathbb{R}\times[\mathrm{TV}_2,+\infty)) \\ \Theta^{\mathrm{FG}} &:= (-\infty,\mathrm{LRV}_1]\times(-\infty,\mathrm{LRV}_2] \end{aligned}$$

That is, Stop is avoided when the true effect for either of the endpoints is above the TV, and Go is avoided when the true effect for either of the endpoints is below the LRV. ◀

Example 2.67 (Either variable, joint distribution)**.** Consider the following description of a policy for a case with one domain and two variables:

- Stop if all variables are Stop (jointly).
- Go if at least one of the variables is Go (jointly).

Let $\alpha^\star_{\text{FS}} := 0.1$, $\alpha^\star_{\text{FG}} := 0.2$, $\alpha := 0.05$, and consider Θ^{FS} and Θ^{FG} defined as in Example 2.66, and M be defined as in (2.60). Then the description above can be implemented as the following function:

$$G^{\text{J}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\tilde{\boldsymbol{\mu}}) = \begin{cases} \text{stop} & M(\Theta^{\text{FS}}; \tilde{\boldsymbol{\mu}}) \leq \alpha^\star_{\text{FS}} \\ \text{go} & M(\Theta^{\text{FG}}; \tilde{\boldsymbol{\mu}}) \leq \alpha^\star_{\text{FG}} \end{cases}$$

A "stop" decision is reached if $M(\Theta^{\text{FS}})$ is small enough, while a "go" decision is reached if $M(\Theta^{\text{FG}})$ is small enough. The parameters $\alpha^\star_{\text{FS}}$ and $\alpha^\star_{\text{FG}}$ may be empirically adjusted in order to obtain the desired FSR and FGR. ◀

Remark 2.68. This approach is an intuitive generalization of the single-variable decision policy. On the one hand, t the probability measure $M(\cdot; \tilde{\boldsymbol{\mu}})$ when $\tilde{\boldsymbol{\mu}} := \hat{\boldsymbol{\mu}}$ amounts to a confidence distribution, which is not well grounded theoretically [Cox06, page 66]. On the other hand, probability measures open up the possibility of implementing Bayesian approaches within the current framework; with the theoretical development in this chapter guaranteeing that the resulting policies are well-behaved (i.e. monotone).

2.9.3 Hotelling's T^2 statistic

In this section an alternative way of building a policy using the joint distribution for multiple endpoints is proposed. Loose theoretical bounds for the FSR and FGR of the resulting policies can be derived as shown in Lemma 2.76.

Given $\tilde{\boldsymbol{\mu}}$, $\Delta \in \mathbb{R}^V$ and $\tilde{\Sigma}_\mu \in \mathbb{R}^{V\times V}$ positive definite, define the following function:

$$T^2[\boldsymbol{\Delta}, \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu] := (\tilde{\boldsymbol{\mu}} - \boldsymbol{\Delta})^\top \left(\tilde{\Sigma}_\mu\right)^{-1} (\tilde{\boldsymbol{\mu}} - \boldsymbol{\Delta})$$

Then $T^2[\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}, \hat{\Sigma}_\mu]$ is Hotelling's T^2 statistic [JW14, §5.2].

Proposition 2.69. *Let $Q(\alpha; c)$ be defined so that:*

$$Q(\alpha; c) := \begin{cases} \frac{V(2N-2)}{2N-V-1} F_{V,2N-V-1,1-\alpha} & c = \text{Unstructured}_N \\ \chi^2_{V,1-\alpha} & c = \text{Theoretical} \end{cases} \tag{2.70}$$

Then:

$$\mathbb{P}(T^2[\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}, \Sigma_\mu] \geq Q(\alpha; \text{Theoretical})) = \alpha$$
$$\mathbb{P}(T^2[\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}, \hat{\Sigma}_\mu] \geq Q(\alpha; \text{Unstructured}_N)) = \alpha$$

Proof. Consider the case where $c = \text{Theoretical}$.

$$\begin{aligned}
&T^2[\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}, \Sigma_\mu] \\
&= (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top \Sigma_\mu^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \\
&= (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top \left((\Sigma_\mu^{1/2})^\top \Sigma_\mu^{1/2} \right)^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) && \Sigma_\mu \text{ pos. def.} \\
&= (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top \Sigma_\mu^{-1/2} (\Sigma_\mu^{-1/2})^\top (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \\
&= \mathbf{z}^\top \mathbf{z} && \mathbf{z} := (\Sigma_\mu^{-1/2})^\top (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \\
&\sim \chi^2_V && z \sim N(\mathbf{0}, I_V)
\end{aligned}$$

For the case $c = \text{Unstructured}_N$, the distribution corresponds to that of a two-sample Hotelling's T^2 statistic, for which a more elaborate argument is required (see [Mui82, §3.2] for a starting point).

□

Definition 2.71 (Policy based on Hotelling's T^2 statistic)**.** Consider $\Theta^{\mathrm{FS}}, \Theta^{\mathrm{FG}} \subseteq \mathbb{R}^V$, and $\alpha^\star_{\mathrm{FS}}, \alpha^\star_{\mathrm{FG}} \in (0, 1)$. The policy G^{Hot} for $\Theta^{\mathrm{FS}}, \Theta^{\mathrm{FG}}$ is defined as follows:

$$G^{\mathrm{Hot}}_{\alpha^\star_{\mathrm{FS}}, \alpha^\star_{\mathrm{FG}}}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) = \begin{cases} \text{stop} & \inf_{\boldsymbol{\Delta} \in \Theta^{\mathrm{FS}}} T^2[\boldsymbol{\Delta}, \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu] \geq Q(\alpha^\star_{\mathrm{FS}}; c) \\ \text{go} & \inf_{\boldsymbol{\Delta} \in \Theta^{\mathrm{FG}}} T^2[\boldsymbol{\Delta}, \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu] \geq Q(\alpha^\star_{\mathrm{FG}}; c) \end{cases}$$

As long as Θ^{FS} and Θ^{FG} are delimited by linear inequalities (for instance, those defined in Example 2.66), computing $\inf_{\boldsymbol{\Delta} \in \Theta^{\mathrm{FS}}} T^2$ and $\inf_{\boldsymbol{\Delta} \in \Theta^{\mathrm{FS}}} T^2$ admits a straightforward reduction to a quadratic programming problem [GI83], for which a solver implementation exists as an R package [`quadprog`].

Remark 2.72 (Hotelling's T^2 policy from hypothesis tests). The policy G^{Hot} is equivalent to performing two statistical inferences. The first pair of hypotheses is:

$$\begin{aligned}
H_0^{\text{stop}} &: \boldsymbol{\Delta} \in \Theta^{\mathrm{FS}} \\
H_{\text{alt}}^{\text{stop}} &: \boldsymbol{\Delta} \notin \Theta^{\mathrm{FS}},
\end{aligned}$$

where rejecting H_0^{stop} at the $\alpha^\star_{\mathrm{FS}}$ level using T^2 results in a "stop" decision. And the second pair of hypotheses is:

$$\begin{aligned}
H_0^{\text{go}} &: \boldsymbol{\Delta} \in \Theta^{\mathrm{FG}} \\
H_{\text{alt}}^{\text{go}} &: \boldsymbol{\Delta} \notin \Theta^{\mathrm{FG}},
\end{aligned}$$

where rejecting H_0^{go} at the $\alpha^\star_{\mathrm{FG}}$ level using the T^2 statistic results in a "go" decision.

Alternatively, the policy G^{Hot} in Definition 2.71 can be defined in terms of the overlap between the False Stop/False Go regions and a confidence set for $\boldsymbol{\mu}$:

Remark 2.73 (Intepretation based on a confidence set). Given $\tilde{\boldsymbol{\mu}}$ and $\tilde{\Sigma}_\mu$, consider the set:

$$\tilde{S}_\alpha(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) := \{\boldsymbol{\Delta} \in \mathbb{R}^V \mid T^2[\boldsymbol{\Delta}, \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu] \leq Q(\alpha; c)\}$$

In particular, $\tilde{S}_\alpha(\hat{\boldsymbol{\mu}}, \hat{\Sigma}_\mu)$ is a $(1-\alpha)$-confidence set (or confidence ellipse) for $\boldsymbol{\mu}$. Then, the policy G^{Hot} can be written as:

$$G^{\text{Hot}}_{\alpha^\star_{\text{FS}}, \alpha^\star_{\text{FG}}}(\tilde{\boldsymbol{\mu}}) = \begin{cases} \text{stop} & \tilde{S}_{\alpha^\star_{\text{FS}}}(\tilde{\boldsymbol{\mu}}) \cap \Theta^{\text{FS}} = \emptyset \\ \text{go} & \tilde{S}_{\alpha^\star_{\text{FG}}}(\tilde{\boldsymbol{\mu}}) \cap \Theta^{\text{FG}} = \emptyset \end{cases}$$

In other words, a "stop" decision is taken if none of the points in Θ^{FS} fall in the $(1-\alpha^\star_{\text{FS}})$-confidence set, and otherwise a "go" decision is taken if none of the points in Θ^{FG} fall inside the $(1-\alpha^\star_{\text{FG}})$-confidence set.

The characterization in Remark 2.73 can be used to prove the monotonicity of the policies based on the T^2 statistic:

Proposition 2.74 (Monotonicity of Hotelling's T^2 policy)**.** *Let Θ^{FS}, $\Theta^{\text{FG}} \subseteq \mathbb{R}^V$ be upwards (respectively downwards) closed. Then G^{Hot} is monotone.*

Proof. Note that:

$$\tilde{S}_\alpha(\tilde{\boldsymbol{\mu}} + \boldsymbol{\delta}) = \tilde{S}_\alpha(\tilde{\boldsymbol{\mu}}) + \boldsymbol{\delta}$$

By Proposition 2.11, the sets $(\Theta^{\text{FS}})^{\complement}$ and $(\Theta^{\text{FG}})^{\complement}$ are downwards and upwards closed, respectively. Therefore, the stop predicate $\tilde{S}_{\alpha^\star_{\text{FS}}}(\tilde{\boldsymbol{\mu}}) \cap \Theta^{\text{FS}} = \emptyset \Leftrightarrow \tilde{S}_{\alpha^\star_{\text{FS}}}(\tilde{\boldsymbol{\mu}}) \subseteq (\Theta^{\text{FS}})^{\complement}$ is downwards, and the go predicate $\tilde{S}_{\alpha^\star_{\text{FG}}}(\tilde{\boldsymbol{\mu}}) \cap \Theta^{\text{FG}} = \emptyset \Leftrightarrow \tilde{S}_{\alpha^\star_{\text{FG}}}(\tilde{\boldsymbol{\mu}}) \subseteq (\Theta^{\text{FG}})^{\complement}$ is upwards. By Definition 2.25, G^{Hot} is a monotone policy. □

For computational purposes, it is useful to phrase G^{Hot} in terms of threshold predicates:

Remark 2.75. Let T^2 be generalized to a set Θ as follows:

$$T^2[\Theta; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu] := \inf_{\boldsymbol{\Delta} \in \Theta} (\boldsymbol{\Delta} - \tilde{\boldsymbol{\mu}})^\top \Sigma_\mu^{-1} (\boldsymbol{\Delta} - \tilde{\boldsymbol{\mu}})$$

By a similar reasoning as in Proposition 2.74:

(i) If Θ is downwards closed, then $T^2[\Theta; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu]$ is a non-increasing function of $\tilde{\boldsymbol{\mu}}$.

(ii) If Θ is upwards closed, then $T^2[\Theta; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu]$ is a non-decreasing function of $\tilde{\boldsymbol{\mu}}$.

Then G^{Hot} can be rewritten as:

$$G^{\text{Hot}}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) = \begin{cases} \text{stop} & Q^{-1}(T^2[\Theta^{\text{FS}}; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu]; c) \leq \alpha^\star_{\text{FS}} \\ \text{go} & Q^{-1}(T^2[\Theta^{\text{FG}}; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu]; c) \leq \alpha^\star_{\text{FG}} \end{cases}$$

As defined in Proposition 2.69, $Q(\alpha; c)$ (and thus $Q^{-1}(\alpha; c)$) are monotonically decreasing functions of α. Therefore, the predicate $Q^{-1}(T^2[\Theta^{\text{FS}}; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}], c) \leq \alpha^\star_{\text{FS}}$ is downwards, and the predicate $Q^{-1}(T^2[\Theta^{\text{FG}}; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}], c) \leq \alpha^\star_{\text{FG}}$ is upwards. This implies that the policy G^{Hot} is monotone.

Unlike the situation with policies defined following the approach in §2.9.2, the T^2 statistic approach yields the following bounds for the FSR and FGR:

Lemma 2.76 (Bounds on FSR and FGR for Hotelling's T^2 policy)**.**

$$\begin{aligned}\mathrm{FSR}(G^{\mathrm{Hot}}) &\leq \alpha^\star_{\mathrm{FS}}\\ \mathrm{FGR}(G^{\mathrm{Hot}}) &\leq \alpha^\star_{\mathrm{FG}}\end{aligned}$$

Proof.

$$\begin{aligned}
\mathrm{FSR}(G^{\mathrm{Hot}}) &= \sup_{\boldsymbol{\Delta}\in\Theta^{\mathrm{FS}}} \mathbb{P}(G^{\mathrm{Hot}}(\hat{\boldsymbol{\mu}}) \ni \mathrm{stop} \mid \mu = \boldsymbol{\Delta}) && \text{by Definition 2.54}\\
&= \sup_{\boldsymbol{\Delta}\in\Theta^{\mathrm{FS}}} \mathbb{P}(\inf_{\tilde{\boldsymbol{\Delta}}\in\Theta^{\mathrm{FS}}} T^2[\tilde{\boldsymbol{\Delta}}, \hat{\boldsymbol{\mu}}] \geq Q(\alpha^\star_{\mathrm{FS}}) \mid \mu = \boldsymbol{\Delta}) && \text{by Definition 2.71}\\
&\leq \sup_{\boldsymbol{\Delta}\in\Theta^{\mathrm{FS}}} \mathbb{P}(T^2[\boldsymbol{\Delta}, \hat{\boldsymbol{\mu}}] \geq Q(\alpha^\star_{\mathrm{FS}}) \mid \mu = \boldsymbol{\Delta}) && \boldsymbol{\Delta} \in \Theta^{\mathrm{FS}}\\
&= \sup_{\boldsymbol{\Delta}\in\Theta^{\mathrm{FS}}} \alpha^\star_{\mathrm{FS}} && \text{by Proposition 2.69}\\
&= \alpha^\star_{\mathrm{FS}}
\end{aligned}$$

Analogously:

$$\begin{aligned}
\mathrm{FGR}(G^{\mathrm{Hot}}) &= \sup_{\boldsymbol{\Delta}\in\Theta^{\mathrm{FG}}} \mathbb{P}(G^{\mathrm{Hot}}(\hat{\boldsymbol{\mu}}) \ni \mathrm{go} \mid \mu = \boldsymbol{\Delta}) && \text{by Definition 2.55}\\
&= \sup_{\boldsymbol{\Delta}\in\Theta^{\mathrm{FG}}} \mathbb{P}(\inf_{\tilde{\boldsymbol{\Delta}}\in\Theta^{\mathrm{FG}}} T^2[\tilde{\boldsymbol{\Delta}}, \hat{\boldsymbol{\mu}}] \geq Q(\alpha^\star_{\mathrm{FG}}) \mid \mu = \boldsymbol{\Delta}) && \text{by Definition 2.71}\\
&\leq \sup_{\boldsymbol{\Delta}\in\Theta^{\mathrm{FG}}} \mathbb{P}(T^2[\boldsymbol{\Delta}, \hat{\boldsymbol{\mu}}] \geq Q(\alpha^\star_{\mathrm{FG}}) \mid \mu = \boldsymbol{\Delta}) && \boldsymbol{\Delta} \in \Theta^{\mathrm{FG}}\\
&= \sup_{\boldsymbol{\Delta}\in\Theta^{\mathrm{FG}}} \alpha^\star_{\mathrm{FG}} && \text{by Proposition 2.69}\\
&= \alpha^\star_{\mathrm{FG}}
\end{aligned}$$

□

In this approach, the FSR and FGR are bounded by $\alpha^\star_{\mathrm{FS}}$ and $\alpha^\star_{\mathrm{FG}}$. However, these bounds are not necessarily tight. In the $V = 1$ case:

Proposition 2.77. *If* $V = 1$*, then* $\mathrm{FSR} = \frac{1}{2}\alpha^\star_{\mathrm{FS}}$ *and* $\mathrm{FGR} = \frac{1}{2}\alpha^\star_{\mathrm{FG}}$ *(assuming* $\alpha^\star_{\mathrm{FS}}, \alpha^\star_{\mathrm{FG}} \in (0, 1)$*).*

Proof. Note that $\Delta \in \Theta^{\mathrm{FS}} \Leftrightarrow \Delta \geq \mathrm{TV}$. Also, for any random variable X, knowing $X^2 \geq \kappa$ for some $\kappa \in \mathbb{R}$ does not give any information about whether $X \leq 0$: the two events are independent.

Consider the case $c =$ Theoretical:

$$\begin{aligned}
\mathrm{FSR}(G^{\mathrm{Hot}}) &= \sup_{\Delta\in\Theta^{\mathrm{FS}}} \mathbb{P}(G^{\mathrm{Hot}}(\hat{\mu}) \ni \mathrm{stop} \mid \mu = \Delta)\\
&= \sup_{\Delta\in\Theta^{\mathrm{FS}}} \mathbb{P}(G^{\mathrm{Hot}}(\Delta + r_\mu) \ni \mathrm{stop})\\
&= \mathbb{P}(G^{\mathrm{Hot}}(\mathrm{TV} + r_\mu) \ni \mathrm{stop})\\
&= \mathbb{P}(r_\mu \geq 0)\mathbb{P}(G^{\mathrm{Hot}}(\mathrm{TV} + r_\mu) \ni \mathrm{stop} \mid r_\mu > 0) +\\
&\qquad \mathbb{P}(r_\mu < 0)\mathbb{P}(G^{\mathrm{Hot}}(\mathrm{TV} + r_\mu) \ni \mathrm{stop} \mid r_\mu \leq 0)\\
&= \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot \mathbb{P}(G^{\mathrm{Hot}}(\mathrm{TV} + r_\mu) \ni \mathrm{stop} \mid r_\mu \leq 0)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\mathbb{P}\left(\inf_{\tilde{\Delta}\geq \mathrm{TV}} T^2[\tilde{\Delta}; \mathrm{TV} + r_\mu] \geq \chi_{1,1-\alpha^\star_{\mathrm{FS}}} \,\middle|\, r_\mu \leq 0\right) \\
&= \frac{1}{2}\mathbb{P}(T^2[\mathrm{TV}; \mathrm{TV} + r_\mu] \geq \chi_{1,1-\alpha^\star_{\mathrm{FS}}} \mid r_\mu \leq 0) \\
&= \frac{1}{2}\mathbb{P}\left(\frac{r_\mu^2}{\sigma_\mu^2} \geq \chi_{1,1-\alpha^\star_{\mathrm{FS}}} \,\middle|\, r_\mu \leq 0\right) \\
&= \frac{1}{2}\mathbb{P}\left(\frac{r_\mu^2}{\sigma_\mu^2} \geq \chi_{1,1-\alpha^\star_{\mathrm{FS}}} \,\middle|\, \frac{r_\mu}{\sigma_\mu} \leq 0\right) \\
&= \frac{1}{2}\mathbb{P}\left(\frac{r_\mu^2}{\sigma_\mu^2} \geq \chi_{1,1-\alpha^\star_{\mathrm{FS}}}\right) \\
&= \tfrac{1}{2}\alpha^\star_{\mathrm{FS}}
\end{aligned}
$$

The proof for FGR is analogous. □

2.10 Evaluating policies

In §2.9 the False Go Risk and the False Stop Risk of a policy are defined. These measure the failure rate of the policy, and are based on the corresponding regions Θ^{FG} and Θ^{FS}. In the univariate case, the decision policy is determined by two parameters, which fix the FSR and FGR. Thus, the desired FSR and FGR determine the behaviour of the policy.

In the multivariate case, there is ample room for variation in the policy even when a bound on these two quantities is given. To evaluate the relative merit of different policies, metrics are needed that assess whether the policy can produce correct decisions, not just avoid incorrect ones. A policy is evaluated on specifically chosen subsets of $\mathbb{R}^V$, which we call scenarios.

Definition 2.78 (Scenario)**.** A scenario on $\mathbb{R}^V$ is a pair of regions $(R^{\mathrm{Go}}, R^{\mathrm{Stop}})$, where:

- The region $R^{\mathrm{Go}} \subseteq \mathbb{R}^V$ is upwards closed, and denotes those points for which a Go decision is desired.
- The region $R^{\mathrm{Stop}} \subseteq \mathbb{R}^V$ is downwards closed, and denotes those points for which a Stop is desired.

Given a scenario and a policy, the following metrics can be defined:

Definition 2.79 (Outcome rates of a policy)**.** Let G be a policy, and let $S = (\hat{\boldsymbol{\mu}}, \hat{\Sigma}_\mu, \mathrm{Unstructured}_N)$ or $S = (\hat{\boldsymbol{\mu}}, \Sigma_\mu, \mathrm{Theoretical})$ be a random element of $\mathcal{S}$ representing the summary of a study with true effect $\boldsymbol{\mu}$. The summary is abbreviated as $(\hat{\boldsymbol{\mu}})$, where the number of patients per arm N and the covariance structure c are considered to be fixed, and the covariance matrix is either a constant (for $c = \mathrm{Theoretical}$) or the appropriate estimator (for $c = \mathrm{Unstructured}_N$). Let $\boldsymbol{\Delta} \in \mathbb{R}^V$. The Go rate and Stop rate of the policy G when the true effect is $\boldsymbol{\Delta}$ are defined as follows:

$$
\begin{aligned}
\mathrm{rGo}(\boldsymbol{\Delta}) &:= \mathbb{P}(G(\hat{\boldsymbol{\mu}}) \text{ is Go} \mid \boldsymbol{\mu} = \boldsymbol{\Delta}) \\
\mathrm{rStop}(\boldsymbol{\Delta}) &:= \mathbb{P}(G(\hat{\boldsymbol{\mu}}) \text{ is Stop} \mid \boldsymbol{\mu} = \boldsymbol{\Delta})
\end{aligned}
$$

Given regions R^{Go} and R^{Stop}, the False Go Rate, False Stop Rate, Correct Go Rate and Correct Stop Rate of the policy G are respectively defined as follows:

$$\begin{aligned}
\text{FGr} &:= \sup_{\boldsymbol{\Delta} \in R^{\text{Stop}}} \text{rGo}(\boldsymbol{\Delta}) \\
\text{FSr} &:= \sup_{\boldsymbol{\Delta} \in R^{\text{Go}}} \text{rStop}(\boldsymbol{\Delta}) \\
\text{CSr} &:= \inf_{\boldsymbol{\Delta} \in R^{\text{Stop}}} \text{rStop}(\boldsymbol{\Delta}) \\
\text{CGr} &:= \inf_{\boldsymbol{\Delta} \in R^{\text{Go}}} \text{rGo}(\boldsymbol{\Delta})
\end{aligned}$$

All else being equal, higher values for the CGr and CSr, and lower values for the FGr and FSr are signs of a better performing policy. The regions R^{Go} and R^{Stop} are to be chosen depending on the outcomes that are considered to be clinically desirable (or undesirable) by the domain experts. Note that the more points that R^{Go} and R^{Stop} contain, the more pessimistic the policy evaluation will be. That is, the FGr and FSr will be higher, and the CGr and CSr will be lower.

Remark 2.80. The False Stop Rate (FSr) and the False Stop Risk (FSR, Definition 2.54) coincide when $R^{\text{Go}} = \Theta^{\text{FS}}$. Similarly, when $R^{\text{Stop}} = \Theta^{\text{FG}}$, the False Go Rate (FGr) is less than or equal than the False Go Risk (FGR, Definition 2.55).

2.10.1 Necessity of metrics

All four of the metrics in Definition 2.79 are necessary in the following sense:

Remark 2.81 (Necessity of metrics). If any of the four metrics in Definition 2.79 is disregarded, it is possible to improve trivially improve another one of the metrics by making the policy less informative. Let G be a policy, and $(P^{\llcorner}_{\text{go}}, P^{\urcorner}_{\text{stop}}) := G$.

- If FSr is disregarded, the CSr metric can be trivially improved without affecting any of the others by using the policy G', where:

  $$G' := (P^{\llcorner}_{\text{go}}, P^{\urcorner}_{\text{stop}} \text{ or not } P^{\llcorner}_{\text{go}}).$$

  This is equivalent to making all "Discuss" decisions of $[G]$ be "Stop" decisions for $[G']$.

- If FGr is disregarded, the CGr metric can be trivially improved by defining a new policy G':

  $$G' := (P^{\llcorner}_{\text{go}} \text{ or not } P^{\urcorner}_{\text{stop}}, P^{\urcorner}_{\text{stop}})$$

  This is equivalent to making all "Discuss" decisions of $[G]$ be "Go" decisions for $[G']$.

- If CGr is disregarded, the FGr metric can be trivially improved by:

  $$G' := (\bot, P^{\urcorner}_{\text{stop}})$$

  This is equivalent to making all "Go" decisions of $[G]$ be "Discuss" decisions for $[G']$.

- If CSr is disregarded, the FSr metric can be trivially improved by:

$$G' := (P^{\llcorner}_{\text{go}} \text{ and not } P^{\urcorner}_{\text{stop}}, \perp)$$

This is equivalent to making all "Stop" decisions of $[G]$ be "Discuss" decisions for $[G']$.

2.10.2 Generalizing metrics from individual true effects

In the single variable case, the evaluation metrics can be computed exactly as follows:

Example 2.82 (Metrics for the univariate decision policy)**.** Consider the scenario $R^{\text{Stop}} := (-\infty, \text{LRV}]$, $R^{\text{Go}} := [\text{TV}, +\infty)$, and the policy $G^{\text{L}}_{\alpha^\star_{\text{FS}}, \alpha^\star_{\text{FG}}}$ as defined in Example 2.30, with c and $\tilde{\sigma}_\mu$ as in case (i) of Remark 2.31.
The CSr is calculated as follows:

$$\begin{aligned}
\text{CSr} &= \inf_{\Delta \leq \text{LRV}} \mathbb{P}(G^{\text{L}}_{\alpha^\star_{\text{FS}}, \alpha^\star_{\text{FG}}}(\hat{\mu}) \text{ is Stop} \mid \mu = \Delta) \\
&= \inf_{\Delta \leq \text{LRV}} \mathbb{P}(G^{\text{L}}_{\alpha^\star_{\text{FS}}, \alpha^\star_{\text{FG}}}(\hat{\mu}) \ni \text{stop} \mid \mu = \Delta) \\
&= \inf_{\Delta \leq \text{LRV}} \mathbb{P}(\hat{\mu} \leq \text{thr}^{\text{stop}} \mid \mu = \Delta) \\
&= \inf_{\Delta \leq \text{LRV}} \Phi\left(\frac{\text{thr}^{\text{stop}} - \Delta}{\sigma_\mu}\right) \\
&= \inf_{\Delta \leq \text{LRV}} \Phi\left(\frac{\text{TV} - \Delta}{\sigma_\mu} + z_{\alpha^\star_{\text{FS}}}\right) \\
&= \Phi\left(\frac{\text{TV} - \text{LRV}}{\sigma_\mu} + z_{\alpha^\star_{\text{FS}}}\right)
\end{aligned}$$

For the FSr, by the same reasoning as in Remark 2.31:

$$\text{FSr} = \alpha^\star_{\text{FS}}$$

For the CGr, assuming $\text{thr}^{\text{go}} > \text{thr}^{\text{stop}}$:

$$\begin{aligned}
\text{CGr} &= \inf_{\Delta \geq \text{TV}} \mathbb{P}(G^{\text{L}}_{\alpha^\star_{\text{FS}}, \alpha^\star_{\text{FG}}}(\hat{\mu}) \text{ is Go} \mid \mu = \Delta) \\
&= \inf_{\Delta \geq \text{TV}} \mathbb{P}(G^{\text{L}}_{\alpha^\star_{\text{FS}}, \alpha^\star_{\text{FG}}}(\hat{\mu}) \ni \text{go and not } G^{\text{L}}_{\alpha^\star_{\text{FS}}, \alpha^\star_{\text{FG}}}(\hat{\mu}) \ni \text{stop} \mid \mu = \Delta) \\
&= \inf_{\Delta \geq \text{TV}} \mathbb{P}(\hat{\mu} \geq \text{thr}^{\text{go}} \text{ and not } \hat{\mu} \leq \text{thr}^{\text{stop}} \mid \mu = \Delta) \\
&= \inf_{\Delta \geq \text{TV}} \mathbb{P}(\hat{\mu} \geq \text{thr}^{\text{go}} \mid \mu = \Delta) \\
&= \inf_{\Delta \geq \text{TV}} \Phi\left(\frac{\Delta - \text{thr}^{\text{go}}}{\sigma_\mu}\right) \\
&= \Phi\left(\frac{\text{TV} - \text{thr}^{\text{go}}}{\sigma_\mu}\right)
\end{aligned}$$

$$= \Phi\left(\frac{\mathrm{TV}-\mathrm{LRV}}{\sigma_\mu} - z_{1-\alpha^\star_{\mathrm{FG}}}\right)$$
$$= \Phi\left(\frac{\mathrm{TV}-\mathrm{LRV}}{\sigma_\mu} + z_{\alpha^\star_{\mathrm{FG}}}\right)$$

And, under the same assumption ($\mathrm{thr}^{\mathrm{go}} > \mathrm{thr}^{\mathrm{stop}}$), and by the same reasoning as Remark 2.31:

$$\mathrm{FGr} = \alpha^\star_{\mathrm{FG}}$$

By contrast, if the assumption $\mathrm{thr}^{\mathrm{go}} > \mathrm{thr}^{\mathrm{stop}}$ does *not* hold, then the last two equations become inequalities:

$$\mathrm{CGr} \leq \Phi\left(\frac{\mathrm{TV}-\mathrm{LRV}}{\sigma_\mu} + z_{\alpha^\star_{\mathrm{FG}}}\right)$$
$$\mathrm{FGr} \leq \alpha^\star_{\mathrm{FG}}$$

◀

In other cases, the computation of the metrics is not straightforward, as the metrics in §2.10 are defined by quantifying over regions R^{Go} and R^{Stop} with uncountably many points. However, if a policy is monotone and the regions R^{Go}, R^{Stop} can be expressed as a finite union of upwards (respectively downwards) cones, then the metrics in §2.10 can be approximated by simulating trials for a finite number of $\boldsymbol{\Delta}_1, ..., \boldsymbol{\Delta}_K \in \mathbb{R}^V$:

Theorem 2.83. *Let G be a monotone policy for V endpoints, and let $\tilde{R}^{\mathrm{Go}}$ and $\tilde{R}^{\mathrm{Stop}} \subseteq \mathbb{R}^V$ be finite sets such that:*

$$R^{\mathrm{Go}} = \bigcup_{\boldsymbol{\Delta}\in\tilde{R}^{\mathrm{Go}}} \llcorner(\boldsymbol{\Delta})$$
$$R^{\mathrm{Stop}} = \bigcup_{\boldsymbol{\Delta}\in\tilde{R}^{\mathrm{Stop}}} \urcorner(\boldsymbol{\Delta})$$

Then:

$$\mathrm{FGr} = \max_{\boldsymbol{\Delta}\in\tilde{R}^{\mathrm{Stop}}} \mathrm{rGo}(\boldsymbol{\Delta}) \tag{2.84}$$
$$\mathrm{FSr} = \max_{\boldsymbol{\Delta}\in\tilde{R}^{\mathrm{Go}}} \mathrm{rStop}(\boldsymbol{\Delta}) \tag{2.85}$$
$$\mathrm{CGr} = \min_{\boldsymbol{\Delta}\in\tilde{R}^{\mathrm{Go}}} \mathrm{rGo}(\boldsymbol{\Delta}) \tag{2.86}$$
$$\mathrm{CSr} = \min_{\boldsymbol{\Delta}\in\tilde{R}^{\mathrm{Stop}}} \mathrm{rStop}(\boldsymbol{\Delta}) \tag{2.87}$$

Proof. For every $\boldsymbol{\Delta} \in R^{\mathrm{Stop}}$, let $\tilde{\boldsymbol{\Delta}} \in \tilde{R}^{\mathrm{Stop}} \subset R^{\mathrm{Stop}}$ denote an element such that $\tilde{\boldsymbol{\Delta}} \geq \Delta$, i.e. $(\boldsymbol{\Delta} - \tilde{\boldsymbol{\Delta}}) \leq 0$. Let $(P^{\urcorner}_{\mathrm{stop}}, P^{\llcorner}_{\mathrm{go}}) := G$. Then:

$$\mathrm{FGr} = \sup_{\boldsymbol{\Delta}\in R^{\mathrm{Stop}}} \mathrm{FGr}(\boldsymbol{\Delta})$$

$$
\begin{aligned}
&= \sup_{\boldsymbol{\Delta} \in R^{\mathrm{Stop}}} \mathbb{P}(G(\hat{\boldsymbol{\mu}}, \hat{\Sigma}_\mu) \text{ is Go} \mid \boldsymbol{\mu} = \boldsymbol{\Delta}) \\
&\quad \textit{by Lemma 2.4:} \\
&= \sup_{\boldsymbol{\Delta} \in R^{\mathrm{Stop}}} \mathbb{P}(G(\hat{\boldsymbol{\mu}} + (\boldsymbol{\Delta} - \tilde{\boldsymbol{\Delta}}), \hat{\Sigma}_\mu) \text{ is Go} \mid \boldsymbol{\mu} = \tilde{\boldsymbol{\Delta}}) \\
&\quad \textit{“}G(\hat{\boldsymbol{\mu}}, \hat{\Sigma}_\mu) \textit{ is Go” is an upwards predicate:} \\
&\leq \sup_{\boldsymbol{\Delta} \in R^{\mathrm{Stop}}} \mathbb{P}(G(\hat{\boldsymbol{\mu}}, \hat{\Sigma}_\mu) \text{ is Go} \mid \boldsymbol{\mu} = \tilde{\boldsymbol{\Delta}}) \\
&\quad \{\tilde{\boldsymbol{\Delta}} \mid \boldsymbol{\Delta} \in R^{\mathrm{Stop}}\} = \tilde{R}^{\mathrm{Stop}}: \\
&= \sup_{\tilde{\boldsymbol{\Delta}} \in \tilde{R}^{\mathrm{Stop}}} \mathbb{P}(G(\hat{\boldsymbol{\mu}}, \hat{\Sigma}_\mu) \text{ is Go} \mid \boldsymbol{\mu} = \tilde{\boldsymbol{\Delta}}) \\
&\quad \tilde{R}^{\mathrm{Stop}} \textit{ is finite:} \\
&= \max_{\tilde{\boldsymbol{\Delta}} \in \tilde{R}^{\mathrm{Stop}}} \mathbb{P}(G(\hat{\boldsymbol{\mu}}, \hat{\Sigma}_\mu) \text{ is Go} \mid \boldsymbol{\mu} = \tilde{\boldsymbol{\Delta}}) \\
&= \max_{\boldsymbol{\Delta} \in \tilde{R}^{\mathrm{Stop}}} \mathrm{rGo}(\boldsymbol{\Delta})
\end{aligned}
$$

This shows that $\mathrm{FGr} \leq \max_{\boldsymbol{\Delta} \in \tilde{R}^{\mathrm{Stop}}} \mathrm{FGr}(\boldsymbol{\Delta})$. Because $\tilde{R}^{\mathrm{Stop}} \subseteq R^{\mathrm{Stop}}$, $\mathrm{FGr} \geq \max_{\boldsymbol{\Delta} \in \tilde{R}^{\mathrm{Stop}}} \mathrm{FGr}(\boldsymbol{\Delta})$. Therefore, Equation 2.84 holds. Equations 2.85, 2.86 and 2.87 follow analogously. □

Example 2.88. If $R^{\mathrm{Go}} := \{\boldsymbol{\Delta} \in \mathbb{R}^2 \mid \boldsymbol{\Delta} \geq \mathbf{0} \text{ and } (\Delta_1 \geq \mathrm{TV}_1 \text{ or } \Delta_2 \geq \mathrm{TV}_2)\}$ and $R^{\mathrm{Stop}} := \{\boldsymbol{\Delta} \in \mathbb{R}^2 \mid \boldsymbol{\Delta} \leq \mathbf{LRV}\}$, then in order to obtain the overall FGr, CSr, CGr and FSr it suffices to calculate each desired metric at $(0, \mathrm{TV}_2)$, $(\mathrm{TV}_1, 0)$ and $(\mathrm{LRV}_1, \mathrm{LRV}_2)$. ◀

Remark 2.80 suggests that the FSR and FGR can also be calculated from a finite number of simulations:

Corollary 2.89 (Computation of False Go and False Stop Risks)**.** *Let $G \equiv (P^{\neg}_{\mathrm{stop}}, P^{\llcorner}_{\mathrm{go}})$ be a monotone policy for V endpoints, and let $\tilde{\Theta}^{\mathrm{FG}}$ and $\tilde{\Theta}^{\mathrm{FS}} \subseteq \mathbb{R}^V$ be finite sets such that:*

$$
\begin{aligned}
\Theta^{\mathrm{FG}} &= \bigcup_{\boldsymbol{\Delta} \in \tilde{\Theta}^{\mathrm{FG}}} \llcorner(\boldsymbol{\Delta}) \\
\Theta^{\mathrm{FS}} &= \bigcup_{\boldsymbol{\Delta} \in \tilde{\Theta}^{\mathrm{FS}}} \neg(\boldsymbol{\Delta})
\end{aligned}
$$

Then:

$$
\begin{aligned}
\mathrm{FGR} &= \max_{\Delta \in \tilde{\Theta}^{\mathrm{FG}}} \mathbb{P}(G(\hat{\boldsymbol{\mu}}) \ni \mathrm{go} \mid \boldsymbol{\mu} = \boldsymbol{\Delta}) \\
\mathrm{FSR} &= \max_{\Delta \in \tilde{\Theta}^{\mathrm{FS}}} \mathbb{P}(G(\hat{\boldsymbol{\mu}}) \ni \mathrm{stop} \mid \boldsymbol{\mu} = \boldsymbol{\Delta})
\end{aligned}
$$

Proof. By the same reasoning as in Theorem 2.83, noting that $G(\hat{\boldsymbol{\mu}}) \ni \mathrm{go} \equiv P^{\llcorner}_{\mathrm{go}}$ is an upwards predicate and $G(\hat{\boldsymbol{\mu}}) \ni \mathrm{stop} \equiv P^{\neg}_{\mathrm{stop}}$ is a downwards predicate. □

2.11 Summary

In order to define a decision policy for a clinical trial, the first step is to determine which endpoints to include in the decision and how to group them into domains. Endpoints which are related, and thus expected to be highly correlated and/or respond similarly to the treatment are assigned to the same domain.

Once the endpoints are chosen, decision policies can be defined either for individual endpoints, or for a small group of endpoints (e.g. all endpoints within the same domain). Policies that do not consider the totality of the endpoints included in the study, but only a subset, may be called *subpolicies*.

A subpolicy that considers an endpoint individually can be defined from the usual decision thresholds (§2.1). A subpolicy that takes into account the covariance structure across multiple endpoints is constructed by first defining regions Θ^{FS} and Θ^{FG} (which generalize the TV and LRV from the univariate framework) and then using either a probability measure (§2.9.2) or Hotelling's T^2 statistic (§2.9.3).

Subpolicies can be combined by using a decision table (§2.7) or by means of logical predicates (§2.8), in order to obtain an overall policy for a study.

Policies can be compared by computing the metrics FSr, FGr, CGr and CSr on a handful of scenarios of interest ($\tilde{R}^{\mathrm{Go}}$,$\tilde{R}^{\mathrm{Stop}}$). The monotonicity of the policies in combination with Theorem 2.83 can be used to generalize these results to entire subsets of $\mathbb{R}^V$ (i.e. R^{Go} and R^{Stop}).

Chapter 3

Policy evaluation

In this chapter a series of simulation-based experiments is run under a range of syncthetic scenarios of interest, with the goal of determining the statistical properties of the policies defined within the framework presented in Chapter 2. The results are used in the case study (Chapter 4) in order to inform which policies to use and which endpoints to consider.

3.1 Scope

The framework presented in Chapter 2 can be applied to a wide variety of situations. In order to obtain actionable insights, the form of the policies and the scope of the analyzed scenarios need to be constrained.

True effects: When evaluating a policy, we focus our attention on cases where, for each endpoint μ_i $(i = 1, ..., V)$, the true effect is in the set $\{0, \text{LRV}_i, \text{TV}_i\} \ni \mu_i$.

True covariance matrix: The true covariance matrix is structured as follows:

$$\Sigma := \begin{pmatrix} \Sigma_{1,1} & \cdots & \Sigma_{1,d_2} & \cdots & \Sigma_{1,D} \\ \vdots & & \vdots & & \vdots \\ \Sigma_{d_1,1} & \cdots & \Sigma_{d_1,d_2} & \cdots & \Sigma_{d_1,D} \\ \vdots & & \vdots & & \vdots \\ \Sigma_{D,1} & \cdots & \Sigma_{D,d_2} & \cdots & \Sigma_{D,D} \end{pmatrix}$$

$$\Sigma_{d_1,d_2} \in \mathbb{R}^{V_{d_1} \times V_{d_2}} \qquad d_1, d_2 \in \{1, ..., D\}$$

$$\Sigma_{d,d} := \begin{pmatrix} \sigma^2_{d,i} & \text{if } i = j \\ \rho\sigma_{d,i}\sigma_{d,j} & \text{if } i \neq j \end{pmatrix}^{V_d}_{i,j=1} \qquad d_1 = d_2 = d$$

$$\Sigma_{d_1,d_2} := \left(\tau\sigma_{d_1,i}\sigma_{d_2,j}\right)^{i=V_{d_1}, j=V_{d_2}}_{i,j=1} \qquad d_1 \neq d_2$$

Each matrix block corresponds to a pair of domains d_1 and d_2. In particular, the blocks along the diagonal are the covariance matrices for variables in the same domain. The constant $\sigma^2_{d,i} > 0$ is the variance of a measurement for the ith variable in domain d (i.e. $\text{Var}(Y^{x,n}_{d[i]})$, for an arm x and patient n). The constant $\rho \in [0, 1)$ is the correlation between variables in the same domain, and the constant $\tau \in [0, 1)$ is the correlation between variables in different domains. Unless otherwise noted, $\rho := 0$ and $\tau := 0$. Because variables in the same domain are supposed to be more related than variables in different domains, it is assumed that $\rho \geq \tau$.

Example 3.1 (Covariance matrix)**.** Consider a study structure with $D = 2$ domains, $V_1 = 2$ variables in the first domain, and $V_2 = 3$ variables in the second domain. The correlation between variables in the same domain is $\rho = 0.8$; variables in different domains are independent ($\tau = 0$). The variance of all variables in the first domain is $\sigma^2_{1,1} = \sigma^2_{1,2} = 1$, and the variance of all variables in the second domain is $\sigma^2_{1,1} = \sigma^2_{1,2} = \sigma^2_{1,3} = 4$. Assume the number of patients in each arm to be $N = 8$. Then the covariance matrix of each vector $\mathbf{y}^{x,n}$ (where $x = 0, 1$ and $n = 1, ..., 8$) is as follows:

$$\Sigma^{\text{dt}} := \begin{pmatrix} 1 & 0.8 & 0 & 0 & 0 \\ 0.8 & 1 & 0 & 0 & 0 \\ 0 & 0 & 4 & 3.2 & 3.2 \\ 0 & 0 & 3.2 & 4 & 3.2 \\ 0 & 0 & 3.2 & 3.2 & 4 \end{pmatrix}$$

And the covariance matrix of $\hat{\boldsymbol{\mu}}$ is:

$$\Sigma^{\text{dt}}_{\mu} = \frac{2}{N}\Sigma^{\text{dt}} = \begin{pmatrix} 0.25 & 0.2 & 0 & 0 & 0 \\ 0.2 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.8 & 0.8 \\ 0 & 0 & 0.8 & 1 & 0.8 \\ 0 & 0 & 0.8 & 0.8 & 1 \end{pmatrix}$$

◀

In the remainder of this section, unless otherwise noted, $\mathbf{TV} := \mathbf{1}$, $\mathbf{LRV} := \mathbf{0.5}$, and $\sigma^2_{d,i} = 1$ for all $d = 1, ..., D$ and $i = 1, ..., V_d$. These choices are made for the sake of example; the methods that we describe in this section can be applied to scenarios where these constants have other values.

Number of patients: The number of patients in each arm is considered to be $N = 17$, yielding a total of $2N = 34$ patients. The number $N = 17$ is chosen as the lowest one which gives a power higher than 0.8 for each variable, where "power" is understood as the probability that a two-sided t-test at significance level $\alpha = 0.05$ will reject the null hypothesis ($\mu_i = 0$) when all of the following hold: (i) $\mu_i = \text{TV}_i = 1$, (ii) the true standard deviation for each variable is $\sigma_i = 1$, and (iii) the estimator of the effect $\hat{\mu}_i$ has the right sign (i.e. $\hat{\mu}_i > 0$).

Policy structure: The framework described in Chapter 2 does not impose a hard distinction between variables in the same domain and variables in different domains when defining and implementing policies. However, in this

Table 3.1: Power for one endpoint depending on the number of patients and its standard deviation, for the combinations of those values used in the simulations in Chapter 3. The power is calculated for a two-sided, two sample t-test at significance level $\alpha := 0.05$ *for an effect of* $\mathrm{TV} = 1$.

No. patients per arm	Standard deviation (σ)	Power (two sided)
17	1	80.7%
17	2	29.3%

chapter, a two level structure is chosen: first a decision for each domain is produced, and then these decisions are be combined into a single overall decision.

3.1.1 False Go Risk and False Stop Risk

A criterion of how well a policy generalizes the Lalonde framework is how well it controls the risk of False Go. Consider a policy G given by a pair of predicates $G \equiv (P^{\neg}_{\text{stop}}, P^{\llcorner}_{\text{go}})$. A False Go is a condition where a policy produces go (that is, the predicate $P^{\llcorner}_{\text{go}}$ is true), but all endpoints are below the LRV. Therefore, the False Go region (Θ^{FG}) is defined as follows:

$$\Theta^{\text{FG}} := \neg(\mathbf{LRV}),$$

By taking $\tilde{\Theta}^{\text{FG}} := \{\mathbf{LRV}\}$, Θ^{FG} fulfills the conditions of Corollary 2.89:

$$\Theta^{\text{FG}} = \bigcup_{\boldsymbol{\Delta} \in \tilde{\Theta}^{\text{FG}}} \neg(\boldsymbol{\Delta}) = \neg(\mathbf{LRV})$$

Thus, the FGR can be obtained by computing the risk of "go" at **LRV**:

$$\begin{aligned}\text{FGR} &= \sup_{\boldsymbol{\Delta} \in \Theta^{\text{FG}}} \mathbb{P}(P^{\llcorner}_{\text{go}}(\hat{\boldsymbol{\mu}}) \mid \boldsymbol{\mu} = \Delta) = \max_{\boldsymbol{\Delta} \in \tilde{\Theta}^{\text{FG}}} \mathbb{P}(P^{\llcorner}_{\text{go}}(\hat{\boldsymbol{\mu}}) \mid \boldsymbol{\mu} = \Delta) \\ &= \mathbb{P}(P^{\llcorner}_{\text{go}}(\hat{\boldsymbol{\mu}}) \mid \boldsymbol{\mu} = \mathbf{LRV})\end{aligned}$$

Similarly, a good policy should also control the risk of False Stop. A False Stop is a condition where a policy produces stop decision even though at least one endpoint reaches the target value (TV). Therefore, a tentative False Stop region (Θ^{FS}) can be defined as follows:

$$\Theta^{\text{FS}} := \{\boldsymbol{\Delta} \in \mathbb{R}^V \mid \Delta_i \geq \mathrm{TV}_i \text{ for at least one of } i = 1, ..., V\}$$

This region is shown in Figure 3.1a for the case $V = 2$. Unlike the case for Θ^{FG}, there is no $\tilde{\Theta}^{\text{FS}}$ fulfilling the conditions of Corollary 2.89. Instead, the following region is used:

$$\Theta_0^{\mathrm{FS}} := \{\boldsymbol{\Delta} \in \mathbb{R}^V \mid \boldsymbol{\Delta} \geq \mathbf{0} \text{ and } \Delta_i \geq \mathrm{TV}_i \text{ for at least 1 of } i = 1, ..., V\}$$

The region Θ_0^{FS} is shown in Figure 3.1c for the case $V = 2$, under the caption $R_{\mathrm{ext}}^{\mathrm{Go}}$. The set:

$$\begin{aligned}\tilde{\Theta}_0^{\mathrm{FS}} &:= \{(\mathrm{TV}_1 \quad 0 \quad \cdots \quad 0), ..., (0 \quad \cdots \quad 0 \quad \mathrm{TV}_V)\} \\ &\equiv \{\mathrm{TV}_1\mathbf{e}^1, ..., \mathrm{TV}_V\mathbf{e}^V\} \subseteq \mathbb{R}^V\end{aligned}$$

witnesses that Θ_0^{FS} fulfills the conditions of Corollary 2.89:

$$\Theta_0^{\mathrm{FS}} = \bigcup_{\boldsymbol{\Delta} \in \tilde{\Theta}_0^{\mathrm{FS}}} \llcorner(\Delta) = \llcorner(\mathrm{TV}_1\mathbf{e}^1) \cup ... \cup \llcorner(\mathrm{TV}_V\mathbf{e}^V)$$

Thus:

$$\begin{aligned}\mathrm{FSR} = \sup_{\boldsymbol{\Delta} \in \Theta_0^{\mathrm{FS}}} \mathbb{P}(P^{\urcorner}_{\mathrm{stop}}(\hat{\boldsymbol{\mu}}) \mid \boldsymbol{\mu} = \Delta) = \max_{\boldsymbol{\Delta} \in \tilde{\Theta}_0^{\mathrm{FS}}} \mathbb{P}(P^{\urcorner}_{\mathrm{stop}}(\hat{\boldsymbol{\mu}}) \mid \boldsymbol{\mu} = \Delta) = \\ = \max\{\mathbb{P}(P^{\urcorner}_{\mathrm{stop}}(\hat{\boldsymbol{\mu}}) \mid \boldsymbol{\mu} = \mathrm{TV}_1\mathbf{e}^1), ..., \mathbb{P}(P^{\urcorner}_{\mathrm{stop}}(\hat{\boldsymbol{\mu}}) \mid \boldsymbol{\mu} = \mathrm{TV}_V\mathbf{e}^V)\}\end{aligned}$$

In other words, the FSR for $\boldsymbol{\mu} \in \Theta_0^{\mathrm{FS}}$ can be obtained by computing the risk of "stop" when $\boldsymbol{\mu} = (\mathrm{TV}_1 \quad 0 \quad \cdots \quad 0)$, $\boldsymbol{\mu} = (0 \quad \mathrm{TV}_2 \quad 0 \quad \cdots \quad 0)$, all the way up to $\boldsymbol{\mu} = (0 \quad \cdots \quad 0 \quad \mathrm{TV}_V)$, and taking the largest such value.

Remark. When defining FSR and FGR, the predicates $P^{\llcorner}_{\mathrm{go}}$ and $P^{\urcorner}_{\mathrm{stop}}$ are considered individually. This means that those cases where the policy produces both "go" and "stop (i.e. both $P^{\llcorner}_{\mathrm{go}}$ and $P^{\urcorner}_{\mathrm{stop}}$ hold), contribute to increase both the FGR and the FSR.

In the one-variable case, the FGR and FSR are determined by the parameters $\alpha^{\star}_{\mathrm{FS}}$ and $\alpha^{\star}_{\mathrm{FG}}$. The extent to which this relationship holds when the number of variables is two or more is quantified in §3.2.1.

The decision presented to stakeholders is either Go or Stop, but not both simultaneously. For this reason, when evaluating the global result of a policy, the metrics FSr and FGr are used. If both "go" and "stop" decisions are produced (i.e. both $P^{\urcorner}_{\mathrm{stop}}$ and $P^{\llcorner}_{\mathrm{go}}$ hold), this only contributes to increase FSr, as "stop" has priority over "go" (Definition 2.20). The details of how these metrics are defined is explained in the following section (§3.1.2). These metrics (together with CGr and CSr) are the ones used for evaluating the global result of a policy (§3.3).

3.1.2 Evaluation metrics

As discussed in Remark 2.81, in order for a policy to be useful, it is not enough for it to avoid producing false decisions, but it should also produce correct decisions. Which combinations of true effects should lead to a Go decision and which to a Stop decision is dependent on the clinical understanding of the drug under study. However, it is unambiguously the case that a true effect of 0 for all endpoints should lead to a Stop decision, and all endpoints reaching the

TV should lead to a Go decision. Therefore, when evaluating policies, these following regions are the first considered:

$$R^{\text{Go}} := \llcorner(\mathbf{TV})$$
$$R^{\text{Stop}} := \neg(\mathbf{0})$$

These regions are depicted in Figure 3.1b for the case $V = 2$. Given the regions R^{Go} and R^{Stop} as defined above, in order to calculate the evaluation metrics for a policy G, it suffices to simulate the outcome of the policy for $\boldsymbol{\mu} = \mathbf{TV}$ and $\boldsymbol{\mu} = \mathbf{0}$ (see Theorem 2.83):

$$\begin{aligned}
\text{FGr} &= \sup_{\boldsymbol{\Delta} \in R^{\text{Stop}}} \text{rGo}(\boldsymbol{\Delta}) &&= \text{rGo}(\mathbf{0}) \\
\text{CSr} &= \inf_{\boldsymbol{\Delta} \in R^{\text{Stop}}} \text{rStop}(\boldsymbol{\Delta}) &&= \text{rStop}(\mathbf{0}) \\
\text{CGr} &= \inf_{\boldsymbol{\Delta} \in R^{\text{Go}}} \text{rGo}(\boldsymbol{\Delta}) &&= \text{rGo}(\mathbf{TV}) \\
\text{FSr} &= \sup_{\boldsymbol{\Delta} \in R^{\text{Go}}} \text{rStop}(\boldsymbol{\Delta}) &&= \text{rStop}(\mathbf{TV}),
\end{aligned}$$

where:

$$\text{rGo}(\boldsymbol{\Delta}) = \mathbb{P}(G(\hat{\boldsymbol{\mu}}) \text{ is Go} \mid \boldsymbol{\mu} = \boldsymbol{\Delta})$$
$$\text{rStop}(\boldsymbol{\Delta}) = \mathbb{P}(G(\hat{\boldsymbol{\mu}}) \text{ is Stop} \mid \boldsymbol{\mu} = \boldsymbol{\Delta})$$

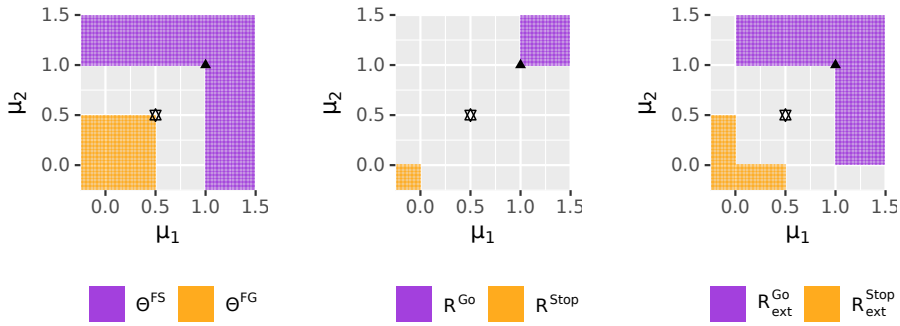
Depending on the policy being evaluated, metrics defined on other regions may also be of interest; for example:

$$R^{\text{Go}}_{\text{ext}} := \{\boldsymbol{\Delta} \in \mathbb{R}^V \mid \Delta_i \geq \text{TV}_i \text{ for some } i = 1, ..., V\} \cap \llcorner(\mathbf{0})$$
$$R^{\text{Stop}}_{\text{ext}} := \{\boldsymbol{\Delta} \in \mathbb{R}^V \mid \Delta_i \leq 0 \text{ for at least } V - 1 \text{ of the } i = 1, ..., V\} \cap \neg(\mathbf{LRV})$$

The region $R^{\text{Go}}_{\text{ext}}$ covers all the cases where the true effect is (i) non-negative for all endpoints and (ii) also exceeds the target value (TV) for at least one of them. The $R^{\text{Stop}}_{\text{ext}}$ region covers those cases where the true effect is non-positive for all but one endpoint, for which it does not exceed the lower reference value (LRV). For a graphical representation of these regions for $V = 2$, see Figure 3.1c.

In the definition of $R^{\text{Go}}_{\text{ext}}$, the intersection with the upwards cone $\llcorner(\mathbf{0})$ is equivalent to only considering cases where the true effect is non-negative for all endpoints. For $R^{\text{Stop}}_{\text{ext}}$, the intersection with the downwards cone $\neg(\mathbf{LRV})$ is equivalent to constraining the evaluation to cases where the true effect is lower than or equal to the LRV for all endpoints. Besides narrowing down the regions to true effects for which a Go (respectively Stop) decision is most desirable, these restrictions have the additional effect of ensuring the existence of sets $\tilde{R}^{\text{Go}}_{\text{ext}}$ and $\tilde{R}^{\text{Stop}}_{\text{ext}}$ fulfilling the conditions of Theorem 2.83:

$$\tilde{R}^{\text{Go}}_{\text{ext}} := \{\text{TV}_1\mathbf{e}^1, ..., \text{TV}_V\mathbf{e}^V\}$$
$$\tilde{R}^{\text{Stop}}_{\text{ext}} := \{\text{LRV}_1\mathbf{e}^1, ..., \text{LRV}_V\mathbf{e}^V\}$$



(a) Regions for defining $G^{\text{meas}}_{(d)}$ *and* $G^{\text{Hot}}_{(d)}$.

(b) A minimal evaluation scenario. $R^{\text{Go}} = \llcorner(\mathbf{TV})$, $R^{\text{Stop}} = \neg(\mathbf{0})$

(c) An evaluation scenario with regions larger than R^{Stop} *and* R^{Go}.

Figure 3.1: Regions used for defining and evaluating the policies in this chapter, for the case $V = 2$. *The point* $\mathbf{LRV}$ *is marked as* ✡, *and the point* $\mathbf{TV}$ *is marked as* ▲. *The reference* FGR *is defined on* Θ^{FG}. *Due to the difficulty of evaluating* FSR *on* Θ^{FS} *(see §3.1.1), the* FSR *is defined on* $\Theta^{\text{FS}}_0 = R^{\text{Go}}_{\text{ext}}$.

Therefore, the metrics can be computed by simulating the policy for a finite number of possible effects:

$$\begin{aligned}
\text{FGr} &= \sup_{\boldsymbol{\Delta}\in R^{\text{Stop}}} \text{rGo}(\boldsymbol{\Delta}) &&= \max_{\boldsymbol{\Delta}\in \tilde{R}^{\text{Stop}}} \text{rGo}(\boldsymbol{\Delta}) \\
\text{CSr} &= \inf_{\boldsymbol{\Delta}\in R^{\text{Stop}}} \text{rStop}(\boldsymbol{\Delta}) &&= \min_{\boldsymbol{\Delta}\in \tilde{R}^{\text{Stop}}} \text{rStop}(\boldsymbol{\Delta}) \\
\text{CGr} &= \inf_{\boldsymbol{\Delta}\in R^{\text{Go}}} \text{rGo}(\boldsymbol{\Delta}) &&= \min_{\boldsymbol{\Delta}\in \tilde{R}^{\text{Go}}} \text{rGo}(\boldsymbol{\Delta}) \\
\text{FSr} &= \sup_{\boldsymbol{\Delta}\in R^{\text{Go}}} \text{rStop}(\boldsymbol{\Delta}) &&= \max_{\boldsymbol{\Delta}\in \tilde{R}^{\text{Go}}} \text{rStop}(\boldsymbol{\Delta})
\end{aligned}$$

Example 3.2. If $V = 3$, then:

$$\begin{aligned}
R^{\text{Go}}_{\text{ext}} &= \llcorner((\text{TV}_1, 0, 0)) \cup \llcorner((0, \text{TV}_2, 0)) \cup \llcorner((0, 0, \text{TV}_3)) \\
R^{\text{Stop}}_{\text{ext}} &= \neg((\text{LRV}_1, 0, 0)) \cup \neg((0, \text{LRV}_2, 0)) \cup \neg((0, 0, \text{LRV}_3))
\end{aligned}$$

and:

$$\begin{aligned}
\text{FGr} &= \max\{\text{rGo}((\text{LRV}_1, 0, 0)),\ \text{rGo}((0, \text{LRV}_2, 0)),\ \text{rGo}((0, 0, \text{LRV}_3))\} \\
\text{CSr} &= \min\{\text{rStop}((\text{LRV}_1, 0, 0)),\ \text{rStop}((0, \text{LRV}_2, 0)),\ \text{rStop}((0, 0, \text{LRV}_3))\} \\
\text{CGr} &= \min\{\text{rGo}((\text{TV}_1, 0, 0)),\ \text{rGo}((0, \text{TV}_2, 0)),\ \text{rGo}((0, 0, \text{TV}_3))\} \\
\text{FSr} &= \max\{\text{rStop}((\text{TV}_1, 0, 0)),\ \text{rStop}((0, \text{TV}_2, 0)),\ \text{rStop}((0, 0, \text{TV}_3))\}
\end{aligned}$$

◀

Policy parameters: Each policy defined in this section is parameterized by $\alpha^\star_{\mathrm{FS}}, \alpha^\star_{\mathrm{FG}} \in (0,1)$. The first parameter affects the probability of a "stop" decision, while the second parameter affects the probability of a "go" decision. The values $\alpha^\star_{\mathrm{FS}} = 0.1$ and $\alpha^\star_{\mathrm{FG}} = 0.2$ are used, following the percentiles suggested in [LKH+07].

Distribution of statistics: In Chapter 2, we assume that the covariance matrix Σ (and thus Σ_μ) are known constants. The policies are thus defined by referring to the probability density of the multivariate normal distribution.

The covariance matrices are in practice estimated from the data. To account for the increased uncertainty, the multivariate normal is replaced by a multivariate generalization of the t distribution.

Following [GB09, equation before (1.4)], and its implementation in the [`mvtnorm`] R package, π_{mvt} is defined as:

$$\pi_{\mathrm{mvt}}(\Delta; \boldsymbol{\mu}, \tilde{\Sigma}, \nu) = \frac{\Gamma(\frac{\nu+V}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\left|\tilde{\Sigma}\right|(\nu\pi)^V}} \left(1 + \frac{(\boldsymbol{\Delta}-\boldsymbol{\mu})^\top \tilde{\Sigma}^{-1}(\boldsymbol{\Delta}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+V}{2}} \tag{3.3}$$

In other words, $\pi_{\mathrm{mvt}}(\boldsymbol{\Delta}; \boldsymbol{\mu}, \tilde{\Sigma}, \nu)$ is the probability density of a multivariate generalization of a t-Student distribution, shifted by the mean vector $\boldsymbol{\mu} \in \mathbb{R}^V$, scaled by the covariance matrix $\tilde{\Sigma} \in \mathbb{R}^{V \times V}$ and having ν degrees of freedom.

Methodology: The rGo, rStop and all metrics defined in terms of these quantities are calculated by simulating $M = 10000$ or more trials. This yields results that accurate up to 1 percentage point. For the sake of performance, $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}_\mu$ are simulated directly from their distributions (Remark 2.3), without computing the intermediate study data.

3.2 Evaluation of domain-level policies

The first part of the evaluation concerns the properties of different strategies for combining the outcomes of multiple endpoints all belonging to the same domain. For each domain $d \in \{1, ..., D\}$, a domain subpolicy $G_{(d)}$ is defined. The subpolicies $G_{(d)}$ are implementations of the informal policy:

> "A domain is Go if at least one endpoint within that domain is Go; a domain is Stop if all endpoints within that domain are Stop". (3.4)

Here are three ways to implement the informal rule above (3.4):

Per-endpoint policy: Let G^{L} be the univariate Lalonde policy as defined in Example 2.30. A policy for a domain d with V_d endpoints can be formulated as follows:

$$G^{\text{ind}}_{(d),\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\tilde{\boldsymbol{\mu}},\tilde{\Sigma}_\mu,c) := \begin{cases} \text{stop} & \text{all of } (G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\tilde{\mu}_{d[1]}) \text{ is Stop}, \\ & \quad ..., \\ & \quad G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\tilde{\mu}_{d[V_d]}) \text{ is Stop}) \\ \text{go} & \text{at least 1 of } (G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\tilde{\mu}_{d[1]}) \text{ is Go}, \\ & \quad ..., \\ & \quad G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\tilde{\mu}_{d[V_d]}) \text{ is Go}) \end{cases}$$

where $d[i]$ is the index of the ith variable in domain d:

$$d[i] := i + \sum_{s=1}^{d-1} V_s \tag{3.5}$$

When $V_d = 2$, this is equivalent to:

$$G^{\text{ind}}_{(d),\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\tilde{\boldsymbol{\mu}}) = T([G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}](\tilde{\mu}_{d[1]}), [G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}](\tilde{\mu}_{d[2]})),$$

where T is the table defined in Example 2.34. The policy $G^{\text{ind}}_{(d)}$ is represented graphically in Figure 3.2a.

Probability measure: From the description (3.4), it is possible to interpret that, for a single domain d:

(i) A "stop" decision for a domain d is a False Stop if *any* of the endpoints in domain d is above the TV.

(ii) A "go" decision for a domain d is a False Go if *all* of the endpoints in the domain are below the LRV.

This interpretation leads to the definition of the following regions of $\mathbb{R}^V$:

$$\Theta^{\text{FS}}_{(d)} := \{\boldsymbol{\Delta} \in \mathbb{R}^{V_d} \mid \Delta_i \geq \text{TV}_i \text{ for at least 1 of } i \in I(d)\} \tag{3.6}$$

$$\Theta^{\text{FG}}_{(d)} := \{\boldsymbol{\Delta} \in \mathbb{R}^{V_d} \mid \Delta_i \leq \text{LRV}_i \text{ for all of } i \in I(d)\}, \tag{3.7}$$

where $I(d)$ are the indices of the variables in domain d:

$$I(d) = \{d[1], ..., d[V_d]\} = \left\{ i + \sum_{s=1}^{d-1} V_s \,\middle|\, i = 1, ..., V_d \right\} \tag{3.8}$$

These regions are depicted in Figure 3.1a for the case $V_d = 2$. For the case $V_d = 3$, the regions can be understood as the following (unions of) cones:

$$\Theta^{\text{FS}}_{(d)} := \llcorner((\text{TV}_1, -\infty, -\infty)) \cup \llcorner((-\infty, \text{TV}_2, -\infty)) \cup \llcorner((-\infty, -\infty, \text{TV}_3))$$
$$\Theta^{\text{FG}}_{(d)} := \neg(\mathbf{LRV})$$

Remark. If we ignore the (zero measure) boundary of $\Theta^{\text{FS}}_{(d)}$, then $\Theta^{\text{FS}}_{(d)} = \mathbb{R}^{V_d} \backslash \neg(\mathbf{TV}_{I(d)})$.

Following §2.9.2, a probability measure M is defined as follows:

$$M_\nu(A; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu) = \int_A \pi_{\text{mvt}}(\boldsymbol{\Delta}; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, \nu) d\boldsymbol{\Delta}, \tag{3.9}$$

where π_{mvt} is as defined in Equation 3.3, and the policy G^{meas} is defined as:

$$G^{\text{meas}}_{(d)}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) := \begin{cases} \text{stop} & M_{\nu(c)}(\Theta^{\text{FS}}_{(d)}; \tilde{\boldsymbol{\mu}}_{I(d)}, (\tilde{\Sigma}_\mu)_{I(d),I(d)}) \leq \alpha^\star_{\text{FS}} \\ \text{go} & M_{\nu(c)}(\Theta^{\text{FG}}_{(d)}; \tilde{\boldsymbol{\mu}}_{I(d)}, (\tilde{\Sigma}_\mu)_{I(d),I(d)}) \leq \alpha^\star_{\text{FG}} \end{cases},$$

where ν is defined as in Equation 2.23. In other words, a "stop" decision (respectively a "go" decision) is produced if, for an estimated mean effect and covariance matrix, the probability of producing an observation in the False Stop region $\Theta^{\text{FS}}_{(d)}$ (respectively in the False Go region $\Theta^{\text{FG}}_{(d)}$) is lower than $\alpha^\star_{\text{FS}}$ (respectively lower than $\alpha^\star_{\text{FG}}$). A graphical representation of $G^{\text{meas}}_{(d)}$ for the case with $V_d = 2$ variables, $N = 17$ patients per arm and $c =$ Theoretical is shown in Figure 3.2b, and the theoretical caveats of this approach are discussed in §2.9.2.

Hotelling's $\mathbf{T^2}$: Consider $\Theta^{\text{FS}}_{(d)}$ and $\Theta^{\text{FG}}_{(d)}$ defined as in (3.6) and (3.7), respectively. Following §2.9.3, let:

$$T^2[\Theta; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu] := \inf_{\boldsymbol{\Delta} \in \Theta} (\boldsymbol{\Delta} - \tilde{\boldsymbol{\mu}})^\top \tilde{\Sigma}_\mu^{-1} (\boldsymbol{\Delta} - \tilde{\boldsymbol{\mu}})$$

Then:

$$G^{\text{Hot}}_{(d)}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) := \begin{cases} \text{stop} & T^2[\Theta^{\text{FS}}_{(d)}; \tilde{\boldsymbol{\mu}}_{I(d)}, (\tilde{\Sigma}_\mu)_{I(d),I(d)}] \geq Q(\alpha^\star_{\text{FS}}; c) \\ \text{go} & T^2[\Theta^{\text{FG}}_{(d)}; \tilde{\boldsymbol{\mu}}_{I(d)}, (\tilde{\Sigma}_\mu)_{I(d),I(d)}] \geq Q(\alpha^\star_{\text{FG}}; c) \end{cases}, \tag{3.10}$$

where Q is defined as in Equation 2.70. As explained in §2.9.3, this policy produces a "stop" decision (respectively a "go" decision) if the confidence region at level $1 - \alpha^\star_{\text{FS}}$ (respectively at level $1 - \alpha^\star_{\text{FG}}$) does not intersect with the False Stop region $\Theta^{\text{FS}}_{(d)}$ (respectively the False Go region $\Theta^{\text{FG}}_{(d)}$).

The policy $G^{\text{Hot}}_{(d)}$ for the case with $V_d = 2$ variables, $N = 17$ patients per arm and $c =$ Theoretical is represented graphically in Figure 3.2c.

3.2.1 Evaluation of unadjusted policies

The policies $G_{(d)}$ are evaluated for the case with one domain ($D = 1$). For each of $G^{\text{ind}}_{(1)}$, $G^{\text{Hot}}_{(1)}$ and $G^{\text{meas}}_{(1)}$, their sensitivity to the number of variables in the first domain ($V_1 = V = 1, 2, 3, 4$) is assessed.

The case for $V = 1$ is summarized in Table 3.2. As explained in Chapter 2, for the policies G^{ind} and G^{meas}, the FSR corresponds to $\alpha^\star_{\text{FS}}$, and the FGR corresponds to the $\alpha^\star_{\text{FG}}$. For the policy G^{Hot}, the FSR corresponds to $\alpha^\star_{\text{FS}}/2$, and the FGR corresponds to $\alpha^\star_{\text{FG}}/2$. The results in Table 3.2 are obtained through simulation, so the correspondences are approximate.

The results when the number of variables $V \geq 2$ are summarized in Table 3.3. The values of the metrics are influenced by the correlation between

Table 3.2: Simulated metrics for selected policies where for $V = 1$ variables, where TV $= 1$, LRV $= 0.5$. *For each combination of true standard deviation (σ) and number of patients per arm (N), $M = 10000$ simulations are run, all starting from the same PRNG seed. The parameters for all policies are set to $\alpha^\star_{\text{FS}} := 0.1$, $\alpha^\star_{\text{FG}} := 0.2$. For the univariate Lalonde policy (G^{L}), the difference between the* FSR *and $\alpha^\star_{\text{FS}}$ and the difference between* FGR *and $\alpha^\star_{\text{FG}}$ are both within the expected margin of error for the simulation. The policies G^{meas} and $G^{\text{Hot.half}}$ (which is a suitably adjusted G^{Hot}, see Equation 3.19) can be seen to be equivalent to G^{L} when $V = 1$.*

	$[\text{TV}, +\infty)$			$(-\infty, \text{LRV}]$		
Policy	FSR	FSr	CGr	FGR	FGr	CSr
$\sigma = 1$, $N = 17$						
G^{L}	0.101	0.101	0.727	0.200	0.200	0.564
G^{Hot}	0.048	0.048	0.564	0.101	0.101	0.412
$G^{\text{Hot.half}}$	0.101	0.101	0.727	0.200	0.200	0.564
G^{meas}	0.101	0.101	0.727	0.200	0.200	0.564
$\sigma = 2$, $N = 17$						
G^{L}	0.101	0.101	0.456	0.200	0.200	0.285
G^{Hot}	0.048	0.048	0.288	0.101	0.101	0.171
$G^{\text{Hot.half}}$	0.101	0.101	0.456	0.200	0.200	0.285
G^{meas}	0.101	0.101	0.456	0.200	0.200	0.285

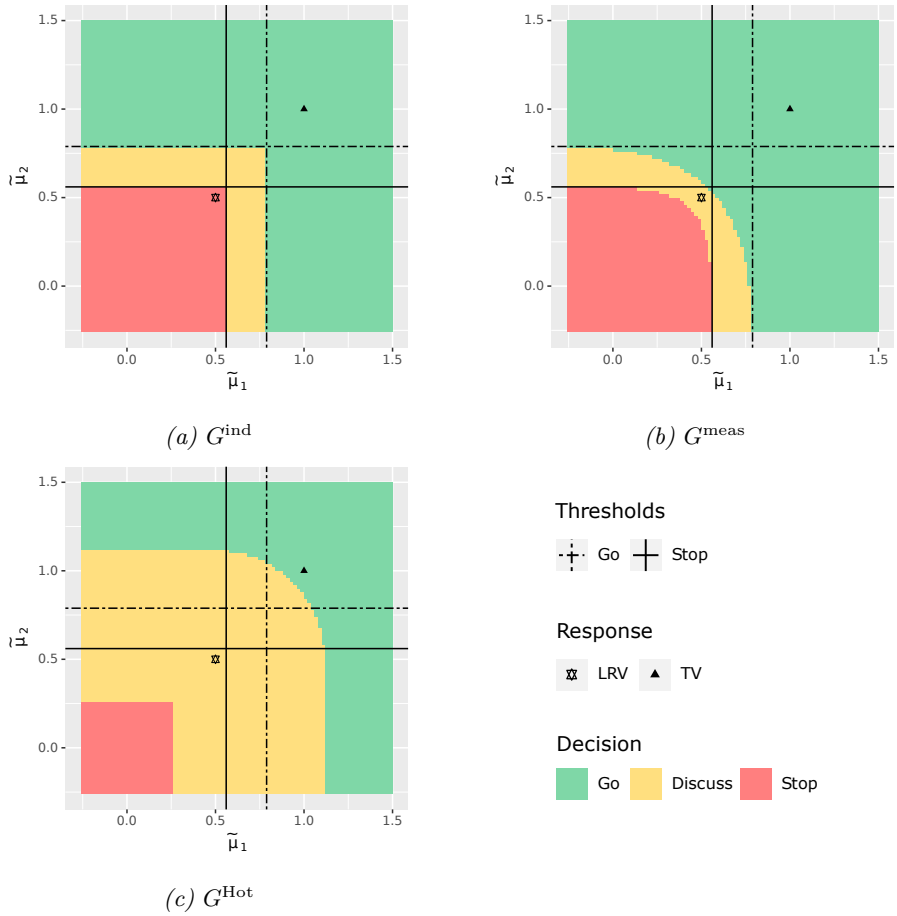
variables in the same domain (ρ) and by the choice of policy (G^{ind}, G^{meas} or G^{Hot}). Relative to the other two policies, the G^{ind} policy produces incorrect decisions with low frequency in the $R^{\text{Stop}}_{\text{ext}}$ and $R^{\text{Go}}_{\text{ext}}$ regions, in which only one of the endpoints is non-zero. An concrete example of this is the case where $V = 4$, $\rho = 0$, FGr $= 0.23 \approx 0.2 = \alpha^\star_{\text{FG}}$ and and FSR $= 0.11 \approx 0.10 = \alpha^\star_{\text{FS}}$. Other combinations of V and ρ perform similarly well. This is due to the fact that endpoints for which the true effect is 0 do seldom affect the result, and thus the situation reduces to the univariate case. On the other hand, when the true effect is LRV for all endpoints, FGR $= 0.27 \geq 0.2 = \alpha^\star_{\text{FS}}$ when $V = 2$ and $\rho = 0.8$. This worsens as the number of variables increases and the correlation decreases; e.g. FGR $= 0.59$ when $V = 4$ and $\rho = 0$. This is due to the multiple comparisons problem, which we elaborate on in §3.2.2. The effect becomes less pronounced as variables become more correlated (e.g. FGR $= 0.35$ when $V = 4$ and $\rho = 0.8$); but this is of little consolation, as such high correlations between endpoints are seldom expected in practice.

The G^{meas} policy fails to bound the FGR on the $\neg(\mathbf{LRV})$ region in general, and specially when the number of variables V grows (e.g FGR $= 0.92$ when $V = 4$ and $\rho = 0$). This behaviour can be ascribed to the lack of a theoretical foundation for the statistical properties of this policy beyond the $V = 1$ case (see Remark 2.68).

Of the three policies, G^{Hot} is the only one where the FSR on Θ^{FS}_0 and the FGR on $\Theta^{\text{FG}} = \neg(\mathbf{LRV})$ are completely dominated by $\alpha^\star_{\text{FS}}$ and $\alpha^\star_{\text{FG}}$ respectively, as proved in Lemma 2.76. The Hotelling T^2-based policy is how-

Table 3.3: Comparison of unadjusted multivariate policies. In bold, the best performing policy among all possible policies (row) for each combination of number of variables (V), correlation (ρ), region (top header) and metric (column). $N = 17$, $\alpha^{\star}_{\mathrm{FS}} = 0.1$, $\alpha^{\star}_{\mathrm{FG}} = 0.2$.

		$\Theta^{\mathrm{FS}} = R^{\mathrm{Go}}_{\mathrm{ext}}$		$R^{\mathrm{Stop}}_{\mathrm{ext}}$		$\llcorner(\mathbf{TV})$	$\neg(\mathbf{LRV})$	$\neg(\mathbf{0})$
	ρ	FSR	CGr	FGr	CSr	CGr	FGR	CSr
V=2								
G^{ind}	0	0.10	0.73	0.22	**0.53**	0.93	0.35	**0.89**
	0.4	0.10	0.73	0.21	**0.55**	0.88	0.32	**0.90**
	0.8	0.10	**0.73**	0.20	**0.56**	0.81	0.27	**0.92**
G^{Hot}	0	**0.01**	0.35	**0.04**	0.16	0.68	**0.09**	0.54
	0.4	**0.02**	0.35	**0.04**	0.20	0.56	**0.08**	0.58
	0.8	**0.02**	0.35	**0.04**	0.22	0.45	**0.06**	0.65
G^{meas}	0	0.09	**0.77**	0.27	0.50	**0.98**	0.52	0.87
	0.4	0.10	**0.74**	0.24	0.53	**0.93**	0.42	0.88
	0.8	0.10	**0.73**	0.20	**0.56**	**0.85**	0.31	0.90
V=3								
G^{ind}	0	0.09	0.73	0.21	**0.50**	0.98	0.49	**0.84**
	0.4	0.10	0.73	0.21	**0.54**	0.93	0.41	**0.86**
	0.8	0.10	**0.72**	0.20	**0.57**	0.85	0.32	**0.90**
G^{Hot}	0	**0.00**	0.22	**0.02**	0.04	0.75	**0.07**	0.21
	0.4	**0.01**	0.21	**0.01**	0.08	0.51	**0.05**	0.31
	0.8	**0.00**	0.21	**0.01**	0.12	0.35	**0.03**	0.43
G^{meas}	0	0.07	**0.81**	0.34	0.44	**1.00**	0.78	0.81
	0.4	0.09	**0.75**	0.27	0.49	**0.98**	0.58	0.82
	0.8	0.10	**0.72**	0.22	0.55	**0.89**	0.38	0.87
V=4								
G^{ind}	0	0.09	0.74	0.23	**0.47**	**1.00**	0.59	**0.79**
	0.4	0.10	0.73	0.22	**0.52**	0.96	0.47	**0.83**
	0.8	0.10	0.72	0.20	**0.56**	0.87	0.35	**0.88**
G^{Hot}	0	**0.00**	0.13	**0.01**	0.01	0.80	**0.06**	0.04
	0.4	**0.00**	0.12	**0.01**	0.03	0.45	**0.03**	0.12
	0.8	**0.00**	0.12	**0.01**	0.06	0.26	**0.02**	0.26
G^{meas}	0	0.06	**0.84**	0.41	0.38	**1.00**	0.92	0.74
	0.4	0.09	**0.76**	0.29	0.46	**0.99**	0.69	0.78
	0.8	0.10	**0.73**	0.22	0.54	**0.92**	0.42	0.85



(a) G^{ind} *(b)* G^{meas}

(c) G^{Hot}

Figure 3.2: Go and Stop regions of unadjusted multivariate policies for $V = 2$*,* $\sigma_1 = \sigma_2 = 1$*,* $N = 17$*. The univariate* Go *and* Stop *thresholds are indicated for reference.*

ever too conservative, specially as the number of variables and the correlation between them increases. In particular, this negatively affects the CGr; for instance when the true effect is $\mu = \mathbf{TV}$, the number of variables is $V_d = 4$ and the correlation is $\rho = 0.4$, CGr $= 0.45$; compare with that obtained when the Lalonde method is applied a single variable whose true effect is TV (see Table 3.2, CGr $= 0.727$). The underlying reason is further explored in §3.2.3.

Overall, the evaluation shows that, as opposed to what happens in the single variable case ($V = 1$), there is no straightforward coupling between the parameter $\alpha^{\star}_{\text{FS}}$ and the empirical FSR, and mutatis mutandis for the parameter $\alpha^{\star}_{\text{FG}}$ and the empirical FGR. This suggests that adjusted versions of the policies need to be developed in which this connection is better realized.

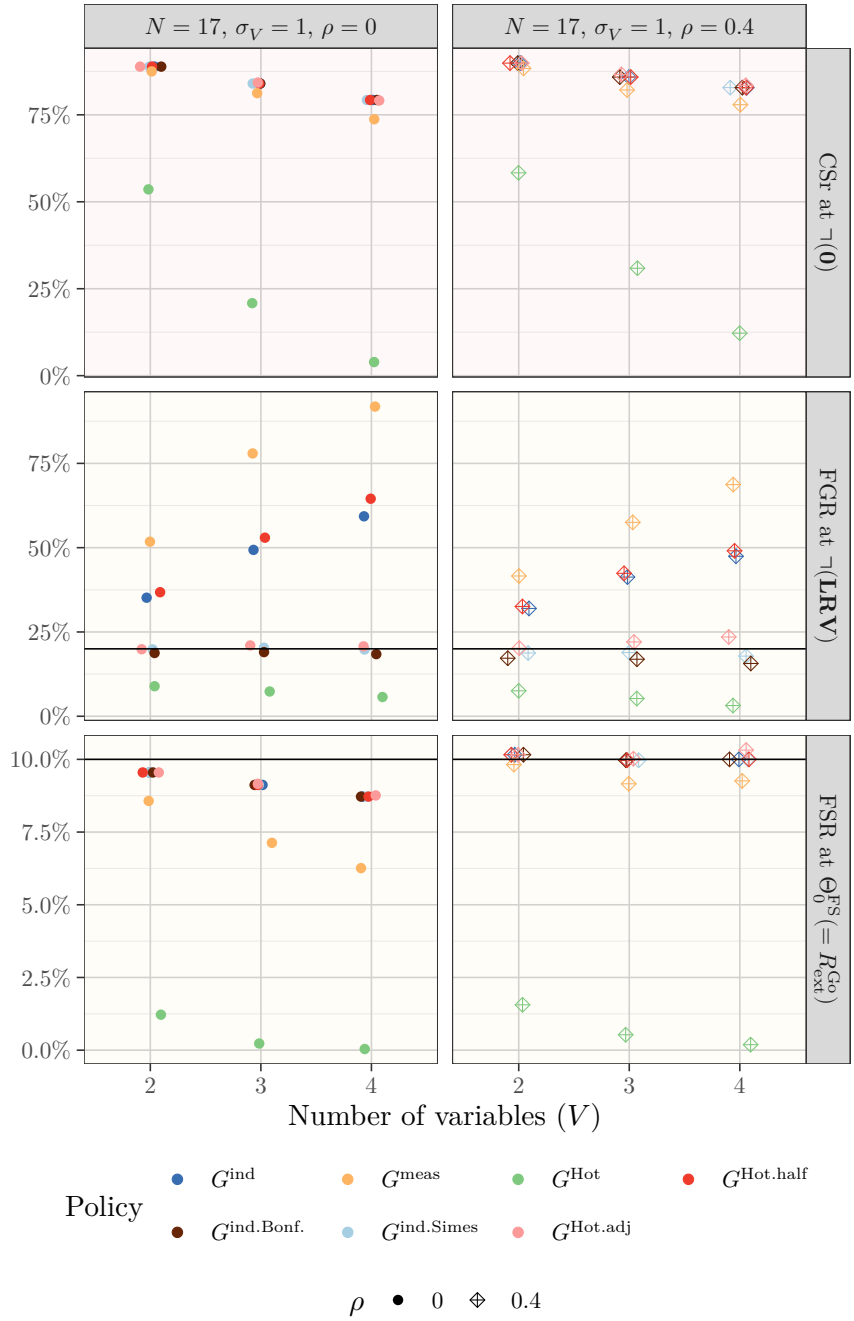


Figure 3.3: Comparison of selected metrics for all policies under consideration for a single domain. Some policies are unadjusted (i.e. G^{ind}, G^{meas} and G^{Hot}; see Table 3.3). Other policies are adjusted (see Table 3.5), either successfully (i.e. $G^{\mathrm{ind.Bonf.}}$, $G^{\mathrm{ind.Simes}}$ and $G^{\mathrm{Hot.adj}}$, for which FGR ≈ 0.2 *and* FSR ≈ 0.1*) or unsuccessfully (i.e. $G^{\mathrm{Hot.half}}$). The background is lightly shaded in the "Stop" color for the Correct Stop rate (*CSr*) metric, and in the "Discuss" colour for the False Go and False Stop Risk metrics (*FGR *and* FSR*, respectively).*

3.2.2 Adjusting policies based on individual variables

In the case of policies combining the results of multiple applications of the univariate Lalonde framework (i.e. G^{ind}), the discrepancy occurs between $\alpha^\star_{\text{FG}}$ and the FGR. This is an instance of the multiple comparisons problem [Tuk53, Hig13]. By formulating the predicate for "go" as a hypothesis test, the same techniques that are used to correct for multiple tests can be applied here.

By the definition of G^{L}, for each variable $d[i]$ in domain d ($i = 1, ..., V_d$), the following equivalences hold:

$$G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\tilde{\mu}_{d[i]}) \ni \text{stop} \Leftrightarrow \tilde{\mu}_{d[i]} \leq \text{thr}^{\text{stop}}_{d[i],\alpha^\star_{\text{FS}}} \Leftrightarrow p^{\text{stop}}_{d[i]} \leq \alpha^\star_{\text{FS}}$$

and:

$$G^{\text{L}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\tilde{\mu}_{d[i]}) \ni \text{go} \Leftrightarrow \tilde{\mu}_{d[i]} \geq \text{thr}^{\text{go}}_{d[i],\alpha^\star_{\text{FG}}} \Leftrightarrow p^{\text{go}}_{d[i]} \leq \alpha^\star_{\text{FG}}$$

where:

$$\begin{aligned}
\text{thr}^{\text{stop}}_{d[i],\alpha^\star_{\text{FS}}} &:= \text{TV}_{d,i} + (\tilde{\Sigma}_\mu)_{d[i],d[i]} t_{\nu(c),\alpha^\star_{\text{FS}}} \\
\text{thr}^{\text{go}}_{d[i],\alpha^\star_{\text{FG}}} &:= \text{LRV}_{d,i} + (\tilde{\Sigma}_\mu)_{d[i],d[i]} t_{\nu(c),1-\alpha^\star_{\text{FG}}} \\
p^{\text{stop}}_{d[i]} &:= F_{t,\nu(c)}\left(\frac{\tilde{\mu}_{d[i]} - \text{TV}_{d[i]}}{(\tilde{\Sigma}_\mu)_{d[i],d[i]}}\right) \\
p^{\text{go}}_{d[i]} &:= F_{t,\nu(c)}\left(\frac{\text{LRV}_{d[i]} - \tilde{\mu}_{d[i]}}{(\tilde{\Sigma}_\mu)_{d[i],d[i]}}\right),
\end{aligned}$$

and $F_{t,\nu}$ is the cumulative distribution function of the Student-t distribution with ν degrees of freedom. The "stop" case requires all hypothesis to hold; so it can be copied verbatim from the unadjusted G^{ind} policy. Only the "go" case needs to be corrected, as this is where the discrepancies between $\alpha^\star_{\text{FG}}$ and the FGR arise. A straightforward approach follows from the Bonferroni inequalities [Bon36, Hay13], where the significance level is divided by the number of hypothesis under test (in this case, V_d):

$$G^{\text{ind.Bonf.}}_{(d),\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) := \begin{cases} \text{stop} & G^{\text{ind}}_{(d),\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) \ni \text{stop} \\ \text{go} & \text{at least 1 of } (p^{\text{go}}_{d[1]} \leq \frac{1}{V_d}\alpha^\star_{\text{FG}}, \\ & \qquad p^{\text{go}}_{d[2]} \leq \frac{1}{V_d}\alpha^\star_{\text{FG}}, \\ & \qquad ..., \\ & \qquad p^{\text{go}}_{d[V_d]} \leq \frac{1}{V_d}\alpha^\star_{\text{FG}}) \end{cases} \tag{3.11}$$

The Bonferroni method can be too conservative, in part because the overall distribution of p-values is disregarded. A method proposed by Simes [Sim86], and later studied and expanded by Benjamini and Hochberg [BH95] has higher power than the Bonferroni method while still having type I error lower than α for the case where all null hypothesis are true. This method is only suitable when the p-values are derived from *positively-correlated* random variables. If negative correlation between variables is considered plausible, the p-values from the Benjamini-Yekutieli method [BY01] may be used instead. To compute the Simes p-values, let $p^{\text{go}}_{[d[1]]} < ... < p^{\text{go}}_{[d[V_d]]}$ be the p-values $p^{\text{go}}_{d[1]}, ..., p^{\text{go}}_{d[V_d]}$ sorted in ascending order. This induces the following policy:

$$G^{\text{ind.Simes}}_{(d),\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\tilde{\boldsymbol{\mu}},\tilde{\Sigma}_\mu,c) := \begin{cases} \text{stop} & G^{\text{ind}}_{(d),\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\tilde{\boldsymbol{\mu}},\tilde{\Sigma}_\mu,c) \ni \text{stop} \\ \text{go} & \text{at least 1 of } (p^{\text{go}}_{[d[1]]} \leq \frac{1}{V_d}\alpha^\star_{\text{FG}}, \\ & \qquad p^{\text{go}}_{[d[2]]} \leq \frac{2}{V_d}\alpha^\star_{\text{FG}}, \\ & \qquad ..., \\ & \qquad p^{\text{go}}_{[d[V_d]]} \leq \frac{V_d}{V_d}\alpha^\star_{\text{FG}} = \alpha^\star_{\text{FG}}) \end{cases} \tag{3.12}$$

Remark 3.13 (Ranked p-values as threshold predicates). Predicates in terms of ranked p-values can also be expressed in terms of thresholds. The predicate:

$$\text{at least 1 of } \left(p^{\text{go}}_{[d[1]]} \leq \frac{1}{V_d}\alpha^\star_{\text{FG}}, p^{\text{go}}_{[d[2]]} \leq \frac{2}{V_d}\alpha^\star_{\text{FG}}, ..., p^{\text{go}}_{[d[V_d]]} \leq \frac{V_d}{V_d}\alpha^\star_{\text{FG}}\right) \tag{3.14}$$

is equivalent to the predicate:

$$\begin{aligned} &\text{at least 1 of } (\tilde{\mu}_{d[i]} \geq \text{thr}^{\text{go}}_{d[i],\frac{1}{V_d}\alpha^\star_{\text{FG}}} \mid i = 1, ..., V_d) && (3.15) \\ &\quad \text{or at least 2 of } (\tilde{\mu}_{d[i]} \geq \text{thr}^{\text{go}}_{d[i],\frac{2}{V_d}\alpha^\star_{\text{FG}}} \mid i = 1, ..., V_d) \\ &\quad \text{or ...} \\ &\quad \text{or at least } (V_d - 1) \text{ of } (\tilde{\mu}_{d[i]} \geq \text{thr}^{\text{go}}_{d[i],\frac{V_d-1}{V_d}\alpha^\star_{\text{FG}}} \mid i = 1, ..., V_d) \\ &\quad \text{or all of } (\tilde{\mu}_{d[i]} \geq \text{thr}^{\text{go}}_{d[i],\alpha^\star_{\text{FG}}} \mid i = 1, ..., V_d) && (3.16) \end{aligned}$$

Note how the first line (3.15) corresponds to performing a Bonferroni correction, while the last line of the predicate (3.16) makes use of the original, uncorrected thresholds.

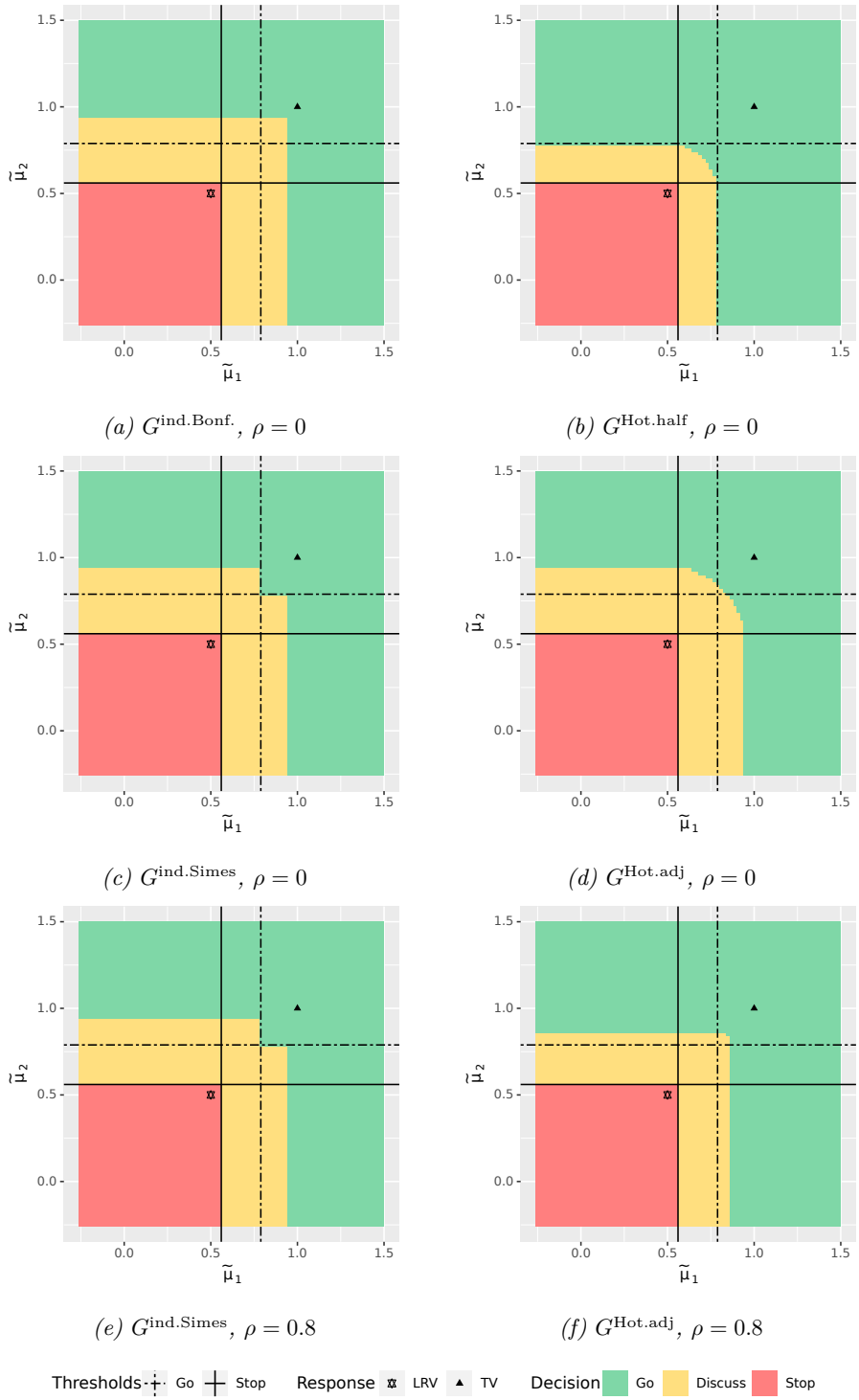
Proposition 3.17. *The policy $G^{\text{ind.Simes}}$ is monotone.*

Proof. By Remark 3.13, Proposition 2.46 and Proposition 2.27. □

Example 3.18 ($G^{\text{ind.Simes}}$ for two variables)**.** In the two variable case, the policy $G^{\text{ind.Simes}}$ adopts the following form:

$$G^{\text{ind.Simes}}_{\alpha^\star_{\text{FS}},\alpha^\star_{\text{FG}}}(\tilde{\boldsymbol{\mu}},\tilde{\Sigma}_\mu,c) := \begin{cases} \text{stop} & \tilde{\mu}_1 \leq \text{thr}^{\text{stop}}_{1,\alpha^\star_{\text{FS}}} \text{ and } \tilde{\mu}_2 \leq \text{thr}^{\text{stop}}_{2,\alpha^\star_{\text{FS}}} \\ \text{go} & \tilde{\mu}_1 \geq \text{thr}^{\text{go}}_{1,\alpha^\star_{\text{FG}}/2} \\ & \text{ or } \tilde{\mu}_2 \geq \text{thr}^{\text{go}}_{2,\alpha^\star_{\text{FG}}/2} \\ & \text{ or } (\tilde{\mu}_1 \geq \text{thr}^{\text{go}}_{1,\alpha^\star_{\text{FG}}} \text{ and } \tilde{\mu}_2 \geq \text{thr}^{\text{go}}_{2,\alpha^\star_{\text{FG}}}) \end{cases}$$

A graphical representation of this policy when $N = 17$ patients per arm and $c =$ Theoretical can be found in Figure 3.4c. ◀



(a) $G^{\text{ind.Bonf.}}$, $\rho = 0$

(b) $G^{\text{Hot.half}}$, $\rho = 0$

(c) $G^{\text{ind.Simes}}$, $\rho = 0$

(d) $G^{\text{Hot.adj}}$, $\rho = 0$

(e) $G^{\text{ind.Simes}}$, $\rho = 0.8$

(f) $G^{\text{Hot.adj}}$, $\rho = 0.8$

Figure 3.4: Go and Stop regions for adjusted multivariate policies with $V = 2$, $N = 17$. *The univariate* Go *and* Stop *thresholds are indicated for reference.*

3.2.3 Adjusting policies based on the T^2 statistic

Policies based on the T^2 statistic (e.g. G^{Hot} and $G^{\text{Hot.half}}$) measure the distance from $\tilde{\mu}$ to the regions Θ^{FS} and Θ^{FG}; that is:

$$T^2[\Theta; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu] := \inf_{\boldsymbol{\Delta} \in \Theta} (\boldsymbol{\Delta} - \tilde{\boldsymbol{\mu}})^\top \tilde{\Sigma}_\mu^{-1} (\boldsymbol{\Delta} - \tilde{\boldsymbol{\mu}})$$

If the regions under consideration are singular points ($\Theta^{\text{FS}} = \{\boldsymbol{\mu}_{\text{Stop}}\}$, $\Theta^{\text{FG}} = \{\boldsymbol{\mu}_{\text{Go}}\}$), the distribution of the statistics $T^2[\Theta^{\text{FS}}; \hat{\boldsymbol{\mu}}]$ and $T^2[\Theta^{\text{FG}}; \hat{\boldsymbol{\mu}}]$ are exactly as described in Proposition 2.69 for $\boldsymbol{\mu} = \boldsymbol{\mu}_{\text{Stop}}$ and $\boldsymbol{\mu} = \boldsymbol{\mu}_{\text{Go}}$, respectively. When the regions contain more points, there are two ways in which the distribution is affected:

(i) Points falling inside the region: If $\hat{\boldsymbol{\mu}} \in \Theta^{\text{FS}}$, then $T^2[\Theta^{\text{FS}}; \hat{\boldsymbol{\mu}}] = 0$. For example, if the region Θ^{FS} is an entire half space, and the true $\boldsymbol{\mu}$ lies on the boundary, then $T^2[\Theta^{\text{FS}}; \hat{\boldsymbol{\mu}}]$ will be 0 with probability $\frac{1}{2}$.

(ii) Decreasing degrees of freedom: With Θ^{FS} as a half space, the distance $T^2[\Theta^{\text{FS}}; \hat{\boldsymbol{\mu}}]$ varies only along an axis perpendicular to the boundary of Θ^{FS}. For example, if the covariance matrix is a known constant ($\tilde{\Sigma}_\mu := \Sigma_\mu$, $c =$ Theoretical) then the distribution of $T^2[\Theta^{\text{FS}}; \hat{\boldsymbol{\mu}}]$ is χ^2_1 (and not χ^2_V, as in the single point case).

Locally and almost everywhere on the boundary, the regions Θ^{FS} and Θ^{FG} can be approximated as a half spaces. This simplification yields the policy $G^{\text{Hot.half}}$:

$$G^{\text{Hot.half}}_{(d)}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) = \begin{cases} \text{Stop} & \frac{1}{2} Q^{-1}_{\text{adj},\delta:=1}(T^2[\Theta^{\text{FS}}; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu]; c) \leq \alpha^\star_{\text{FS}} \\ \text{Go} & \frac{1}{2} Q^{-1}_{\text{adj},\delta:=1}(T^2[\Theta^{\text{FG}}; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu]; c) \leq \alpha^\star_{\text{FG}} \end{cases}, \quad (3.19)$$

where $Q_{\text{adj},\delta}(\alpha; c)$ is defined as follows:

$$Q_{\text{adj},\delta}(\alpha; c) := \begin{cases} \frac{\delta(2N-2)}{2N-\delta-1} \text{F}_{\delta, 2N-\delta-1, 1-\alpha} & c = \text{Unstructured}_N \\ \chi^2_{\delta, 1-\alpha} & c = \text{Theoretical} \end{cases} \quad (3.20)$$

The constant δ is an estimate of the effective degrees of freedom for the T^2 statistic when minimized over a region Θ. In the equation (3.19), $\delta := 1$.

Remark. Given $Q_{\text{adj},\delta}$ as in Equation 3.20, $Q^{-1}_{\text{adj},\delta}$ is as follows:

$$Q^{-1}_{\text{adj},\delta}(x; c) := \begin{cases} 1 - F_{\text{F},\delta,2N-\delta-1}\left(\frac{2N-\delta-1}{\delta(2N-2)} x\right) & c = \text{Unstructured}_N \\ 1 - F_{\chi^2,\delta}(x) & c = \text{Theoretical} \end{cases}, \quad (3.21)$$

where $F_{\text{F},\delta_1,\delta_2}$ is the cumulative distribution function of the F distribution with δ_1 and δ_2 degrees of freedom, and $F_{\chi^2,\delta}$ is the c.d.f. of the χ^2 distribution with δ degrees of freedom.

The simplification underlying $G^{\text{Hot.half}}$ holds exactly for the case $V = 1$, $c =$ Theoretical (Proposition 2.77). However, this approximation fails to obtain the desired FGR in the corner case **LRV**. This can be seen in Figure 3.3,

where the FGR for region $\neg(\mathbf{LRV})$ and policy $G^{\text{Hot.half}}$ deviates from $\alpha^{\star}_{\text{FG}} = 0.2$ progressively more as the number of variables increases. This could be remedied by better addressing the situations described in (i) and (ii).

To correct for points falling inside a region Θ (situation (i)), one can estimate the minimum probability that $\hat{\boldsymbol{\mu}} \in \Theta$ given that $\boldsymbol{\mu} \in \Theta$. For a region Θ that fulfills the premises in Theorem 2.83 and a known Σ_{μ}:

$$A[\Theta] := \inf_{\boldsymbol{\Delta} \in \Theta} \mathbb{P}(\hat{\boldsymbol{\mu}} \in \Theta \mid \boldsymbol{\mu} = \boldsymbol{\Delta}), \tag{3.22}$$

If the region Θ is a half-plane, then $A = \frac{1}{2}$. Otherwise, if Θ is a finite union of upwards (or downwards) cones:

$$\Theta = \bigcup_{\boldsymbol{\Delta} \in \tilde{\Theta}} \neg(\boldsymbol{\Delta}) \quad \text{or} \quad \Theta = \bigcup_{\boldsymbol{\Delta} \in \tilde{\Theta}} \llcorner(\boldsymbol{\Delta}) \qquad (\tilde{\Theta} \subseteq \mathbb{R}^V, \tilde{\Theta} \text{ finite}),$$

the following value for A can be used:

$$A[\Theta; \tilde{\Sigma}_{\mu}, c] = \min_{\boldsymbol{\Delta} \in \tilde{\Theta}} M_{\nu(c)}(\Theta; \boldsymbol{\Delta}, \tilde{\Sigma}_{\mu}), \tag{3.23}$$

where M is defined as in Equation 3.9.

To correct for the decrease in the degrees of freedom (situation (ii)), the degrees of freedom δ may be estimated based on the local shape of the regions Θ^{FG} and Θ^{FS}. Consider the case $V = 2$, with $\Sigma_{\mu} = I_{2\times 2}$. This means that the T^2-statistic coincides with the Euclidean distance between $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{\mu}$. For $\Theta^{\text{FG}} = \neg(\mathbf{LRV})$, the maximum value of A is $A = 0.25$, which is reached at $\mathbf{LRV}$. Consider four quadrants centred at the point $\mathbf{LRV}$:

- In the quadrant $[\text{LRV}, +\infty) \times [\text{LRV}, +\infty)$, the squared distance to Θ^{FG} is the same as the squared distance to $\mathbf{LRV}$, which is proportional to the sum of the squares of the distances on both axes. This gives approximately 2 degrees of freedom.
- In the quadrants $(-\infty, \text{LRV}_1] \times [\text{LRV}_2, +\infty)$ and $[\text{LRV}_1, +\infty) \times (-\infty, \text{LRV}_2]$, the squared T^2-distance to Θ^{FG} is the proportional to the squared distance along one of the axes. This gives approximately 1 degree of freedom.
- The remaining quadrant is inside Θ^{FS}, which is already accounted for in the A factor.

The average degrees of freedom for the quadrants not contained in Θ^{FG} (i.e. the first three quadrants mentioned) is:

$$\delta = \frac{1}{3}(2 + 1 + 1) = \frac{4}{3}$$

It can be conjectured that this case analysis generalizes to numbers of variables $V > 2$ and other values of $\tilde{\Sigma}_{\mu}$. More specifically, let:

$$D := -\log_2(A)$$
$$\delta(A) := \frac{A}{1 - A} \sum_{d=0}^{D} d \binom{D}{d}$$

Table 3.4: Approximate degrees of freedom for $G^{\text{Hot.adj}}$, with A as defined in (3.22)

Area (A)	$D := -\log_2(A)$	$\delta(A)$
$1 = 2^0$	0	0
$0.5 = 2^{-1}$	1	1
$0.25 = 2^{-2}$	2	$\frac{1}{3}\left(1\binom{2}{1} + 2\binom{2}{2}\right) = \frac{4}{3} \approxeq 1.333$
$0.125 = 2^{-3}$	3	$\frac{1}{7}\left(1\binom{3}{1} + 2\binom{3}{2} + 3\binom{3}{3}\right) = \frac{12}{7} \approxeq 1.714$
$0.0625 = 2^{-4}$	4	$\frac{1}{15}\left(1\binom{4}{1} + 2\binom{4}{2} + 3\binom{4}{3} + 4\binom{4}{4}\right) = \frac{32}{15} \approxeq 2.133$

The values of δ for $D = 0, 1, 2, 3, 4$ are given in Table 3.4. When $-\log_2(A)$ is not a natural number, spline interpolation between the points $(d, \delta(2^{-d}))$ is performed.

$$G^{\text{Hot.adj}}_{(d)}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu, c) = \begin{cases} \text{stop} & A_{\text{FS}} Q^{-1}_{\text{adj},\delta_{\text{FS}}}(T^2[\Theta^{\text{FS}}; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu]; c) \leq \alpha^\star_{\text{FS}} \\ \text{go} & A_{\text{FG}} Q^{-1}_{\text{adj},\delta_{\text{FG}}}(T^2[\Theta^{\text{FG}}; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu]; c) \leq \alpha^\star_{\text{FG}} \end{cases},$$

where $A_{\text{FS}} := A[\Theta^{\text{FS}}; \tilde{\Sigma}_\mu, c]$ $A_{\text{FG}} := A[\Theta^{\text{FG}}; \tilde{\Sigma}_\mu, c]$ (see Equation 3.23), $\delta_{\text{FS}} := \delta(A_{\text{FS}})$ and $\delta_{\text{FG}} := \delta(A_{\text{FG}})$. The values of A_{FS}, A_{FG}, δ_{FS} and δ_{FG} depend only on $\tilde{\Sigma}_\mu$, and not on $\tilde{\boldsymbol{\mu}}$. If $\tilde{\Sigma}_\mu := \Sigma_\mu$, they are constants; otherwise, if $\tilde{\Sigma}_\mu := \hat{\Sigma}_\mu$, they are random variables independent from $\hat{\boldsymbol{\mu}}$. In any of these cases, the monotonicity of $G^{\text{Hot.half}}$ and $G^{\text{Hot.adj}}$ follows by the same reasoning as that of G^{Hot}.

The validity of this conjecture (that is, the ability of the policy $G^{\text{Hot.adj}}$ to control the FGR and the FSR) is validated empirically in §3.2.4.

3.2.4 Evaluation of adjusted policies

In Table 3.5 the adjusted policies are evaluated in the same conditions as the policies in Table 3.3. Those adjusted policies which successfully control the FGR are also evaluated when the standard deviation of a single variable is doubled (thus decreasing its power), in order to check whether their suitability persists when variances are heterogeneous. The results are summarized in Figure 3.5.

As expected from the analysis preceding the definition of $G^{\text{Hot.adj}}$, the $G^{\text{Hot.half}}$ policy underestimates the degrees of freedom in corner cases. When the true effect is **LRV**, the FGR is much higher than $\alpha^\star_{\text{FG}}$ (if $V = 4$ and $\rho = 0$, $\text{FGR} = 0.64 \gg 0.2 = \alpha^\star_{\text{FG}}$). Therefore, the $G^{\text{Hot.half}}$ policy will be ignored in the remainder of the analysis.

The policies $G^{\text{ind.Simes}}$ and $G^{\text{ind.Bonf.}}$ (and to a limited extent, $G^{\text{Hot.adj}}$) all succeed at bounding the FSR at $\Theta^{\text{FS}}_0 = R^{\text{Go}}_{\text{ext}}$. The difference between the numbers in the FSR column for effects in the Θ^{FS}_0 region and the parameter $\alpha^\star_{\text{FS}}$ is within the estimated error for the simulation. As for the False Go Risk, the FGR at $\neg(\mathbf{LRV})$ stays below the the $\alpha^\star_{\text{FG}}$ for $G^{\text{ind.Simes}}$ and $G^{\text{ind.Bonf.}}$, although they are too conservative when the correlation is non-zero ($\rho = 0.4, 0.8$). This is specially true for the latter policy. The policy $G^{\text{Hot.adj}}$ produces an FGR in the $\neg(\mathbf{LRV})$ region which is above the $\alpha^\star_{\text{FG}}$ by

Table 3.5: Comparison of adjusted policies. In bold, the best values for each combination of number of variables (V), correlation (ρ), region (top header) and metric (column). $N = 17$, $\alpha^{\star}_{\text{FS}} = 0.1$, $\alpha^{\star}_{\text{FG}} = 0.2$, and $\sigma_1 = ... = \sigma_V = 1$.

		$\Theta_0^{\text{FS}} = R_{\text{ext}}^{\text{Go}}$		$R_{\text{ext}}^{\text{Stop}}$		$\llcorner(\mathbf{TV})$	$\neg(\mathbf{LRV})$	$\neg(\mathbf{0})$
	ρ	FSR	CGr	FGr	CSr	CGr	FGR	CSr
V=2								
$G^{\text{ind.Simes}}$	0	**0.10**	0.57	0.11	**0.53**	0.83	0.20	**0.89**
	0.4	**0.10**	0.56	**0.10**	**0.55**	0.77	0.19	**0.90**
	0.8	**0.10**	0.56	**0.10**	**0.56**	0.71	0.17	**0.92**
$G^{\text{ind.Bonf.}}$	0	**0.10**	0.57	0.11	**0.53**	0.81	**0.19**	**0.89**
	0.4	**0.10**	0.56	**0.10**	**0.55**	0.74	**0.17**	**0.90**
	0.8	**0.10**	0.56	**0.10**	**0.56**	0.67	**0.15**	**0.92**
$G^{\text{Hot.half}}$	0	**0.10**	**0.73**	0.22	**0.53**	**0.94**	0.37	**0.89**
	0.4	**0.10**	**0.73**	0.21	**0.55**	**0.89**	0.33	**0.90**
	0.8	**0.10**	**0.73**	0.20	**0.56**	**0.81**	0.27	**0.92**
$G^{\text{Hot.adj}}$	0	**0.10**	0.56	**0.10**	**0.53**	0.85	0.20	**0.89**
	0.4	**0.10**	0.60	0.12	**0.55**	0.79	0.20	**0.90**
	0.8	**0.10**	0.65	0.15	**0.56**	0.75	0.20	**0.92**
V=3								
$G^{\text{ind.Simes}}$	0	**0.09**	0.47	0.07	**0.50**	0.89	0.20	**0.84**
	0.4	**0.10**	0.47	**0.07**	**0.54**	0.80	0.19	0.86
	0.8	**0.10**	0.46	**0.07**	**0.57**	0.70	0.16	**0.90**
$G^{\text{ind.Bonf.}}$	0	**0.09**	0.47	0.07	**0.50**	0.85	**0.19**	**0.84**
	0.4	**0.10**	0.47	**0.07**	**0.54**	0.75	**0.17**	0.86
	0.8	**0.10**	0.46	**0.07**	**0.57**	0.62	**0.13**	**0.90**
$G^{\text{Hot.half}}$	0	**0.09**	**0.74**	0.22	**0.50**	**0.99**	0.53	**0.84**
	0.4	**0.10**	**0.73**	0.21	**0.54**	**0.94**	0.42	0.86
	0.8	**0.10**	**0.72**	0.20	**0.57**	**0.85**	0.32	**0.90**
$G^{\text{Hot.adj}}$	0	**0.09**	0.43	**0.06**	**0.50**	0.92	0.21	**0.84**
	0.4	**0.10**	0.53	0.09	0.53	0.83	0.22	**0.87**
	0.8	**0.10**	0.61	0.13	0.56	0.77	0.22	**0.90**
V=4								
$G^{\text{ind.Simes}}$	0	**0.09**	0.41	0.06	**0.47**	0.92	0.20	**0.79**
	0.4	**0.10**	0.41	0.06	**0.52**	0.80	0.18	**0.83**
	0.8	**0.10**	0.41	**0.05**	**0.56**	0.69	0.15	0.88
$G^{\text{ind.Bonf.}}$	0	**0.09**	0.41	0.06	**0.47**	0.88	**0.18**	**0.79**
	0.4	**0.10**	0.41	**0.05**	**0.52**	0.76	**0.16**	**0.83**
	0.8	**0.10**	0.41	**0.05**	**0.56**	0.59	**0.11**	0.88
$G^{\text{Hot.half}}$	0	**0.09**	**0.74**	0.24	**0.47**	**1.00**	0.65	**0.79**
	0.4	**0.10**	**0.73**	0.22	**0.52**	**0.96**	0.49	**0.83**
	0.8	**0.10**	**0.72**	0.20	**0.56**	**0.87**	0.35	0.88
$G^{\text{Hot.adj}}$	0	**0.09**	0.35	**0.04**	**0.47**	0.95	0.21	**0.79**
	0.4	**0.10**	0.48	0.08	**0.52**	0.85	0.24	**0.83**
	0.8	**0.10**	0.59	0.12	0.55	0.77	0.22	**0.89**

up to 4 percentage points for the case $G^{\text{Hot.adj}}$ (e.g. for $V = 4$, $\rho = 0.4$ the FGR $= 0.24 > 0.2 = \alpha^{\star}_{\text{FG}}$). This points to an underestimation of the degrees of freedom of the statistics used to build the policy. A better understanding of the statistical properties of $T^2[\Theta^{\text{FS}}]$ and $T^2[\Theta^{\text{FG}}]$, on which $G^{\text{Hot.adj}}$ is based, could remove this difference.

When it comes to the CGr at $\llcorner(\mathbf{TV})$, $G^{\text{ind.Simes}}$ consistently outperforms $G^{\text{ind.Bonf.}}$ by around 4 percentage points. In the case of $R^{\text{Go}}_{\text{ext}}$, the performance of $G^{\text{ind.Simes}}$ is similar to that of $G^{\text{ind.Bonf.}}$, as in both cases there is a single endpoint which has a reasonable chance of producing an effect above the threshold for a Go decision.

As for $G^{\text{Hot.adj}}$, its metrics are remarkably consistent regardless of the actual covariance structure of the variables involved, and even improve on $G^{\text{ind.Simes}}$ when the correlation between endpoints is high. For example, in Table 3.5, when $N = 17$, $\rho = 0.8$ and $V = 4$, the CGr at $R^{\text{Go}}_{\text{ext}}$ is 18 percentage points higher for the former policy (0.59 for $G^{\text{Hot.adj}}$ vs. 0.41 for $G^{\text{ind.Simes}}$). When the correlation is low (e.g. $\rho = 0$) the policy $G^{\text{ind.Simes}}$ has a small advantage over $G^{\text{Hot.adj}}$ (for example, if $N = 17$, $V = 4$ and $\rho = 0$, the CGr at $R^{\text{Go}}_{\text{ext}}$ is 6 percentage points higher for $G^{\text{ind.Simes}}$; 0.41 vs. 0.35 for $G^{\text{Hot.adj}}$). For $\llcorner(\mathbf{TV})$, the difference is milder but favours $G^{\text{ind.Simes}}$.

The ability of these three policies to control the FSR and FGR is preserved even in scenarios where some variables have lower power than others (see Table 3.6).

In view of these results, both $G^{\text{ind.Simes}}$ and $G^{\text{Hot.adj}}$ are reasonable candidates for a domain-level policy. Regarding $G^{\text{Hot.adj}}$, this policy looks particularly promising for high levels of within-domain correlation, (e.g. $\rho = 0.8$), but such high levels of correlation may be unlikely to arise in practice. Additionally, the statistical properties of the Simes p-values used in $G^{\text{ind.Simes}}$ are better understood, specially compared to those of the ad-hoc multiplicity adjustment that is used in $G^{\text{Hot.adj}}$ (§3.2.3). Finally, as opposed to $G^{\text{Hot.adj}}$ (and like the simpler $G^{\text{ind.Bonf.}}$) the $G^{\text{ind.Simes}}$ policy can be formulated in terms of Go and Stop thresholds (Remark 3.13), which may make the policy more understandable to all stakeholders. Based on these results, the policy $G^{\text{ind.Simes}}$ will be used as a domain-level policy for the remaining of the analysis.

3.2.5 Sensitivity of domain-level decisions to the addition of a low-powered variable

When designing a study, it may be necessary to decide whether to include an additional, low-powered variable into an existing domain. In order to assess the effect of such a decision, a study with a single domain $D = 1$ and $V = \nu$ variables with 0.8 power is considered, and the effect of adding an additional variable with smaller power is analyzed (for a total of $V = \nu + 1$ endpoints). As Figure 3.6 shows, adding an extra variable with low power results in a large decrease in the probability of a correct Stop decision in the case where there is no effect ($\boldsymbol{\mu} = \mathbf{0}$).

As for those cases where a Go decision is desired, due to the multiple comparisons adjustment that is needed to keep the FGR bounded, the probability of Go when $\boldsymbol{\mu} = \mathbf{TV}$ will decrease if the power of the new variable is low enough (see the $\mathbf{TV}$ column), compared to the case where the variable is not

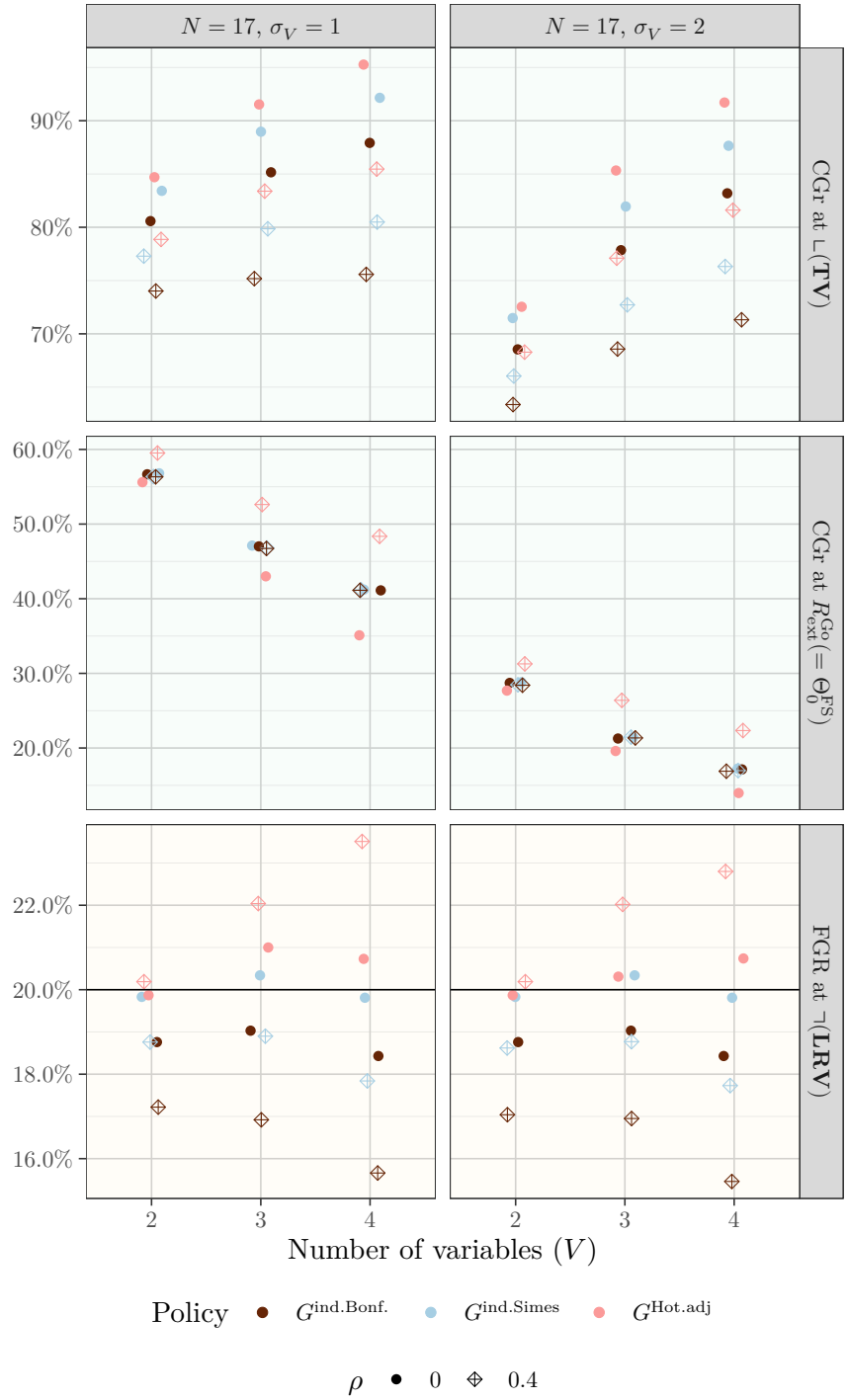


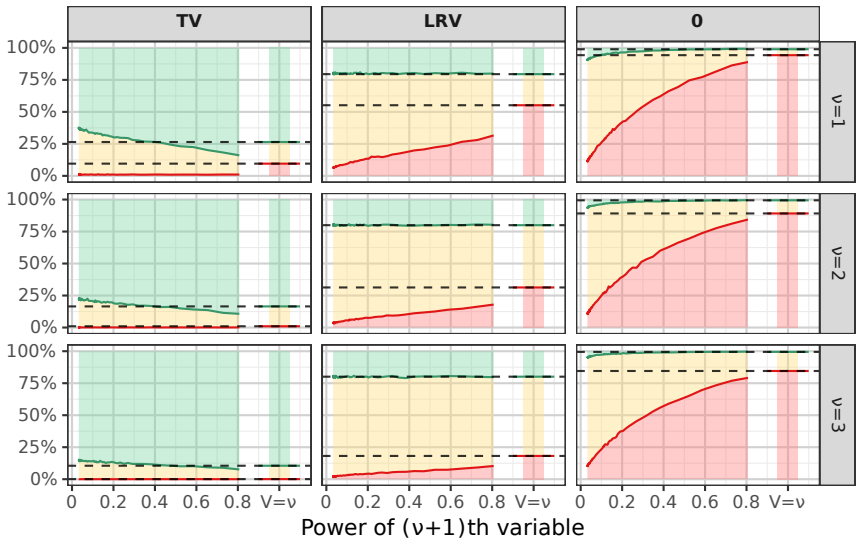
Figure 3.5: Comparison of Go *rates for the adjusted policies, for a single domain.* $N = 17$, $\alpha^{\star}_{\mathrm{FS}} = 0.1$, $\alpha^{\star}_{\mathrm{FG}} = 0.2$. *The background is shaded according to the expected decision for the Correct Go rate (*CGr*) metrics, and in the "Discuss" color for the False Go Risk (*FGR*). For the latter metric, the 20% threshold is marked with a horizontal line.*

Table 3.6: Comparison of selected adjusted policies in the presence of one underpowered variable. In bold, the best performing policy for each combination of number of variables (V), the assumed correlation (ρ), the region for the metric (top header) and the metric (column). $N = 17$, $\alpha^{\star}_{\text{FS}} = 0.1$, $\alpha^{\star}_{\text{FG}} = 0.2$, $\sigma_1 = ... = \sigma_{V-1} = 1$, $\sigma_V = 2$.

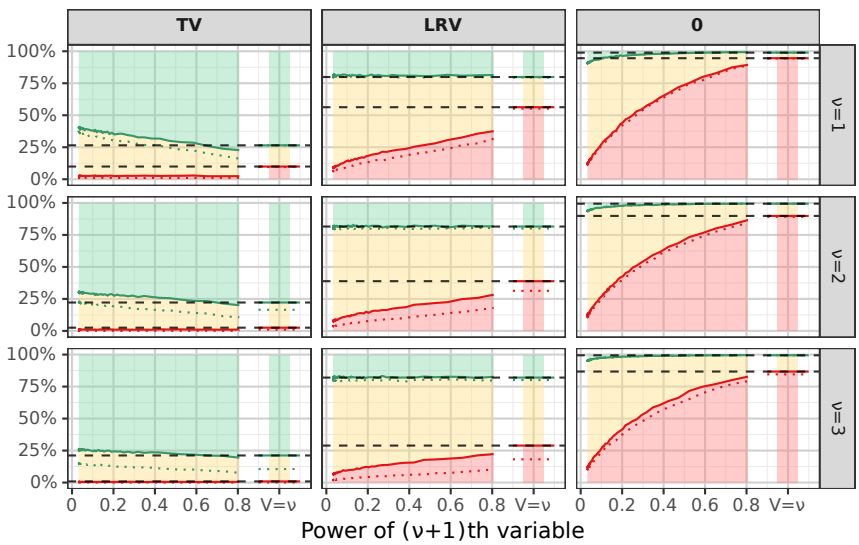
		$\Theta_0^{\text{FS}} = R_{\text{ext}}^{\text{Go}}$		$R_{\text{ext}}^{\text{Stop}}$		$\llcorner(\mathbf{TV})$	$\neg(\mathbf{LRV})$	$\neg(\mathbf{0})$
	ρ	FSR	CGr	FGr	CSr	CGr	FGR	CSr
V=2								
$G^{\text{ind.Simes}}$	0	**0.09**	**0.29**	**0.12**	**0.26**	0.71	0.20	**0.53**
	0.4	**0.10**	0.28	0.12	**0.28**	0.66	0.19	**0.55**
	0.8	**0.10**	0.28	0.11	**0.28**	0.61	0.17	**0.56**
$G^{\text{ind.Bonf.}}$	0	**0.09**	**0.29**	**0.12**	**0.26**	0.69	**0.19**	**0.53**
	0.4	**0.10**	0.28	**0.11**	**0.28**	0.63	**0.17**	**0.55**
	0.8	**0.10**	0.28	**0.10**	**0.28**	0.58	**0.15**	**0.56**
$G^{\text{Hot.adj}}$	0	**0.09**	0.28	**0.12**	**0.26**	**0.73**	0.20	**0.53**
	0.4	**0.10**	**0.31**	0.13	**0.28**	**0.68**	0.20	**0.55**
	0.8	**0.10**	**0.37**	0.15	**0.28**	**0.67**	0.20	**0.56**
V=3								
$G^{\text{ind.Simes}}$	0	**0.09**	**0.21**	0.08	**0.26**	0.82	0.20	**0.50**
	0.4	**0.10**	0.21	**0.08**	**0.28**	0.73	0.19	0.53
	0.8	**0.10**	0.21	**0.07**	**0.29**	0.63	0.16	0.56
$G^{\text{ind.Bonf.}}$	0	**0.09**	**0.21**	0.08	**0.26**	0.78	**0.19**	**0.50**
	0.4	**0.10**	0.21	**0.08**	**0.28**	0.69	**0.17**	0.53
	0.8	**0.10**	0.21	**0.07**	**0.29**	0.58	**0.12**	0.56
$G^{\text{Hot.adj}}$	0	**0.09**	0.20	**0.07**	0.25	**0.85**	0.20	**0.50**
	0.4	**0.10**	**0.26**	0.11	0.27	**0.77**	0.22	**0.54**
	0.8	**0.10**	**0.33**	0.13	0.28	**0.72**	0.20	**0.57**
V=4								
$G^{\text{ind.Simes}}$	0	**0.08**	**0.17**	0.07	**0.25**	0.88	0.20	0.47
	0.4	**0.10**	0.17	**0.06**	**0.28**	0.76	0.18	**0.53**
	0.8	**0.10**	0.17	0.06	**0.29**	0.64	0.15	**0.56**
$G^{\text{ind.Bonf.}}$	0	**0.08**	**0.17**	0.06	**0.25**	0.83	**0.18**	0.47
	0.4	**0.10**	0.17	**0.06**	**0.28**	0.71	**0.15**	**0.53**
	0.8	**0.10**	0.17	**0.05**	**0.29**	0.56	**0.11**	**0.56**
$G^{\text{Hot.adj}}$	0	**0.08**	0.14	**0.05**	0.24	**0.92**	0.21	**0.48**
	0.4	**0.10**	**0.22**	0.08	0.27	**0.82**	0.23	**0.53**
	0.8	**0.10**	**0.31**	0.12	0.28	**0.75**	0.22	**0.56**

Figure 3.6: Impact of adding a variable of power ≤ 0.8 on the domain-level decision probabilities. For each combination of $\nu = 1, 2, 3$ and effect ($\boldsymbol{\mu} = \mathbf{TV}, \mathbf{LRV}, \mathbf{0}$): on the right, a bar with the probabilities of each decision with $V = \nu$ variables with 0.8 power; on the left, the probabilities of each decision when a variable is added (for a total of $V = \nu + 1$ variables), depending on the power of said variable.

(a) Correlation among variables: $\rho = 0$



(b) Correlation $\rho = 0.4$. Values from (a) redrawn as the guide "$\rho = 0$".



included. In all cases, the relative advantages and disadvantages of adding a new, low-powered variable become less pronounced when the correlation increases (compare Figure 3.6a and Figure 3.6b). Adding a variable may be justified if it has higher power than other variables in the same domain, or if it covers an aspect of the disease that is not sufficiently reflected in other endpoints in the same domain.

The impact on the an overall policy of adding a new, low-power variable is explored in §3.3.7. If an endpoint is important enough to be considered on its own, without regards to the effects of multiple comparisons, it may be suitable to include this endpoint in a new domain of its own. The impact of this decision is explored in §3.3.6.

3.3 Evaluation on multiple domains

In §3.2, different ways of combining multiple endpoints within the same domain to achieve a Go, Discuss or Stop decision for that domain are presented. Among them, the subpolicy $G_{(d)}^{\text{ind.Simes}}$ emerges as a suitable alternative, due to its relatively familiar formulation, ability to reliably bound the FGR and FSR, and higher Correct Go rate (CGr) compared to simpler approaches such as $G^{\text{ind.Bonf.}}$.

In this section we address how the decisions obtained for the different domains can be combined into a single overall decision. Using the subpolicies $G_{(d)}$ for each domain $d = 1, ..., D$, two overall policies are defined. The first policy considers all domains with equal weight, and can be paraphrased as follows:

Definition 3.24 ("All domains equal" policy)**.** Let the Go status of each domain d be determined according to a subpolicy $G_{(d)}$. Then, for $x > z \geq 0$:

(i) The policy produces Stop if at most z domains are Go.

(ii) The policy produces Go if at least x domains are Go and no variable shows a statistically significant, negative effect.

The "all domains equal" policy is implemented as a policy with three parameters (x, z and α):

$$G_{x,z,\alpha}^{\text{all.eq}}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu) := \begin{cases} \text{stop} & \text{at most } z \text{ of } (G_{(d)}(\tilde{\boldsymbol{\mu}}_{I(d)}, (\tilde{\Sigma}_\mu)_{I(d),I(d)}) \text{ is Go} \\ & \qquad \mid d = 1, ..., D) \\ \text{go} & \text{at least } x \text{ of } (G_{(d)}(\tilde{\boldsymbol{\mu}}_{I(d)}, (\tilde{\Sigma}_\mu)_{I(d),I(d)}) \text{ is Go} \\ & \qquad \mid d = 1, ..., D) \\ & \text{and none of } (P_{(i),\alpha}^{\neg,\text{neg}}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu) \mid i = 1, ..., V), \end{cases} \tag{3.25}$$

where the predicate $P_\alpha^{\neg,\text{neg}}$ is defined as in Example 2.28, and $I(d)$ is the set of variable indices corresponding to endpoints in domain d (Equation 3.8).

In other situations there may be a domain which is of particular significance, and could on its own determine whether the drug candidate should go ahead into the next phase. In this case, a hierarchical policy such as the following may be suitable:

Definition 3.26 ("Hierarchical domains" policy)**.** Let the Go/Stop status of each domain d be determined according to a subpolicy $G_{(d)}$. Then:

(i) The policy produces Stop if the most important domain is Stop, and at least 2 other domains are Stop.

(ii) The policy produces Go if the most important domain is Go or at least x (of the other) domains are Go; and also no variable shows a statistically significant, negative effect.

If both conditions hold, the Stop decision has priority.

Without loss of generality, it can be assumed that the most important domain is the first domain. Then the "hierarchical domains" policy can be implemented as follows:

$$G^{\text{hier}}_{x,\alpha}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu) := \begin{cases} \text{stop} & G_{(1)}(\tilde{\boldsymbol{\mu}}_{I(1)}, (\tilde{\Sigma}_\mu)_{I(1),I(1)}) \text{ is Stop and at least 2} \\ & \text{of } (G_{(d)}(\boldsymbol{\mu}_{I(d)}, (\tilde{\Sigma}_\mu)_{I(d),I(d)}) \text{ is Stop} \mid d = 2, ..., D) \\ \text{go} & \begin{pmatrix} G_{(1)}(\tilde{\boldsymbol{\mu}}_{I(1)}, (\tilde{\Sigma}_\mu)_{I(1),I(1)}) \text{ is Go or} \\ \text{at least } x \text{ of } (G_{(d)}(\tilde{\boldsymbol{\mu}}_{I(d)}, (\tilde{\Sigma}_\mu)_{I(d),I(d)}) \text{ is Go} \\ \mid d = 2, ..., D) \end{pmatrix} \\ & \text{and none of } (P^{\neg,\text{neg}}_{(i),\alpha}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}_\mu) \mid i = 1, ..., V) \end{cases} \tag{3.27}$$

Following the conclusions of §3.2.4, the subpolicy $G_{(d)}$ is defined as $G_{(d)} := G^{\text{ind.Simes}}_{(d),\alpha^\star_{\text{FS}}:=0.1,\alpha^\star_{\text{FG}}:=0.2}$. This means that the FSR of $G_{(d)}$ at each individual domain d coincides with $\alpha^\star_{\text{FS}} = 0.1$, and the FGR of $G_{(d)}$ at each individual domain d coincides $\alpha^\star_{\text{FG}} = 0.2$. The rates of Go and Stop for the policy implementations $G^{\text{all.eq}}_{x,z,\alpha}$ and $G^{\text{hier}}_{x,\alpha}$ depend on the chosen values for x, z and α, and can be computed using simulations.

3.3.1 Scope

The analysis evaluates the policies with parameters set to $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ and $G^{\text{hier}}_{x:=2,\alpha:=0.05}$. These values were chosen so that the resulting policies are applicable to all scenarios with $D \geq 2$ ($D \geq 3$ in the case of G^{hier}).

In the performed simulations, the number of domains is restricted to $D \in \{2, 3, 4, 5\}$; in the case of $G^{\text{hier}}_{x:=2,\alpha:=0.05}$, only $D \in \{3, 4, 5\}$ are considered. The number of variables in a domain is restricted to vary between 1 and 4, with a maximum of 10 total variables. Therefore, the total number of variables is $V \in \{D, ..., \min(4D, 10)\}$. For each D and V, all combinations of V_1, ..., V_D such that $(V_1 + V_2 + ... + V_D = V)$ and $1 \leq V_d \leq 4$ for all $d = 1, ..., D$ are simulated (where V_d is the number of variables in domain d).

For the sake of performance, a single representative is chosen for those combinations that differ only in the ordering of V_2,...,V_d; as such scenarios are symmetrical. Note that the value of V_1 must be considered separately due to its distinct handling in $G^{\text{hier}}_{x:=2,\alpha:=0.05}$.

The remaining constraints for the scenario are set as in §3.1: The number of patients per arm is fixed at $N = 17$, and $\sigma_1 = ... = \sigma_V = 1$; which means that

Table 3.7: Results for the policy $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$, for $N = 17$. For each combination of number of domains (D), number of variables (V) region (top header) and metric (column), the worst case for the value of that metric among all possible arrangements of variables per domain ($V_1 + ... + V_D = V$) is shown; together with, in a smaller font, the difference relative to the best case.

(a) No correlation between endpoints: $\rho = \tau = 0$

	⌞(**TV**)		¬(**LRV**)		¬(**0**)	
V	CGr	FSr	FGr	CSr	CSr	FGr
D=2						
2	0.53 +0.00	0.07 -0.00	0.04 -0.00	0.64 -0.00	0.98 +0.00	0.00 -0.00
5	0.66 +0.09	0.02 -0.01	0.04 -0.00	0.65 -0.01	0.98 +0.01	0.00 -0.00
8	0.85 +0.00	0.00 -0.00	0.04 -0.00	0.64 -0.00	0.99 +0.00	0.00 -0.00
D=3						
3	0.82 +0.00	0.02 -0.00	0.11 -0.00	0.51 -0.00	0.97 +0.00	0.00 -0.00
7	0.93 +0.01	0.00 -0.00	0.11 -0.00	0.52 -0.01	0.98 +0.01	0.00 -0.00
10	0.97 +0.01	0.00 -0.00	0.11 -0.01	0.51 -0.01	0.98 +0.00	0.00 -0.00
D=4						
4	0.94 +0.00	0.00 -0.00	0.17 -0.00	0.41 -0.00	0.96 +0.00	0.00 -0.00
7	0.97 +0.01	0.00 -0.00	0.19 -0.01	0.41 -0.01	0.96 +0.01	0.00 -0.00
10	0.99 +0.01	0.00 -0.00	0.18 -0.01	0.42 -0.02	0.97 +0.01	0.00 -0.00
D=5						
5	0.98 +0.00	0.00 -0.00	0.26 -0.00	0.33 -0.00	0.95 +0.00	0.00 -0.00
7	0.99 +0.00	0.00 -0.00	0.27 -0.02	0.34 -0.02	0.95 +0.01	0.00 -0.00
10	0.99 +0.00	0.00 -0.00	0.27 -0.01	0.34 -0.02	0.96 +0.01	0.00 -0.00

(b) Correlation inside domain: $\rho = 0.4$. Correlation between domains: $\tau = 0.2$.

	⌞(**TV**)		¬(**LRV**)		¬(**0**)	
V	CGr	FSr	FGr	CSr	CSr	FGr
D=2						
2	0.55 +0.00	0.10 -0.00	0.05 -0.00	0.66 -0.00	0.98 +0.00	0.00 -0.00
5	0.62 +0.03	0.08 -0.01	0.06 -0.01	0.68 -0.01	0.98 +0.00	0.00 -0.00
8	0.68 +0.00	0.07 -0.00	0.06 -0.00	0.70 -0.00	0.99 +0.00	0.00 -0.00
D=3						
3	0.79 +0.00	0.04 -0.00	0.13 -0.00	0.55 -0.00	0.97 +0.00	0.00 -0.00
7	0.83 +0.01	0.03 -0.00	0.13 -0.01	0.59 -0.02	0.98 +0.01	0.00 -0.00
10	0.85 +0.01	0.03 -0.00	0.12 -0.00	0.60 -0.00	0.98 +0.00	0.00 -0.00
D=4						
4	0.89 +0.00	0.02 -0.00	0.20 -0.00	0.47 -0.00	0.96 +0.00	0.00 -0.00
7	0.90 +0.01	0.02 -0.00	0.20 -0.01	0.50 -0.02	0.96 +0.01	0.00 -0.00
10	0.92 +0.01	0.02 -0.01	0.20 -0.01	0.52 -0.02	0.97 +0.01	0.00 -0.00
D=5						
5	0.95 +0.00	0.01 -0.00	0.28 -0.00	0.41 -0.00	0.95 +0.00	0.00 -0.00
7	0.95 +0.00	0.01 -0.00	0.27 -0.00	0.42 -0.01	0.95 +0.01	0.00 -0.00
10	0.95 +0.01	0.01 -0.00	0.26 -0.01	0.45 -0.02	0.96 +0.01	0.00 -0.00

Table 3.8: Metrics for the policy $G^{\text{hier}}_{x:=2,\alpha:=0.05}$. For each arrangement of V variables into D domains, the worst value for the metric is shown, together with, in a smaller font, the difference relative to the best value.

(a) Metrics for the policy $G^{\text{hier}}_{x:=2,\alpha:=0.05}$, for $\rho = 0$, $\tau = 0$, $N = 17$.

	$\llcorner(\mathbf{TV})$		$\neg(\mathbf{LRV})$		$\neg(\mathbf{0})$	
V	CGr	FSr	FGr	CSr	CSr	FGr
D=3						
3	0.87 +0.00	0.00 -0.00	0.23 -0.00	0.17 -0.00	0.85 +0.00	0.01 -0.00
5	0.90 +0.04	0.00 -0.00	0.23 -0.00	0.06 -0.00	0.75 +0.00	0.01 -0.01
8	0.95 +0.03	0.00 -0.00	0.23 -0.01	0.01 -0.00	0.63 +0.01	0.01 -0.00
10	0.97 +0.01	0.00 -0.00	0.23 -0.01	0.00 -0.00	0.56 +0.01	0.00 -0.00
D=4						
4	0.95 +0.00	0.00 -0.00	0.27 -0.00	0.34 -0.00	0.93 +0.00	0.01 -0.00
6	0.97 +0.01	0.00 -0.00	0.29 -0.01	0.22 -0.12	0.84 +0.09	0.01 -0.00
8	0.98 +0.01	0.00 -0.00	0.29 -0.01	0.13 -0.08	0.78 +0.13	0.01 -0.00
10	0.99 +0.01	0.00 -0.00	0.29 -0.01	0.06 -0.04	0.77 +0.12	0.01 -0.01
D=5						
5	0.98 +0.00	0.01 -0.00	0.30 -0.00	0.44 -0.00	0.94 +0.00	0.01 -0.00
8	0.99 +0.01	0.00 -0.00	0.33 -0.01	0.35 -0.27	0.80 +0.15	0.01 -0.01
10	1.00 +0.00	0.00 -0.00	0.34 -0.02	0.25 -0.19	0.79 +0.16	0.01 -0.01

(b) Metrics for the policy $G^{\text{hier}}_{x:=2,\alpha:=0.05}$, for $\rho = 0.4$, $\tau = 0.2$, $N = 17$.

V	CGr	FSr	FGr	CSr	CSr	FGr
D=3						
3	0.85 +0.00	0.00 -0.00	0.24 -0.00	0.23 -0.00	0.85 +0.00	0.01 -0.00
5	0.86 +0.02	0.00 -0.00	0.24 -0.02	0.14 -0.01	0.78 +0.01	0.01 -0.01
8	0.87 +0.02	0.00 -0.00	0.23 -0.02	0.07 -0.01	0.70 +0.01	0.01 -0.01
10	0.89 +0.00	0.00 -0.00	0.22 -0.02	0.05 -0.01	0.66 +0.01	0.01 -0.00
D=4						
4	0.91 +0.00	0.01 -0.00	0.29 -0.00	0.37 -0.00	0.93 +0.00	0.01 -0.00
6	0.92 +0.01	0.00 -0.00	0.28 -0.01	0.31 -0.11	0.84 +0.08	0.01 -0.01
8	0.93 +0.01	0.00 -0.00	0.28 -0.02	0.23 -0.08	0.82 +0.08	0.01 -0.01
10	0.93 +0.01	0.00 -0.00	0.28 -0.03	0.19 -0.07	0.80 +0.09	0.01 -0.01
D=5						
5	0.95 +0.00	0.01 -0.00	0.31 -0.00	0.44 -0.00	0.95 +0.00	0.01 -0.00
8	0.95 +0.01	0.01 -0.01	0.33 -0.02	0.40 -0.20	0.83 +0.11	0.01 -0.01
10	0.96 +0.01	0.00 -0.00	0.32 -0.02	0.35 -0.17	0.83 +0.11	0.01 -0.01

the power of each individual variable is 0.8. A correlation for variables within a domain of either $\rho := 0$ or $\rho := 0.4$ is chosen, as being representative of the range that can be realistically expected in practice. The correlation between variables in different domains is set to either $\tau := 0$ or $\tau := 0.2$ (subject to the assumption that $\tau < \rho$). The policies are initially evaluated at the following true effects and metrics:

(i) For $\boldsymbol{\mu} \in \llcorner(\mathbf{TV})$, the CGr and FSr metrics.

(ii) For $\boldsymbol{\mu} \in \neg(\mathbf{LRV})$, the FGR metric.

(iii) For $\boldsymbol{\mu} \in \neg(\mathbf{0})$, the CSr and FGr metrics.

The results of the simulations are summarized in Table 3.7 and Table 3.8. In each cell, the worst case for that metric among all possible combinations of $V_1 + ... + V_D = V$ is shown in a larger font; and the difference relative to the best combination is shown in a smaller font. A difference of ± 0.02 or less indicates that the metric is largely determined by the number of domains (D) and total number of endpoints (V), regardless of how the variables are distributed among the different domains. The contents of these tables are dissected in the following sections.

3.3.2 Sensitivity of Go rates to the number of domains and variables

In Figure 3.7, the FGr and CGr of $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ and $G^{\text{hier}}_{x:=2,\alpha:=0.05}$ are compared for increasing numbers of domains and variables. The figure shows that the variation due to the distribution of variables across domains is minor, and is only apparent for $G^{\text{hier}}_{x:=2,\alpha:=0.05}$ when the number of endpoints and domains is large.

The main driver of differences in terms of the CGr and FGr is in the number of domains under consideration. As the number of domains increases, so do the CGr and FGr. This is an instance of the multiple comparisons problem: as the number of domains increases, the probability that any domain will individually result on a Go decision increases. However, systematically applying a multiple comparisons correction here might not be advisable; for example, a Go decision for a treatment that reaches an effect close to the **LRV** across multiple domains may not be undesirable enough to warrant reducing the chance of a positive trial result when a single one of the endpoints reaches the TV.

An increase in the number of endpoints has a very minor effect on the FGr, due to the multiplicity correction performed by the $G^{\text{ind.Simes}}$ policy. By contrast, increasing the number of endpoints invariably results in an improvement in the CGr. The improvement diminishes if the correlation between variables in different domains is higher, due to each new endpoint contributing less new information overall.

As for the differences between policies, the $G^{\text{hier}}_{x:=2,\alpha:=0.05}$ policy has a consistently higher FGr than the $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ policy, presumably due to the fact that $G^{\text{hier}}_{x:=2,\alpha:=0.05}$ only requires a Go decision in a single domain (the most important one) for an overall Go; compared to the minimum of two decisions required for $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$. By contrast, the improvement in the CGr is

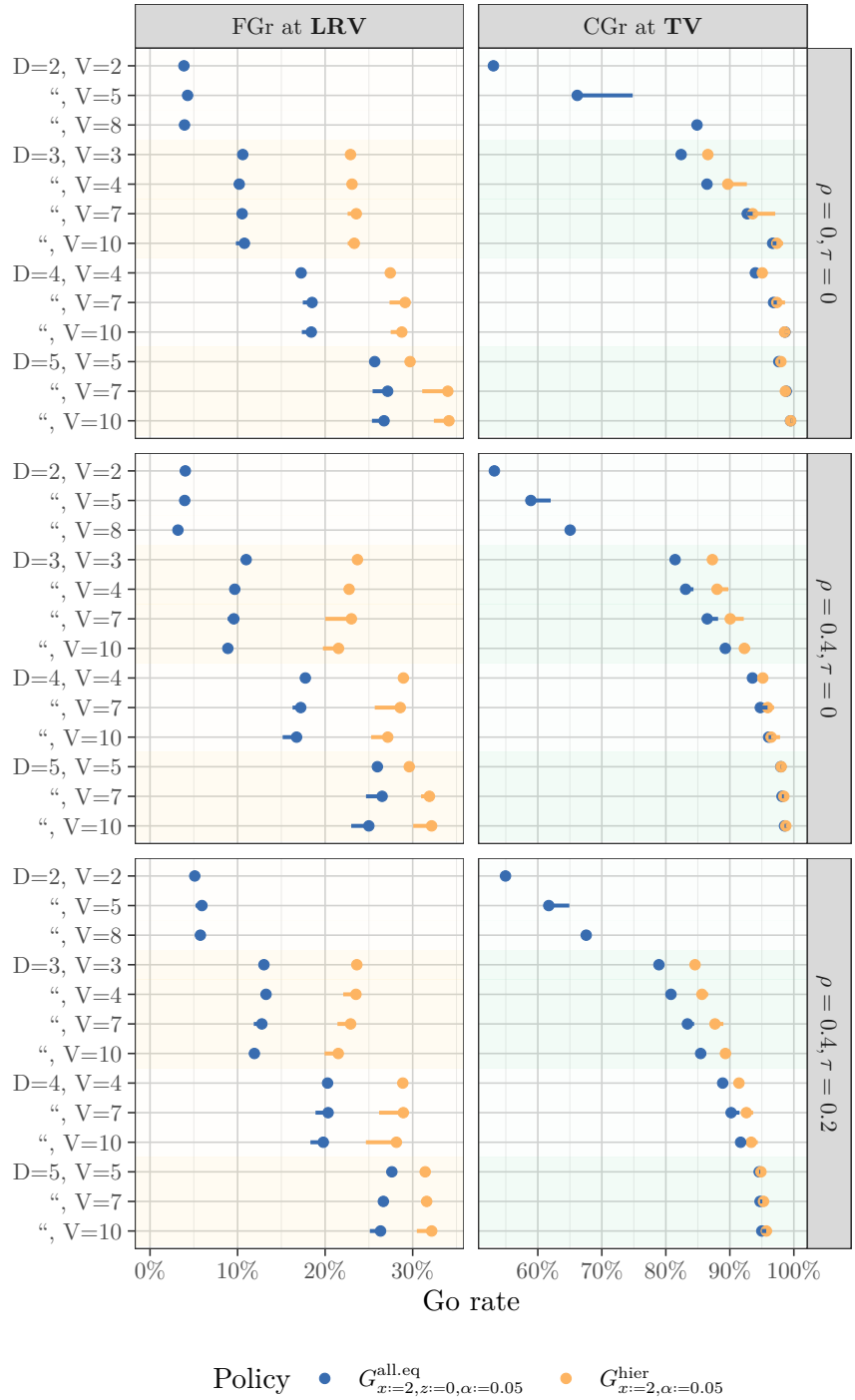


*Figure 3.7: False Go rate (*FGr*) and Correct Go rate (*CGr*) for $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ and $G^{\text{hier}}_{x:=2,\alpha:=0.05}$ depending on the number of domains (D), the total number of variables (V), the within-domain correlation (ρ) and the inter-domain correlation (τ). The horizontal segments cover the possible* FGr/CGr *values depending on the distribution of the V variables across the D domains. The dot is the most "pessimistic" value: the largest one for the* FGr*, and the smallest one for the* CGr*.*

minimal, possibly because all the endpoints reach the TV, so there is not so much advantage in focusing on only one of them. The circumstances under which the $G^{\text{hier}}_{x:=2,\alpha:=0.05}$ policy may be advantageous from a statistical perspective are explored in §3.3.9.

3.3.3 Sensitivity of Stop rates to the number of domains and variables

In Figure 3.8, the FSr and CSr for the two policies is visualized. The situation compared to the Go rate (§3.3.2) is reversed: the criteria used by $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ are less conservative than those of $G^{\text{hier}}_{x:=2,\alpha:=0.05}$, resulting in a comparatively high FSr when the number of domains is low ($D = 2$). This difference is reduced as the number of domains increases. If the correlation between domains is low, this difference ultimately becomes negligible.

As for the CSr, the $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ policy can reliably produce a Stop decision in the absence of an effect ($\boldsymbol{\mu} = \mathbf{0}$), even as the number of domains increases. By contrast, the policy $G^{\text{hier}}_{x:=2,\alpha:=0.05}$ produces Stop with diminishing frequency as the number of endpoints increases, due to the fact that this also decreases the probability of Stop at the domain-level policies (c.f. Table 3.5: $G^{\text{ind.Simes}}$, row $\rho = 0.4$, region $\neg(\mathbf{0})$, column CSr; compare 0.90 for $V = 2$ vs. 0.83 for $V = 4$).

3.3.4 Sensitivity to the arrangement of variables

Given a number of domains and variables, there may be many possible ways of distributing the variables across the domains. In Table 3.7 and Table 3.8, the difference in a metric among arrangements for a fixed D and V is shown in a smaller font next to each metric. The policy $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ is largely insensitive to how the variables are divided across domains. In the case of $G^{\text{hier}}_{x:=2,\alpha:=0.05}$, given a number of variables V and a number of domains D, the policy is also largely unaffected by how the variables are arranged in domains. However, the value of the CSr metric for the case where $\boldsymbol{\mu} = \mathbf{0}$, can vary up to 15 percentage points depending on how the variables are grouped (see Table 3.8a, $D = 5$, $V = 8$, $\neg(\mathbf{0})$, where CSr = 0.80 $_{+0.15}$). It can be seen in Figure 3.9 that the Stop rate is highest when the number of variables in the most important domain $V_1 = 1$, and decreases as V_1 becomes larger. This can be explained by the fact that, according to $G^{\text{hier}}_{x:=2,\alpha:=0.05}$, a Stop decision in the most important domain is always necessary for a combined Stop decision; and this decision is more likely the fewer variables that there are in said domain (see Table 3.5). The effect of the number of variables in the first domain is less pronounced when the number of domains is smaller, as at least two of the remaining domains must also be Stop, and this becomes harder if the other domains have more variables.

3.3.5 Sensitivity to correlation between endpoints

In this analysis we consider three cases for the correlation between endpoints: $\rho = \tau = 0$ (all endpoints are uncorrelated), $\rho = 0.4, \tau = 0$ (only endpoints in the domains are correlated), and $\rho = 0.4, \tau = 0.2$ (all endpoints are correlated,

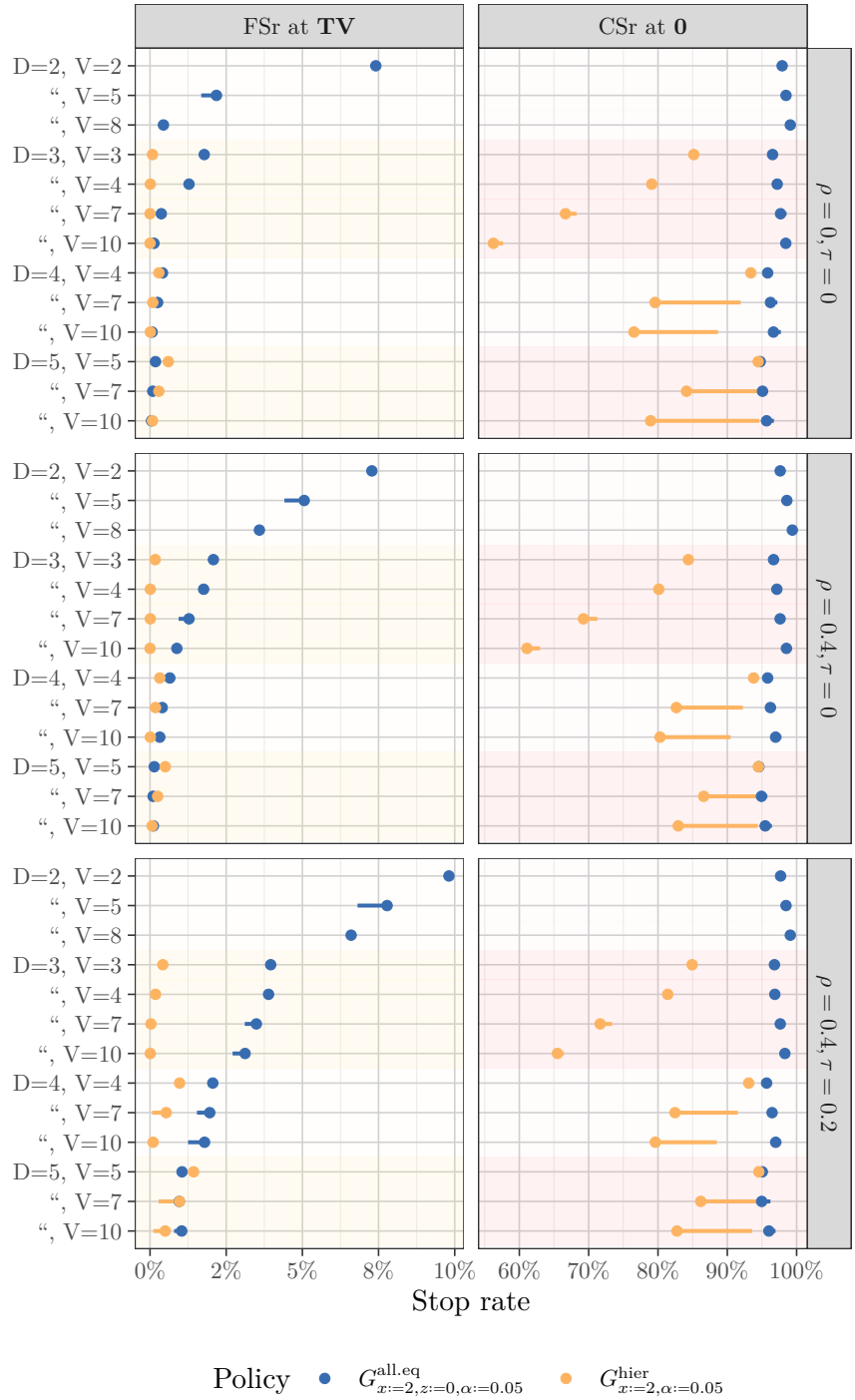


*Figure 3.8: False Stop rate (*FSr*) at* $\boldsymbol{\mu} = \mathbf{TV}$ *and Correct Stop rate (*CSr*) when* $\boldsymbol{\mu} = \mathbf{0}$*, for the policies* $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ *and* $G^{\text{hier}}_{x:=2,\alpha:=0.05}$*, depending on the number of domains (*D*) and variables (*V*), the within-domain correlation (*ρ*) and the inter-domain correlation (*τ*). The horizontal segments cover the possible* FSr*/*CSr *values depending on the distribution of the* V *variables across the* D *domains. The point is the most “pessimistic” value: the largest one for the* FSr*, and the smallest one for the* CSr*.*

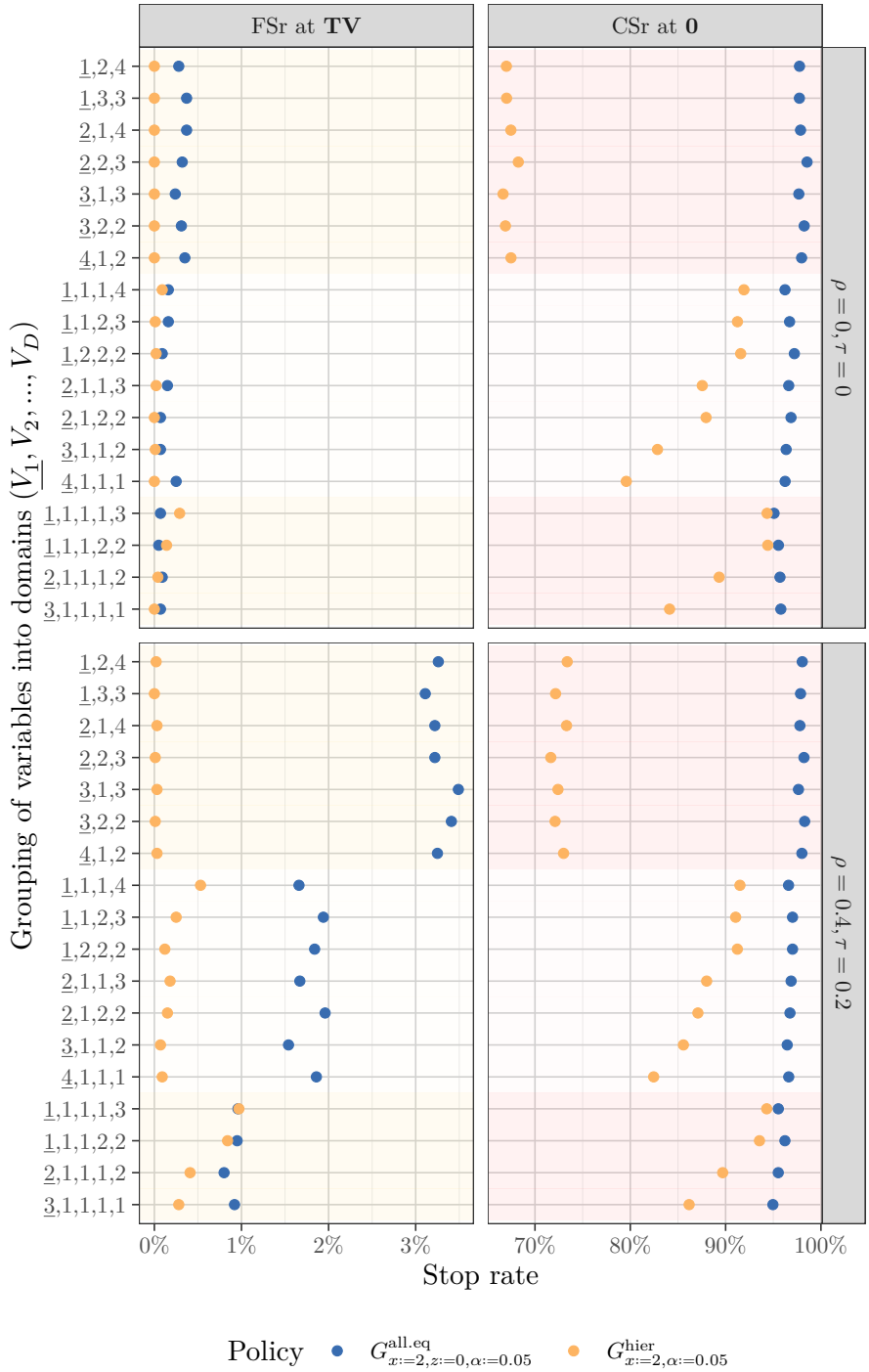


Figure 3.9: Detail of Figure 3.8 for $D = 3, 4, 5$ and $V = 10$. For policies $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ and $G^{\text{hier}}_{x:=2,\alpha:=0.05}$, the value of the metric FSr *when* $\boldsymbol{\mu} = \mathbf{TV}$ *at the value of the metric* CSr *when* $\boldsymbol{\mu} = \mathbf{0}$ *is shown. The number of variables in the most important domain for the $G^{\text{hier}}_{x:=2,\alpha:=0.05}$ policy is underlined.*

with the within-domain correlation ρ being higher than the between-domain correlation τ). The correlation between endpoints impacts metrics in various ways, although at least two trends can be observed.

The first observed trend is that policy decisions that require *all* elements of a set of domains or endpoints to read above or below a given threshold become more likely when the correlation between them is higher. One example of this is the CSr at **0** for $G^{\text{hier}}_{x:=2,\alpha:=0.05}$ when $D = 3$ (Figure 3.8), where the Stop rate increases with both ρ and τ; or, in the same figure, the FSr at **TV** for $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$. Another example is the CGr at **TV** for $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ when $D = 2$ (Figure 3.7), where the Go rate increases with the between-domain correlation (τ). This is because the policy $G^{\text{all.eq}}_{x:=2,z:=0}$ effectively requires all 2 domains to be Go. In both cases, having high correlation between the domains reduces the odds that the observations of any one endpoint will be abnormally low.

The second trend is that policy decisions that only require *some* of the elements of the set to read above a certain (not too extreme) threshold may become more likely when correlation is lower. For instance, observe that the CGr at **TV** for $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ when $D \geq 3$ (Figure 3.7) decreases both when the within-domain correlation (ρ) increases, and also when the between-domain correlation (τ) increases. A possible explanation is that the criterion for a Go decision only requiring one endpoint in each of two domains to yield a "go" decision. If the measurements are uncorrelated, the odds are lower that the measurements for all endpoints will be simultaneously too low to produce a Go decision.

Given that the impact of correlation among endpoints on the decision probabilities can be in either direction, and may depend not only on the metrics being evaluated, but also on other factors such as the number of domains, these trends should not be strongly relied upon. Instead, a sensitivity analysis should be performed that assess the potential impact of within- and between-domain correlation on the probability of a correct/incorrect decision.

3.3.6 Impact of adding a variable with low power into a separate domain

As explained in the introduction (§1.2), the number of patients in an early stage trial (Phase II) might be too low for the main endpoint of interest (Phase III) to have enough power to produce a meaningful decision on its own. However, it could be the case that including this endpoint as part of the policy still gives valuable information that can aid in the early stage decision.

In order to analyze whether such endpoints are worth including, studies where all variables have power 0.8 are simulated ($N = 17$, $\sigma_1 = ... = \sigma_V = 1$, TV = 1), and the resulting metrics are compared to those obtained when an additional variable of power 0.3 is included ($\sigma_{V+1} = 2$, TV = 1). The within-domain correlation is $\rho = 0.4$, and the between-domain correlation is $\tau = 0.2$. The results of this comparison for policy $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ are collected in Table 3.9a, and those for policy $G^{\text{hier}}_{x:=2,\alpha:=0.05}$ are collected in Table 3.9b.

Impact of adding a low-powered variable in a separate domain when using an "all domains equal" policy: As shown in Table 3.9a, when us-

Table 3.9: The effect of including ("inc.low" column) an additional variable of low power in its own domain ($\sigma = 2$, power 0.3) when all the existing variables have high power ($\sigma = 1$, power 0.8), compared to ignoring this variable ("exc.low" column). Thus, the "exc.low" column involves V variables across D domains, while the "inc.low" column involves $V + 1$ variables across $D + 1$ domains. The $(D + 1)th$ domain always has 1 variable (the low power one). The maximum range of variation depending on the distribution of V across $V_1, ..., V_D$ is indicated in a smaller font. The simulated studies have $N = 17$ patients per arm, $\rho = 0.4$ whithin-domain correlation, and $\tau = 0.2$ between-domain correlation.

(a) Effect for an "all domains equal" policy ($G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$)

	CSr at $\neg(\mathbf{0})$		FGr at $\neg(\mathbf{LRV})$		CGr at $\llcorner(\mathbf{TV})$	
V	exc.low	inc.low	exc.low	inc.low	exc.low	inc.low
D=2						
2	**0.98 +0.00**	0.92 +0.00	**0.06 -0.00**	0.14 -0.00	0.56 +0.00	**0.69 +0.00**
4	**0.98 +0.00**	0.93 +0.00	**0.06 -0.00**	0.13 -0.00	0.60 +0.02	**0.72 +0.01**
6	**0.99 +0.00**	0.94 +0.00	**0.06 -0.00**	0.12 -0.00	0.65 +0.01	**0.75 +0.00**
8	**0.99 +0.00**	0.93 +0.00	**0.05 -0.00**	0.12 -0.00	0.68 +0.00	**0.76 +0.00**
D=3						
3	**0.97 +0.00**	0.91 +0.00	**0.14 -0.00**	0.20 -0.00	0.79 +0.00	**0.85 +0.00**
6	**0.97 +0.01**	0.92 +0.00	**0.13 -0.01**	0.20 -0.00	0.82 +0.02	**0.86 +0.02**
9	**0.98 +0.00**	0.93 +0.00	**0.12 -0.00**	0.19 -0.01	0.84 +0.01	**0.88 +0.01**
D=4						
4	**0.96 +0.00**	0.90 +0.00	**0.20 -0.00**	0.27 -0.00	0.89 +0.00	**0.92 +0.00**
6	**0.96 +0.00**	0.91 +0.00	**0.20 -0.00**	0.27 -0.00	0.90 +0.01	**0.92 +0.00**
9	**0.97 +0.01**	0.92 +0.01	**0.19 -0.01**	0.26 -0.01	0.91 +0.01	**0.93 +0.01**

(b) Effect for a hierarchical policy ($G^{\text{hier}}_{x:=2,\alpha:=0.05}$)

	CSr at $\neg(\mathbf{0})$		FGr at $\neg(\mathbf{LRV})$		CGr at $\llcorner(\mathbf{TV})$	
V	exc.low	inc.low	exc.low	inc.low	exc.low	inc.low
D=3						
3	0.85 +0.00	**0.89 +0.00**	**0.23 -0.00**	0.30 -0.00	0.85 +0.00	**0.89 +0.00**
6	0.75 +0.01	**0.79 +0.05**	**0.23 -0.02**	0.28 -0.02	0.86 +0.02	**0.89 +0.01**
9	0.68 +0.01	**0.75 +0.04**	**0.22 -0.02**	0.27 -0.01	0.87 +0.02	**0.91 +0.01**
D=4						
4	0.93 +0.00	0.93 +0.00	**0.30 -0.00**	0.31 -0.00	0.92 +0.00	**0.93 +0.00**
6	0.85 +0.06	**0.86 +0.07**	**0.29 -0.02**	0.33 -0.01	0.92 +0.01	**0.93 +0.00**
9	0.81 +0.09	**0.82 +0.09**	**0.28 -0.02**	0.32 -0.02	0.93 +0.01	**0.94 +0.01**

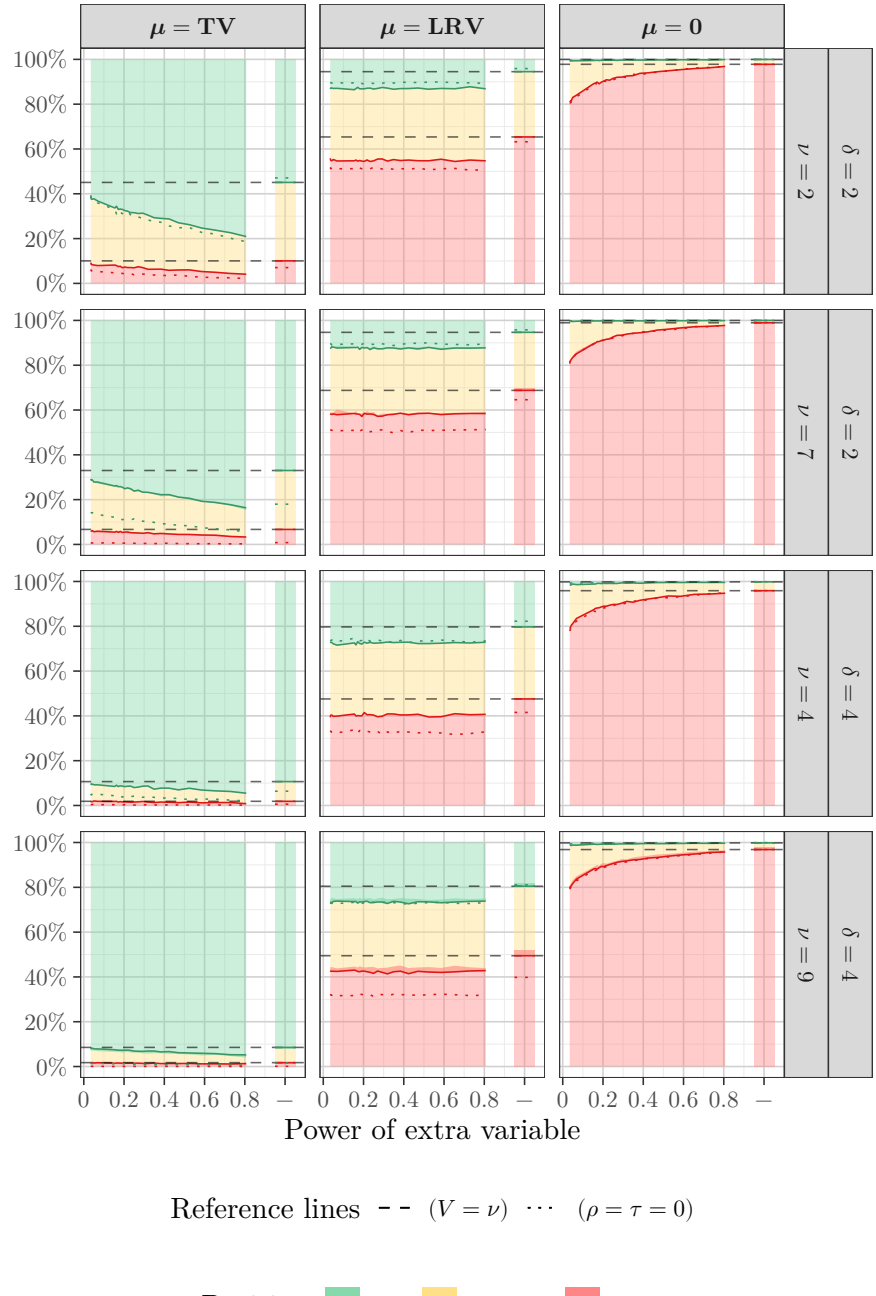


Figure 3.10: Effect of adding a variable with lower power in its own domain when using the "all domains equal" policy. On the right side of each subplot, a column (—) showing the decision probabilities for $D = \delta$ domains and $V = \nu$ variables, within-domain correlation $\rho = 0.4$, between-domain correlation $\tau = 0.2$, true effect $\boldsymbol{\mu}$, and power 0.8. Left of the column, the probabilities when the new variable is added (yielding $D = \delta + 1$ domains and $V = \nu + 1$ variables), depending on its power (≤ 0.8). In a darker shade, the eventual variation due to the arrangement of variables into domains.

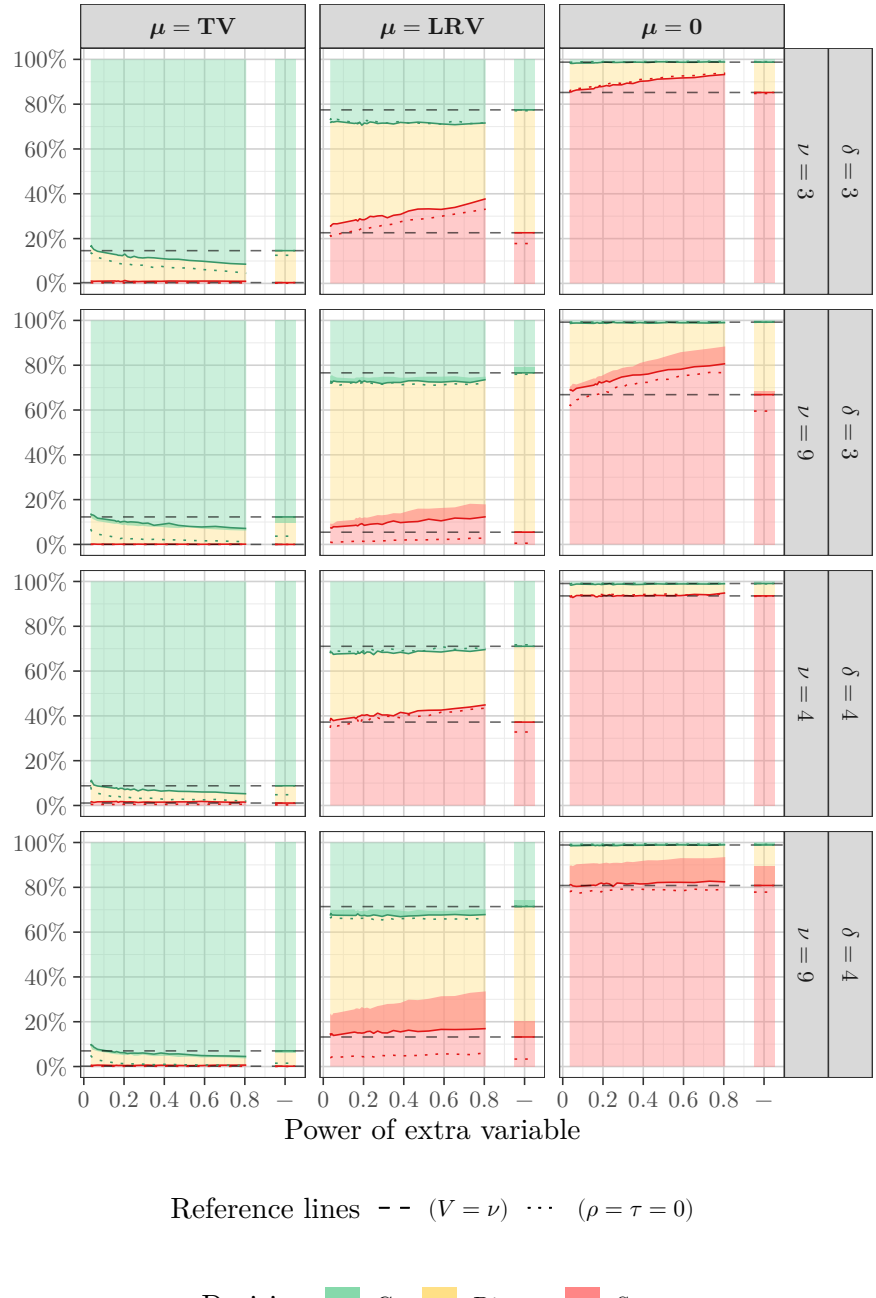


Figure 3.11: Impact of adding a variable with lower power in its own domain when using the "hierarchical domains" policy. On the right side of each subplot, a column (—) showing the decision probabilities for $D = \delta$ domains and $V = \nu$ variables, within-domain correlation $\rho = 0.4$, between-domain correlation $\tau = 0.2$, true effect $\boldsymbol{\mu}$, and power 0.8. Left of the column, the probabilities when the new variable is added (yielding $D = \delta + 1$ domains and $V = \nu + 1$ variables), depending on its power (≤ 0.8). In a darker shade, the eventual variation due to the arrangement of variables into domains.

ing the policy $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$, including the additional low powered variable (column "inc.low") has the largest impact on increasing correct "Go" decisions when the number of domains is 2, and the number of endpoints is low (e.g. for $D = 2$, $V = 2$, when adding a new variable in its own domain (yielding 3 domains and 3 variables) the CGr when $\boldsymbol{\mu} \geq \mathbf{TV}$ increases by 13 p.p., from 0.56 to 0.69). Note that this comes at the expense of increasing the FGr when $\boldsymbol{\mu} \leq \mathbf{LRV}$ (in that same scenario, 0.06 vs 0.14), and decreasing the CSr when $\boldsymbol{\mu} \leq \mathbf{0}$ (0.98 vs 0.92). The benefits of including the additional low-powered endpoint diminish as the number of high-powered endpoints and variables increases (e.g. with 3 domains and 9 endpoints: $D = 3$, $V = 9$), adding an additional low powered endpoint (for a total of 4 domains and 10 endpoints) increases the CGr when $\boldsymbol{\mu} \geq \mathbf{TV}$ from 0.84 to 0.88 (4 p.p.), but also increases the FGr when $\boldsymbol{\mu} \leq \mathbf{LRV}$ by 7 p.p (0.12 vs. 0.19) and decreases the CSr when $\boldsymbol{\mu} \leq \mathbf{0}$ by 5 p.p (0.98 vs. 0.93).

How the resulting decision probabilities vary depending on the power of the new endpoint is shown in Figure 3.10. Both the negative impact of adding a new endpoint on the Stop rate when $\boldsymbol{\mu} = \mathbf{0}$, and on the Go rate when $\boldsymbol{\mu} = \mathbf{LRV}$ are similar regardless of the number of existing, well-powered domains and variables; and the impact in the case when $\boldsymbol{\mu} = \mathbf{LRV}$ is also independent of the power of the new endpoint. However, the benefit of the extra variable decreases as its power decreases and the number of pre-existing endpoints (ν) increases. Therefore, adding a new endpoint in a separate domain is most appropriate when the new endpoint has high power, or when the number of well-powered endpoints already included in the study is low.

Impact of adding a low-powered variable in a separate domain when using a "hierarchical domains" policy: Compared to the situation for the "all domains equal" policy, the impact of adding an additional, low powered endpoint in a separate domain is much more moderate, not exceeding 4 p.p. (compare columns "inc.low" and "exc.low" for the CGr at $\llcorner(\mathbf{TV})$ in Table 3.9b). This is consistent with decisions in the hierarchical domains policy being largely driven by the outcome in the most important domain, and the fact that there are already at least 3 well-powered domains in the base case, which decreases the impact of adding a fourth or fifth domain. The impact of adding the extra variable as the power changes is shown in Figure 3.11. Due to the way the policy is defined, the additional variable increases the rate of Go when $\boldsymbol{\mu} = \mathbf{TV}, \mathbf{LRV}$, and of Stop when $\boldsymbol{\mu} = \mathbf{LRV}, \mathbf{0}$ across the board, specially when the variable has high power and the number of existing well-powered endpoints is low. The magnitude of changes in the decision probabilities is strongly dependent on how the variables are arranged across domains, although the trend is the same in all cases.

3.3.7 Impact of adding a variable with low power into an existing domain

In other situations, there may be several ways of measuring the effect of a treatment on the endpoints in a given clinical domain. Setting aside issues of cost and convenience, the endpoint with the higher powered is preferred. However, it may be the case that measuring the lower-powered endpoint still

yields meaningful data that can be used to obtain a better decision. In this section we assess the effect of adding a variable into an existing domain. The initial setup is the same as in §3.3.6: the new variable has power 0.3 (which would be considered low), while all other variables in the study have power 0.8.

Impact of adding a low-powered variable into an existing domain when using "all domains equal" policy: The results of the evaluation when using $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ are collected in Table 3.10a. Due to the multiple comparisons adjustment, adding this new variable produces a moderate decrease across all metrics. The low power of the variable means that any measurements are not much more than noise which tends to uniformly increase the Go probability across all true effects. The variation as the power of the extra variable changes is shown in Figure 3.12. It is shown that adding an additional variable has a moderately positive effect on the Go rate when $\boldsymbol{\mu} = \mathbf{TV}$ and the power of the new variable is high enough (e.g. when starting from 2 variables in 2 domains, the power must be higher than 0.6 to see any benefit from adding a third variable). The improvement on the Go rate of one additional variable decreases as the number of domains and variables increases, while the negative impact on the Stop rate when $\boldsymbol{\mu} = \mathbf{0}$ increases with the number of domains, and decreases with the number of endpoints. Note that adding an additional variable of relatively low power (< 0.6) may still be justified if it measures an aspect of the disease that is not adequately covered by other endpoints.

Impact of adding a low-powered variable into the most important domain when using the "hierarchical domains" policy: For the hierarchical domains policy, the effect of adding a variable can be expected to be strongest when this variable is added into the most important domain. The impact of adding a variable with low power into the most important domain when using $G^{\text{hier}}_{x:=2,\alpha:=0.05}$ is shown in Table 3.10b. The effect on the Go rate is mild, similarly to the $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ case (Table 3.10a); however, the CSr when $\boldsymbol{\mu} \leq \mathbf{0}$ worsens dramatically (e.g. for $D = 4$ and $V = 9$, from 0.83 to 0.52), as the Stop decision thus hinges on a variable that is largely uninformative.

The variation as the power of the extra variable changes is shown in Figure 3.13. Compared to the $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ case, the new variable needs much higher power to produce an improvement on the probability of Go when $\boldsymbol{\mu} = \mathbf{TV}$, and also avoid producing incorrect Go decisions when $\boldsymbol{\mu} = \mathbf{0}$. Only if the additional endpoint has enough power does the Stop rate when $\boldsymbol{\mu} = \mathbf{0}$ remain relatively unaffected. In this section the focus is on the effect of adding a new endpoint into the most important domain; the impact of adding the variable into another domain can be expected to be lower.

Table 3.10: The impact of including ("inc.low" columns) an additional variable of low power ($\sigma = 2$, power 0.3) in an existing domain *when all the existing variables have high power ($\sigma = 1$, power 0.8), compared to ignoring this variable ("exc.low" column). Including the variable results in $V + 1$ variables across D domains, with $V_1 + 1$ variables in the first domain. The maximum range of variation depending on the distribution of the number of variables V (or $V + 1$) variables across the D domains is indicated in a smaller font. There are $N = 17$ patients per arm, and the correlation between endpoints is $\rho = 0.4$ (within-domain) and $\tau = 0.2$ (between-domain).*

(a) Impact for an "all domains equal" policy ($G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$).

	CSr at $\neg(\mathbf{0})$		FGr at $\neg(\mathbf{LRV})$		CGr at $\llcorner(\mathbf{TV})$	
V	exc.low	inc.low	exc.low	inc.low	exc.low	inc.low
D=2						
2	**0.98 +0.00**	0.97 +0.00	0.06 -0.00	0.06 -0.00	**0.56 +0.00**	0.52 +0.00
4	**0.98 +0.00**	0.97 +0.01	**0.06 -0.00**	0.06 -0.01	**0.60 +0.02**	0.56 +0.03
7	**0.99 +0.00**	0.98 +0.00	0.05 -0.00	0.05 -0.00	**0.67 +0.00**	0.64 +0.00
D=3						
3	**0.97 +0.00**	0.95 +0.00	0.14 -0.00	**0.12 -0.00**	**0.79 +0.00**	0.76 +0.00
6	**0.98 +0.00**	0.96 +0.01	0.13 -0.00	**0.12 -0.01**	**0.82 +0.02**	0.79 +0.03
9	**0.98 +0.00**	0.97 +0.01	**0.12 -0.00**	0.12 -0.01	**0.85 +0.01**	0.82 +0.02
D=4						
4	**0.96 +0.00**	0.94 +0.00	0.20 -0.00	**0.19 -0.00**	**0.89 +0.00**	0.88 +0.00
7	**0.97 +0.01**	0.95 +0.01	0.20 -0.01	0.20 -0.01	**0.90 +0.01**	0.89 +0.01
9	**0.97 +0.01**	0.96 +0.01	**0.19 -0.01**	0.20 -0.01	**0.91 +0.01**	0.90 +0.01
D=5						
5	**0.94 +0.00**	0.93 +0.00	0.27 -0.00	0.27 -0.00	**0.94 +0.00**	0.93 +0.00
7	**0.95 +0.00**	0.94 +0.01	**0.26 -0.01**	0.27 -0.01	0.94 +0.00	0.94 +0.00
9	**0.96 +0.01**	0.95 +0.01	0.27 -0.01	**0.26 -0.01**	**0.95 +0.01**	0.94 +0.00

(b) Impact for a "hierarchical domains" policy ($G^{\text{hier}}_{x:=2,\alpha:=0.05}$), with the most important domain being the first domain (where the new variable is included).

	CSr at $\neg(\mathbf{0})$		FGr at $\neg(\mathbf{LRV})$		CGr at $\llcorner(\mathbf{TV})$	
V	exc.low	inc.low	exc.low	inc.low	exc.low	inc.low
D=3						
3	**0.85 +0.00**	0.50 +0.00	0.23 -0.00	0.23 -0.00	**0.85 +0.00**	0.82 +0.00
6	**0.75 +0.01**	0.46 +0.02	0.23 -0.01	**0.22 -0.01**	**0.86 +0.02**	0.83 +0.03
9	**0.68 +0.01**	0.43 +0.00	0.22 -0.02	**0.22 -0.01**	**0.87 +0.02**	0.85 +0.03
D=4						
4	**0.93 +0.00**	0.55 +0.00	0.30 -0.00	**0.27 -0.00**	**0.92 +0.00**	0.90 +0.00
7	**0.85 +0.07**	0.52 +0.02	0.28 -0.01	**0.27 -0.01**	**0.92 +0.00**	0.90 +0.01
9	**0.83 +0.07**	0.52 +0.02	0.28 -0.01	**0.27 -0.01**	**0.93 +0.01**	0.92 +0.02
D=5						
5	**0.94 +0.00**	0.55 +0.00	**0.30 -0.00**	0.33 -0.00	**0.95 +0.00**	0.94 +0.00
7	**0.86 +0.08**	0.52 +0.03	**0.32 -0.01**	0.33 -0.02	**0.95 +0.00**	0.94 +0.01
9	**0.86 +0.08**	0.52 +0.02	0.32 -0.01	**0.31 -0.01**	0.95 +0.01	0.95 +0.01

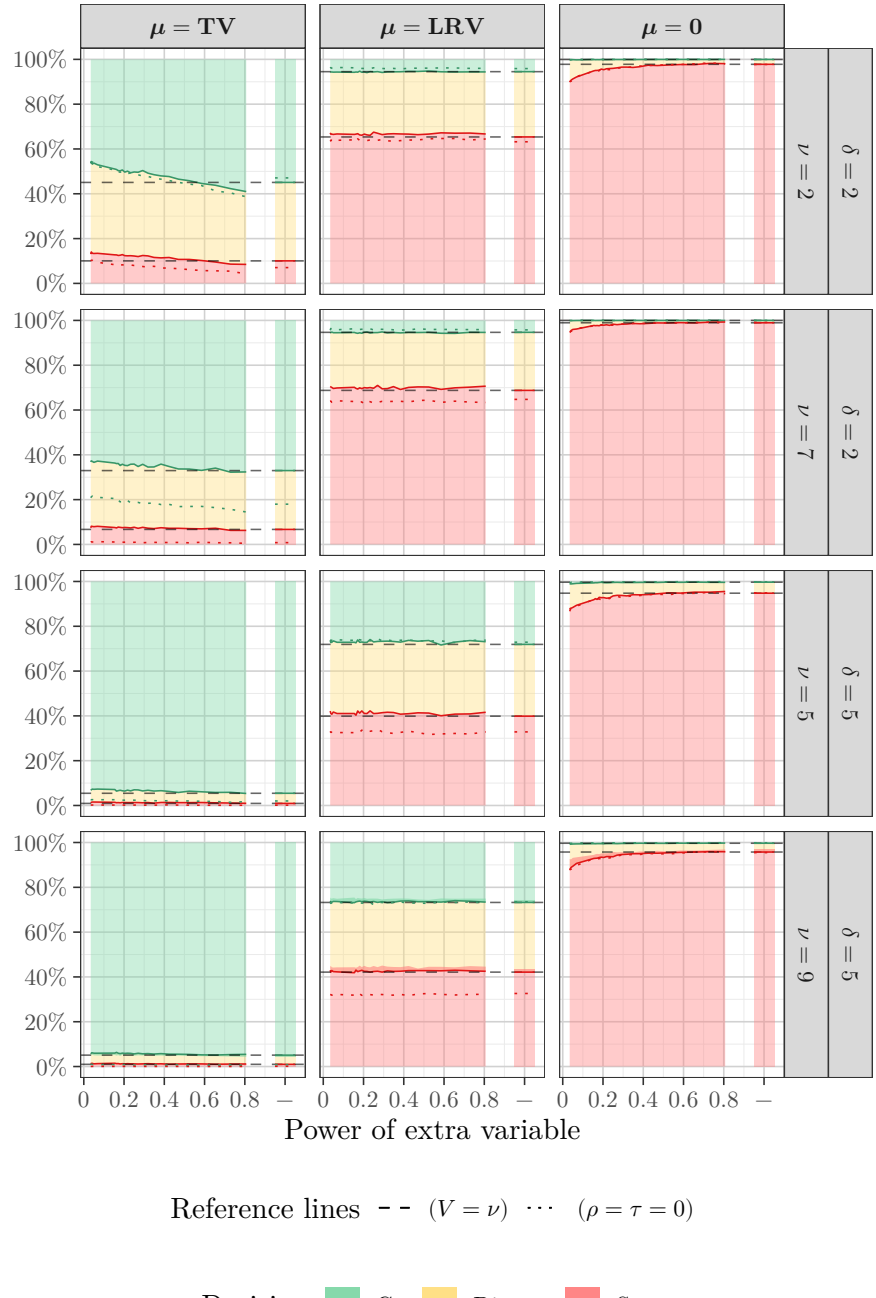


Figure 3.12: Impact of adding a variable with lower power into an existing domain when using the "all domains equal" policy. On the right side of each subplot, a column (—) showing the decision probabilities for $D = \delta$ domains and $V = \nu$ variables, within-domain correlation $\rho = 0.4$, between-domain correlation $\tau = 0.2$, true effect $\boldsymbol{\mu}$, and power 0.8. Left of the column, the probabilities when the new variable is added (yielding $D = \delta$ domains and $V = \nu + 1$ variables), depending on its power (≤ 0.8). In a darker shade, the eventual variation due to the arrangement of variables into domains.

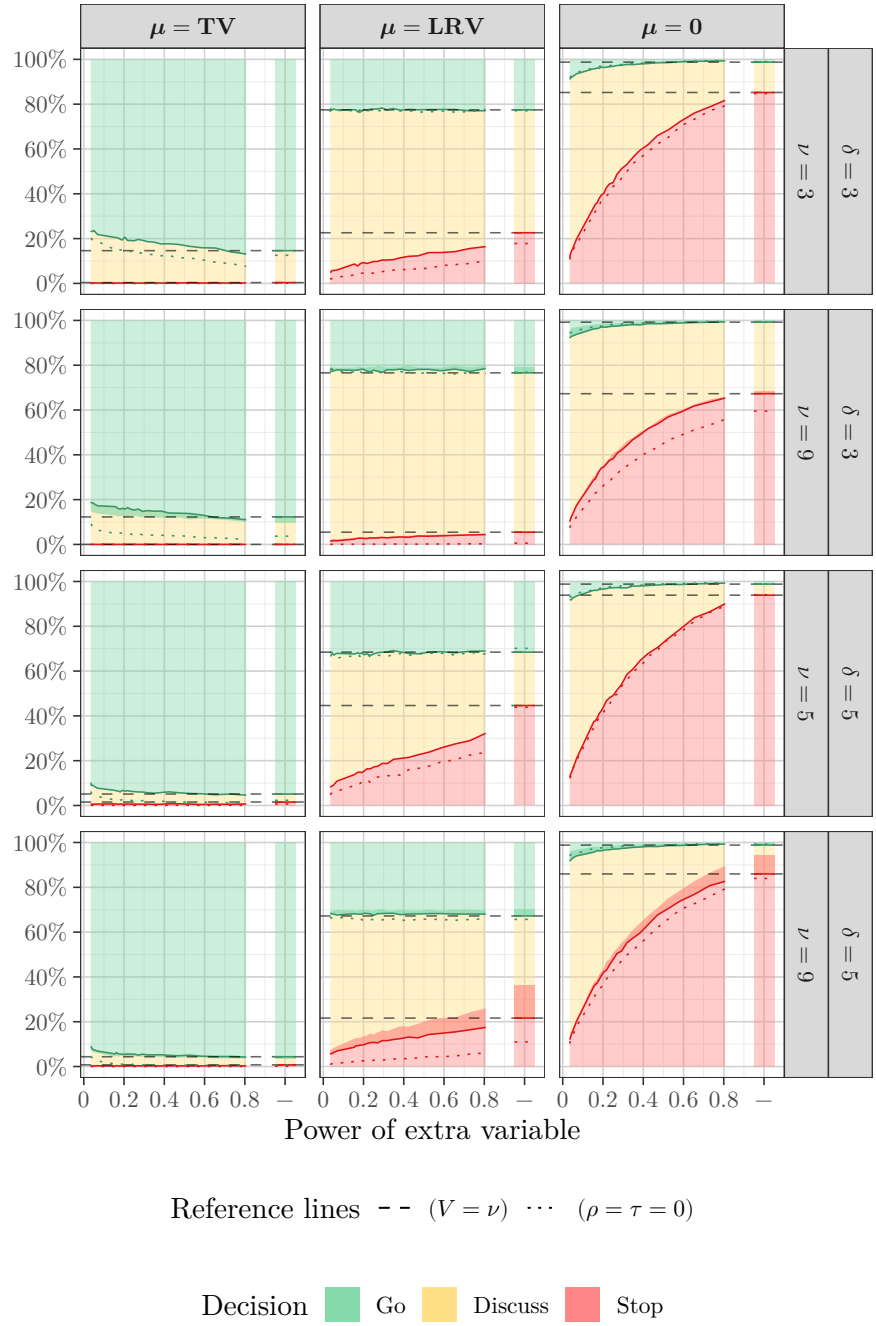


Figure 3.13: Impact of adding a variable with lower power into the most important domain when using the "hierarchical domains" policy. On the right side of each subplot, a column (—) showing the decision probabilities for $D = \delta$ domains and $V = \nu$ variables, within-domain correlation $\rho = 0.4$, between-domain correlation $\tau = 0.2$, true effect $\boldsymbol{\mu}$, and power 0.8. Left of the column, the probabilities when the new variable is added (yielding $D = \delta$ domains and $V = \nu + 1$ variables), depending on its power (≤ 0.8). In a darker shade, the eventual variation due to the arrangement of variables into domains.

3.3.8 Impact of the safety condition

When performing a Phase II study, it is assumed that the drug will either have no effect or a positive effect each endpoint. However, there could still be deleterious effects that were missed in earlier trials. To guard against this possibility, stakeholders will often prescribe adding a safety condition that prevents a Go result when there is any endpoint that is statistically significant in the wrong direction. This was done when defining the policies evaluated in this chapter, namely $G^{\text{all.eq}}$ (Equation 3.25) and G^{hier} (Equation 3.27). In this section the impact of the safety condition is assessed by comparing the first of the aforementioned policies (Equation 3.25) with a corresponding policy that dispenses with the condition (Equation 3.28).

$$G^{\text{all.eq.w/o.cond}}_{x,z}(\boldsymbol{\mu}, \tilde{\Sigma}_\mu) := \begin{cases} \text{stop} & \text{at most } z \text{ of} \\ & \quad (G_{(d)}(\boldsymbol{\mu}_{I(d)}, (\tilde{\Sigma}_\mu)_{I(d),I(d)}) \text{ is Go} \\ & \quad \mid 1 \le d \le D) \\ \text{go} & \text{at least } x \text{ of} \\ & \quad (G_{(d)}(\boldsymbol{\mu}_{I(d)}, (\tilde{\Sigma}_\mu)_{I(d),I(d)}) \text{ is Go} \\ & \quad \mid 1 \le d \le D) \end{cases} \tag{3.28}$$

Because metrics related to the Stop decision (CSr and FSr) are unaffected, the analysis is focused on the FGr and the CGr. For the cases which have been studied so far in this chapter, in which the treatment affects all endpoints in the same way ($\boldsymbol{\mu} = \mathbf{TV}$, $\boldsymbol{\mu} = \mathbf{LRV}$ or $\boldsymbol{\mu} = \mathbf{0}$) the effect of this condition is limited (Table 3.11a). One may conjecture that the effect of the safety condition will be larger when the treatment has no effect on some of the domains. To test this hypothesis, the following additional regions are considered:

$$\begin{aligned} R^{\text{Go}}_{\text{tv.}-1} &:= \{\boldsymbol{\Delta} \in \mathbb{R}^V \mid \Delta_{I(d)} \ge \mathbf{TV} \text{ for at least } D-1 \text{ of } d \in 1, ..., D\} \cap \llcorner(\mathbf{0}) \\ R^{\text{NoGo}}_{\text{lrv.1}} &:= \{\boldsymbol{\Delta} \in \mathbb{R}^V \mid \Delta_{I(d)} \le \mathbf{0} \text{ for at least } D-1 \text{ of } d \in 1, ..., D\} \cap \neg(\mathbf{LRV}) \\ R^{\text{NoGo}}_{\text{lrv.2}} &:= \{\boldsymbol{\Delta} \in \mathbb{R}^V \mid \Delta_{I(d)} \le \mathbf{0} \text{ for at least } D-2 \text{ of } d \in 1, ..., D\} \cap \neg(\mathbf{LRV}), \end{aligned}$$

where $I(d)$ are indices of the endpoints in domain d.

Example 3.29. For $D = 4$, $V_1 = V_2 = V_3 = V_4 = 1$, the regions $R^{\text{Go}}_{\text{tv.2}}$, $R^{\text{NoGo}}_{\text{tv.1}}$ and $R^{\text{NoGo}}_{\text{lrv.2}}$ are as follows:

$$\begin{aligned} R^{\text{Go}}_{\text{tv.}-1} &= \llcorner((\text{TV}_1, \text{TV}_2, \text{TV}_3, 0)) \cup \llcorner((\text{TV}_1, \text{TV}_2, 0, \text{TV}_4)) \cup \\ &\quad \llcorner((\text{TV}_1, 0, \text{TV}_3, \text{TV}_4)) \cup \llcorner((0, \text{TV}_2, \text{TV}_3, \text{TV}_4)) \\ R^{\text{NoGo}}_{\text{lrv.1}} &= \neg((\text{LRV}_1, 0, 0, 0)) \cup \neg((0, \text{LRV}_2, 0, 0)) \cup \\ &\quad \neg((0, 0, \text{LRV}_3, 0)) \cup \neg((0, 0, 0, \text{LRV}_4)) \\ R^{\text{NoGo}}_{\text{lrv.2}} &= \neg((\text{LRV}_1, \text{LRV}_2, 0, 0)) \cup \neg((\text{LRV}_1, 0, \text{LRV}_3, 0)) \cup \\ &\quad \neg((\text{LRV}_1, 0, 0, \text{LRV}_4)) \cup \neg((0, \text{LRV}_2, \text{LRV}_3, 0)) \cup \\ &\quad \neg((0, \text{LRV}_2, 0, \text{LRV}_4)) \cup \neg((0, 0, \text{LRV}_3, \text{LRV}_4)) \end{aligned}$$

◀

The results of evaluating the $G^{\text{all.eq.w/o.cond}}_{x:=2,z:=0}$ policy on $R^{\text{Go}}_{\text{tv.}-1}$, $R^{\text{NoGo}}_{\text{lrv.1}}$ and $R^{\text{NoGo}}_{\text{lrv.2}}$ are shown in Table 3.11b. In cases where the true effect is 0 for some

*Table 3.11: Simulation results for the policy with and without the safety condition: $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ ("with.cond" column) vs. $G^{\text{all.eq.w/o.cond}}_{x:=2,z:=0}$ ("w/o.cond" column). $N = 17$, $\rho = 0.4$, $\tau = 0.2$. For each number of variables (V), number of domains (D), metric (*FGr *and* CGr*), region and policy, the worst value of the metric among all possible combinations of $V_1, ..., V_D$ with $V_1 + ... + V_D = V$ is shown; and, in a smaller font size, the difference between the worst and best value of the metric depending on the arrangement $V_1, ..., V_D$. In bold, the best performing policy of the two for that combination of D, V, metric and region.*

(a) Regions with homogeneous true effect.

	FGr at $\neg(\mathbf{0})$		FGr at $\neg(\mathbf{LRV})$		CGr at $\llcorner(\mathbf{TV})$	
V	with.cond	w/o.cond	with.cond	w/o.cond	with.cond	w/o.cond
D=3						
3	0.00 -0.00	0.00 -0.00	0.13 -0.00	0.13 -0.00	0.79 +0.00	0.79 +0.00
7	0.00 -0.00	0.00 -0.00	0.13 -0.01	0.13 -0.01	0.83 +0.01	**0.83 +0.01**
10	0.00 -0.00	0.00 -0.00	0.12 -0.00	0.12 -0.00	0.85 +0.01	**0.85 +0.01**
D=4						
4	0.00 -0.00	0.00 -0.00	0.20 -0.00	0.20 -0.00	0.89 +0.00	0.89 +0.00
7	0.00 -0.00	0.00 -0.00	0.20 -0.01	0.20 -0.01	0.90 +0.01	0.90 +0.01
10	0.00 -0.00	0.00 -0.00	**0.20 -0.01**	0.20 -0.01	0.92 +0.01	0.92 +0.01
D=5						
5	0.00 -0.00	0.00 -0.00	**0.28 -0.00**	0.28 -0.00	0.95 +0.00	0.95 +0.00
7	0.00 -0.00	0.00 -0.00	0.27 -0.00	0.27 -0.00	0.95 +0.00	0.95 +0.00
10	**0.00 -0.00**	0.00 -0.00	**0.26 -0.01**	0.26 -0.01	0.95 +0.01	0.95 +0.01

(b) Regions with heterogeneous true effect.

	FGr at $R^{\text{NoGo}}_{\text{lrv.1}}$		FGr at $R^{\text{NoGo}}_{\text{lrv.2}}$		CGr at $R^{\text{Go}}_{\text{tv.}-1}$	
V	with.cond	w/o.cond	with.cond	w/o.cond	with.cond	w/o.cond
D=3						
3	**0.01 -0.00**	0.01 -0.00	**0.06 -0.00**	0.06 -0.00	0.54 +0.00	**0.55 +0.00**
7	**0.01 -0.00**	0.01 -0.00	**0.06 -0.00**	0.06 -0.00	0.52 +0.06	**0.58 +0.05**
10	0.01 -0.00	0.01 -0.00	**0.06 -0.00**	0.06 -0.00	0.58 +0.02	**0.65 +0.02**
D=4						
4	**0.01 -0.00**	0.01 -0.00	**0.07 -0.00**	0.07 -0.00	0.76 +0.00	**0.79 +0.00**
7	**0.01 -0.00**	0.01 -0.00	**0.07 -0.01**	0.07 -0.00	0.68 +0.08	**0.78 +0.04**
10	**0.01 -0.00**	0.01 -0.00	**0.06 -0.01**	0.07 -0.01	0.71 +0.05	**0.82 +0.03**
D=5						
5	**0.02 -0.00**	0.02 -0.00	**0.07 -0.00**	0.08 -0.00	0.85 +0.00	**0.89 +0.00**
7	**0.02 -0.00**	0.02 -0.00	**0.07 -0.00**	0.07 -0.00	0.79 +0.04	**0.90 +0.01**
10	**0.01 -0.00**	0.02 -0.00	**0.07 -0.01**	0.07 -0.00	0.77 +0.07	**0.90 +0.01**

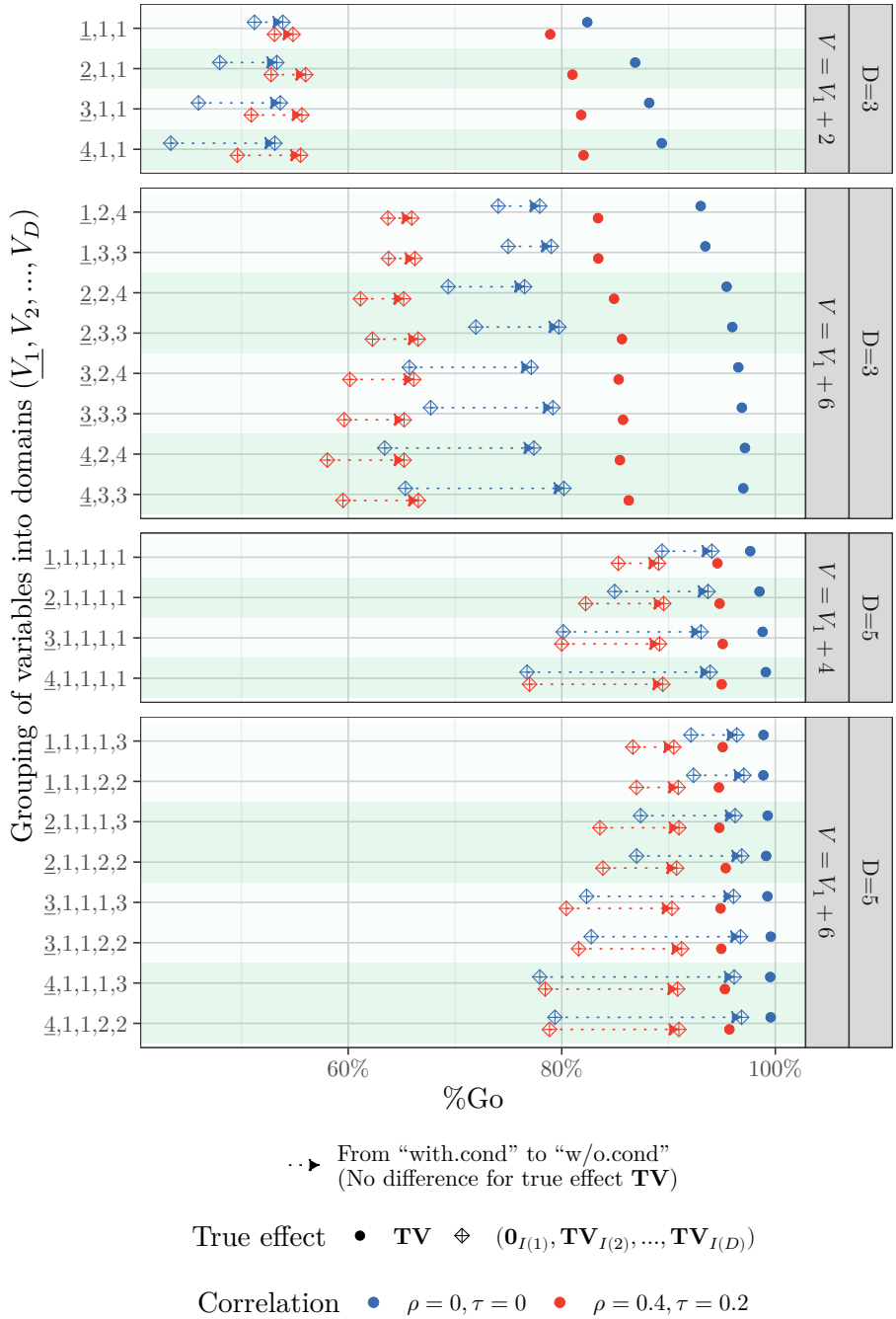


Figure 3.14: Effect of a safety condition in the absence of an effect in one domain. Probability of Go *for policies* $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ *(with.cond) vs.* $G^{\text{all.eq.w/o.cond}}_{x:=2,z:=0}$ *(w/o.cond) depending on the number of variables* $(V_1, ..., V_D)$ *and the correlation between endpoints in the same domain* (ρ) *and in different domains* (τ)*. The true effect is either* **TV***, for which the safety condition has almost no effect; or* 0 *for endpoints in the first domain and* TV *for the remaining ones, in which case the difference is shown as an arrow between the* Go *rates for the two policies.*

endpoints with enough of the other endpoints reach the TV to make a Go decision desirable, the safety condition strongly reduces the probability of Go. This impact becomes stronger as the number of variables increases. The safety condition seems to reduce the Go rate the most in those cases where a Go decision is warranted (e.g. all endpoints reach the target value in all but 1 of the domains; region $R^{\text{Go}}_{\text{tv.}-1}$). The condition has almost no effect in borderline cases where Go is not desired; e.g. when some endpoints are near the LRV (regions $R^{\text{NoGo}}_{\text{lrv.1}}$, $R^{\text{NoGo}}_{\text{lrv.2}}$).

In Figure 3.14, the effect of the safety condition on the Correct Go rate (CGr) is analyzed in detail. It can be seen that the decrease in CGr when the safety condition is added occurs regardless of the distribution of the number of variables across each domain.

In conclusion, assuming that the treatment has no deleterious effects for the endpoints under consideration, the impact of the safety condition on quality of the decisions either negative of non-existent. If such a condition is a hard requirement from other stakeholders, one should aim to either minimize the number of endpoints for which an effective drug could have no effect, or simply accept lower probabilities of Go for certain combinations of true effects (perhaps increasing the sample size to compensate).

3.3.9 Impact of giving more importance to one domain

Under some circumstances, clinicians may deem one domain to be of higher clinical importance when deciding whether to go ahead with a Phase III study. From a purely statistical point of view, a hierarchical policy that gives more preeminence to one domain such as $G^{\text{hier}}_{x,\alpha}$ is generally inferior to a non-hierarchical one like $G^{\text{all.eq}}_{x,z,\alpha}$ (§3.3.2). However, in all the scenarios studied in §3.3.2, the effect of the treatment was homogeneous across endpoints (either all reach the TV, or none show any effect). It is possible that for certain combinations of underlying true effects, a hierarchical policy may produce accurate decisions more often than a non-hierarchical one. Consider the following regions:

$$
\begin{aligned}
R^{\text{Go}}_{\text{hier}} := \{\boldsymbol{\Delta} \in \mathbb{R}^V \mid\ & \boldsymbol{\Delta}_{I(1)} \geq \mathbf{TV}_{I(1)} \\
& \text{and } \boldsymbol{\Delta}_{I(d)} \geq \mathbf{TV}_{I(d)} \text{ for at least 1 of } d = 2, ..., D \\
& \text{and } \boldsymbol{\Delta} \geq \mathbf{0}\} \\
R^{\text{Stop}}_{\text{hier}} := \{\boldsymbol{\Delta} \in \mathbb{R}^V \mid\ & \boldsymbol{\Delta}_{I(1)} \leq \mathbf{0} \text{ and } \boldsymbol{\Delta} \leq \mathbf{LRV}\},
\end{aligned}
$$

where $I(d)$ are the indices of variables in domain d. In $R^{\text{Go}}_{\text{hier}}$, some of the domains may show no effect, but at least one of the domains that has an effect is the most important domain. Similarly, in $R^{\text{Stop}}_{\text{hier}}$, some domains may have a clinically relevant effect ($\mathbf{LRV}$), but the most important domain has no effect.

Remark. The region $R^{\text{Go}}_{\text{hier}}$ contains $\llcorner(\mathbf{TV})$, and $R^{\text{Stop}}_{\text{hier}}$ contains $\neg(\mathbf{0})$. This means that the CGr for $\boldsymbol{\mu} \in R^{\text{Go}}_{\text{hier}}$ is a lower bound for the CGr when $\boldsymbol{\mu} \geq \mathbf{TV}$, and the FGr for $\boldsymbol{\mu} \in R^{\text{Stop}}_{\text{hier}}$ is an upper bound for the FGr when $\boldsymbol{\mu} \leq \mathbf{0}$

Table 3.12: Comparison between $G^{\text{hier.w/o.cond}}_{x:=2}$ *("hier" column) and* $G^{\text{all.eq.w/o.cond}}_{x:=2,z:=0}$ *("all.eq" column) with heterogeneous true effects for which a* Go ($R^{\text{Go}}_{\text{hier}}$) *or* Stop ($R^{\text{Stop}}_{\text{hier}}$) *decision is desired. In bold, the best performing policy of the two for that combination of number of domains (D) and variables (V). The potential variation due to the arrangement of variables into domains is shown in a smaller font.*

(a) Metrics for $\rho = 0$, $\tau = 0$, $N = 17$.

	CSr at $R^{\text{Stop}}_{\text{hier}}$		FGr at $R^{\text{Stop}}_{\text{hier}}$		CGr at $R^{\text{Go}}_{\text{hier}}$	
V	all.eq	hier	all.eq	hier	all.eq	hier
D=3						
3	**0.63 +0.00**	0.31 +0.00	**0.04 -0.00**	0.05 -0.00	0.53 +0.00	**0.73 +0.00**
7	**0.62 +0.02**	0.03 +0.12	**0.05 -0.01**	0.06 -0.01	0.61 +0.13	**0.72 +0.20**
10	**0.64 +0.01**	0.01 +0.02	**0.04 -0.00**	0.05 -0.00	0.77 +0.05	**0.84 +0.08**
D=4						
4	0.51 +0.00	**0.56 +0.00**	0.11 -0.00	0.11 -0.00	0.53 +0.00	**0.73 +0.00**
7	**0.50 +0.01**	0.22 +0.25	0.11 -0.01	**0.11 -0.00**	0.53 +0.14	**0.73 +0.19**
10	**0.51 +0.01**	0.08 +0.22	**0.11 -0.01**	0.12 -0.01	0.54 +0.23	**0.72 +0.20**
D=5						
5	0.40 +0.00	**0.73 +0.00**	0.19 -0.00	**0.12 -0.00**	0.54 +0.00	**0.72 +0.00**
7	0.40 +0.01	**0.57 +0.10**	0.19 -0.01	**0.14 -0.03**	0.54 +0.11	**0.72 +0.16**
10	**0.40 +0.01**	0.29 +0.24	0.19 -0.01	**0.17 -0.03**	0.54 +0.16	**0.72 +0.20**

(b) Metrics for $\rho = 0.4$, $\tau = 0.2$, $N = 17$.

	CSr at $R^{\text{Stop}}_{\text{hier}}$		FGr at $R^{\text{Stop}}_{\text{hier}}$		CGr at $R^{\text{Go}}_{\text{hier}}$	
V	all.eq	hier	all.eq	hier	all.eq	hier
D=3						
3	**0.64 +0.00**	0.34 +0.00	0.06 -0.00	0.06 -0.00	0.56 +0.00	**0.72 +0.00**
7	**0.67 +0.02**	0.11 +0.11	**0.06 -0.01**	0.07 -0.01	0.58 +0.06	**0.72 +0.08**
10	**0.68 +0.01**	0.08 +0.03	0.06 -0.01	0.06 -0.01	0.65 +0.02	**0.77 +0.04**
D=4						
4	0.55 +0.00	**0.56 +0.00**	**0.13 -0.00**	0.14 -0.00	0.56 +0.00	**0.72 +0.00**
7	**0.55 +0.02**	0.34 +0.18	0.14 -0.01	0.14 -0.01	0.55 +0.06	**0.72 +0.08**
10	**0.57 +0.03**	0.23 +0.16	0.13 -0.01	0.13 -0.01	0.55 +0.09	**0.72 +0.08**
D=5						
5	0.46 +0.00	**0.70 +0.00**	0.21 -0.00	**0.16 -0.00**	0.56 +0.00	**0.72 +0.00**
7	0.46 +0.02	**0.59 +0.07**	0.21 -0.01	**0.17 -0.01**	0.56 +0.04	**0.72 +0.07**
10	**0.47 +0.04**	0.42 +0.15	0.20 -0.02	**0.19 -0.02**	0.55 +0.07	**0.72 +0.09**

Example 3.30. For $D = 3$, $V_1 = V_2 = V_3 = 1$, the regions $R^{\text{Go}}_{\text{hier}}$ and $R^{\text{Stop}}_{\text{hier}}$ are as follows:

$$R^{\text{Go}}_{\text{hier}} = \llcorner((\text{TV}_1, \text{TV}_2, 0)) \cup \llcorner((\text{TV}_1, 0, \text{TV}_3))$$
$$R^{\text{Stop}}_{\text{hier}} = \neg((0, \text{LRV}_2, \text{LRV}_3))$$

Note that the cone $\llcorner((\text{TV}_1, \text{TV}_2, \text{TV}_3))$ is contained in $R^{\text{Go}}_{\text{hier}}$, and the cone $\neg((0, 0, 0))$ is contained in $R^{\text{Stop}}_{\text{hier}}$. ◀

As shown in §3.3.8, the inclusion of a check for variables significant in the negative direction could impact the results when the true effect for some endpoints is 0. Therefore, the policies $G^{\text{all.eq.w/o.cond}}_{x:=2,z:=0}$ and $G^{\text{hier.w/o.cond}}_{x:=2}$ are used for this comparison. In table Table 3.12a and Figure 3.15 it is shown that, in these scenarios, the hierarchical policy ($G^{\text{hier.w/o.cond}}_{x:=2}$) produces correct Go decisions more often than the one where all domains are treated equal ($G^{\text{all.eq.w/o.cond}}_{x:=2,z:=0}$), at least as the number of endpoints and domains increases. The advantage persists when the correlation between domains increases (Table 3.12b). As for Stop decisions, the hierarchical policy is advantageous when the number of domains is large (provided again that the true effect is in the $R^{\text{Stop}}_{\text{hier}}$ region). T This means that when designing a policy with many domains, a hierarchical policy may make for a higher rate of correct Stop decisions. The advantage of the hierarchical policy in this case diminishes as the correlation between endpoints increases.

Additionally, a large degree of variation in the correct Go rate can be observed in Figure 3.15, depending on how the endpoints are distributed across domains. The detailed view in Figure 3.16 shows that the more endpoints belonging to the first domain (where the true effect is consistent with the desired decision), the higher the CGr and CSr. This trend affects both the hierarchical policy and the "all domains equal" policy. This suggests that, when attempting to increase the correct Go rate of a study by measuring additional endpoints, it is better to choose new endpoints that are in the domain that the policy considers to be the most important.

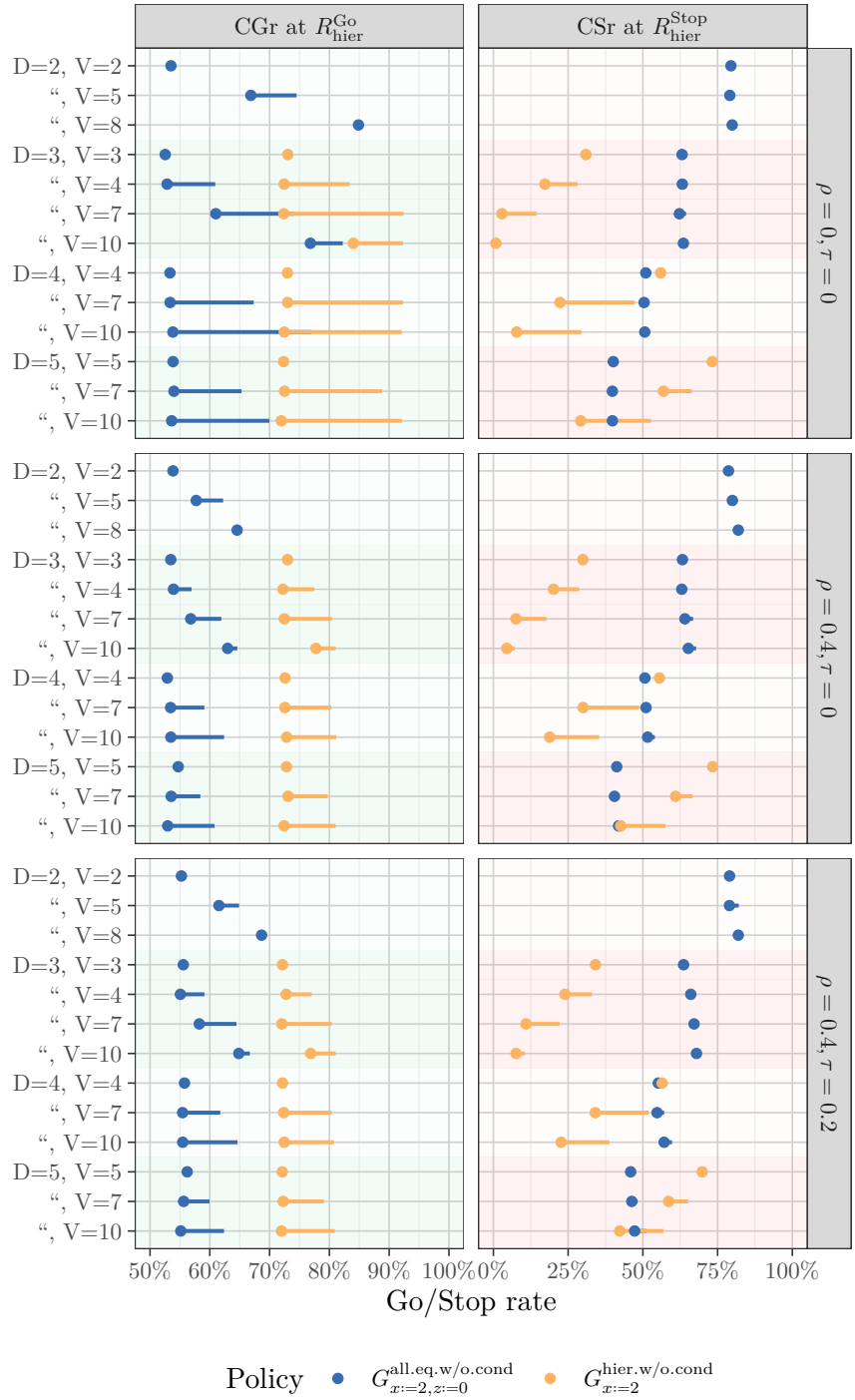


Figure 3.15: Correct Go *and* Stop *rates of policies* $G^{\text{all.eq.w/o.cond}}_{x:=2,z:=0}$ *vs.* $G^{\text{hier.w/o.cond}}_{x:=2}$ *for true effects in the regions* $R^{\text{Go}}_{\text{hier}}$ *and* $R^{\text{Stop}}_{\text{hier}}$*, respectively; and varying within- (*ρ*) and between-domain (*τ*) correlations. The horizontal segments cover the possible* CGr/CSr *values depending on the distribution of the* V *variables across the* D *domains. The point is the most “pessimistic” value of the range.*

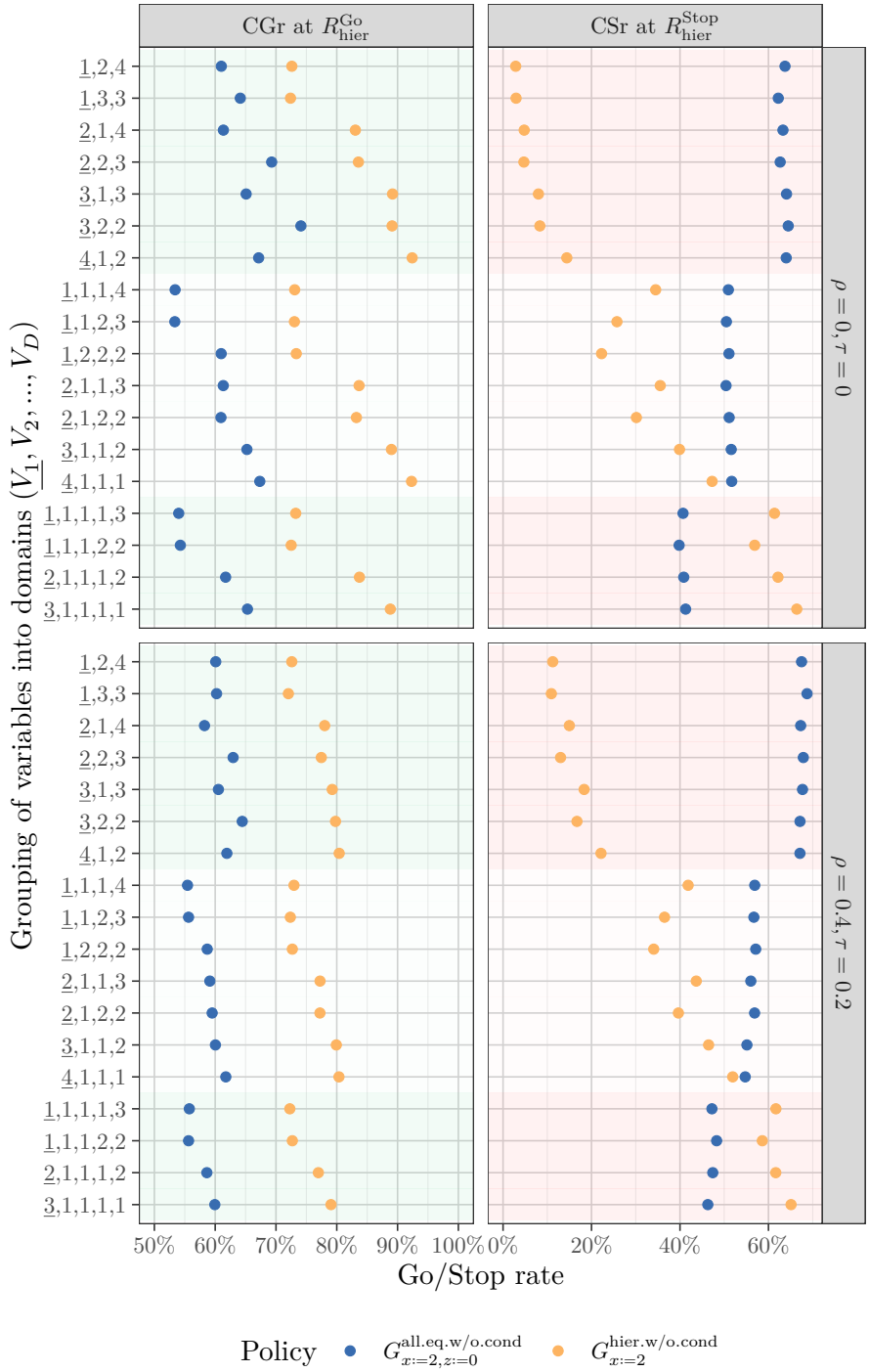


Figure 3.16: Detail of Figure 3.15 for $D = 3, 4, 5$ and $V = 7$. Correct Go *and* Stop *rates of policies $G^{\text{all.eq.w/o.cond}}_{x:=2,z:=0}$ vs. $G^{\text{hier.w/o.cond}}_{x:=2}$ for true effects in the regions $R^{\text{Go}}_{\text{hier}}$ and $R^{\text{Stop}}_{\text{hier}}$, respectively.*

3.4 Summary

In this chapter the rate of correct and false Go and Stop decisions has been analyzed across a range of synthetic scenarios, both within a single domain and when combining the decisions on multiple domains into an overall decision. Decisions at the domain level are sensitive to the multiple comparisons problem, leading to a high rate of False Go decisions. By adjusting the decision thresholds using the Simes method (a particular case of Benjamini-Hochberg procedure), it is possible to bound the False Go Risk at a chosen level ($\alpha^{\star}_{\text{FG}}$). Compared to other approaches such as the Bonferroni method, this method consistently produces higher rates of correct Go decisions at the domain-level, of around 4 percentage points, with the only requirement of assuming that that the endpoints are not negatively correlated. From these results, the Simes correction can be recommended when combining endpoints from the same domain.

Adding endpoints to a policy is in general advantageous, and results in more frequent correct Go decisions, as long as these endpoints have sufficient power. However, if the new endpoint has low power, its addition will have the effect of increasing the frequency of Discuss or incorrect decisions, as it becomes essentially a source of noise for the policy. The negative impact of a low-powered endpoint can be mitigated by assigning it to its own domain, which works specially well when the policy does not rely on the absence of Go decisions to produce a Stop decision. A phenomenon of diminishing returns is observed, as adding new endpoints increases the False Go Rate by a similar amount regardless of the benefit they provide, and this marginal benefit becomes smaller as the number of existing, well-powered endpoints increases. Thus, lower power endpoints should preferably be added in their own domain, used with policies that do not rely on absence of Go to produce Stop, and only be added if the study has a relatively small number of existing high-power endpoints.

Regarding the two policies treated in the chapter, the "hierarchical domains" policy is advantageous in situations where Go (respectively Stop) is desired, and the most important endpoint reaches the target value (respectively exhibits no effect), but the endpoints in some of the other domains do not exhibit an effect (respectively exhibit a subclinical effect). On the other hand the "all domains equal" policy produces correct decisions with higher frequency when the endpoint are uniformly affected by the policy (e.g. all endpoints reach the TV, or all endpoints show no effect). For both families of policies, stakeholders often require that a Go decision is not produced if any of the endpoints is statistically significant in the wrong direction. This condition has a very small effect when endpoints are uniformly affected by the policy, but results in fewer correct Go decisions in cases where the treatment only produces the target effect in some domains. Although this condition can in practice not be removed, care should be taken to not introduce too many redundant endpoints for which the true effect is plausibly 0, as they may result in false determinations of the existence of a negative true effect.

Regardless of how suitable the chosen policy is, a study may sometimes be inconclusive; i.e. neither yield a Go nor a Stop decision. This may be either by chance, or because the underlying true effect is neither close to $\mathbf{0}$ nor to the

TV. This situation is conventionally understood to call for further discussion on how to proceed. In the simulations in this chapter it has been shown that (i) high correlation between endpoints, (ii) the existence of low-powered endpoints and (iii) the safety condition can all turn decisions that would be a Go or a Stop into a "Discuss". In such borderline cases, and depending on the endpoints concerned, these circumstances may be taken into account when deciding a course of action.

Last but not least, the metrics obtained from the above policies are sensitive to correlation between endpoints. The way in which correlation affects the metrics depend on both the policy itself, and on the number of domains and endpoints to which its applied. Although some general rules of thumb can be derived regarding the specific way that correlation affects the results (§3.3.5), a sensitivity analysis with different correlation levels will be pertinent when designing a study using a domain approach.

Chapter 4

Case study

In this chapter the framework developed in Chapter 2 and evaluated in Chapter 3 is used in the design of a hypothetical early stage (Phase II) study with two arms: treatment and control. The study evaluates the viability of a hypothetical heart medication in order to decide whether to "Go" ahead with a Ph3 study, or "Stop" development of the treatment. The Phase III endpoint is the ratio of the number of MACE (Major Adverse Cardiovascular Events) in the treatment arm relative to the control arm. The number of events expected to be observed among the number of patients in a Phase II trial is low, so the decision of whether to go ahead with a Phase III trial must be based disease markers with higher statistical power.

The main goal of the case study is to determine, for an actual Phase II trial, which endpoints to include and how many patients need to be enrolled, so that Go and Stop decisions are produced with the desired probabilities.

The first step in the case study is to ensure that the endpoints fulfill the assumptions in the Chapter 2 framework by means of appropriate transformations. Then, based on input from practitioners, decision policies and requirements regarding the probabilities of Go and Stop under different true effects of interest are specified. Finally, simulations are used to calculate the minimum number of patients needed to fulfill the requirements.

4.1 Endpoint properties

A range of endpoints related to heart disease are listed in Table 4.1. The target values (TV), lower reference values (LRV) and standard deviations for each of these endpoints have been elicited from practitioners involved in early-stage clinical trials, and collected in Table 4.2. Endpoints are assumed to be normally distributed unless otherwise indicated. The correlation between endpoints within the same domain is assumed to be $\rho = 0.4$, and the correlation between variables in different domains is taken to be $\tau = 0.2$.

4.1.1 Measurement of variables

For each endpoint i, each arm $x = 0, 1$, and each patient $n = 1, ..., N$ in arm x, the value of the underlying variable is measured at the start of the study,

Table 4.1: Endpoints for a case study for an early stage (Phase II) clinical trial to assess effectiveness of a hypothetical heart drug. The study includes 9 endpoints across 5 different domains. In an eventual Phase III trial for cardiovascular outcomes (CVOT), MACE is the endpoint that would determine approval.

Biomarker: Plasma levels of a certain molecule that is correlated with the disease.	
NT-proBNP	*N-terminal prohormone of B-type natriuretic peptide* A precursor of BNP, a hormone produced by the heart in response to internal changes in pressure. High levels of NT-proBNP are a sign of heart failure [BWM04].
Exercise: Patient's exercise capacity.	
6MWD	*6-minute walking distance* Distance walked back-and-forth by the patient in a continuous 30m corridor in a 6 minute interval.
VO2max	*Volume of molecular oxygen (maximum)* Maximal oxygen consumption by the patient during exercise of increasing intensity.
Well-being: Subjective well-being as reported by the patient.	
KCCQ-TSS	*Kansas City Cardiomyopathy Questionnaire – Total Symptom Score* A questionaire to assess the patients well-being in areas connected to heart function [GPBS00].
Imaging: Cardiac structure and function, measured by medical imaging.	
GLS	*Global Longitudinal Strain* The change in the length of the heart muscle in the left ventricle during systole relative to its relaxed, baseline length. The GLS is a negative percentage. More negative GLS (i.e. larger GLS in absolute value) correspond to a decrease in mortality [AFB$^+$18].
LAVI	*Left-atrial volume index* The volume of the left atrium divided by the surface area of the patients body. An increased LAV is a predictor of heart-failure-related hospitalization and mortality [RAWS08].
LVMI	*Left-ventricular mass index* The (estimated) weight of the patients left ventricle, divided by the surface are of the patients body. An increased LVM is a sign of cardiovascular disease [TGS$^+$15].
LVEF	*Left-ventricular ejection fraction* The percentage of the volume of blood collected in the left ventricle during diastole that is ejected during systole. A decrease LVEF is a predictor of heart-failure-related hospitalization and mortality [RAWS08].
Events: Adverse events which occur during the trial.	
MACE (HR)	*Major Adverse Cardiovascular Events (Hazard Ratio)* The ratio of stroke, myocardial infarction events, hospitalization for heart failure, cardiac deaths etc. in the treatment arm (numerator) relative to the control arm (denominator).

Table 4.2: Data for the endpoints in the case study as elicited from practitioners. The SD is reported for the change from baseline (CFB). Endpoints are assumed to be normally distributed unless otherwise noted.

Domain	Variable	TV	LRV	SD	Change (unit)
Biomarker	NT-proBNP	-15%	-5%	0.8[a]	Relative, geometric mean
Exercise	6MWD	20	12	70	Absolute (meters)
“	VO2max	1	0.7	2	Absolute (mℓ/min/kg)
Well-being	KCCQ-TSS	5	2	20	Absolute (points)
Imaging	GLS	-0.75	-0.25	2.5	Absolute (percentage points)
“	LAVI	-2	-0.5	7	Absolute (mℓ/m^2)
“	LVMI	-8	-4	12	Absolute (g/m^2)
“	LVEF	4	2	10	Absolute (percentage points)
Events	MACE (HR)	0.8	0.9	—[b]	

[a] This endpoint is assumed to follow a log-normal distribution. The given standard deviation concerns the logarithm of the variable.

[b] The distribution of the logarithm of the hazard ratio can be approximated as a normal distribution. The variance of this quantity depends on the number of events (Equation 4.1). For this approximation, it is assumed that ca. 5% of patients enrolled in the trial will suffer such an event.

producing the baseline ($A_i^{x,n,\mathrm{B}}$), and also measured at the end of the treatment ($A_i^{x,n,\mathrm{E}}$). How these variables are interpreted as an endpoint depends on whether they are considered in absolute or relative terms, or whether they instead are a hazard ratio (HR).

Absolute endpoints: In the case of “Absolute” endpoints (all except for NT-proBNP and MACE) the measurement considered for each patient is the difference between these two values, i.e. the change from baseline:

$$C_i^{x,n} = A_i^{x,n,\mathrm{E}} - A_i^{x,n,\mathrm{B}}$$

The $C_i^{x,n}$ are assumed to be identically distributed random variables across all patients and arms with expected value $E[C_i^{x,n}] = c_i^x$, and the standard deviation given in the “SD” column. The endpoint is the difference between the expected values for the two arms:

$$c_i := c_i^1 - c_i^0$$

The C_i cannot be observed directly; they are instead estimated as:

$$\hat{c}_i = \frac{1}{N}\left(\sum_{n=1}^{N} C_i^{1,n}\right) - \frac{1}{N}\left(\sum_{n=1}^{N} C_i^{0,n}\right)$$

The TV and LRV are possible values of c_i, and the SD refers to the standard deviation of $C_i^{x,n}$.

Relative endpoints: In the case of a relative endpoint i (in this case, only NT-proBNP), the variable of interest for each patient is the relative change with respect to the initial measurement:

$$C_i^x = \frac{A_i^{x,n,\mathrm{E}}}{A_i^{x,n,\mathrm{B}}}$$

The $C_i^{x,n}$ are assumed to be identically distributed random variables such that $E[\log(C_i^{x,n})] = \log(c_i^x)$. The "SD" in the table refers to the standard deviation of $\log(C_i^{x,n})$, and the endpoint is the ratio between these values for the two arms:

$$c_i := \frac{c_i^1}{c_i^0}$$

A consistent estimator for c_i is the ratio of geometric means:

$$\hat{c}_i = \frac{\exp\left(\frac{1}{N}\sum_{n=1}^{N}\log\left(C_i^{1,n}\right)\right)}{\exp\left(\frac{1}{N}\sum_{n=1}^{N}\log\left(C_i^{0,n}\right)\right)}$$

The TV and LRV are given as percentage changes; i.e. as possible values of $((c_i - 1)\cdot 100\%)$. This is estimated from the study data by $((\hat{c}_i - 1)\cdot 100\%)$.

Hazard ratio: The hazard ratio (HR) is taken as the number of events (MACE) in the treatment arm divided by the number of events in the control arm. The yearly event rate is assumed to be around 10%. Over a 6 month follow-up period, the number of events in each arm is approximately 5% of the number of participants in that arm. If the number of participants in each arm is N, an approximation for the standard error of the logarithm of the HR is as follows [MCP06, Equation 3.8]:

$$\mathrm{SE}_{\log(\mathrm{HR})} \approx \sqrt{\frac{4}{2N\cdot 0.05}} \tag{4.1}$$

This approximation can be derived by assuming that the number of events in each arm is a Poisson-distributed with mean $0.05N$ and applying the delta method to the logarithm of these random variables. The resulting approximations to the expected value and variance are then used as the mean and variance parameters of a normal distribution.

Following the same rationale as in §4.1.2, because the logarithm of the hazard ratio is an endpoint for which a decrease is desired, the negation of this value is used as the endpoint in our framework. The approximation of the standard error is equally suitable for the negated endpoint ($\mathrm{SE}_{\text{-}\log(\mathrm{HR})} = \mathrm{SE}_{\log(\mathrm{HR})}$).

4.1.2 Transformed endpoints

The framework in Chapter 2 imposes some constraints regarding the distribution of the endpoints, the target value (TV) and the lower reference value (LRV). This means that some transformations need to be applied before proceeding with the case study.

Normally distributed endpoints: If an endpoint i can be assumed to be normally distributed, then the measurement of variable i for patient is the change from baseline $Y_i^{x,n} := C_i^{x,n}$, so that its distribution is assumed to be $Y_i^{x,n} \sim N(c_i^x, \sigma_i^2)$. The standard deviation (σ_i) and the reference values TV_i and LRV_i are as given in Table 4.2. These values must fulfill $0 < \text{LRV}_i < \text{TV}_i$. How to handle the cases where this does not hold is described below.

Decreasing endpoints: The framework in Chapter 2 imposes that, for each endpoint, $0 < \text{LRV}_i < \text{TV}_i$. To ensure that this is the case in endpoints where a decrease in the measured value is desired, (i.e. $0 > \text{LRV}_\text{a} > \text{TV}_\text{a}$), where LRV_a and TV_a are the values given in the table for that endpoint, one can take $\text{LRV}_i := -\text{LRV}_\text{a}$, $\text{TV}_i := -\text{TV}_\text{a}$, and $Y_i^{x,n} := -C_{i,n}$. In this case, with $Y_i^{x,n} \sim N(-c_i^x, \sigma_i)$, where the standard deviation (σ_i) is as given in Table 4.2.

Log-normal distributed endpoints: The framework imposes that the endpoints are normally distributed. If an endpoint i is assumed log-normal distributed (a common approach for variables which are strictly positive and skewed to the right, which is often the case for relative percentage changes), then the logarithm of the endpoint is assumed to be normally distributed. In this case, the measurement is taken as $Y_i^{x,n} := \log(C_i^{x,n})$. It follows that $Y_i^{x,n} \sim N(\log(c_i^x), \sigma_i)$, with σ_i being the SD for the logarithm of the random variable. Note that the assumed distribution for $Y_i^{x,n}$ does not imply that $E[\exp(\widehat{\mu}_i^x)] = c_i^x$ (or, for that matter, that $E[\exp(\widehat{\mu}_i)] = c_i$); this holds only asymptotically as $N \to \infty$.

If the TV and LRV are given as a relative percentage increases, then the reference values are taken as $\text{TV}_i := \log(1 + \text{TV}_\text{a})$ and $\text{LRV}_i := \log(1 + \text{LRV}_\text{a})$, where TV_a and LRV_a are the desired relative changes. For instance, if $\text{TV}_i := \log(1.2)$, this means that it is desired that (at least asymptotically) the mean value of the change from baseline in the treatment group is 20% larger than the value for the control group.

If the endpoint is such that a decrease is desired (as is the case with "NT-proBNP"), then $Y_i^{x,n} := -\log(C_i^{x,n})$, $\mu_i^x := -\log(c_i^x)$, $\text{TV}_i := -\log(1 + \text{TV}_\text{a})$ and $\text{LRV}_i := -\log(1 + \text{LRV}_\text{a})$, with TV_a and LRV_a as given in Table 4.2. In this case, $Y_i^{x,n} \sim N(-\log(c_i^x), \sigma_i)$ with σ_i being the SD of the logarithm of the change from baseline, as given in the same table.

Transformed values and covariance matrix: Following the Chapter 2 framework, this results in a model with $D = 5$ domains and $V = 9$ variables, with $V_1 = V_3 = V_5 = 1$, $V_2 = 2$, $V_4 = 4$. The transformed effects are shown in Table 4.3. The covariance matrix Σ follows the structure outlined in §3.1 ("True effects", page 39). Taking $\rho = 0.4$ (within-domain correlation) and $\tau = 0.2$ (between-domain correlation) yields the following covariance matrix

for the variables:

$$\Sigma := \begin{pmatrix} 0.64 & 11.20 & 0.32 & 3.20 & 0.40 & 1.12 & 1.92 & 1.60 & 0.72 \\ 11.20 & 4900.00 & 56.00 & 280.00 & 35.00 & 98.00 & 168.00 & 140.00 & 62.61 \\ 0.32 & 56.00 & 4.00 & 8.00 & 1.00 & 2.80 & 4.80 & 4.00 & 1.79 \\ 3.20 & 280.00 & 8.00 & 400.00 & 10.00 & 28.00 & 48.00 & 40.00 & 17.89 \\ 0.40 & 35.00 & 1.00 & 10.00 & 6.25 & 7.00 & 12.00 & 10.00 & 2.24 \\ 1.12 & 98.00 & 2.80 & 28.00 & 7.00 & 49.00 & 33.60 & 28.00 & 6.26 \\ 1.92 & 168.00 & 4.80 & 48.00 & 12.00 & 33.60 & 144.00 & 48.00 & 10.73 \\ 1.60 & 140.00 & 4.00 & 40.00 & 10.00 & 28.00 & 48.00 & 100.00 & 8.94 \\ 0.72 & 62.61 & 1.79 & 17.89 & 2.24 & 6.26 & 10.73 & 8.94 & 20.00 \end{pmatrix}$$

The diagonal elements of Σ are σ_i^2 $(i = 1, ..., 9)$ the remaining elements in the blocks along the diagonal are $\rho\sigma_i\sigma_j$ $(i, j = 1, ..., 9)$, and the remaining elements are $\tau\sigma_i\sigma_j$ $(i, j = 1, ..., 9)$. With N patients per arm, the covariance matrix for the estimator of the endpoint effects $(\hat{\mu})$ is:

$$\Sigma_\mu := \frac{2}{N}\Sigma$$

Remark 4.2. The value of σ_9 is chosen so that the standard error for that variable $(\sqrt{(\Sigma_\mu)_{9,9}})$ coincides with approximation for the hazard ratio from Equation 4.1:

$$\sqrt{(\Sigma_\mu)_{9,9}} = \sqrt{\frac{2}{N}\sigma_9^2} = \sqrt{\frac{2}{N} \cdot \frac{1}{0.05}} = \sqrt{\frac{4}{2N \cdot 0.05}}$$

Table 4.3: Transformed endpoints such that $0 < \text{LRV}_i < \text{TV}_i$*. The resulting transformed endpoints are assumed to be normally distributed. The variables follow the same order as in Table 4.2. Domain names are abbreviated, and the transformation applied to each original variable is indicated in the variable name.*

d	Domain	i	Variable/Endpoint	TV_i	LRV_i	σ_i
1	Biomarker	1	−log(NT.proBNP)	0.163[a]	0.051[b]	0.8
2	Exercise	2	6MWD	20	12	70
2	Exercise	3	VO2max	1	0.7	2
3	Well-being	4	KCCQ.TSS	5	2	20
4	Imaging	5	−GLS	0.75	0.25	2.5
4	Imaging	6	−LAVI	2	0.5	7
4	Imaging	7	−LVMI	8	4	12
4	Imaging	8	LVEF	4	2	10
5	Events	9	−log(MACE.HR)	0.223[c]	0.105[d]	4.47[e]

[a] $\text{TV}_1 := -\log(1 - 0.15)$
[b] $\text{LRV}_1 := -\log(1 - 0.05)$
[c] $\text{TV}_9 := -\log(0.8)$
[d] $\text{LRV}_9 := -\log(0.9)$
[e] $\sigma_9 := \sqrt{1/0.05} \approx 4.47$ (see Remark 4.2).

4.2 Policy requirements

The criteria used to produce a decision given the outcome of a study are defined based on the knowledge of the medical experts, who take into account the expected mechanism of action of the drug and the characteristics of the patient population, among many other factors.

For this case study, the policy is formulated first in terms of each of the five domains under consideration; and then those domain-level decisions are combined into a single overall decision.

At the domain level, a domain should produce a Go decision if any of its endpoints produces a Go decision; and a Stop decision if all the endpoints in that domain produce a Stop decision. The domain-level decisions are combined according to one of the following two policies.

1. An "all domains equal" policy, where:
   (i) The decision is Go if at least 2 domains are Go,
   (ii) The decision is Stop if no domains are Go.
2. A "hierarchical domains" policy, where:
   (i) The decision is Go if either the Exercise domain is Go, or at least 2 of the other domains are Go.
   (ii) The decision is Stop if the Exercise domain is Stop *and* at least two of the other domains are Stop.

Both at the domain level, and for each of the two policies, a Stop has priority over the Go decision. Furthermore, if a statistically significant effect in the opposite direction to the TV/LRV is observed for any of the endpoints, an overall Go decision is precluded. Under the assumption $0 < \text{LRV}_i < \text{TV}_i$, this means that a Go decision is precluded (i.e. eventually becomes "Discuss") if an statistically significant, negative effect is observed for any of the endpoints. The significance level is taken as $\alpha = 0.05$ (one-sided).

Remark. In the notation of Chapter 3, the two policies above correspond to $G^{\text{all.eq}}_{x:=2,z:=0,\alpha:=0.05}$ and $G^{\text{hier}}_{x:=2,\alpha:=0.05}$, respectively; with the caveat that, for the latter, the most important domain is the second one (in contrast to Equation 3.27, where the most important domain is the first).

4.2.1 Decision probabilities

For a study to be worthwhile, both financially and ethically, it should have a high probability of success. Success is understood as producing a non-erroneous, actionable decision. This would be either to Go ahead with an effective treatment; or to Stop development of an ineffective one. Following consultations with practitioners, the following criteria to what constitutes an "acceptably high" probability are defined:

(CT) When $\boldsymbol{\mu} \geq \mathbf{TV}$ (that is, all the endpoints reach the target value), a Go decision should be produced with at least 90% probability.

(CTc) When $\boldsymbol{\mu}_{I(d)} \geq \mathbf{TV}_{I(d)}$ for both $d = 5$ and at least 3 of $d = 1, 2, 3, 4$ (that is, for at least 4 domains, one of which is the Events domain, all endpoints in those domains reach the target value), a Go decision should be produced with at least 80% probability.

(C0) When all endpoints have no effect ($\boldsymbol{\mu} \leq \mathbf{0}$), a Stop decision should be reached with at least 80% probability.

(C0') Under the same circumstances as (C0), a Go decision should be reached with less than 5% probability.

(C0c) When $\boldsymbol{\mu} \in R_{\text{case}}^{\text{Stop}}$ (i.e. four out of five domains have no effect, including Events, and the remaining one has an effect lower than the LRV for all endpoints), a Stop decision should be reached with at least 70% probability.

(C0c') Under the same circumstances as (C0c), a Go decision should be reached with less than 5% probability.

Note that all these criteria refer to the true mean effect ($\boldsymbol{\mu}$), not to the observed values in a particular study ($\hat{\boldsymbol{\mu}}$). Due to random variability these two will differ almost surely.

4.3 Experimental setup

The aim of the following experiments is to determine the number of patients per arm that is needed to fulfill the criteria in §4.2.1. The probabilities associated with these criteria are calculated using simulations. For simplicity, only two arms are considered (treatment and control).

True effects: When evaluating a policy, only a finite number of scenarios can be simulated. For this case study, the first scenarios under consideration are those where the true effects are consistent across endpoints: either all endpoints reach the TV ($\llcorner(\mathbf{TV})$), all endpoints are below the LRV ($\neg(\mathbf{LRV})$) or all endpoints have zero or negative effect ($\neg(\mathbf{0})$).

It is possible that an effective treatment will not affect all of the endpoints across all of the domains in the study. Therefore, the policies are also evaluated in scenarios where possibly only three of the disease markers (plus the Phase III endpoint, i.e. the HR for adverse events; $d = 5$) reach the target value:

$$R_{\text{case}}^{\text{Go}} := \{\boldsymbol{\Delta} \in \mathbb{R} \mid \boldsymbol{\Delta}_{I(d)} \geq \mathbf{TV}_{I(d)} \text{ for } d = 5 \text{ and for 3 of } d = 1, 2, 3, 4\} \cap \llcorner(\mathbf{0})$$

Similarly, a treatment may show a minimal, clinically relevant effect in one of the domains, even if it is ultimately ineffective. Therefore, true effects where only three of the disease markers (plus the Phase III endpoint, $d = 5$) show no effect are considered:

$$R_{\text{case}}^{\text{Stop}} := \{\boldsymbol{\Delta} \in \mathbb{R} \mid \boldsymbol{\Delta}_{I(d)} \leq \mathbf{0} \text{ for } d = 5 \text{ and for 3 of } d = 1, 2, 3, 4\} \cap \neg(\mathbf{LRV})$$

Note that for all the true effects in $R^{\text{Go}}_{\text{case}}$, the Phase III endpoint reaches the TV. Conversely, for all true effects in $R^{\text{Stop}}_{\text{case}}$, the treatment has no effect (or negative effect) on the Phase III endpoint. This is consistent with the fact that whether a Go or Stop decision is desirable ultimately depends on whether the endpoint for an eventual Phase III trial reaches its target value.

Per-arm effects: In this setting, the true effects under consideration are potential values for the endpoints ($\boldsymbol{\mu}$). The trial model defines $\boldsymbol{\mu} := \boldsymbol{\mu}^1 - \boldsymbol{\mu}^0$, with $\boldsymbol{\mu}^1$ being the expected change-from-baseline (CFB) in the treatment arm, and $\boldsymbol{\mu}^0$ being the expected CFB in the control arm. When simulating, the effect in the control arm is fixed as $\boldsymbol{\mu}^0 := \mathbf{0}$, with $\boldsymbol{\mu}^1 = \boldsymbol{\mu}$. Because all inferences are done on $\boldsymbol{\mu}$ and Σ_μ and the estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}_\mu$ are simulated directly from their theoretical distributions, this choice is ultimately immaterial.

Number of patients: Due to financial and ethical constraints, the number of participants in a typical Phase II study will not exceed the hundreds. Therefore, for the purposes of this case study, it suffices to simulate numbers of patients per arm between 10 and 600. The range is explored in increments of 5, which means that the number of patients required for a study with sufficient probability of success may be overestimated by up to 4 patients per arm (i.e. 8 total patients).

Number of simulations: The probabilities of each decision for a given combination true effect and number of patients are estimated using M simulations, with $M = 50000$. This choice of M bounds the error in the estimated probabilities to less than 1 percentage point with >95% confidence.

4.4 Policies and simulation results

The policies are first defined at the domain level. For each endpoint $i = 1, ..., 9$, and for a given significance level α, consider the following thresholds, the first two of which are derived from the decision framework (§2.1):

$$\begin{aligned}
\text{thr}^{\text{go}}_{i,\alpha} &:= \text{LRV}_i + \hat{\sigma}_{\mu,i} t_{\nu,1-\alpha} \\
\text{thr}^{\text{stop}}_{i,\alpha} &:= \text{TV}_i + \hat{\sigma}_{\mu,i} t_{\nu,\alpha} \\
\text{thr}^{\text{neg}}_{i,\alpha} &:= \hat{\sigma}_{\mu,i} t_{\nu,\alpha},
\end{aligned}$$

where $\hat{\sigma}_{\mu,i}$ is the estimator for the standard error of the ith endpoint (for N patients per arm, $\sqrt{2/N}\hat{\sigma}_i$); ν is the degrees of freedom for this estimator (for N patients per arm and 2 arms, $\nu = 2N - 2$), and $t_{\nu,\alpha}$ is the α quantile of a Student-t distribution with ν degrees of freedom.

Let $\alpha^\star_{\text{FG}}$ be the maximum desired False Go rate, and $\alpha^\star_{\text{FS}}$ the maximum tolerated False Stop rate (per convention, $\alpha^\star_{\text{FG}} := 0.2$ and $\alpha^\star_{\text{FS}} := 0.1$). The arguments $\tilde{\boldsymbol{\mu}}$, $\tilde{\Sigma}$ are fixed to the random variables $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$, and no structure is imposed on the covariance matrix ($c = \text{Unstructured}_N$). Under these conditions, the per-domain subpolicies $G_{(1)}, ..., G_{(5)}$ are implemented as follows:

$$G_{(1)} := \begin{cases} \text{stop} & \hat{\mu}_1 \leq \text{thr}^{\text{stop}}_{1,\alpha^\star_{\text{FS}}} \\ \text{go} & \hat{\mu}_1 \geq \text{thr}^{\text{go}}_{1,\alpha^\star_{\text{FG}}} \end{cases}$$

$$G_{(2)} := \begin{cases} \text{stop} & \widehat{\mu}_2 \leq \text{thr}^{\text{stop}}_{2,\alpha^\star_{\text{FS}}} \text{ and } \widehat{\mu}_3 \leq \text{thr}^{\text{stop}}_{3,\alpha^\star_{\text{FS}}} \\ \text{go} & \widehat{\mu}_2 \geq \text{thr}^{\text{go}}_{2,(\alpha^\star_{\text{FG}}/2)} \text{ or } \widehat{\mu}_3 \geq \text{thr}^{\text{go}}_{3,(\alpha^\star_{\text{FG}}/2)} \\ & \text{or } (\widehat{\mu}_2 \geq \text{thr}^{\text{go}}_{2,\alpha^\star_{\text{FG}}} \text{ and } \widehat{\mu}_3 \geq \text{thr}^{\text{go}}_{3,\alpha^\star_{\text{FG}}}) \end{cases}$$

$$G_{(3)} := \begin{cases} \text{stop} & \widehat{\mu}_4 \leq \text{thr}^{\text{stop}}_{4,\alpha^\star_{\text{FS}}} \\ \text{go} & \widehat{\mu}_4 \geq \text{thr}^{\text{go}}_{4,\alpha^\star_{\text{FG}}} \end{cases}$$

$$G_{(4)} := \begin{cases} \text{stop} & \text{all of } \left(\widehat{\mu}_i \leq \text{thr}^{\text{stop}}_{i,\alpha^\star_{\text{FS}}} \mid i = 5,6,7,8\right) \\ \text{go} & \text{at least 1 of } \left(\widehat{\mu}_i \geq \text{thr}^{\text{go}}_{i,(\alpha^\star_{\text{FG}}/4)} \mid i = 5,6,7,8\right) \\ & \text{or at least 2 of } \left(\widehat{\mu}_i \geq \text{thr}^{\text{go}}_{i,(2\alpha^\star_{\text{FG}}/4)} \mid i = 5,6,7,8\right) \\ & \text{or at least 3 of } \left(\widehat{\mu}_i \geq \text{thr}^{\text{go}}_{i,(3\alpha^\star_{\text{FG}}/4)} \mid i = 5,6,7,8\right) \\ & \text{or all of } \left(\widehat{\mu}_i \geq \text{thr}^{\text{go}}_{i,\alpha^\star_{\text{FG}}} \mid i = 5,6,7,8\right) \end{cases}$$

$$G_{(5)} := \begin{cases} \text{stop} & \widehat{\mu}_9 \leq \text{thr}^{\text{stop}}_{9,\alpha^\star_{\text{FS}}} \\ \text{go} & \widehat{\mu}_9 \geq \text{thr}^{\text{go}}_{9,\alpha^\star_{\text{FG}}} \end{cases}$$

Policies $G_{(2)}$ and $G_{(4)}$ use the Simes procedure (a particular case of the Benjamini-Hochberg procedure) in order to compensate for multiple comparisons. As explained and demonstrated in §3.2.2, this means that, for all the domains (including, in particular, $d = 2$ and $d = 4$), the FSR when any endpoint read above the TV is less than $\alpha^\star_{\text{FS}}$, and the FGR when all endpoints read below the LRV is less than $\alpha^\star_{\text{FG}}$.

Consider, for example, $N = 155$ patients per arm. (The relevance of $N = 155$ will become apparent when discussing the required size of the study; §4.4.1 and §4.4.2). The thresholds for $N = 155$ patients per arm are shown in Figure 4.1. For a Stop decision there is a single threshold under which all observed mean effects must fall for the domain to be Stop. For a Go decision, whether an observed value for an endpoint is high enough to declare a Go for the domain, depends on how high the values of the other endpoints are. This is due to use of the Simes procedure when performing multiple comparisons on variables within the same domain. Using this method, if all observed values are above the lowest of the Go thresholds (which corresponds to the threshold from the single-variable decision framework) a Go decision will be produced. But it also suffices if the observed value for a single one of the endpoints is above the highest of these thresholds, which corresponds to dividing the associated significance level by the number of endpoints in the domain; i.e. the Bonferroni correction. As explained in Figure 4.3, other combinations of number of endpoints and thresholds are also allowed.

This adjustment for multiple comparisons is overall more conservative than not performing such an adjustment, but not uniformly so. As exemplified in Figure 4.2, in some circumstances, an observed value for an endpoint which is low-enough to be a "Stop" can contribute towards a domain-level "Go" decision when the observed values for the other domains are high enough. This is consistent with the specification of the decision framework, where the probability of producing Go when all endpoints are below the LRV is bounded by $\alpha^\star_{\text{FG}} = 0.2$.

Regarding the Phase III endpoint (MACE), the thresholds are far away

Figure 4.1: Graphical representation of the thresholds on the observed effects entailed by the domain-level policies $G_{(1)}, ..., G_{(5)}$ for $N = 155$ patients per arm, with $\alpha^{\star}_{\mathrm{FG}} = 0.2$ and $\alpha^{\star}_{\mathrm{FS}} = 0.1$. The background of each plot is shaded according to the decision produced by the single-variable policy. For a Stop *decision at the domain level, the threshold coincides with that of the single-variable decision framework ($\mathrm{thr}^{\mathrm{stop}}_{\alpha^{\star}_{\mathrm{FS}}}$): all observed values for all endpoints must be below $\mathrm{thr}^{\mathrm{stop}}_{\alpha^{\star}_{\mathrm{FS}}}$ for the domain* Stop *decision to be produced. For a* Go *decision at the domain level, each threshold has an associated required number of endpoints. If an endpoint is above its "all" threshold (which coincides with $\mathrm{thr}^{\mathrm{go}}_{\alpha^{\star}_{\mathrm{FG}}}$ in the univariate decision framework) then it suffices for all other endpoints to also be above the "all" threshold to obtain a* Go*. If an endpoint is above the "all minus 1" threshold, then it suffices for all but one of the other endpoints to also be above the "all minus 1" threshold to obtain a* Go*. For a domain with 3 endpoints, it is sufficient for 1 endpoint to be above the "all minus 2" threshold for the whole domain to be declared* Go *($3 - 2 = 1$). As a visual aid, the middle bar is shaded according to these thresholds, with intermediate colors between* Go *and* Discuss*/*Stop *when a* Go *for that endpoint does not automatically imply a* Go *for the entire domain. Note that if the observed effect for an endpoint is in both the* Stop *and the* Go *region, it could contribute to a* Go *decision if other endpoints in the same domain are* Go*.*

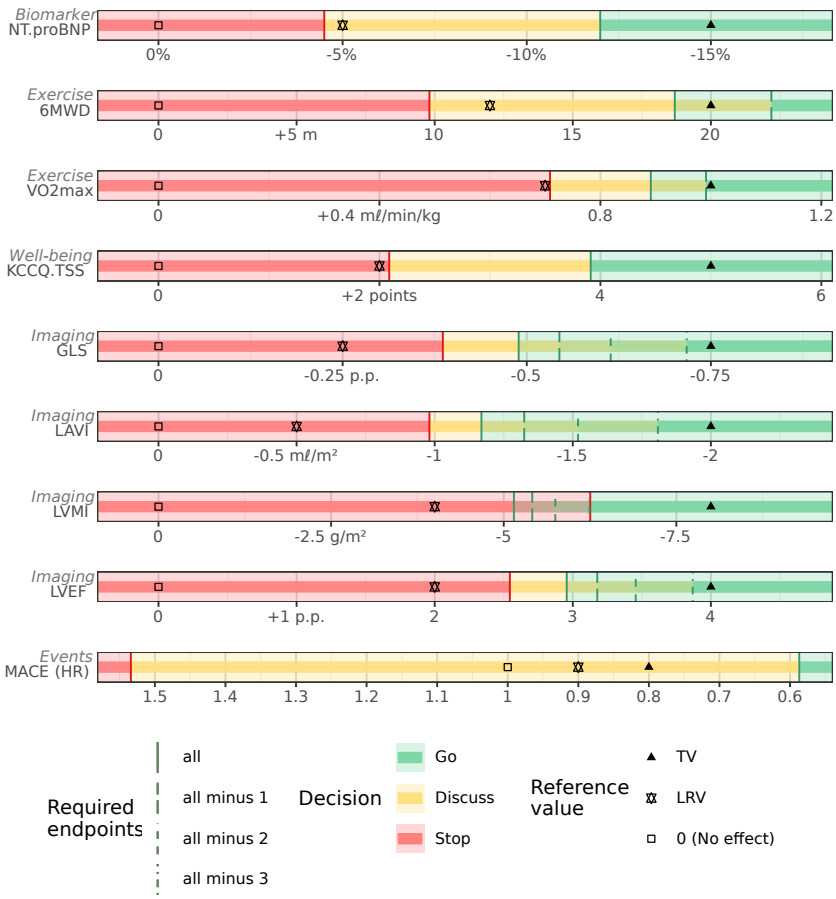
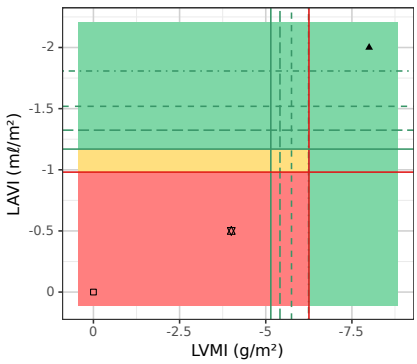
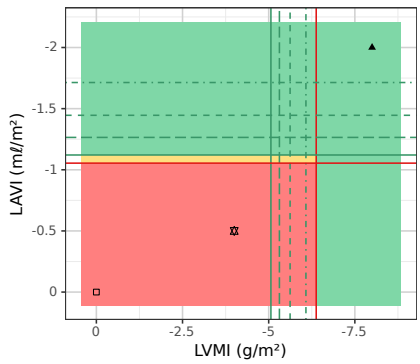


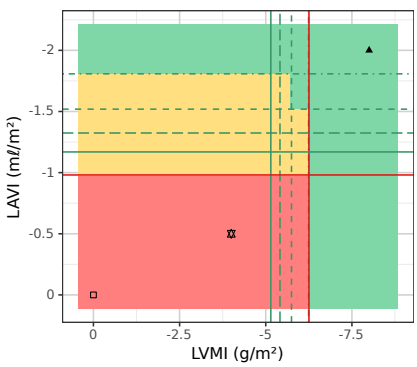
Figure 4.2: Domain level decision for the "Imaging" domain depending on the observed values for LVMI and LAVI, with $N = 155,180$ patients per arm. The observed effect for LVEF and GLS is 0, which is well below Stop *threshold for these endpoints. Subfigures (a) and (b) represent the decisions with an unadjusted policy (an instance of G^{ind}). In such a policy, given endpoint-level decisions based on the univariate decision framework (§2.1), a* Go *decision for a domain is produced when any of its endpoints are* Go*, and a* Stop *decision for a domain is produced when all of its endpoints are* Stop*. Subfigures (c) and (d) represent the domain-level policy that is used in the case study for the "Imaging" endpoints. The meaning of the thresholds is as described in Figure 4.1. Note that in Subfigures (a) and (c), the "all minus 3" "*go*" threshold for LVMI happens to overlap with the "*stop*" threshold.*



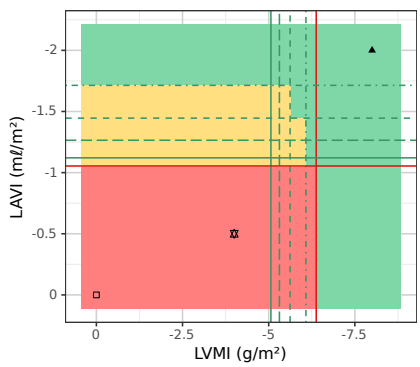
(a) G^{ind}, $N = 155$



(b) G^{ind}, $N = 180$



(c) $G_{(4)}$ ($G^{\text{ind.Simes}}$), $N = 155$



(d) $G_{(4)}$ ($G^{\text{ind.Simes}}$),$N = 180$

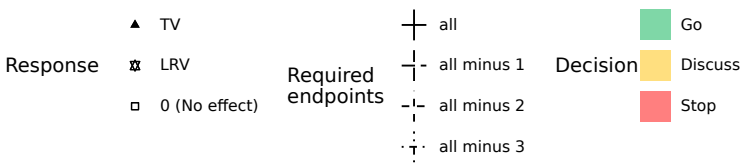
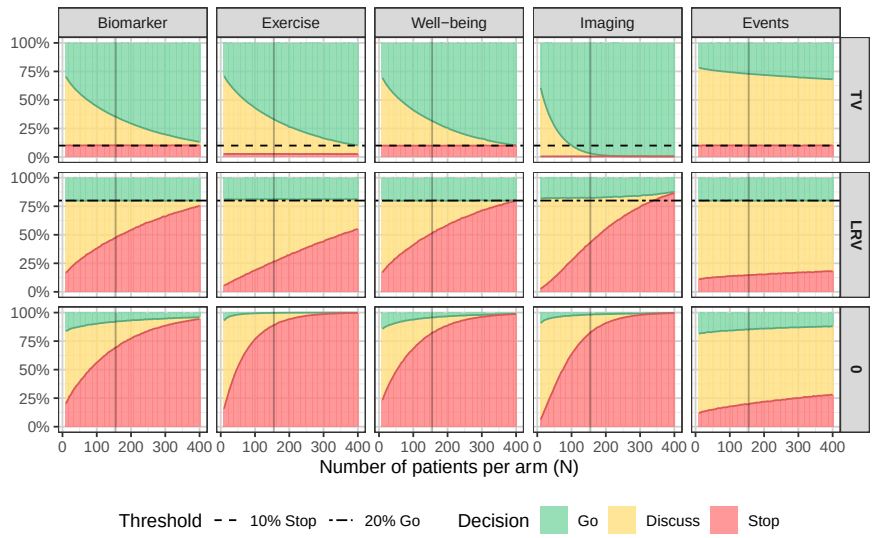


Figure 4.3: Decision distribution of the domain subpolicies ($\alpha^{\star}_{\text{FS}} = 0.1$, $\alpha^{\star}_{\text{FG}} = 0.2$). For each number of patients per arm (N), the probabilities of a Go, Stop *and* Discuss *decision are shown. When the true effect is* **TV**, *the rate of* Stop *is always less than 10%. For those domains with more than one endpoint, this rate is lower, due to the fact that all endpoints in the domain need to be under their* Stop *threshold for the decision to be* Stop. *When the true effect is* **LRV**, *the* Go *rate is bounded by 20%, as expected from the multiple comparisons adjustment that is performed. As the number of patients increases, this rate becomes lower, due to the fact that* Stop *has priority over* Go. *For domains with more than one endpoint, this rate is also decreased due to the fact that the multiple comparisons adjustment is too conservative in the presence of correlation between endpoints (Table 3.5 and Figure 3.5).*



from the reference values (see "MACE (HR)" in Figure 4.1). These high thresholds are consistent with the low power of the endpoint (see Table 4.4). If taken at face value, they would indicate that the treatment is either implausibly beneficial, or an acutely toxic drug; but an observed value surpassing them could just as well be the result of random variation. The high Go and Stop thresholds for the MACE endpoint also manifest themselves in the distribution of decisions for the corresponding domain. As shown in Figure 4.3, the probabilities of each decision for the Phase III endpoint (i.e. the "Events" domain) are relatively similar across the three true effects tested ($\mu_9 = 0, \text{LRV}_9, \text{TV}_9$) for all the numbers of patients per arm evaluated. Intuitively, this means that, for plausible true effects, observing a Go or Stop decision for the "Events" domain is not particularly informative of what the true effect actually is. The low usefulness of this endpoint underscores the need for the other disease markers.

Table 4.4: Power of each endpoint in the case study. The power is calculated for a two-sided, two sample t-test with $\alpha = 0.05$ and $N = 155$ values in each sample. The effect size is the target value for the corresponding endpoint.

Domain	Endpoint	deg.f.	SE	Power
Biomarker	-log(NT.proBNP)	308	0.091	0.430
Exercise	6MWD	308	7.951	0.708
Exercise	VO2max	308	0.227	0.992
Well-being	KCCQ.TSS	308	2.272	0.593
Imaging	-GLS	308	0.284	0.749
Imaging	-LAVI	308	0.795	0.708
Imaging	-LVMI	308	1.363	1.000
Imaging	LVEF	308	1.136	0.939
Events	-log(MACE.HR)	308	0.508	0.064

4.4.1 "All domains equal" policy

The domain level decisions may be combined in different ways, depending on which disease markers are considered most relevant for the treatment under study. This choice will affect the probabilities of the different outcomes for the overall policy. In this section, a policy where all domains are given the same importance is considered.

The "all domains equal" policy for the case study is an instance of Equation 3.25 with $x := 2$, $z := 0$ and $\alpha := 0.05$:

$$G^{\text{all.eq}}_{\text{case}} := \begin{cases} \text{stop} & \text{none of } \left(G_{(d)} \text{ is Go} \mid d = 1,2,3,4,5\right) \\ \text{go} & \text{at least 2 of } (G_{(d)} \text{ is Go} \mid d = 1,2,3,4,5) \\ & \text{and none of } \left(\hat{\mu}_i \leq \text{thr}^{\text{neg}}_{i,0.05} \mid i = 1,...,9\right) \end{cases}, \tag{4.3}$$

The results of evaluating the $G^{\text{all.eq}}_{\text{case}}$ policy are given in Table 4.5. Fulfilling (CT) requires $N = 145$ patients per arm, while fulfilling (C0) requires $N = 260$, depending on the desired value. Other criteria become impractical to fulfil; for instance (CTc) requires $N = 365$ patients per arm.

The required number of patients may be brought down to a more plausible amount by being more selective with which endpoints are included in the study. The number of included endpoints may be reduced in several ways:

Ignoring a low-powered domain: The Events domain is ultimately decisive in an eventual Phase III trial. However, even when the number of patients is high, the probability of each of the possible decisions for that domain (Go, Stop or Discuss) remains largely unchanged, regardless of the true effect (Figure 4.3). This is likely due to the low power of that endpoint (Table 4.4). This means that it could be advisable to disregard the MACE endpoint for assessing effectiveness (§3.3.6).

The policy $G^{\text{all.eq}}_{\text{w/o.E}}$ defined below ignores this domain for the effectiveness conditions ($d = 5$), but still includes a safety condition for the endpoint in

Table 4.5: For the $G^{\text{all.eq}}_{\text{case}}$, $G^{\text{all.eq}}_{\text{w/o.E}}$ and $G^{\text{all.eq}}_{\text{w/o.E,6M}}$ and $G^{\text{all.eq.w/o.cond}}_{\text{w/o.E,6M}}$, the required number of patients per arm (N) to achieve certain probabilities of a correct decision, depending on the true effect. For each combination of true effects for the domains, the minimum number of patients per arm (N) required to achieve the given rate of Go *or* Stop *decisions is shown. The parameters used are $\alpha^{\star}_{\text{FG}} = 0.2, \alpha^{\star}_{\text{FS}} = 0.1$. The number of patients has a resolution of* 5.

(a) Sets of true effects where a Go *decision is desirable.*

	Biomarker	Exercise	Well-being	Imaging	Events	$G^{\text{all.eq}}$	$G^{\text{all.eq}}_{\text{w/o.E}}$	$G^{\text{all.eq}}_{\text{w/o.E,6M}}$	$G^{\text{all.eq.w/o.cond}}_{\text{w/o.E,6M}}$
all ≥TV									
>90% Go	TV	TV	TV	TV	TV	145	155	150	135
Events and 3 others ≥TV, remaining one ≥0									
>80% Go	TV	TV	TV	0	TV	365	385	385	205
	TV	TV	0	TV	TV	140	160	155	125
	TV	0	TV	TV	TV	165	185	155	130
	0	TV	TV	TV	TV	135	150	145	120

(b) Sets of true effects where a Stop *decision is desirable.*

	Biomarker	Exercise	Well-being	Imaging	Events	$G^{\text{all.eq}}$	$G^{\text{all.eq}}_{\text{w/o.E}}$	$G^{\text{all.eq}}_{\text{w/o.E,6M}}$	$G^{\text{all.eq.w/o.cond}}_{\text{w/o.E,6M}}$
all ≤0									
>70% Stop	0	0	0	0	0	80	20	20	20
>80% Stop	0	0	0	0	0	260	75	70	70
<5% Go	0	0	0	0	0	105	30	30	30
Events and 3 others ≤0, remaining one ≤LRV									
>70% Stop	0	0	0	LRV	0	270	80	75	75
	0	0	LRV	0	0	400	70	65	65
	0	LRV	0	0	0	405	110	125	125
	LRV	0	0	0	0	335	45	40	40
<5% Go	0	0	0	LRV	0	355	105	105	105
	0	0	LRV	0	0	390	65	60	65
	0	LRV	0	0	0	450	135	135	145
	LRV	0	0	0	0	265	50	40	45

that domain ($i = 9$):

$$G^{\text{all.eq}}_{\text{w/o.E}} := \begin{cases} \text{stop} & \text{none of } \left(G_{(d)} \text{ is Go} \mid d = 1,2,3,4\right) \\ \text{go} & \text{at least 2 of } (G_{(d)} \text{ is Go} \mid d = 1,2,3,4) \\ & \text{and none of } \left(\hat{\mu}_i \leq \text{thr}^{\text{neg}}_{i,0.05} \mid i = 1,...,9\right). \end{cases} \tag{4.4}$$

As shown in Table 4.5a, excluding the "Events" endpoint from the main go/Stop criteria reduces the number of patients required to achieve an acceptable Stop rate (Table 4.5b, $G^{\text{all.eq}}_{\text{w/o.E}}$ column), with a limited increase in the number of patients required to fulfill the criteria for the Go rate (Table 4.5a, $G^{\text{all.eq}}_{\text{w/o.E}}$ column).

Ignoring a low-powered endpoint: As observed in §3.2.5, in the presence of correlation between endpoints in the same domain, including variables of lower power into an existing domain makes the rate of Go decisions go down. As seen in Table 4.4, the power of the 6MWD endpoint is low compared to that of VO2max. Furthermore, due to the exertion required by the patient to perform the required task (6 minutes of walking), measuring this endpoint may require additional clinical visits, those increasing the cost of the trial and potentially hindering patient retention.

Assuming that the treatment affects both endpoints in the Exercise domain, and that these endpoints measure the same underlying change (i.e. either they both exhibit no effect, they both reach the LRV, or they both reach the TV), the 6MWD endpoint can be removed from the analysis. It is then assumed that this endpoint is not measured at all, so it is also excluded from the non-negativity check:

$$G^{\text{all.eq}}_{\text{w/o.E,6M}} := \begin{cases} \text{stop} & \text{none of } \left(G'_{(d),\alpha^\star_{\text{FG}}} \text{ is Go} \mid d = 1,2,3,4\right) \\ \text{go} & \text{at least 2 of } (G'_{(d),\alpha^\star_{\text{FG}}} \text{ is Go} \mid d = 1,2,3,4) \\ & \text{and none of } \left(\hat{\mu}_i \leq \text{thr}^{\text{neg}}_{i,0.05} \mid i = 1,3,4,...,9\right) \end{cases} \tag{4.5}$$

$$G'_{(d)} := G_{(d)} \text{ for } d = 1,3,4 \tag{4.6}$$

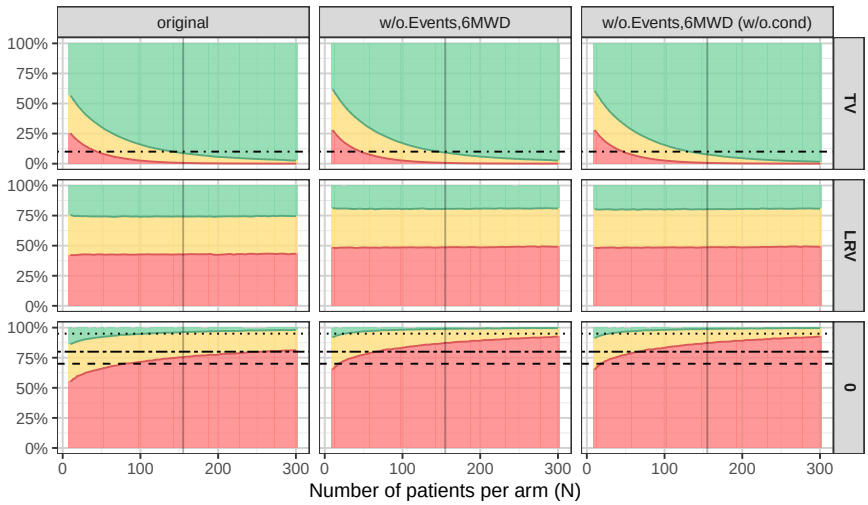
$$G'_{(2)} := \begin{cases} \text{stop} & \hat{\mu}_3 \leq \text{thr}^{\text{stop}}_{3,\alpha^\star_{\text{FS}}} \\ \text{go} & \hat{\mu}_3 \geq \text{thr}^{\text{go}}_{3,\alpha^\star_{\text{FG}}}. \end{cases} \tag{4.7}$$

Simulations show that removing the 6MWD endpoint from the analysis has a small effect in the number of patients required (compare columns $G^{\text{all.eq}}_{\text{w/o.E}}$ and $G^{\text{all.eq}}_{\text{w/o.E,6M}}$ in Table 4.5a). Furthermore, the fact that this endpoint does not need to be measured can lower the cost of performing the trial, all else being equal. However, as mentioned before, this improvement might no longer hold if the effect of the treatment on the Exercise domain was only on the 6MWD endpoint, and not on the VO2max.

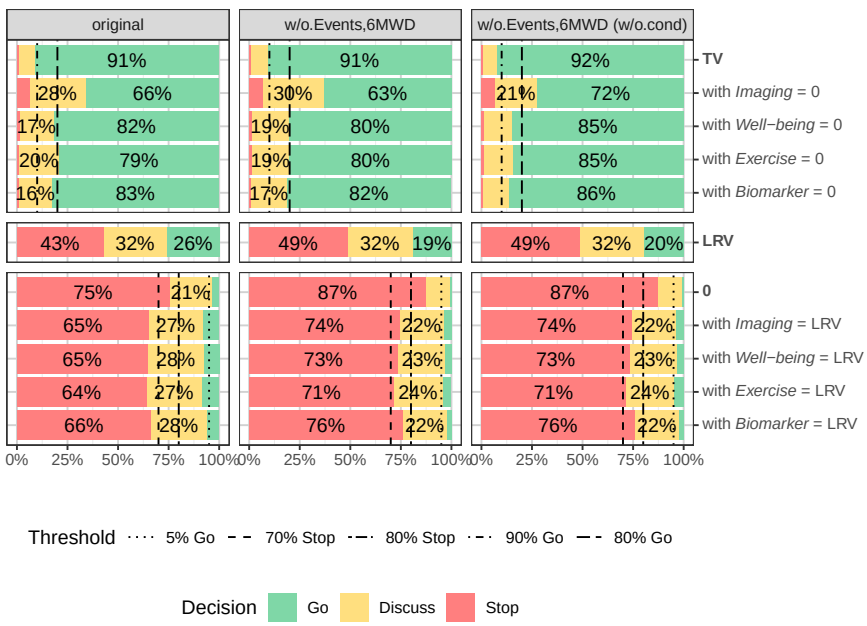
Effect of the safety condition: As a safety check, the policies include an additional condition that prevents a Go decision when any of the observed mean effects is statistically significant in the wrong direction. However, as shown in Table 4.5a (column $G^{\text{all.eq}}_{\text{w/o.E,6M}}$), in cases where the treatment has no

Figure 4.4: For the $G^{\text{all.eq}}_{\text{case}}$ *("original"), and* $G^{\text{all.eq}}_{\text{w/o.E,6M}}$ *("w/o.Events,6MWD"), and* $G^{\text{all.eq.w/o.cond}}_{\text{w/o.E,6M}}$ *("w/o.Events,6MWD (w/o.cond)") with* $\alpha^{\star}_{\text{FG}} = 0.2, \alpha^{\star}_{\text{FS}} = 0.1$ *, the probability of each possible decision is shown depending on the true effect* $(\boldsymbol{\mu} = \mathbf{TV}, \mathbf{LRV}, \mathbf{0})$*. Some probability thresholds indicated.*

(a) Probability of Go,Stop *and* Discuss *depending on the number of patients per arm (N). The vertical line indicates* $N = 155$ *patients per arm.*



(b) Probability of Go,Stop *and* Discuss *for* $N = 155$ *patients per arm. Under* $\boldsymbol{\mu} = \mathbf{TV}$*, the probabilities when each one of the domains has true effect* $\mathbf{0}$ $(\boldsymbol{\mu} \in R^{\text{Go}}_{\text{case}})$*. are shown. Similarly, under* $\boldsymbol{\mu} = \mathbf{0}$*, the probabilites when one of the domains reaches the* **LRV** *are shown.*



effect on any of the endpoints in the Imaging domain (i.e. the effect at the four endpoints is 0) but reaches the target value in all other endpoints, the number of patients required to obtain the desired Go decision at least 80% of the time is impractically high. This may be due to the way the safety condition is affected by multiple comparisons problem. For the sole purpose of quantifying the impact of this condition, we consider a policy that omits it:

$$G^{\text{all.eq.w/o.cond}}_{\text{w/o.E,6M}} := \begin{cases} \text{stop} & \text{none of } \left(G'_{(d),\alpha^\star_{\text{FG}}} \text{ is Go} \mid d = 1,2,3,4\right) \\ \\ \text{go} & \text{at least 2 of } (G'_{(d),\alpha^\star_{\text{FG}}} \text{ is Go} \mid d = 1,2,3,4) \end{cases} \tag{4.8}$$

Comparing $G^{\text{all.eq}}_{\text{w/o.E,6M}}$ and $G^{\text{all.eq.w/o.cond}}_{\text{w/o.E,6M}}$ (Table 4.5) shows that the safety condition nearly doubles the number of patients required per arm (205 vs. 385) in order to achieve the desired probability for the case where the treatment has no effect on endpoints in the "Imaging" domain. As the safety condition is arguably necessary, one course of action is to avoid measuring endpoints for which an effect is not expected, or which are redundant relative to other endpoints (see Figure 4.4b).

Conclusion: Despite its caveats, the safety condition must be included, which precludes the $G^{\text{all.eq.w/o.cond}}_{\text{w/o.E,6M}}$ policy. Therefore, based on the numbers in Table 4.5, the policy $G^{\text{all.eq}}_{\text{w/o.E,6M}}$ is chosen for the hypothetical study, with a study size of $N = 155$ patients per arm. The probabilities of each decision for the resulting design are shown in Figure 4.4b. The criteria (CT), (C0), (C0c), (C0') and (C0c') are fulfilled. Criterion (CTc) is not fulfilled in the case where the effect in Imaging is 0. The probability achieved is 63%, which is 17 percentage points short of the desired 80%.

4.4.2 Hierarchical domains policy

The "hierarchical domains" policy is a variation of the policy in Equation 3.27 with $x := 2$ and $\alpha := 0.05$, taking the second domain ($d = 2$) as the most important domain:

$$G^{\text{hier}}_{\text{case}} := \begin{cases} \text{stop} & G_{(2),\alpha^\star_{\text{FG}}} \text{ is Stop and} \\ & \text{at least 2 of } (G_{(d)} \text{ is Stop} \mid d = 1,3,4,5) \\ \\ \text{go} & \begin{pmatrix} G_{(2),\alpha^\star_{\text{FG}}} \text{ is Go or} \\ \text{at least 2 of } \left(G_{(d)} \text{ is Go} \mid d = 1,3,4,5\right) \end{pmatrix} \\ & \text{and none of } \left(\hat{\mu}_i \leq \text{thr}^{\text{neg}}_{i,0.05} \mid i = 1,...,9\right) \end{cases} \tag{4.9}$$

Following the same reasoning as in §4.4.1, a policy is defined that ignores the "Events" domain (4.10), while still taking it into consideration when checking

Table 4.6: For the $G^{\text{hier}}_{\text{case}}$, $G^{\text{hier}}_{\text{w/o.E}}$, $G^{\text{hier}}_{\text{w/o.E,6M}}$ *and* $G^{\text{hier.w/o.cond}}_{\text{w/o.E,6M}}$, *the required number of patients per arm (N) to achieve certain probabilities of a correct decision, depending on the true effect. For each combination of true effects for the domains, the minimum number of patients per arm (N) required to achieve the given rate of* Go *or* Stop *decisions is shown. The policies use the parameters* $\alpha^{\star}_{\text{FG}} = 0.2, \alpha^{\star}_{\text{FS}} = 0.1$. *The number of patients has a resolution of* 5.

(a) Sets of true effects where a Go *decision is desirable.*

	Biomarker	Exercise	Well-being	Imaging	Events	G^{hier}	$G^{\text{hier}}_{\text{w/o.E}}$	$G^{\text{hier}}_{\text{w/o.E,6M}}$	$G^{\text{hier.w/o.cond}}_{\text{w/o.E,6M}}$
all ≥TV									
>90% Go	TV	TV	TV	TV	TV	140	150	145	130
Events and 3 others ≥TV, remaining one ≥0									
>80% Go	TV	TV	TV	0	TV	330	340	330	160
	TV	TV	0	TV	TV	135	155	145	115
	TV	0	TV	TV	TV	175	185	155	130
	0	TV	TV	TV	TV	125	145	140	110

(b) Sets of true effects where a Stop *decision is desirable.*

	Biomarker	Exercise	Well-being	Imaging	Events	G^{hier}	$G^{\text{hier}}_{\text{w/o.E}}$	$G^{\text{hier}}_{\text{w/o.E,6M}}$	$G^{\text{hier.w/o.cond}}_{\text{w/o.E,6M}}$
all ≤0									
>70% Stop	0	0	0	0	0	135	135	110	110
>80% Stop	0	0	0	0	0	165	170	140	140
<5% Go	0	0	0	0	0	105	45	35	35
Events and 3 others ≤0, remaining one ≤LRV									
>70% Stop	0	0	0	LRV	0	165	175	155	155
	0	0	LRV	0	0	155	165	145	145
	0	LRV	0	0	0	590	590	295	295
	LRV	0	0	0	0	145	150	130	130
<5% Go	0	0	0	LRV	0	215	110	105	105
	0	0	LRV	0	0	190	85	65	70
	0	LRV	0	0	0	>600	>600	>600	>600
	LRV	0	0	0	0	155	60	45	50

for statistically significant observed negative effects:

$$G^{\text{hier}}_{\text{w/o.E}} := \begin{cases} \text{stop} & G_{(2),\alpha^\star_{\text{FG}}} \text{ is Stop and} \\ & \text{at least 2 of } (G_{(d)} \text{ is Stop} \mid d = 1,3,4) \\ \text{go} & \begin{pmatrix} G_{(2),\alpha^\star_{\text{FG}}} \text{ is Go or} \\ \text{at least 2 of } \left(G_{(d)} \text{ is Go} \mid d = 1,3,4\right) \end{pmatrix} \\ & \text{and none of } \left(\hat{\mu}_i \leq \text{thr}^{\text{neg}}_{0.05} \mid i = 1,...,9\right). \end{cases} \tag{4.10}$$

A policy is also defined that completely ignores the "6MWD" endpoint (also when checking for negative significance, as the assumption is that this endpoint would not be measured at all):

$$G^{\text{hier}}_{\text{w/o.E,6M}} := \begin{cases} \text{stop} & G'_{(2),\alpha^\star_{\text{FG}}} \text{ is Stop and} \\ & \text{at least 2 of } (G_{(d)} \text{ is Stop} \mid d = 1,3,4) \\ \text{go} & \begin{pmatrix} G'_{(2),\alpha^\star_{\text{FG}}} \text{ is Go or} \\ \text{at least 2 of } \left(G_{(d)} \text{ is Go} \mid d = 1,3,4\right) \end{pmatrix} \\ & \text{and none of } \left(\hat{\mu}_i \leq \text{thr}^{\text{neg}}_{0.05} \mid i = 1,3,4,...,9\right), \end{cases} \tag{4.11}$$

with $G'_{(2)}$ as in (4.7).

Finally, as in §4.4.1, the impact of the check for statistically significant, negative observed effects is assessed:

$$G^{\text{hier.w/o.cond}}_{\text{w/o.E,6M}} := \begin{cases} \text{stop} & G'_{(2),\alpha^\star_{\text{FG}}} \text{ is Stop and} \\ & \text{at least 2 of } (G_{(d)} \text{ is Stop} \mid d = 1,3,4) \\ \text{go} & G'_{(2),\alpha^\star_{\text{FG}}} \text{ is Go or} \\ & \text{at least 2 of } \left(G_{(d)} \text{ is Go} \mid d = 1,3,4\right). \end{cases} \tag{4.12}$$

The decision probabilities as the number of patients N varies are displayed in Figure 4.5a. Based on the numbers in Table 4.6, a study size of $N = 155$ is chosen. As corroborated by Figure 4.5b, the chosen number of patients per arm suffices to fulfill all the criteria in §4.2.1, provided that the following two situations are considered implausible:

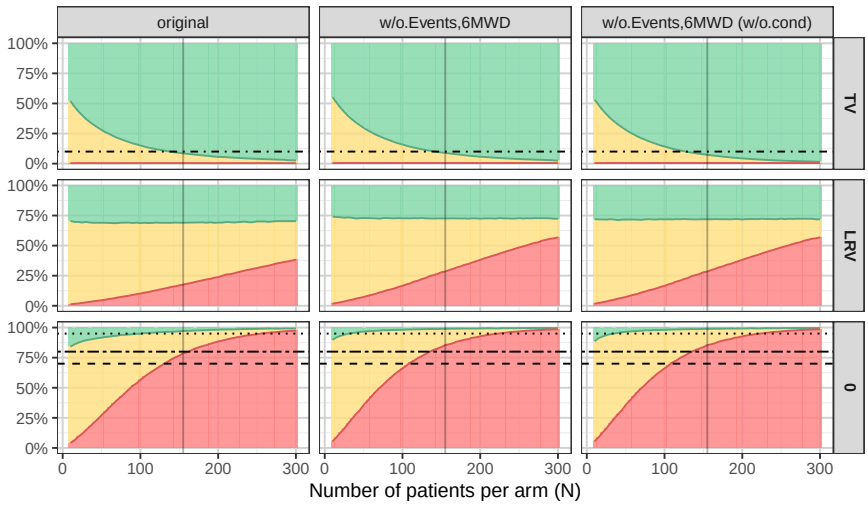
(i) The true effect of a treatment reaches the target value on the Phase III endpoint (MACE.HR in the "Events" domain), but does not reach the target value on any of the endpoints the "Imaging" domain. (i.e. $\boldsymbol{\mu}_{I(5)} \geq \mathbf{TV}_{I(5)}$ and $\boldsymbol{\mu}_{I(4)} \leq \mathbf{0}$).

(i) A treatment has no true effect on the Phase III endpoint, but has a non-zero effect on the most important domain (i.e. $\boldsymbol{\mu} = (0, \mu_{I(2)}, 0, 0, 0)$ with $0 < \mu_{I(2)} \leq \mathbf{LRV}$).

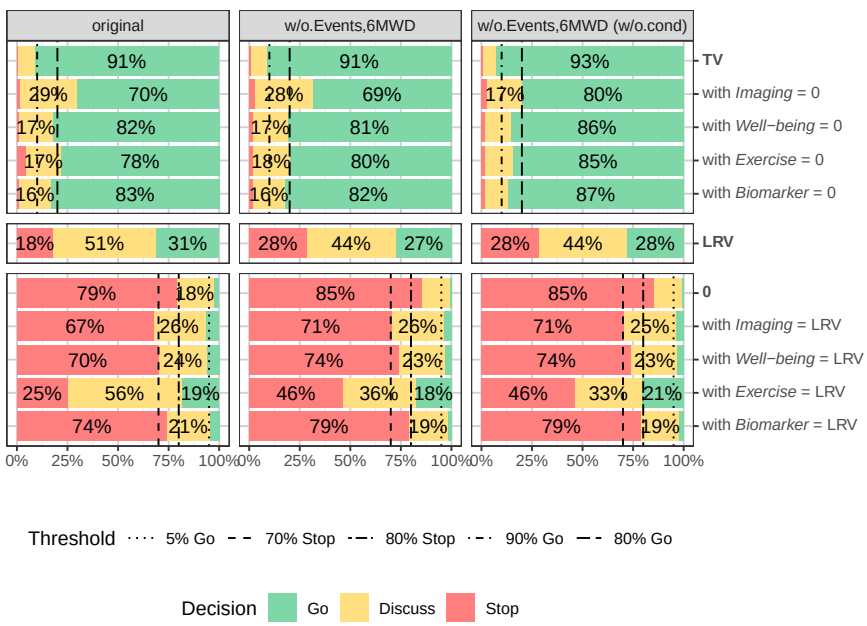
The last situation is directly linked with the fact that Exercise is designated as the most important domain. By the design of the policy, and as Figure 4.5b shows, even in the absence of an effect of any of the other domains and the presence of a safety condition, having a small clinically-relevant effect in the most important domain suffices for the FGR (i.e. the probability of "go", regardless of "stop") to almost equal $\alpha^\star_{\text{FG}} = 0.2$.

Figure 4.5: For the $G^{\text{hier}}_{\text{case}}$ *("original"), and* $G^{\text{hier}}_{\text{w/o.E,6M}}$ *("w/o.Ev,6MWD") and* $G^{\text{hier.w/o.cond}}_{\text{w/o.E,6M}}$ *("w/o.Events,6MWD (w/o.cond)"), the probability of each outcome is indicated. The parameters are* $\alpha^{\star}_{\text{FG}} = 0.2$, $\alpha^{\star}_{\text{FS}} = 0.1$.

(a) Probability of each decision depending on the number of patients per arm (N). The vertical line indicates $N = 155$ *patients per arm.*



(b) Probability of Go,Stop *and* Discuss *for* $N = 155$ *patients per arm. Under* $\boldsymbol{\mu} = \mathbf{TV}$, *the probabilities when each one of the domains has true effect* $\mathbf{0}$ *(*$\boldsymbol{\mu} \in R^{\text{Go}}_{\text{case}}$*). are shown. Similarly, under* $\boldsymbol{\mu} = \mathbf{0}$, *the probabilites when one of the domains reaches the* **LRV** *are shown.*



4.4.3 Generalizability of the simulation results

The simulations in this case study cover specific numbers of patients and a finite number of combinations of true effects. However, when running an actual study, it is unreasonable to expect that the number of patients recruited will be exactly as specified, or that the true effect will be one of the few discrete points that have been simulated.

Generalizability of the number of patients: The framework given in Chapter 2 does not provide any guarantees that increasing the number of patients will increase the probability of a correct decision. However, the empirical trends displayed in Figure 4.3 Figure 4.4a and Figure 4.5a, indicate that, at least when the treatment affects all endpoints in each domain uniformly, the probability of a correct decision does increase with the number of patients in each arm. Furthermore, the patient counts in Table 4.5 and Table 4.6 are such that the desired probability is achieved not only for the given N, but also for all larger N' that have been simulated (i.e. for all $N' = N, N+5, N+10, ..., 600$). This means that, if the table says that $N = 105$ are required to achieve a probability of a given decision, then $N = 135$ patients per arm also suffice.

Generalizability of the decision probabilities: On the other hand, the monotonicity property of policies (Definition 2.25) means that the probabilities calculated for discrete effects generalize to a whole region. Together with the observation in the previous paragraph, this is why, for example, for an "all domains equal" policy (e.g. $G^{\text{all.eq}}_{\text{w/o.E,6M}}$) the fact that when $N = 70$ and true effect is $\mathbf{0}$ the desired probabilities of Go ($> 80\%$) and Stop ($< 5\%$) are achieved (Table 4.5b, "all ≤ 0"), means that the same probabilities (or better) will also be achieved when the true effect for all endpoints is 0 or less (C0), (C0'). The same rationale relates the number of patients per arm for the other true effects to the remaining criteria in §4.2.1.

The monotonicity property of policies also applies to the probabilities in Figure 4.4b and Figure 4.5b. For instance, consider the case with $N = 155$ patients per arm, the policy $G^{\text{hier}}_{\text{w/o.E,6M}}$ (Figure 4.5b, "w/o.Events,6MWD"), and a true effect equal to the TV for all endpoints except for those in the "Imaging" domain, for which it is 0. This gives a probability of Go is 69%, and the probability of Stop is 3%. This in turn means that if the true effect is instead the LRV for some (or all) of the endpoints in the "Imaging" domain, the probability of Go is guaranteed to be at least 69%, and the probability of Stop at most 3%.

4.5 Summary

In a Phase II trial with multiple endpoints divided across several domains, it can be expected that the treatment will not necessarily have a strong effect in all of them, even if it is ultimately effective. In order to fulfill the desired criteria with a reasonable number of patients, the chosen policies ignore certain endpoints which either have very low power (MACE.HR), or which are both costly to measure and are in the same domain as another one which has higher

power (6MWD vs. VO2max). For these policies ($G^{\text{all.eq}}_{\text{w/o.E,6M}}$ and $G^{\text{hier}}_{\text{w/o.E,6M}}$), $N = 155$ patients per arm suffice for the trial to achieve 90% or more probability of Go when all endpoints reach the target value (CT); and when the treatment has no effect, 80% or more probability of Stop (C0), and less than 5% probability of Go (C0').

However, for both of the chosen policies, the probability of Go is lower than the desired 80% in those cases where the treatment shows no effect in the Imaging domain even if it reaches the target value in all other domains, thus only partially fulfilling (CTc). Whether this is an issue depends on the coupling between the true effect in the Imaging domain and the chances of reaching the target value for the Phase III endpoint.

The $G^{\text{all.eq}}_{\text{w/o.E,6M}}$ policy does produce a Stop decision with 70% or more probability, and Go with 5% or less probability, for for all true effects in the $R^{\text{Stop}}_{\text{case}}$ region; thus fulfilling (C0c) and (C0c'). As for the hierarchical policy, if the most important domain reaches the LRV, a Stop decision is produced only with 46% probability ($< 70\%$), while a Go decision is produced with 18% probability ($> 5\%$), even if the treatment has no effect in any other domain. This may be acceptable as long as the choice of policy reflects the actual clinical importance of the Exercise domain.

For both policies, the safety condition has a high impact on the probability of a Go decision, specially when multiple endpoints on which the treatment has no effect are involved. This indicates that if the treatment has potentially no effect in some domain, the number of endpoints in that domain should be minimized. In this case, if a null true effect in the "Imaging" domain is plausible, clinicians should be consulted to assess whether any of the endpoints in that domain is redundant and can be disregarded. Of course, if the effect is thought to be potentially negative on an endpoint, then the endpoint *should* be included despite the drawbacks. Safety criteria involve trade-offs that need to be discussed with clinicians, in order prevent counterproductive treatments from progressing without unnecessarily delaying the development of promising ones.

Regarding the number of patients, requiring a high probability of Go decisions for all clinically-promising combinations of true effects may increase the required number of patients to an impractical level. Special care needs to be taken to adjust the requirements to those combinations of true effects that are most relevant given the treatment under study.

Chapter 5

Conclusion

In the previous chapters, an extension of the Lalonde framework is defined and its properties are evaluated. The use of the framework is then exemplified by a case study. The results and insights gained in the process are summarized here, followed by a discussion of the limitations of the framework and directions for future development.

5.1 Main results

In the decision framework proposed by Lalonde, the probability of obtaining a "go" decision when the true effect is below the smallest clinically-relevant value (LRV) is the False Go Risk (FGR); and the probability of obtaining a "stop" decision when the true effect is above the target value for a marketable drug (TV) is the False Stop Risk (FSR). The FGR and the FSR completely determine the thresholds on the observed effect for the "go" and "stop" decisions, and thus the probability of each decision under all possible values for the true effect.

This report extends the Lalonde framework to multiple endpoints. In the policies under study, clinically-related endpoints are grouped into "domains".

Extension to multiple endpoints: In the extension presented in this report, the pair of Go and Stop thresholds from the Lalonde framework are generalized to a pair of predicates (i.e. a policy implementation) that fulfill a monotonicity condition, on which the remaining results in this report rely. A number of building blocks and rules for combining them are provided in order to help practitioners define new policies. The resulting policy implementations are guaranteed to be monotone.

False Go and False Stop Risk for monotone policies: In the presence of multiple endpoints, the LRV and TV are generalized vectors **LRV**, **TV**, with one component for each endpoint. For endpoints in the same domain, the FGR is the upper bound on the probability of obtaining a "go" decision when the true effect for *all* endpoints is below the corresponding LRV, while the FSR is the upper bound probability of obtaining a "stop" when some endpoint is

above its target value. Because of the monotonicity property, these upper bounds can be calculated by simulating the study on a finite number of true effects (e.g. for calculating the FGR, it is enough to simulate the study when the true effect is **LRV**).

Bounds on the per-domain False Go and False Stop risks: Deciding a domain level "go" when any of the endpoints would individually be a go in the decision framework results in a rising FGR as the number of endpoints in the domain increases, due to the multiple comparisons problem.

By using a procedure to adjust for multiple comparisons, such as the Bonferroni correction or the Simes/Benjamini-Hochberg procedure, an upper bound can be given for the FGR. In both cases, the resulting policy is expressible as a finite number of thresholds on the individual endpoints. More specifically, the Bonferroni correction produces one threshold per endpoint, and the Simes procedure as many thresholds per endpoint as there are endpoints in the domain. When comparing the two approaches, the higher complexity of the latter was compensated by an improvement to the correct Go rate (CGr) of up to around 4 percentage points with respect to the former. The extent of the improvement depends on the specific combination of true effects. Even an improvement of this magnitude can be meaningful, specially considering that the cost of analyzing the data with a new the statistical technique is low compared to the stakes involved in a clinical study.

Another approach for adjusting for multiple comparisons is to use Hotelling's T^2 statistic. Such a policy does not admit a straightforward representation as a finite number of thresholds, can be too conservative (unless the degrees of freedom are suitably adjusted), and is only meaningfully better than the Simes/Benjamini-Hochberg procedure when the correlation between endpoints in the domain is rather high (e.g. $\rho \geq 0.8$). Therefore, policies with per-endpoint thresholds are preferred.

Evaluation metrics for monotone policies: The variety of policies that can be built when taking into account multiple endpoints means that, unlike the single endpoint case, the FGR and FSR do not by themselves determine the probability of a correct decision. Several metrics are defined, of which the following three have been analyzed with special attention:

(i) The CGr (Correct Go rate), which is a lower bound for the probability of an overall Go when the true effect for the endpoints is large enough for the drug to be potentially marketable (e.g. $\boldsymbol{\mu} \geq \mathbf{TV}$).

(ii) The CSr (Correct Stop rate), which is a lower bound on the probability of Stop when the treatment has either no effect, a subclinical effect, or a negative effect on all endpoints (e.g. $\boldsymbol{\mu} \leq \mathbf{0}$).

(iii) The FGr (False Go rate), which is an upper bound on the probability of Go for the same true effects for which the CSr is defined.

These metrics can be considered either at the level of a single domain, or at the level of the overall policy decision.

All the metrics are implicitly calculated over a region of true effects (in the above examples, the CGr is defined at $\llcorner(\mathbf{TV})$, and the CSr and FGr at

$\neg(\mathbf{0})$). If it is plausible that a drug for which a Go decision is desired will have no effect on some endpoints; or that a drug for which development should be stopped will have a small effect on some of the endpoints; then these regions can be extended to include these true effects.

Given a region of true effects, the CGr and CSr can be understood as the "power" of the study for Go and Stop decisions (respectively) when the true effects are in the corresponding region. Conversely, the FGr can be understood as a Type I error rate for Go decisions when the drug has no effect. If desired, the FSr (False Stop rate) can also be considered, which would be a Type I error rate for Stop decisions.

Due to the monotonicity property of policies, these lower bounds can be calculated by simulating the result of the policy on a finite number of true effects (in the examples, $\boldsymbol{\mu} = \mathbf{TV}$ and $\boldsymbol{\mu} = \mathbf{0}$).

Simultaneous stop **and** go**:** In the single endpoint case, if an observed effect fulfills the conditions for both a "stop" and a "go" decision, then the "stop" decision prevails. Once the potential overlap has been considered, a "go" decision in the absence of a "stop" decision is denoted "Go", while a "stop" decision is denoted "Stop". If neither the condition for "go" or "stop" is fulfilled, the decision is "Discuss". In the extended framework, this prioritization happens at the domain level, and not at the endpoint level. This means that it is still possible for an endpoint, which in framework proposed by Lalonde would be an unambiguous Stop, to contribute towards a domain-level "Go". This may happen provided that there is sufficient evidence that the true effects for all the endpoints in the domain are not all below their LRV. Allowing for this possibility increases the CGr of the study (i.e. the "power" of Go decisions) while preserving the domain-level bounds on the FGR and FSR.

Effect of the number of variables in each domain: For a fixed number of domains, adding new endpoints increases the CGr. The effect on the CSr depends on how the policy combines the domain-level decisions to produce a Stop decision.

- If a Stop decision is produced in the absence of domain-level Go decisions, then having more endpoints in a domain makes a Go decision in the case when the true effect is $\mathbf{0}$ less likely (and thus makes the CSr higher).
- If a Stop decision is produced in the presence of domain-level Stop decisions, adding endpoints makes a Stop decision less likely (and thus lowers the CSr).

These rules of thumb are valid when all endpoints are affected by the treatment in the same way and have the same power. If, for instance, only one of the endpoints in the domain reaches the target value, adding more endpoints to that domain will decrease the CGr due to the multiple comparisons correction required to keep the FGR bounded.

Effect of adding a low-powered endpoint to an existing domain: As mentioned above, the improvement resulting from adding a new endpoint is contingent on the new variable having high-enough power: endpoints with low

power will in fact decrease the CGr. The power needed for the addition of a new endpoint to improve the CGr increases with the correlation between the endpoints in that domain. The CSr at the domain level is in all cases negatively impacted by the addition of a new endpoint; the lower the power, the more negative this impact. Therefore, endpoints included in an existing domain should be well-powered (at least $> 60\%$, ideally $>$ 70-80%) with respect to the TV, specially if the policy is predicated on domain-level Stop decisions. If the endpoint has lower power than that, a good case should be made that the new endpoint is not clinically redundant with respect to the endpoints already included in the domain.

Effect of adding new endpoints into a separate domain: The policies under consideration perform a multiple comparisons adjustment within a domain, but not for the number of domains. Therefore, adding a new variable into its own domain will invariably improve the CGr. This improvement will only be meaningful if the variable is adequately powered; and will become negligible as the power becomes closer to 0.

Due to the lack of multiplicity adjustment, adding a new variable in its own domain also increases the overall FGr, and this increase is independent of the power of the new endpoint. Furthermore, in the case of policies that rely on the absence of domain-level Go to produce a Stop, the rate of Stop may decrease by more than 15 percentage points as the power approaches 0, regardless of the number of domains and endpoints already in the study and the correlation between them. Therefore, when considering adding a new endpoint in a separate domain, this endpoint should have a power high-enough to compensate the risk of misleading or inconclusive decisions when the treatment has little or no effect. If the study has only two domains and the current CGr is low (around 50%), a new domain with an endpoint of only 40% power may be beneficial. If the CGr is already high (more than 80%), the new endpoint should have a power close to 80% to be helpful.

Effect of the arrangement of variables across domains: The CGr and CSr of a policy depend not only on the total number of domains and variables, but also potentially on how these variables are arranged across domains. For policies that treat all domains equally, the CGr and CSr vary with the total number of endpoints, but are relatively unaffected by how many endpoints each specific domain has. For a hierarchical policy in which one domain is more important than the others, the CSr decreases when the number endpoints in the most important domain increases, although this effect diminishes as the correlation of endpoints within the same domain increases. Therefore, when using a hierarchical policy in which an overall Stop decision is predicated on a Stop decision for the most important domain, the number of endpoints in this domain should be as small as possible.

Hierarchical vs. "all domains equal" policies: The analysis in this report is mainly focused on scenarios where effects are homogeneous: either the true treatment effect for all endpoints reaches the TV, or the true treatment effect is 0 for all endpoints. However, this may not hold in practice.

For instance, a treatment which has an effect on the Phase III endpoint may also have an effect in all but one of the domains, perhaps one which is not considered of strong clinical importance. In this case, a hierarchical policy which prioritizes another domain that is more clinically relevant will obtain a correct Go decision more often than a policy which treats all domains equally.

On the other hand a treatment that has no effect on the Phase III endpoint may have a small, clinically-relevant true effect (LRV) on the domain deemed to be most important. As shown in the case-study, this makes the probability of Stop low, even if the effect on all other domains is 0. Whether this is an issue depends on how plausible such a scenario is in light of the clinical relation between the Phase III endpoint and the most important domain.

Ultimately, whether to treat all domains equally or use a hierarchical policy is decided based on clinical and not statistical considerations. However, it is important to account for the statistical implications of this choice when designing the study.

Effect of non-negativity conditions: As indicated when discussing hierarchical vs "all domains equal" policies, a treatment that makes the Phase III endpoint reach the TV might have 0 effect on some disease markers. It is common to avoid going ahead with a Phase III study (Go) if any of the endpoints in the study shows a significant effect in the wrong direction. Due to the multiple comparisons problem, if there are many endpoints in a domain for which the treatment has no effect, this can result in a meaningful decrease in the probability of Go. For moderate levels of correlation between endpoints ($\rho = 0.4$, $\tau = 0.2$), a decrease of more than 10 percentage points is possible.

The inclusion of non-negativity safety condition is often a hard requirement, and due to the stakes involved can seldom be softened. Whether the effect of including this condition is an issue depends on which combinations of true effects are clinically plausible, and how high probability of Go is required for the study to be beneficial. The impact can be mitigated by reducing the number of clinically-redundant endpoints.

Evaluation on regions vs. on individual effects: The metrics for regions defined in §2.10 are useful for evaluating a range of study designs in the abstract, as is done in Chapter 3. However, when dealing with a case study with concrete endpoints, exploring the probabilities of each decision for individual true effects, as is done in Chapter 4, can lead to more fruitful discussions with other stakeholders. When studying individual effects, the probability of a Go decision is a lower (or upper) bound on the probability of Go for true effects that are uniformly larger (respectively smaller) than the one under study. Conversely, the probability of Stop is a lower (or upper) bound on the probability of Stop for true effects that are uniformly smaller (respectively larger).

Considerations in the event of a "Discuss**" decision:** When designing a study, agreeing with the stakeholders on a decision policy that prescribes a course of action on every eventuality can be an unending task. In fact, it is in the nature of the framework proposed by Lalonde that, under some circumstances, the observed effect from the study may neither be large enough

to justify a Go decision, or low enough to justify a Stop decision. In the univariate case, the natural next step is to look at other endpoints; in the multivariate setting these other endpoints are presumably already included. This means that a decision needs to be made based on the endpoints that were already part of the policy.

(i) The criteria for a domain being Stop can be stringent, specially if any of the endpoints in the domain are low-powered. A combination of Stop and Discuss may be sufficient to declare a domain to be Stop.

(ii) The policies under study do not fully specify what to do when only one domain is Go (unless it is, for instance, the most important domain). For example, if many endpoints in an important domain are Go, and some of the remaining endpoints have low power and are not Stop, this could be sufficient grounds to cautiously recommend an overall Go decision. Similarly, when using a hierarchical policy, cases where the domains are a combination of Stop and Discuss may be sufficient grounds for an overall Stop decision.

(iii) The per-domain multiplicity correction may be too conservative in the presence of a high correlation between endpoints. When the observed effects for a domain are very close to the the "go" thresholds, and the sample within-domain correlation is high ($\gg 0.4$), a Go decision can be taken for that domain without counterfactually increasing the FGR.

The aim should in any case be to include as many of these considerations as possible into the original policy so that the probabilities calculated before the study is run reflect the actual decisions that will be made. The framework is flexible enough to express all of the examples criteria while guaranteeing the required monotonicity property.

5.2 Designing a study step by step

Based on the results summarized in §5.1, a procedure for designing a policy for a study and determining the number of required patients is suggested. An important step that influences both the probabilities of a successful study and the cost of the trial is determining which endpoints to include. The endpoint selection process is formulated as a step-wise backward selection procedure (Procedure 1), with the selection of the next endpoint to exclude happening in Procedure 2. The backwards nature of the procedure has the added advantage that the probability of each decision given that an endpoint is included can be used as one of the reasons to justify its exclusion.

In addition to the step-wise backwards selection, the procedure described below includes a mechanism that allows for marking which subset of the endpoints should be considered "tentative", and thus be candidates for exclusion. The tentative endpoints are considered from lowest to highest power, and, the first one fulfilling the exclusion conditions is marked as "excluded". Then the study simulations are rerun to assess whether the desired criteria concerning the probabilities of each decision under different scenarios are achieved. If the effects of excluding an endpoint are not satisfactory, this endpoint can

be marked as "included", and will thus not be a candidate for exclusion any longer. At the end of the procedure, both "included" and "tentative" endpoints are considered for the final policy.

Procedure 1 (Designing a policy for a study with multiple endpoints, and determining the study size)**.**

Input: For each endpoint, the LRV, the TV, the standard deviation for measurements of that endpoint in the study population, and the correlation with other endpoints.

Output: A policy considering some or all of the endpoints, the required number of patients per arm, and the decision probabilities for different combinations of true effects.

Step 1a. Determine main therapeutic objective. *Example:* Reduce hospitalizations.

Step 1b. Determine candidate endpoints related to the objective.

Step 2a. Determine true effects of interest. *Example:* All endpoints reach the target value (**TV**), the drug has no effect on any of the endpoints (**0**), all endpoints reach the lower reference value (**LRV**).

Step 2b. Are there combinations of domains for which an effective drug could plausibly show no effect?

→ Yes: Define the region of true effects R^{Go}, and the associated $\tilde{R}^{\text{Go}}$.
Example: If there are three domains with one endpoint each, and an effective drug could plausibly have no effect on either the second or third domains:

$$\tilde{R}^{\text{Go}} := \{(\text{TV}_1, \text{TV}_2, 0), (\text{TV}_1, 0, \text{TV}_3)\}$$
$$R^{\text{Go}} := \llcorner(\tilde{R}^{\text{Go}})$$

→ No:

$$\tilde{R}^{\text{Go}} := \{\mathbf{TV}\}$$
$$R^{\text{Go}} := \llcorner(\mathbf{TV})$$

Continue to Step 2c.

Step 2c. Are there combinations of domains for which an ineffective drug could plausibly show a clinically relevant effect?

→ Yes: Define a region R^{Stop}, and the corresponding $\tilde{R}^{\text{Stop}}$.
Example: If there are three domains with one endpoint each, and an ineffective drug could plausibly have an effect on the second domain:

$$\tilde{R}^{\text{Stop}} := \{(0, \text{LRV}_2, 0)\}$$
$$R^{\text{Stop}} := \neg(\tilde{R}^{\text{Stop}})$$

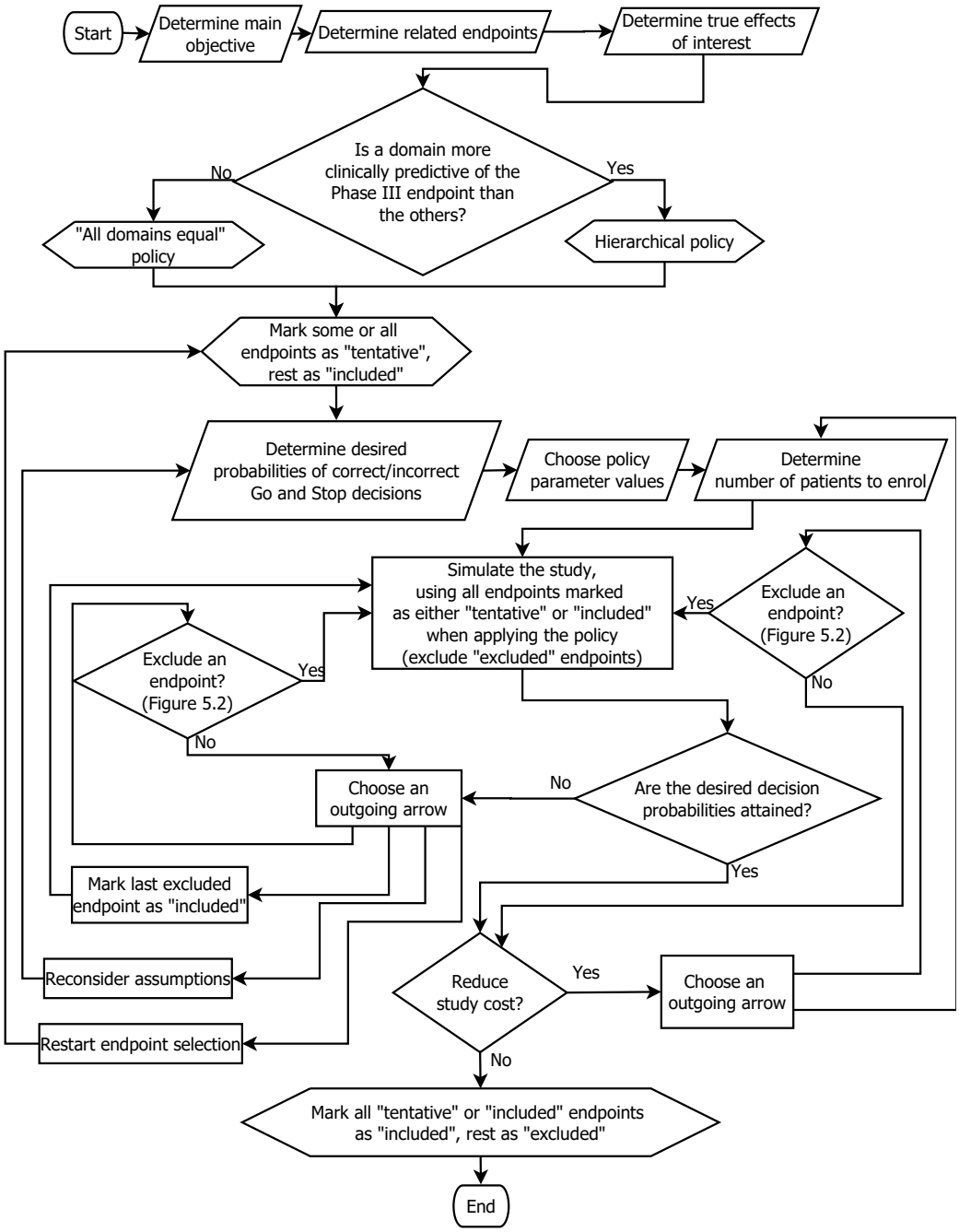


Figure 5.1: Procedure to design a study, determining a policy and the number of patients required. This is a summary of Procedure 1.

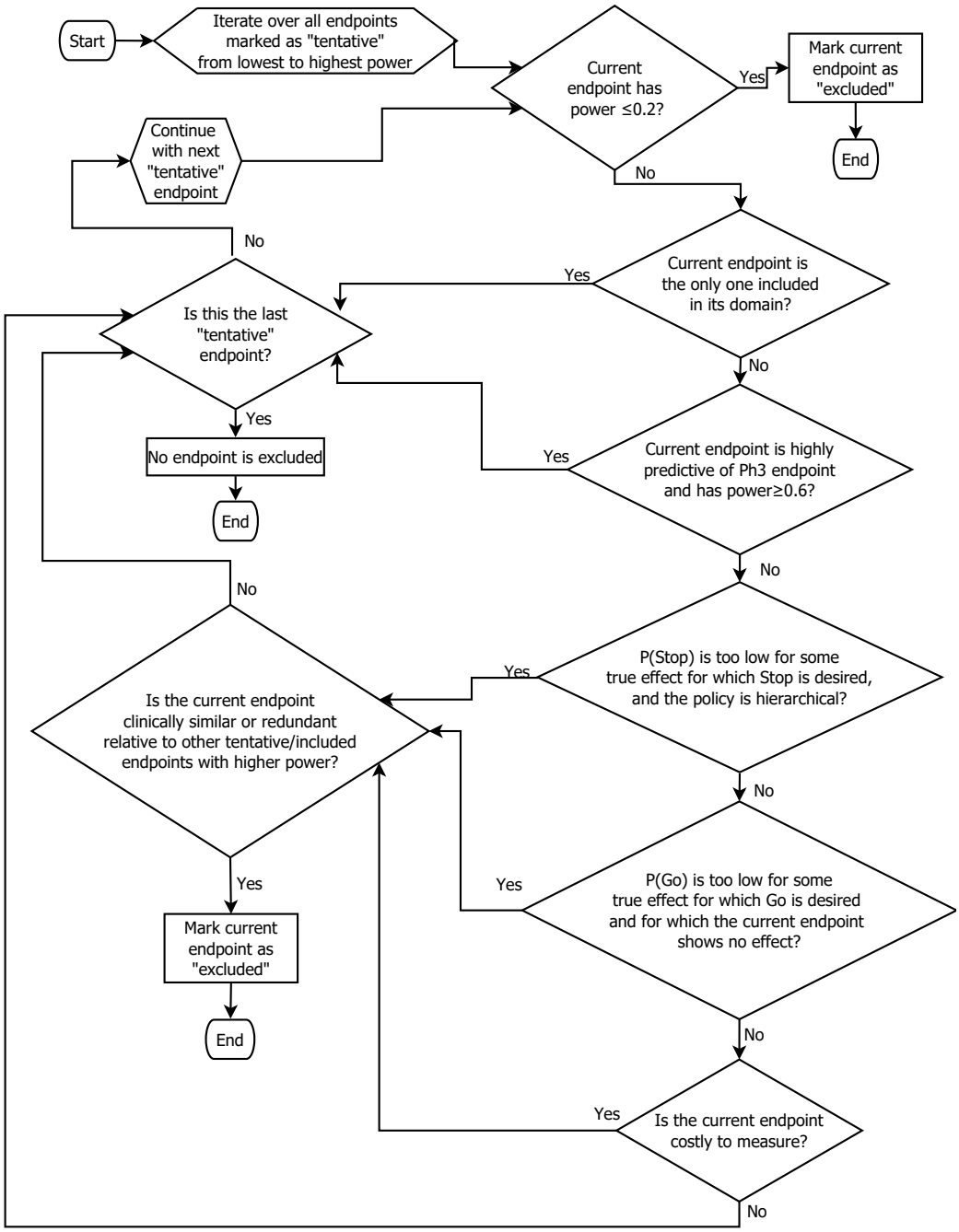


Figure 5.2: Procedure to choose a single endpoint for exclusion. This is a summary of Procedure 2.

→ No:

$$\tilde{R}^{\text{Stop}} := \{\mathbf{0}\}$$
$$R^{\text{Stop}} := \neg(\mathbf{0})$$

Continue to Step 3.

Step 3. Choose the type of policy.
Is there a domain such that:

- ☐ The domain is clinically more important than others.
- ☐ In particular, a drug with no effect on the Phase III endpoint is unlikely to have a relevant effect (LRV) on this endpoint.
- ☐ And (ideally) one or more of the following hold:
  - ○ The domain has one or two endpoints with high power ($\gg 80\%$).
  - ○ The total number of domains is large (4—5).

→ Yes: Consider a hierarchical policy with that domain as the most important domain ($G^{\text{hier}}_{x,\alpha}$).

→ No: Consider an "all domains equal" policy ($G^{\text{all.eq}}_{x,z,\alpha}$).

Notes:

- The parameters x (number of domain-level Go decisions sufficient for an overall Go) and z (number of domain-level Stop decisions sufficient for an overall Stop) are chosen in consultation with the stakeholders. Suggestion: $x = 2$, and, if applicable, $z = 0$.
- In both cases, a custom policy may be considered instead. The policy should fulfill the required monotonicity property.

Step 4. The candidate endpoints are labelled as either "included", "tentative" or "excluded". Included endpoints will definitely be in the final policy, while excluded endpoints will be definitely excluded.
Example: Mark all endpoints as "tentative".

Step 5. In consultation with stakeholders, determine the minimal desired for each decision and effect. The following criteria can be a reasonable starting point:

- ☐ Is the CGr at $\llcorner(\mathbf{TV})$ high enough? *Example:* CGr $\geq 90\%$.
- ☐ Is the CSr at $\neg(\mathbf{0})$ high enough? *Example:* CSr $\geq 80\%$.
- ☐ Is the FGr at $\neg(\mathbf{0})$ low enough? *Example:* FGr $\leq 5\%$.
- ☐ Is the CGr at R^{Go} high enough? *Example:* CGr $\geq 80\%$.
- ☐ Is the CSr at R^{Stop} high enough? *Example:* CSr $\geq 70\%$.
- ☐ Is the FGr at R^{Stop} low enough? *Example:* FGr $\leq 5\%$.

Note: If the risk of a False Stop is of particular concern, bounds on the overall FSr can also be established.

Step 6. In consultation with clinicians and other stakeholders, determine the upper bounds on the False Stop and False Go risks for the domain-level policies ($\alpha^\star_{\text{FS}}$, $\alpha^\star_{\text{FG}}$).
Example: $\alpha^\star_{\text{FS}} := 0.1$, $\alpha^\star_{\text{FG}} := 0.2$.

Step 7. In consultation with clinicians and other stakeholders, determine an acceptable number of patients per arm (N).
Example: $N = 150$.
Calculate the power of each endpoint ($\alpha = 0.05$, two-sided) when the true effect is the TV.

Step 8. Simulate the study with the chosen N, $\boldsymbol{\mu} = \mathbf{TV}, \mathbf{LRV}, \mathbf{0}$, and, if applicable, $\boldsymbol{\mu} \in \tilde{R}^{\text{Go}}$ and $\boldsymbol{\mu} \in \tilde{R}^{\text{Stop}}$.

Only included and tentative endpoints should be used when determining the Go/Discuss/Stop status of each domain. However, a safety condition should comprise all of the following:

(i) The included endpoints.

(ii) The tentative endpoints.

(iii) The Phase III endpoint.

Step 9. Are the criteria decided in Step 5 fulfilled according to the simulation in Step 8?

→ Yes: Then either:

(a) Reduce the cost of the study by enrolling fewer of patients (Step 7).

(b) Reduce the cost of the study by removing an endpoint (Step 10).

(c) Accept the design (Step 11).

→ No: Then either:

(a) Reconsider some assumptions; for instance the parameters of the policy (Step 6) or the required probabilities of Go/Stop (Step 5) for each true effect of interest.

(b) Enrol more patients (Step 7).

(c) Exclude an endpoint (Step 10).

(d) Re-include a previously excluded endpoint, marking this inclusion as final (i.e. marking the endpoint as "included").

(e) Reset the endpoints labels (Step 4).

If any changes were made, rerun the simulation (Step 8).
Otherwise, consider another course of action (Step 9).

Step 10. Does Procedure 2 (see below) suggest excluding an endpoint?

→ Yes: Mark the endpoint as "excluded", go back to Step 8.

→ No: Go back to Step 9, and consider another course of action.

Step 11. Done! All "included" and remaining "tentative" endpoints are included in the resulting policy. Discuss the resulting policy, the included and excluded endpoints, the required number of patients per arm and the decision probabilities with other stakeholders.

◀

Every endpoint included in a study gives additional information about the safety and effectiveness of the drug which may not be apparent from the other endpoints. Therefore, all else being equal, an endpoint should be included if possible. Except for endpoints with very low power, the exclusion of an endpoint should be motivated clinically, even if its statistical properties are not optimal.

Example 5.1 (General criteria for excluding an endpoint)**.** These are some potential reasons to exclude an endpoint in a study:

- ○ The endpoint has low power, i.e. $\ll 0.2$.
- ○ The endpoint is relatively costly or inconvenient to measure, thus increasing the cost of the study.
- ○ The endpoint is clinically similar to other endpoints with higher power; that is, it is expected to be affected by the drug to a similar degree.

◀

Procedure 2 (Determine an endpoint to exclude)**.**

Input: The tentative endpoints for exclusion, together with their power (calculated in Step 7 of Procedure 1), and the probabilities of each decision under different scenarios (calculated in Step 8 of Procedure 1).

Output: A decision to either exclude an endpoint, or to not exclude any endpoints.

To choose an endpoint for exclusion, iterate over the tentative endpoints in increasing order of power (as calculated in Step 7), and apply the following steps to each endpoint:

Step 1. Does the endpoint have low power (≤ 0.2)?

→ Yes: Exclude the current endpoint.

→ No: Continue with the next step.

Step 2. Do all of the following hold?

- ☐ The endpoint is strongly connected with the presumed mechanism of action of the treatment, or is otherwise judged by clinicians to be a better predictor of the Phase III endpoint than any of the other endpoints in the same domain.

☐ The endpoint has relatively high power (e.g. ≥ 0.6).

→ Yes: Skip the current endpoint (Step 6).

→ No: Continue with the next step.

Step 3. Do all of the following hold?:

☐ An effective drug may have 0 effect on this domain.

☐ When the endpoint is included, and the true effect is in R^{Go}, but the true effect of the drug is 0 for the domain to which this endpoint belongs; the probability of Go is too low.

☐ The current endpoint is clinically similar to other tentative or included endpoints with higher power.

→ Yes: Exclude the current endpoint.

→ No: Continue with the next step.

Step 4. Do all of the following hold?:

☐ A policy that takes into account domain-level Stop decisions is used (for instance, the "hierarchical domains" policy).

☐ When this endpoint is included, and the true effect is in R^{Stop}, the overall probability of Stop is too low.

☐ The current endpoint is clinically similar to other tentative or included endpoints with higher power.

→ Yes: Exclude the current endpoint.

→ No: Continue with the next step.

Step 5. Do all of the following hold?

☐ The endpoint is clinically similar to other endpoints in the study.

☐ The endpoint is particularly costly to measure.

→ Yes: Exclude the current endpoint.

→ No: Skip this endpoint (Step 6).

Step 6. Continue with the next endpoint (if any). Otherwise no endpoint is excluded.

NB: An excluded endpoint may still be considered as part of a safety condition (see Step 8 in Procedure 1). ◀

Procedure 1 and Procedure 2 are summarized in Figure 5.1 and Figure 5.2, respectively. These procedures are intended to be a starting point for designing policies within the framework in this report. The reader is encouraged to adapt and expand the procedures based on the input of all the stakeholders involved.

5.3 Limitations and future research

The framework has been evaluated on a range of purely theoretical study designs, and on a synthetic case study with realistic endpoints. A number of assumptions and simplifications were made in order to delimit the scope of the report. The resulting limitations in the applicability of the results are discussed below, together with mitigation strategies and possibilities for improvement.

Non-normally distributed endpoints: The framework as described is restricted to normally distributed variables. This also covers log-normal distributions by means of an appropriate transformation, and, under certain assumptions about their variance, endpoints with other distributions such as hazard ratios can be modelled in this way. Endpoints with particularly high skewness, such as time-to-event, are less straightforward. The framework itself does not preclude such endpoints, but further development is needed to see how they can be made to fit with Lemma 2.4, on which the fulfilment of the monotonicity requirement relies.

Heterogeneous correlation across endpoints: Our evaluation assumes that the correlations between all endpoints are given by two parameters: one for endpoints in the same domain, and one for endpoints in different domains. The policies under study do not themselves rely on this information, assuming instead an unstructured covariance matrix. However, the sensitivity of the evaluations done in Chapter 3 and Chapter 4 to these assumptions has not been assessed.

Safety endpoints: The report focuses on the case where the true effects are all non-negative; that is, the change from baseline in the treatment group is of the same sign as the target value. A condition such as no endpoints being significant in the wrong direction can be formulated in the framework, and its effect on hindering a Go decision in cases when they are warranted has been explored. The converse; that is, its effectiveness at preventing a Go decision when the treatment is detrimental has not been evaluated. Additionally, conditions which are triggered by an endpoint falling within a range of observed values are outside the framework, as the resulting predicate is not monotone. In those cases, one possibility is to consider the absolute deviation from the "normal" value as the endpoint, and adequately transform this difference so that its distribution is in line with the normality assumptions.

Non axis-parallel true effect regions: The regions in the space of true effects that are used to define the statistical properties and the evaluation metrics are all defined in terms of thresholds on individual variables (e.g. the true effect at each endpoint is less than 0, or the true effect at at least one endpoint is greater than the target value). Conditions involving combinations of multiple endpoints (e.g. the weighted sum of these endpoints is larger than this target value) may result in regions with non-axis-parallel boundaries, which do not in general fulfill the preconditions of Theorem 2.83. This means that the FGR, FSR and the evaluation metrics cannot be reduced to their values

on a finite number of true effects. A possibility is to include such a linear combination as an additional endpoint, with the caveat that, if all the constituent endpoints are also included, the resulting covariance matrix may be singular. In such a situation, both inference and simulation become more complicated, and the conclusions of this report less applicable.

5.4 Concluding remarks

The framework proposed by Lalonde can be generalized to multiple endpoints with relatively few additional complications while preserving its essential statistical properties. When designing a policy in a multivariate framework, including an additional endpoint is not always beneficial, and can lower the probability of obtaining a conclusive, correct decision from the study. Whether an endpoint should be included depends on its power, on whether it is expected to add new information with respect to the endpoints already in the study, and on the strength of its clinical relationship with the eventual Phase III endpoint.

Bibliography

[AFB+18] Kumar Ashish, Mohammed Faisaluddin, Dhrubajyoti Bandyopadhyay, Adrija Hajra, and Eyal Herzog. Prognostic value of global longitudinal strain in heart failure subjects: A recent prototype. *International Journal of Cardiology. Heart & vasculature*, 22:48–49, Dec 2018. doi:10.1016/j.ijcha.2018.11.009.

[BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. doi:10.1111/j.2517-6161.1995.tb02031.x.

[Bon36] Carlo Emilio Bonferroni. *Teoria statistica delle classi e calcolo delle probabilità*. Reale Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 1936.

[BWM04] Vikas Bhalla, Scott Willis, and Alan S. Maisel. B-Type Natriuretic Peptide: The level and the drug – Partners in the diagnosis and management of congestive heart failure. *Congestive Heart Failure*, 10(s1):3–27, January 2004. doi:10.1111/j.1527-5299.2004.03310.x.

[BY01] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, August 2001. doi:10.1214/aos/1013699998.

[CL13] Shein-Chung Chow and Jen-Pei Liu. *Design and Analysis of Clinical Trials*. John Wiley and Sons, Inc., 3 edition, 2013.

[Cox06] D. R. Cox. *Principles of Statistical Inference*. Cambridge University Press, 2006.

[Den19] David Abraham Deniz. Multidimensional decision making in early phase clinical trials. Master's thesis, University of Gothenburg, 2019. Unpublished.

[FDA04] Food and Drug Administration, U.S. Department of Health and Human Services. Challenge and opportunity on the critical path to new medical technologies, 2004. URL `https://www.who.int/intellectualproperty/documents/en/FDAproposals.pdf`.

[FDA18] Food, U.S. Department of Health Drug Administration, and Human Services. Clinical research phase studies, 2018. URL `https://www.fda.gov/patients/drug-development-process/step-3-clinical-research`.

[FMWM16] P. Frewer, P. Mitchell, C. Watkins, and J. Matcham. Decision-making in early clinical drug development. *Pharmaceutical Statistics*, 15(3):255–263, May 2016. doi:10.1002/pst.1746.

[GB09] Alan Genz and Frank Betz. *Computation of Multivariate Normal and* t *probabilities*. Springer, 2009. doi:10.1007/978-3-642-01689-9.

[GI83] D. Goldfarb and A. Idnani. A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, 27:1–33, September 1983. doi:10.1007/BF02591962.

[GPBS00] C. Patrick Green, Charles B. Porter, Dennis R. Bresnahan, and John A. Spertus. Development and evaluation of the Kansas City Cardiomyopathy Questionnaire: a new health status measure for heart failure. *Journal of the American College of Cardiology*, 35(5):1245–1255, 2000. doi:10.1016/S0735-1097(00)00531-3.

[Hay13] Winston Haynes. *Bonferroni Correction*, page 154. Springer New York, 2013. doi:10.1007/978-1-4419-9863-7_1213.

[Hig13] Roger Higdon. *Multiple Hypothesis Testing*, pages 1468–1469. Springer New York, 2013. doi:10.1007/978-1-4419-9863-7_1211.

[HTC+14] Michael Hay, David W. Thomas, John L. Craighead, Celia Economides, and Jesse Rosenthal. Clinical development success rates for investigational drugs. *Nature Biotechnology*, 32(1):40–51, Jan 2014. doi:10.1038/nbt.2786.

[JW14] Richard Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Education Limited, 6th edition, 2014.

[LKH+07] R. L. Lalonde, K. G. Kowalski, M. M. Hutmacher, W. Ewy, D. J. Nichols, P. A. Milligan, B. W. Corrigan, P. A. Lockwood, S. A. Marshall, L. J. Benincosa, T. G. Tensfeldt, K. Parivar, M. Amantea, P. Glue, H. Koide, and R. Miller. Model-based drug development. *Clinical Pharmacology & Therapeutics*, 82(1):21–32, Jul 2007. doi:10.1038/sj.clpt.6100235.

[MCP06] David Machin, Yin Bun Cheung, and Mahesh KB Parmar. *Survival analysis: A practical approach*. John Wiley & Sons, Ltd, 2nd edition, 2006.

[Mui82] Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Inc., 1982. doi:10.1002/9780470316559.

[RAWS08] Bryan Ristow, Sadia Ali, Mary A. Whooley, and Nelson B. Schiller. Usefulness of left atrial volume index to predict

heart failure hospitalization and mortality in ambulatory patients with coronary heart disease and comparison to left ventricular ejection fraction (from the Heart and Soul study). *The American Journal of Cardiology*, 102(1):70–76, Jul 2008. doi:10.1016/j.amjcard.2008.02.099.

[SBBE14] Aylin Sertkaya, Anna Birkenbach, Ayesha Berlink, and John Eyraud. Examination of clinical trial costs and barriers for drug development, 2014. URL `https://aspe.hhs.gov/reports/examination-clinical-trial-costs-barriers-drug-development-0`.

[Sim86] Robert John Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986. doi:10.1093/biomet/73.3.751.

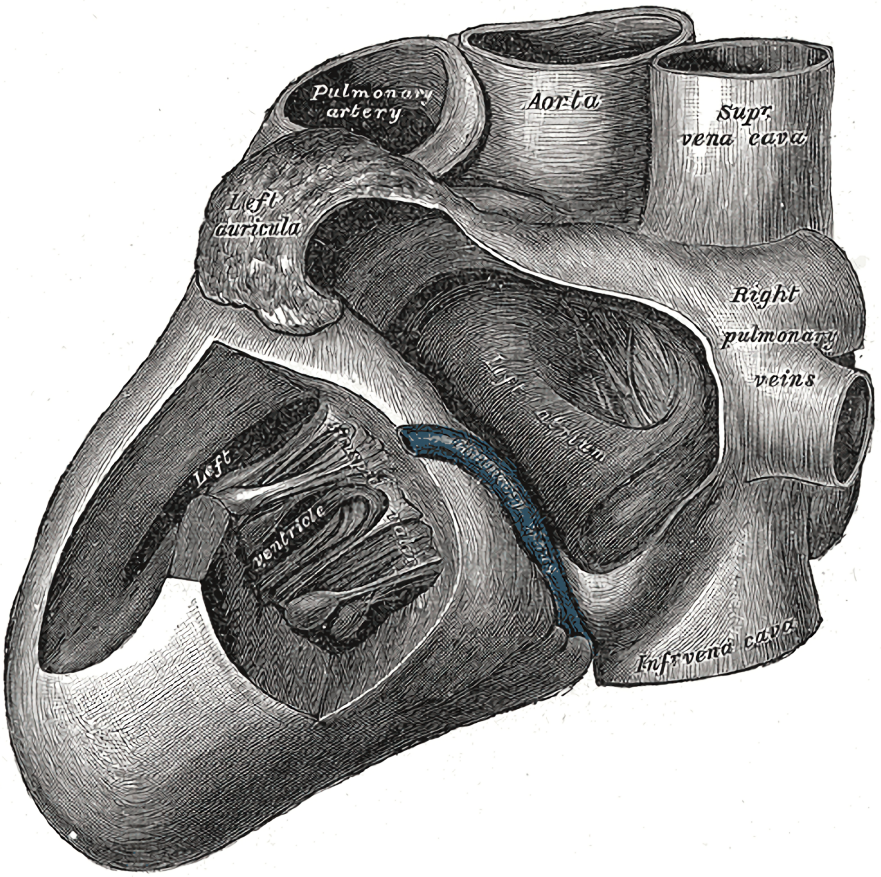
[gonogo] Víctor López Juan. *gonogo: A multivariate extension of the Lalonde framework*, 2022. URL `https://lopezjuan.com/project/gonogo`. R package version 0.1.

[mvtnorm] Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, and Torsten Hothorn. *mvtnorm: Multivariate Normal and t Distributions*, 2021. URL `https://CRAN.R-project.org/package=mvtnorm`. R package version 1.1-3.

[quadprog] Berwin A. Turlach, Andreas Weingessel, and Cleve Moler. *quadprog: Functions to Solve Quadratic Programming Problems*, 2019. URL `https://CRAN.R-project.org/package=quadprog`. R package version 1.5-8.

[TGS+15] Connie W. Tsao, Philimon N. Gona, Carol J. Salton, Michael L. Chuang, Daniel Levy, Warren J. Manning, and Christopher J. O'Donnell. Left ventricular structure and risk of cardiovascular events: A Framingham heart study – cardiac magnetic resonance study. *Journal of the American Heart Association*, 4(9):e002188–e002188, Sep 2015. doi:10.1161/JAHA.115.002188.

[Tuk53] John W. Tukey. The problem of multiple comparisons, 1953.



Anatomy of the heart from the left.
Henry Gray. Anatomy of the Human Body, 1924.
Public domain.