A Bayesian multivariate framework for analysis and decision-making with binary outcome variables

Kavelaars, X.M.

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

A Bayesian multivariate framework for analysis and decision-making with binary outcome variables

Xynthia Kavelaars

The research in all chapters was funded by the Dutch Organization for Scientific Research (NWO) through a Research Talent Grant (406.18.505).

Author: Xynthia Kavelaars
Cover design: Alexe Spaans
Printed by: Gildeprint
ISBN: 978-94-6419-803-4

A Bayesian multivariate framework for analysis and decision-making with binary outcome variables

Proefschrift ter verkrijging van de graad van doctor aan
Tilburg University op gezag van de rector magnificus,
prof. dr. W.B.H.J. van de Donk, in het openbaar te verdedigen
ten overstaan van een door het college voor promoties
aangewezen commissie in de Aula van de Universiteit op

vrijdag 07 juli 2023 om 10.00 uur

door

Xynthia Miquitte Kavelaars
geboren te Gouda

Promotor:	prof. dr. M. C. Kaptein (Tilburg University)
Copromotor:	dr. ir. J. Mulder (Tilburg University)
Leden promotiecommissie:	prof. dr. I. Klugkist (Utrecht University) prof. dr. D. van Ravenzwaaij (Rijksuniversiteit Groningen) prof. dr. A. G. de Waal (Tilburg University) dr. M. van Smeden (UMC Utrecht)

Contents

Chapter 1

Introduction

In medical research, Randomized Controlled Trials (RCTs) are considered the gold standard by which the effects of new treatments, therapies, and interventions are evaluated (Evans, 2003; Grol & Grimshaw, 2003; Harbour & Miller, 2001). The standard RCT compares the experimental treatment to a control condition (usually the standard treatment or a placebo) on one or multiple pre-selected outcomes of interest, which are often aspects of (relative) effectiveness and safety. The control condition provides direct comparison regarding superiority or inferiority of the treatment, and randomization allows us to draw causal conclusions. The associated statistical analysis plan provides control over the quality of superiority and inferiority decisions in terms of decision error rates. That is, decisions with sufficient statistical power are targeted to select more effective treatments (i.e., avoiding Type II errors), while controlling the rate of false superiority and inferiority conclusions (i.e., controlling Type I errors). In the RCT setting, a priori estimates of the treatment effect often form a basis to calculate how much data are needed for decisions with prespecified error rates. The results of RCTs are primarily used in two ways. First, the information collected in the various phases of treatment evaluation impacts approval by regulatory bodies such as the Food and Drug Administration and the European Medicine Agency. Second, if the treatment stands these tests and is approved for a broader rollout, RCT outcomes will also guide treatment prescription in clinical practice.

Although a robust standard with favorable properties, RCT methodology is challenged by the personalization of medicine (Hamburg & Collins, 2010; Mirnezami et al., 2012; Ng et al., 2009; Schork, 2015). Personalization refers to the ideas that a) patients with different characteristics respond differently to treatments and b) better treatments can be prescribed to patients if these characteristics are taken into account (Goldberger & Buxton, 2013). The implementation of these ideas demands RCTs to answer new, more complex research questions and to provide additional information that clinicians need to support treatment prescriptions to individual patients with adequate evidence. More specifically, treatment prescriptions ideally originate from a trade-off between patient-specific risks and benefits, thus reflecting a multivariate decision procedure that considers several outcomes in relation to each other. In such a procedure, clinicians a) weigh advantages and disadvantages and b)

determine how many side effects are acceptable given the effectiveness of the treatment (Murray et al., 2016). Many RCTs have data available to support these decisions with evidence as they measure aspects of both safety and efficacy (Biswas et al., 2009). Unfortunately, the statistical procedures of RCTs often ignore the multivariate nature of clinical decision-making, and trial conduct is separated from clinical practice (Murray et al., 2016). As a result, analyses remain largely univariate, relations between outcome variables are neglected, and superiority and inferiority conclusions are based on a single primary outcome variable (Food and Drug Administration, 2010; Murray et al., 2016; Oliveira & Teixeira-Pinto, 2015).

In addition to these more complex research questions, personalization potentially affects characteristics of datasets that are relevant for the choice of analysis techniques. The structure of datasets can be affected in three major ways. First, sample sizes can be smaller due to limited eligibility for participation because newly developed interventions are more and more targeted at specific patient groups that are, for example, defined by a biomarker, a disease subtype, a personal characteristic, or (combinations of) specific characteristics. Trials to evaluate these interventions are subject to stringent inclusion and exclusion criteria that make the recruitment of a sufficient number of patients potentially challenging (Renfro & Sargent, 2017). The challenge to recruit sufficient participants leads to small samples, which can have important consequences: Analyses are underpowered, decisions regarding superiority remain inconclusive, and new treatments are left unable to demonstrate their potential. In a worst-case scenario, patients keep being exposed to suboptimal treatments because there are not enough data to rigorously test promising treatments among subjects with the given patient profile. Second, datasets often contain information from multiple subgroups of patients with potentially different (i.e., heterogeneous) treatment effects. This is the case when trials evaluate treatments that target a more diverse group of patients. Here, RCTs should provide insights into the effects on subpopulations of relatively similar patients, in addition to the general effect among the broader, more diverse target population. Treatment heterogeneity is frequently ignored, and the focus of trials is typically on average treatment among the study

population, preventing researchers from distinguishing between patients who are expected to benefit from the treatment from those who do not (Thall, 2020). Subgroups of patients who are actually adversely affected by the treatment can be overlooked and unintentionally prescribed a harmful treatment. Third, datasets more often consist of clusters of patients more similar to each other than to patients from other clusters. Trials are increasingly conducted among multiple treatment centers and countries, as expanding the (geographical) recruitment region can a) help to speed up enrollment and b) answer substantive questions regarding treatment effects in different recruitment locations (Gallo, 2000; Lin, 1999). Datasets with such a so-called multilevel structure complicate analysis as they require analysis methods that correct for the clustered structure (Gelman & Hill, 2007; Hox et al., 2017). Ignoring the multilevel structure compromises statistical validity, potentially up to the point where treatments are incorrectly approved or dismissed due to inflated decision error rates.

Datasets with combinations of these three characteristics are becoming more standard in contemporary medical research. In an overview of several hundred trials, a multivariate and multilevel structure was commonly observed (Biswas et al., 2009): More than half of the trials were executed in multiple treatment centers, and the majority of these trials assessed multiple outcome variables. Further, treatment heterogeneity is increasingly targeted, for example, via subgroup analysis or via so-called master protocols that conduct several sub trials in parallel (Simon, 2010; Thall, 2020; Woodcock & LaVange, 2017). Examples of such protocols are umbrella or basket trials, which distinguish subtrials by disease subtype or biomarker respectively (Renfro & Sargent, 2017; Renfro & Mandrekar, 2017). Unfortunately, we lack statistical methodology to accommodate these changing data features and to answer more comprehensive research questions in a flexible way. In this dissertation, we address this gap and develop methods to target the changing nature of RCTs in the personalization of medicine.

1.1 Motivating example

We will illustrate potential consequences of the introduced complexities in more detail with the Cognition and Radiation Study B (CAR-Study B; Schimmel et al., 2018). CAR-Study B is an RCT that investigated the effects on the cognitive performance of cancer patients with multiple ($11 - 20$) brain metastases after these intracranial tumors have been treated with one of the two radiation-based interventions used for comparison. The target population of this study has a poor prognosis: The expected survival time without treatment is less than three months (Khalsa et al., 2013; Niranjan et al., 2012). For a long time, radiation of the whole brain (Whole Brain Radiation Therapy; WBRT) has been the standard treatment for these patients (Eisuke & Aoyama, 2011). WBRT may improve intracranial tumor control and survival, but comes with severe side effects in the form of irreparable damage to cognitive functioning (Brown et al., 2016; Ellis et al., 2012; Habets et al., 2015). The decline in memory, concentration, and attention can be invalidating as these functions are highly important to navigate through daily life and profoundly influence the quality of life (Schimmel et al., 2018).

At the start of CAR-Study B, local treatment of individual metastases (Gamma Knife Radiosurgery; GKRS) was increasingly used as an alternative treatment to WBRT among patients with multiple brain metastases (Yamamoto et al., 2014). Local radiation spares more healthy brain tissue, is expected to better preserve cognitive functioning, and has been adopted as an initial treatment option for patients with a smaller number of brain metastases (Chang et al., 2009; Eisuke & Aoyama, 2011; Schimmel et al., 2018). CAR-Study B aimed to evaluate whether local radiation indeed causes less damage to the cognitive functions of the specific group of cancer patients with multiple brain metastases. Despite the earlier promising results of GKRS among related subpopulations, testing the treatment among patients with multiple brain metastases was challenging. The researchers struggled with a troublesome inclusion process, had to deal with a substantial dropout that is inherent to the population, and managed to enroll less than one-thirds of the required number of patients in over a year.

1.2 Goals of the current dissertation

CAR-Study B exemplifies how increasing personalization of medicine can make it hard to find enough eligible participants. Inspired by the promising development of personalized medicine and motivated by the cumbersome inclusion process of CAR-Study B, novel methods are needed a) to create more extensive overviews of treatment effects among a range of diverse patient populations and/or b) to reduce the required number of participants without compromising decision error rates. Specifically, we argue that trials can improve the value of collected data and can target both goals simultaneously when using available information differently. As noted earlier, trials usually have more information available than included in the analyses. Attempts to incorporate these resources are often inefficient, and hence they are treated in isolation. This is unfortunate, since sharing information more efficiently within trials has two major advantages: It helps to answer more complex research questions, and it is known to improve the statistical power of decision-making thereby allowing for sample size reduction (Biswas et al., 2009; Leon-Novelo et al., 2012). For example, analyzing multiple dependent variables simultaneously provides insight into their co-occurrences. In contrast, their common disjoint analysis can paradoxically require larger samples to achieve the same power as preselecting a single outcome variable (Food and Drug Administration, 2010; Senn & Bretz, 2007; Sozu et al., 2010). A similar problem is seen with heterogeneous treatment effects. Modeling heterogeneity directly reveals the relation between subpopulations and their treatment effects and is more powerful than the common subgroup analysis (i.e., stratified analysis; Food and Drug Administration, 2010; Kaptein, 2014; Kaptein et al., 2015; Thall, 2020). Thus, sharing information between outcome variables and subpopulations can greatly improve the value of RCTs in personalized medicine since it a) borrows strength from other variables to improve the efficiency of clinical trial methodology, b) enables more refined decisions thereby facilitating alignment of trial conduct and clinical decision-making, and c) creates more comprehensive insights into the way treatment effects vary over related but different subpopulations.

We implemented the idea of information sharing in a Bayesian multivariate framework for

RCT data with multiple correlated binary outcome variables. The focus on outcome variables of a binary measurement level was motivated by their frequent use in medical research and medical practice. Central to the framework are three components:

1. a multivariate analysis model for multiple binary outcome variables to benefit from the correlation between outcome variables;
2. a transformation procedure to make the resulting model parameters interpretable in terms of (multivariate) success probabilities and the differences between them;
3. a decision procedure to make treatment comparisons and draw conclusions regarding superiority and inferiority with prespecified frequentist error rates.

Together, these three components form a comprehensive framework for statistical analysis and decision-making with multiple (correlated) binary outcome variables.

The Bayesian approach was adopted for several reasons. First, Bayesian analysis naturally works with distributions and posterior samples of model parameters. Transformation of these posterior samples to other parametrizations provides point estimates, spread, and other distributional characteristics of the transformed parameters. The latter were convenient for the transformation and decision procedures of our framework as they enabled decision-making and inference with the transformed parameters while controlling frequentist decision error rates (Marsman & Wagenmakers, 2016; Schönbrodt et al., 2017). Second, working with a posterior sample of transformed parameters also provided much flexibility in defining decision criteria, such as superiority and inferiority, and consequently, in the formalization of decision rules. The desirability of different outcome scenarios can be specified in a natural way (Berger, 2010) while maintaining good frequentist properties, e.g., Type I errors (Food and Drug Administration, 2010). Finally, information external to the trial can be integrated via prior information. When desired, new data can be combined and weighed with available historical data in different ways (e.g., Ibrahim & Chen, 2000).

1.3 Overview of the current dissertation

Throughout the dissertation, three increasingly complex variations of the framework are presented:

1. modeling multiple binary outcome variables and the relation between them;
2. modeling multiple binary outcome variables, the relation between them, and the relation with observed covariates;
3. modeling multiple binary outcome variables, the relation between them, and the relation with observed covariates in clustered data.

The core of this dissertation is written as four separate articles that can be read independently from each other. We chose to preserve the writing of the original articles where possible, which induces some inconsistency in notation and cross-referencing throughout the dissertation.

Chapter 2 provides a non-technical introduction to the multivariate analysis and treatment comparison procedure with multiple correlated binary outcome variables. Intuitive explanations are given of the underlying multinomial model and the choice of prior parameters, the meaning of the resulting posterior (multinomial) joint response probabilities is discussed, and the transformation to (multivariate) success probabilities and treatment differences are explained. Further, multiple suggestions to define superiority in the multivariate context are given and an additional decision rule is proposed that a) is suitable to weigh different outcome variables differently and b) naturally has a compensatory mechanism that allows small negative treatment differences to be compensated for by larger positive treatment differences. The framework is introduced in the context of adaptive trial methodology where the concept of interim monitoring is explained. The chapter is supplemented with an R Shiny app to perform the analysis, which is explained in a tutorial.

Chapter 3 extends Chapter 2 with a technical introduction. The multivariate analysis procedure is introduced as a conjugate combination of a multivariate Bernoulli likelihood and a Dirichlet prior distribution. The model relies on a multinomial parametrization and results in a posterior distribution with a known functional form. We demonstrate that a multivariate

analysis procedure provides better decision error control compared to multiple univariate analyses and allows for more flexible and more efficient options for the choice of a decision rule. An extensive simulation study demonstrated frequentist operational characteristics of the framework a) with various multivariate decision rules, b) in adaptive as well as fixed designs with a priori estimated sample sizes, and c) with several non-informative and informative prior specifications.

Chapter 4 presents the framework with a different model that can analyze multiple correlated binary outcome variables and heterogeneous treatment effects simultaneously. This is useful for trials that assess multiple dichotomous treatment effects and include different but related groups of patients. Although the modeling procedure in Chapters 2 and 3 can be used to perform independent subgroup analyses, these analyses treat subgroups in isolation, are quite inefficient in terms of sample size, and can be less informative. Therefore, a more efficient alternative is presented that uses the relation between patient groups: a multivariate logistic regression model to account for observable treatment heterogeneity. This model allows sharing of information between outcome variables and borrowing strength from other subgroups, resulting in more comprehensive insights into the relation between treatment effects and subpopulations and in more efficient multivariate treatment comparisons among subpopulations of the study population while including information from the entire study sample. We included an illustration with real-world data from the International Stroke Trial (International Stroke Trial Collaborative Group, 1997).

In **Chapter 5**, the multivariate logistic regression model from Chapter 4 is extended to the multilevel context. These days, trials are increasingly executed in collaboration with different treatment centers and/or in different countries. In these situations, research subjects from one center or country could well be more similar to each other than to subjects from another center or country, giving the data a clustered (or hierarchical or multilevel) structure and implying that subjects within such a cluster are not independent. Such a clustered data structure requires specific analysis techniques that can deal with non-independent data whilst naively applying standard techniques result in inaccurate decision error rates. The proposed

multilevel multivariate logistic regression model is suitable for estimation and inference among different subpopulations and different clusters. We illustrated this model in a re-analysis of data from the third International Stroke Trial (The International Stroke Trial-3 Collaborative Group, 2012).

In **Chapter 6**, a general discussion on the presented framework is provided. First, the proposed framework is briefly summarized, followed by a discussion of various openings for critical evaluation of unexplored aspects and extensions. Directions for future research and implementation in practice are presented as well. Finally, we address several topics that are worth debating for the advancement of the medical field in an era of personalization.

Chapter 2

Going multivariate in clinical trial studies: A Bayesian framework for multiple binary outcomes

Based on Kavelaars, X. (2020). Going multivariate in clinical trial studies: A Bayesian framework for multiple binary outcomes. In R. Van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners*. Routledge. https://doi.org/10.4324/9780429273872

Abstract

In an era where medicine is increasingly personalized, clinical trials often suffer from small samples. As a consequence, treatment comparison based on the data of these trials may result in inconclusive decisions. Efficient decision-making strategies are highly needed so decisions can be made with smaller samples without increasing the risk of errors. The current chapter centers around one such strategy: Including information from multiple outcomes in the decision, thereby focusing on data from binary outcomes. Key elements of the approach are (1) criteria for treatment comparison that are suitable for two outcomes, and (2) a multivariate Bayesian technique to analyze multiple binary outcomes simultaneously. The conceptual discussion of these elements is complemented with software to implement the approach. To further facilitate trials with small samples, the chapter also outlines how interim analyses may result in more efficient decisions compared to the traditional sample size estimation before data collection.

2.1 Introduction

Clinical trials often compare a new treatment to standard care or a placebo. If the collected data provide sufficient evidence that the new treatment is better than the control treatment, the new treatment is declared superior. Since these superiority decisions ultimately contribute to a decision about treatment adoption, proper error control is crucial to ensure that better treatments are indeed selected. Key to regulating decision errors is collecting sufficient information: A quantity that is often expressed in terms of a minimum number of participants, or required sample size.

Recruiting sufficiently large samples can be challenging, however. This is especially true in an era in which medicine is increasingly personalized (Hamburg & Collins, 2010; Ng et al., 2009). Personalization of medicine refers to the targeting of treatments at specific patient and/or disease characteristics under the assumption that patients with different (disease) characteristics respond differently to treatments (Goldberger & Buxton, 2013). Since personalization limits the target population of the treatment, inclusion and exclusion criteria for trials become more stringent and the eligible number of participants decreases. This inherently decreases the sample size of studies conducted with the same resources. Consequences of small samples may be substantial: Trials may be left underpowered and decisions about superiority might remain inconclusive.

The problem associated with small sample sizes due to stringent inclusion criteria is illustrated by CAR-Study B (Schimmel et al., 2018). CAR-Study B aims to improve treatment for cancer patients with 11–20 metastatic brain tumors (i.e., tumors that originate from another site in the body and have spread to the brain). These patients have a life expectancy of one or two months and are currently treated with whole-brain radiation therapy. However, whole-brain radiation has adverse side effects: The treatment damages brain tissue and results in severe cognitive impairment. Local radiation of the individual tumors (stereotactic surgery) is a promising alternative that spares healthy tissue and prevents cognitive decline without increasing mortality. The protective effect on cognition has been demonstrated in a related population of patients with fewer brain tumors (Chang

et al., 2009; Yamamoto et al., 2014). However, investigating whether local radiation reduces side effects in the current target population is difficult: Clinicians are reluctant to prescribe the alternative treatment and not all referred patients are eligible for participation, leaving the researchers unable to recruit the required sample.

To improve decision-making with limited samples, studies such as CAR-Study B might combine information from multiple outcomes. The current chapter introduces a Bayesian decision-making framework to combine two binary outcomes. Since superiority with two outcomes can be defined in multiple ways, several criteria to evaluate treatments are discussed in the Decision rules section. Evaluation of these decision rules requires a statistical analysis procedure that combines the outcomes. The Data analysis section outlines such a multivariate approach for Bayesian analysis of binary outcomes. The proposed decision-making strategy is illustrated in the Computation in practice section, which introduces an online app to analyze real data[1]. Since trials with limited access to participants aim for the smallest sample possible, the chapter continues with Sample size considerations to explain how interim analyses during the trial may improve efficiency compared to traditional sample size estimation before running the trial. The Concluding remarks section highlights some extensions of the framework. Throughout the chapter, the comparison of local and whole -brain radiation in CAR-Study B serves as an example with cognitive functioning and quality of life as the outcomes under consideration.

2.2 Decision rules

A key element of decision-making is the decision rule: A procedure to decide whether a treatment is considered superior. When dealing with two outcomes, superiority can be defined in several ways (Food and Drug Administration, 2017), such as a favorable effect on:

1. The most important outcome ("Single outcome rule")
2. Both outcomes ("All rule")

[1]http://utrecht-university.shinyapps.io/multiple_binary_outcomes/ – for the annotated R code and potential newer versions go to http://www.github.com/XynthiaKavelaars

3. Any of the outcomes ("Any rule")
4. The sum of outcomes ("Compensatory rule")

Each of these decision rules weighs the effects of the two outcomes differently. The Single outcome rule evaluates the data from one outcome and ignores the other outcome in the decision procedure. In CAR-Study B, local radiation would be the treatment of preference if it impairs cognitive functioning less than whole brain radiation, irrespective of the effects on quality of life. The All rule evaluates both outcomes, and requires favorable effects on each of them. Compared to whole brain radiation, more patients should maintain both cognitive functioning and quality of life after local radiation. The Any rule requires a beneficial effect on at least one outcome and ignores any result on the other outcome. Local radiation would be considered superior if fewer patients experience cognitive side effects, a lower quality of life, or both. The Compensatory rule also requires at least one favorable treatment effect, but the compensatory mechanism poses a restriction on the second outcome. The new treatment may perform better, similarly or even worse than the control treatment on this outcome, but the rule takes the size of the treatment differences into account to weigh beneficial and adverse effects. A net advantage on the sum of outcomes is required, such that several outcome combinations would result in a preference for local radiation. Superiority is concluded as long as favorable effects on cognitive functioning outweigh unfavorable effects on quality of life or vice versa.

The aforementioned decision rules ultimately lead to a conclusion about the treatment difference: The new treatment is considered superior if the difference between the new and the control treatment is larger than zero according to the decision rule of interest. For each of the decision rules, the corresponding superiority region is plotted in Figure 2.1. These superiority regions graphically represent how the treatment differences on both individual outcomes should be related to result in superiority: If the probability that the treatment difference falls in the marked area is sufficiently large, the treatment would be declared superior.

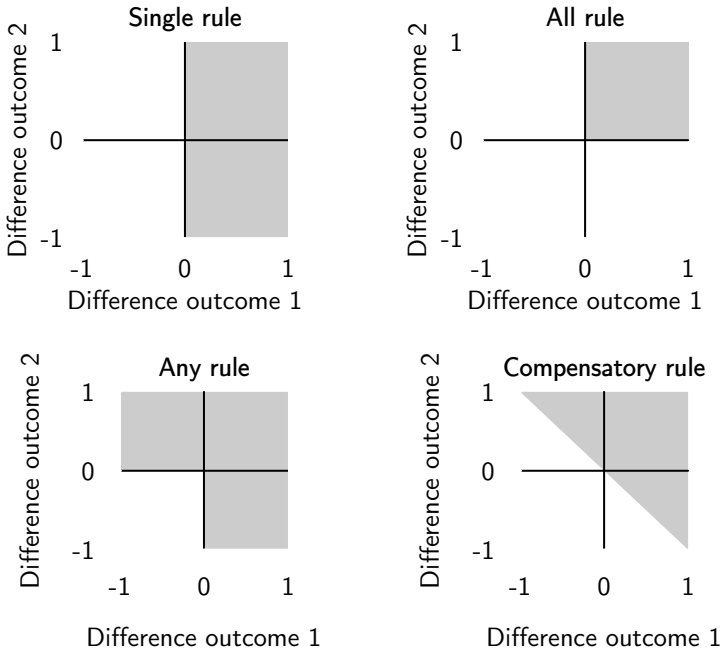


Figure 2.1: Superiority regions (shaded areas) for different decision rules.

2.2.1 Selecting a decision rule

The choice for a decision rule should be guided by the researcher's standard for superiority. To illustrate this, consider the following situations (see Figure 2.2 for a graphical representation):

1. Local radiation performs better on cognitive functioning as well as quality of life
2. Local radiation performs better on cognitive functioning and similarly on quality of life
3. Local radiation performs much better on cognitive functioning and slightly worse on quality of life
4. Local radiation performs slightly better on cognitive functioning and much worse on quality of life

If outcomes are equally important, most researchers would either (a) set a high standard and consider local radiation superior if both outcomes demonstrate an advantage (situation 1), or (b) balance outcomes and consider local radiation superior if advantages outweigh disadvantages (situations 1–3). Situation 4 is unlikely to result in a preference for local radiation, unless cognitive functioning is much more important than quality of life.

While the All rule applies to the high standard and differentiates situation 1 (superior)

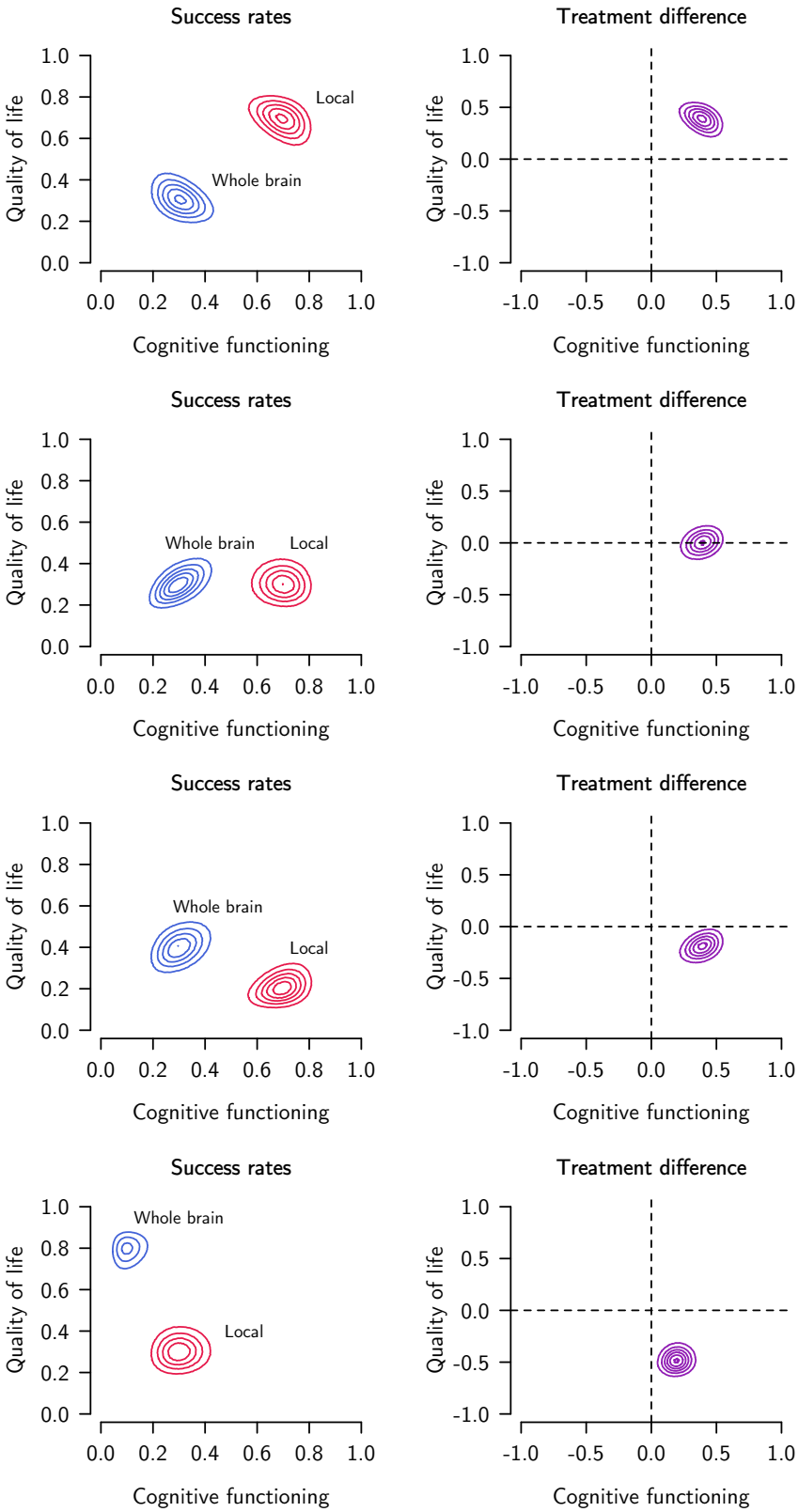


Figure 2.2: Example posterior distributions (left panels) and distributions of the treatment difference (right panels) for four different potential treatment differences (local radiation-whole brain radiation) in CAR-Study B.

from situations 2–4 (not superior), the Compensatory rule balances results and distinguishes situations 1–3 (superior) from situation 4 (not superior). The Single and Any rules do not meet these standards and would conclude that local radiation performs better in all situations, including the fourth. These rules should be used only when unfavorable effects can safely be ignored in the presence of a specific (Single rule) or any (Any rule) favorable effect.

2.3 Data analysis

To evaluate the decision rules discussed in the previous section, treatment comparison requires a procedure to quantify evidence in favor of the new treatment. The current section introduces the elements of a Bayesian approach to analyze data from two binary outcomes: likelihood, prior, and posterior distributions.

2.3.1 Description of the data and specification of the likelihood

Binary data have two values, traditionally labeled as 1 for success and 0 for failure. In general, success refers to improvement or absence of decline, and failure indicates the opposite: decline or absence of improvement respectively. Considering two outcomes together results in two binary responses per participant that can take four different combinations (see Table 2.1). The patient can have successes on both outcomes (x_{11}^{obs}); a success on one outcome, but not on the other (x_{10}^{obs} or x_{01}^{obs}); or failures on both outcomes (x_{00}^{obs}). The total number of successes on a particular outcome equals the sum of simultaneous and separate successes on that outcome, such that $x_1^{\text{obs}} = x_{11}^{\text{obs}} + x_{10}^{\text{obs}}$, etc.

Table 2.1: Response combinations for two binary outcomes

	Outcome 2		
Outcome 1	Success	Failure	Total
Success	x_{11}	x_{10}	x_1
Failure	x_{01}	x_{00}	$n - x_1$
Total	x_2	$n - x_2$	n

The multivariate likelihood of the outcomes is based on the four response frequencies.

These four response frequencies reflect (a) the individual success rates, and (b) the relation between outcomes. The latter serves as an additional source of information that may contribute to more efficient decision-making (Food and Drug Administration, 2010).

2.3.2 Specification of prior information

Prior information represents prior beliefs about success rates of individual treatments as well as the difference between treatments. These prior beliefs can, for example, incorporate information from comparable studies into the current one. Prior beliefs about two binary outcomes are quantified by four prior frequencies, expressed as x_{11}^{prior}, x_{10}^{prior}, x_{01}^{prior} and x_{00}^{prior}(Olkin & Trikalinos, 2015). Each of these individual prior frequencies incorporates information about one of the response frequencies in the data (x_{11}^{obs}, x_{10}^{obs}, x_{01}^{obs} and x_{00}^{obs}). Conveniently, one can think of these prior observations as an extra dataset, where the total number of observations in this prior dataset reflects the strength of the prior beliefs. Strong prior beliefs are translated to many prior observations, whereas weak prior beliefs can be expressed through small numbers of prior observations. An uninformative prior specification for the analysis of two binary outcomes would be a half observation for each response combination, such that the total number of prior observations equals two (Berger et al., 2015). This specification is also called Jeffreys prior and conveys virtually no information about the success rates of individual outcomes or the correlation between outcomes. If both treatments have this specification, no prior information about the treatment difference is provided either.

2.3.3 The posterior distribution

The posterior distribution reflects prior beliefs after they have been updated with the data and indicate the posterior success rates of individual outcomes in relation to each other; see also Miočević et al. (2020b), Miočević et al. (2020a), and Van de Schoot et al. (2020). The posterior response frequencies equal the sum of prior and observed frequencies, such that $x_{11}^{\text{post}} = x_{11}^{\text{prior}} + x_{11}^{\text{obs}}$, etc. Examples of posterior distributions for treatment effects with two

outcomes are graphically presented in Figure 2.2.

Comparison of the two posterior distributions allows for decision-making about treatment superiority, by quantifying evidence for a relevant treatment difference as a posterior probability. This posterior probability depends on the definition of superiority as defined via the decision rule and allows for two decisions. If the posterior probability exceeds a prespecified-specified threshold (often .95 or .99 in clinical trials (Food and Drug Administration, 2010), evidence is strong enough to consider the treatment superior. If the posterior probability is lower than the threshold, there is not sufficient evidence to conclude superiority.

2.4 Computation in practice

The online supplement offers a Shiny app to analyze real data using the framework proposed in the previous sections. If the researcher enters the prior (x_{11}^{prior}, x_{10}^{prior}, x_{01}^{prior}, x_{00}^{prior}) and observed (x_{11}^{obs}, x_{10}^{obs}, x_{01}^{obs}, x_{00}^{obs}) response frequencies for two treatments, the application:

(a) Computes the posterior probability of a treatment difference given the introduced decision rules

(b) Plots the posterior treatment distributions

(c) Plots the posterior distribution of the treatment difference

(d) Computes the prior, observed and posterior correlations between outcomes.

The Shiny app including user guide can be found at http://utrecht-university.shinyapps.io/multiple_binary_outcomes/ – for the annotated R code and potential newer versions go to http://www.github.com/XynthiaKavelaars.

The method and app are illustrated with artificial data from two treatment distributions with two negatively correlated binary outcome variables ($n = 100$ cases per treatment). The true success probabilities of the experimental and control treatments were .60 and .40 on both outcomes respectively, such that the experimental treatment performs better on both individual outcomes. The data were used to quantify evidence in favor of the experimental treatment according to the different decision rules (Single, Any, All, Compensatory).

User guide | Data | Prior | Treatment distributions | Treatment difference

Observed frequencies

Experimental treatment

	Success outcome 1	Failure outcome 1	Total outcome 2
Success outcome 2	32	29	61
Failure outcome 2	32	7	39
Total outcome 1	64	36	100

Control treatment

	Success outcome 1	Failure outcome 1	Total outcome 2
Success outcome 2	6	28	34
Failure outcome 2	33	33	66
Total outcome 1	39	61	100

Observed correlations

Treatment	Correlation
Experimental	-0.30
Control	-0.31

Figure 2.3: Screenshot of *Data* tab

The observed response frequencies were entered in the four upper-left cells of the table in the *Data* tab (see Figure 2.3). The app subsequently computed the total observed successes and failures in the margins as well as the observed correlations.

User guide | Data | Prior | Treatment distributions | Treatment difference

Prior frequencies

Experimental treatment

	Success outcome 1	Failure outcome 1	Total outcome 2
Success outcome 2	0.5	0.5	1
Failure outcome 2	0.5	0.5	1
Total outcome 1	1	1	2

Control treatment

	Success outcome 1	Failure outcome 1	Total outcome 2
Success outcome 2	0.5	0.5	1
Failure outcome 2	0.5	0.5	1
Total outcome 1	1	1	2

Prior correlations

Treatment	Correlation
Experimental	0.00
Control	0.00

Figure 2.4: Screenshot of *Prior* tab

Without any prior knowledge about the treatments or treatment differences, Jeffreys prior served as a prior distribution, such that each response category was assigned a half observation. After entering the prior frequencies in the *Prior* tab, the app provided the successes and failures per outcome and the prior correlation between outcomes (Figure 2.4).

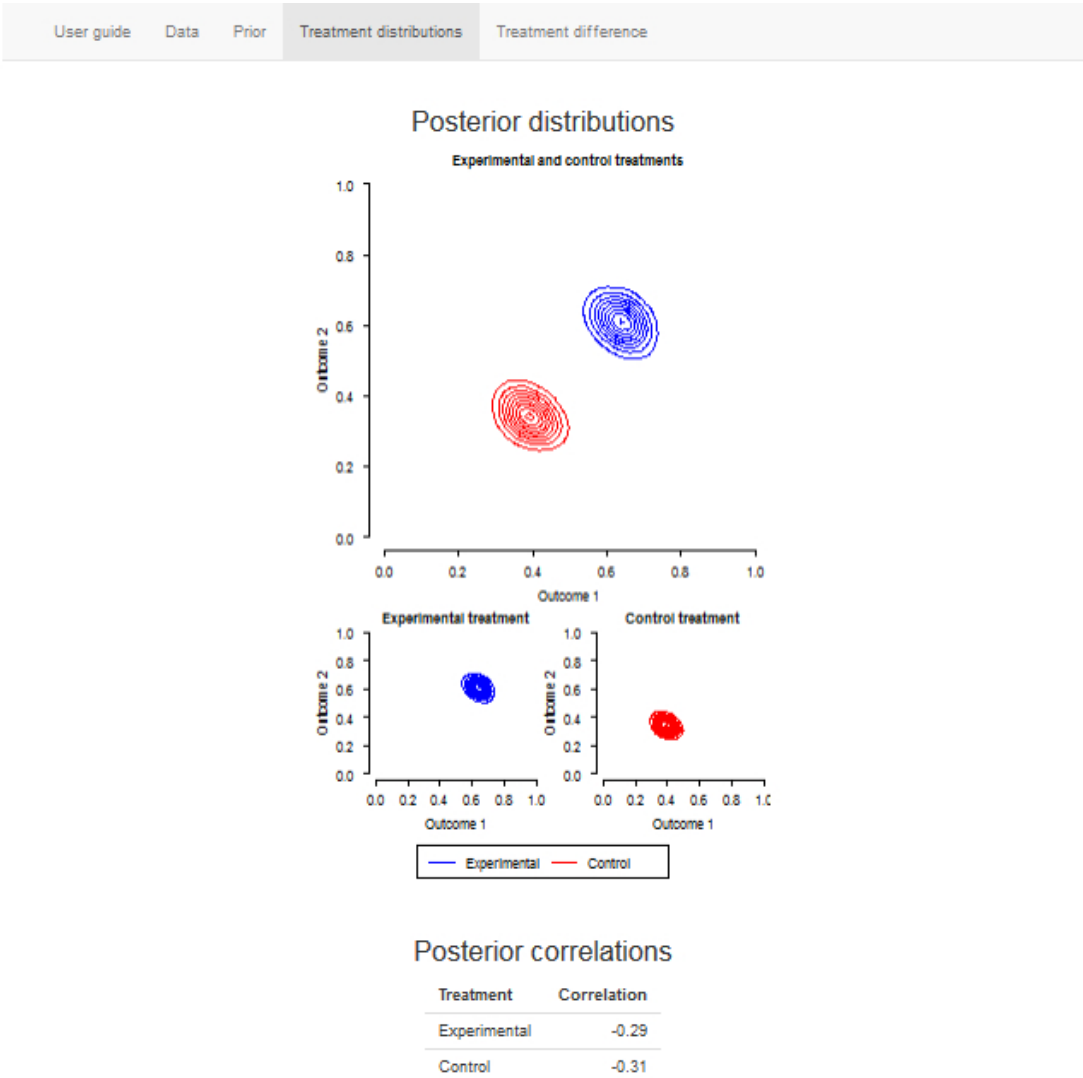


Figure 2.5: Screenshot of *Treatment distributions* tab

The *Treatment distributions* tab showed the posterior treatment distributions and posterior correlations of both treatments (Figure 2.5).

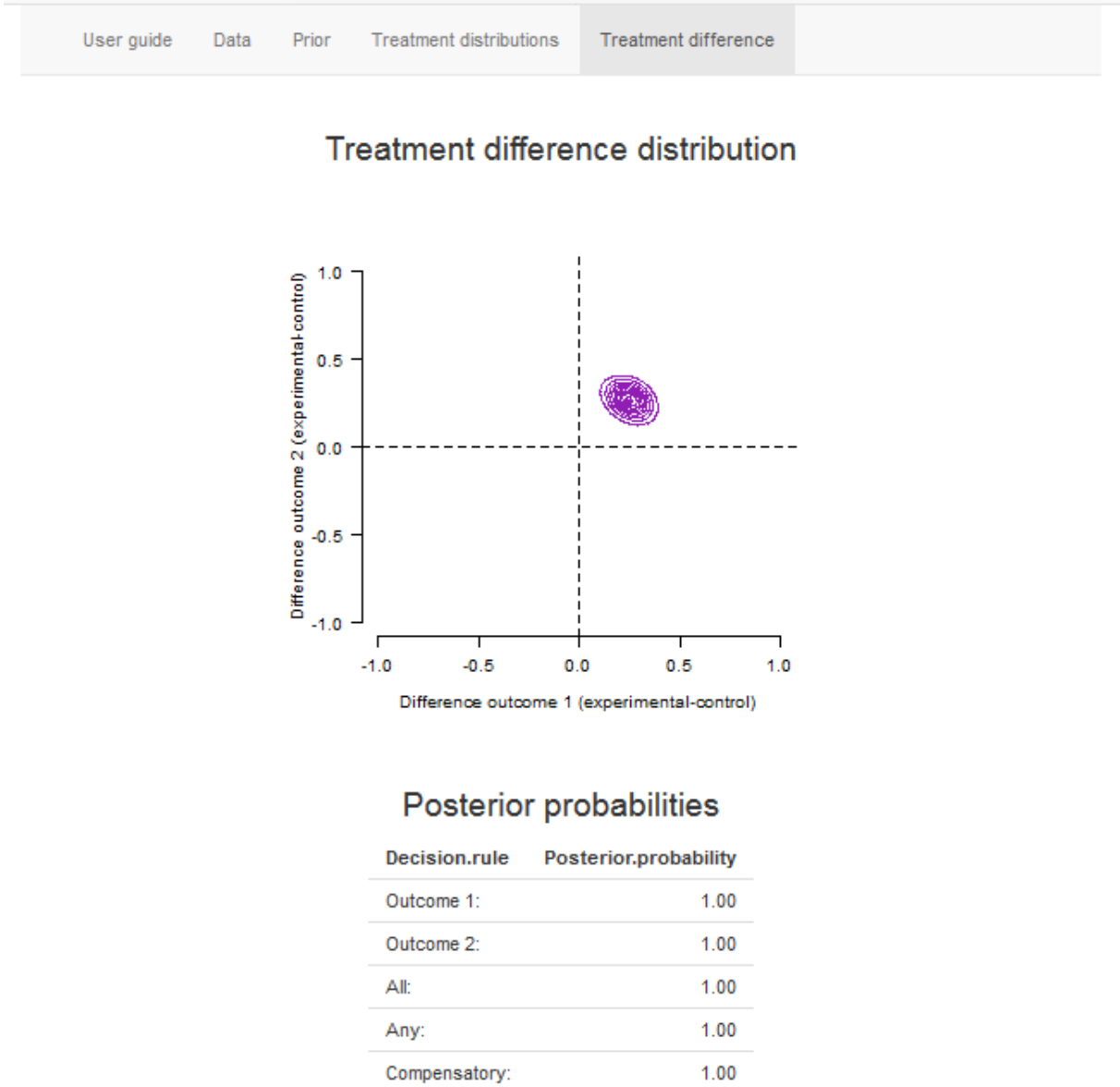


Figure 2.6: Screenshot of *Treatment difference* tab.

The *Treatment difference* tab (Figure 2.6) presented the distribution of the posterior treatment difference and the evidence in favor of the experimental treatment according to the proposed decision rules.

2.5 Sample size considerations

When the availability of participants is limited, a highly relevant question is how much data are minimally needed to make a sufficiently powerful decision. Since the sample size traditionally determines when to stop data collection, researchers often estimate the required number of participants before running the trial. Efficient a priori sample size estimation is difficult due to uncertainty about one or multiple treatment differences, regardless of the number of outcomes, since treatment differences are unknown in advance and need to be estimated. However, small inaccuracies in their estimation may have important consequences. Overestimating a treatment difference results in too small a sample to make a powerful decision, while (limited) underestimation needlessly extends the trial.

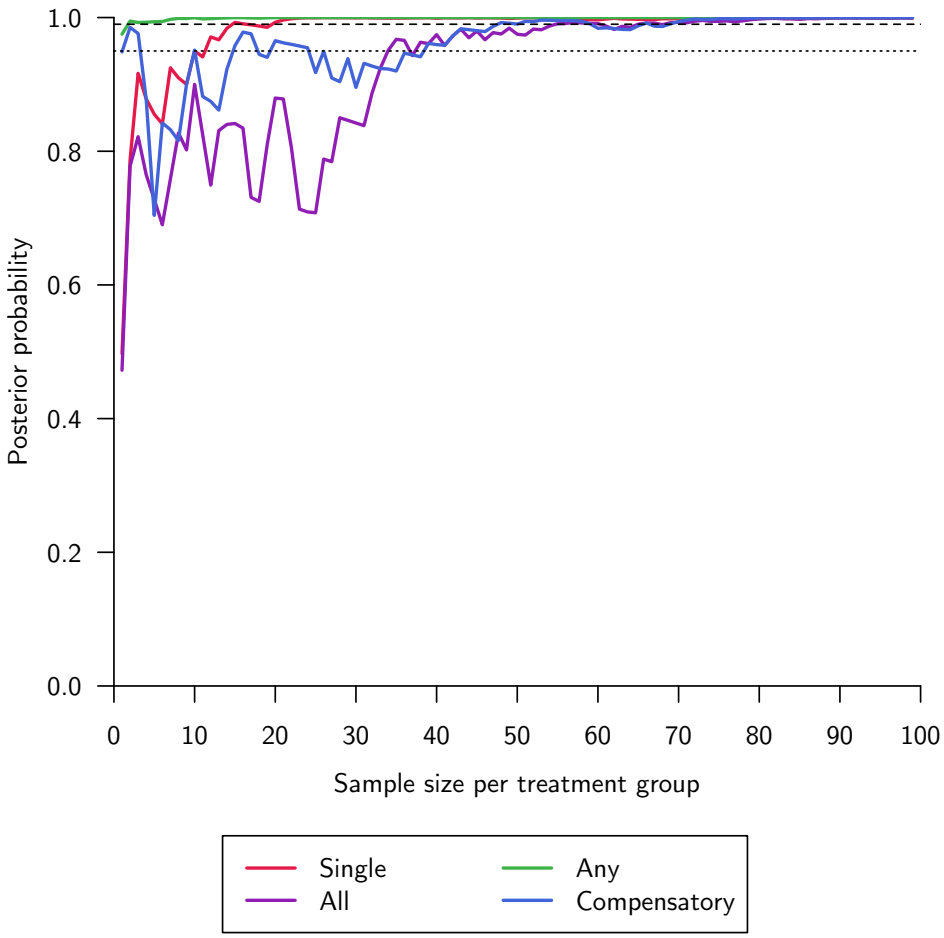


Figure 2.7: Example of evidence collection as data accumulate for different decision rules and two different decision criteria (dots = .95; dashes = .99)

In trials with multiple outcomes, the required sample size also depends on the decision rule as illustrated in Figure 2.7. The figure shows how evidence in favor of the decision rule under consideration changes for the example data from the Computation in practice section, while increasing the sample size in steps of one observation per group. Although the posterior probabilities of all decision rules ultimately approach one and conclude superiority as the data accumulate, different decision rules require different numbers of observations to arrive at that conclusion. With the data presented in Figure 2.7, the Any rule requires fewest observations to cross decision thresholds, followed by the Compensatory and Single outcome rules. The All rule requires the largest sample.

The relative efficiency of decision rules displayed in Figure 2.7 is specific to the particular scenario, since different relations between outcomes require different sample sizes to evaluate a specific decision rule (Food and Drug Administration, 2010). To provide an idea of the influence of the correlation between the outcomes, posterior treatment distributions for three correlation structures are displayed in Figure 2.8. This influence affects the proportion of overlap between the distribution of the posterior treatment difference and the superiority region of a decision rule, such that evidence in favor of the new treatment (i.e., posterior probability) as well as the required sample size to reach the decision threshold differ.

Figure 2.9 illustrates how the amount of evidence for each decision rule depends on the correlation when treatment differences are identical. The Single rule is not sensitive to the correlation: The proportion of the difference distribution that overlaps with the superiority

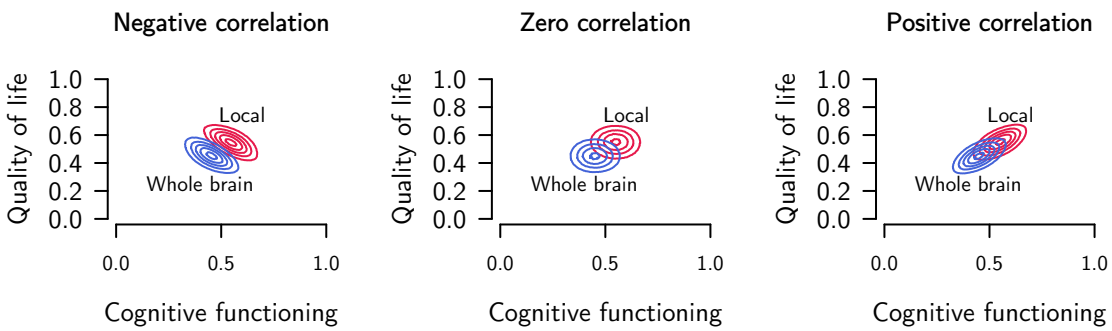


Figure 2.8: The influence of the correlation between outcomes on posterior treatment distributions.

region is similar for each correlation structure. The required sample size to conclude superiority will be the same. The All rule has a (slightly) larger proportion of overlap between the distribution of the difference and the superiority region when the correlation is positive. Compared to negatively correlated outcomes, the same amount of evidence can thus be obtained with a smaller sample. The Any and Compensatory rules demonstrate the relationship between the correlation structure and sample size more clearly. The distribution of the treatment difference falls completely in the superiority region when outcomes are negatively correlated (implying a posterior probability of one), while uncorrelated or positively correlated data result in a part of the distribution outside the superiority region (i.e., a posterior probability below one). The sample size will be smallest with negatively correlated outcomes.

In summary, several sources of uncertainty complicate a priori sample size estimation in trials with multiple outcomes: Treatment differences on individual outcomes, the correlation between outcomes, and the decision rule influence the required number of observations. The difficulty of accurately estimating the sample size interferes with the potential efficiency gain of multiple outcomes, such that a priori sample size estimation may be inadequate with small samples and multiple outcomes (Rauch & Kieser, 2015).

2.5.1 Adaptive trial design

To reduce the impact of unknown information on the efficiency of trials the sample size can be estimated while running the trial, using a method called adaptive stopping (Berry et al., 2010). Adaptive stopping performs one or multiple interim analyses and stops the trial as soon as evidence is conclusive, such that efficiency is optimized. Compared to a priori sample size estimation, adaptive stopping may result in early trial termination if the treatment difference is larger than expected (i.e., underestimated). If the treatment difference appears smaller than anticipated (i.e., overestimated) and evidence remains inconclusive, the trial may be extended beyond the planned sample size. Adaptive stopping thus forms a flexible alternative that embraces the uncertainties of the traditional a priori estimated sample size (Bauer et al.,

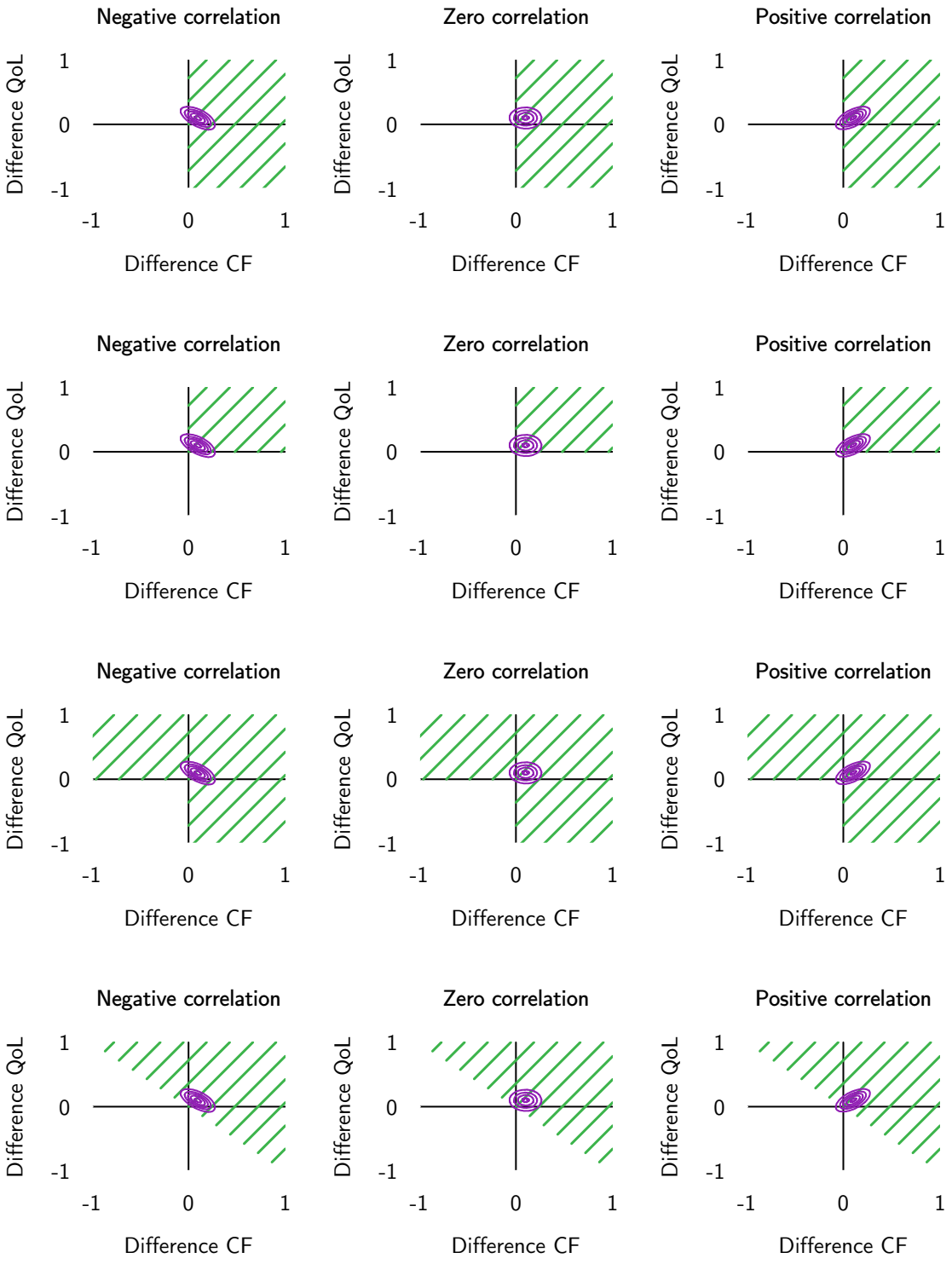


Figure 2.9: The influence of the correlation on the evidence for various decision rules. A larger proportion of overlap between the distribution of the treatment difference and the superiority region (shaded area) indicate more evidence. CF = cognitive functioning; QoL = Quality of Life.

2016; Thorlund et al., 2018).

Although interim analyses form an attractive approach to improve efficiency, adaptive trials must be designed carefully (Food and Drug Administration, 2010; Sanborn & Hills, 2014). The final decision about superiority potentially requires several interim decisions to evaluate whether evidence is strong enough to draw a conclusion. Without properly adjusting the design to repeated decision-making, the risk of falsely concluding superiority (i.e., Type I error) over all decisions is larger than anticipated, as shown in Figure 2.10 (Sanborn & Hills, 2014). To keep the Type I error risk over all decisions acceptable, the Type I error rate for individual decisions must be adjusted (Jennison & Turnbull, 1999). A 5% Type I error risk over multiple decisions consequentially results in individual decisions that have a Type I error risk below 5%. The size of the adjustment depends on the number of interim decisions: More decisions require a larger adjustment of the Type I error rate for individual decisions (see Figure 2.10).

A key element in Type I error control is the decision threshold: the lower limit for the posterior probability to conclude superiority. The decision threshold equals $1 - \alpha$, where α is the maximum Type I error probability (Marsman & Wagenmakers, 2016). A 5% risk of an incorrect superiority decision ($\alpha = .05$) results in a minimal posterior probability of .95. A very high threshold might be attractive to minimize Type I errors, but does not contribute to efficient decision-making: A larger sample size is required to regulate the chance to detect a true treatment difference (i.e., to protect power). The decision threshold thus relates the Type I error and required sample size via the number of interim analyses (Shi & Yin, 2019). Limiting the number of decisions is key to efficiently designing an adaptive trial (Jennison & Turnbull, 1999). To this end, the Food and Drug Administration (2010) recommends balancing the number of interim analyses with decision error rates, by carefully choosing three design parameters:

1. The sample size to look at the data for the first time (n_{min})
2. The number of added participants if the previous analysis did not provide sufficient evidence (interim group size)
3. The sample size to stop the trial if evidence is not strong enough to conclude superiority

(n_{max}).

The sample size at the first interim analysis (n_{min}) should not be too small for two reasons. First, a small interim sample size could detect unrealistically large treatment effects only and needlessly increases the number of interim analyses. Second, very small samples increase the probability of falsely concluding superiority (Schönbrodt et al., 2017). As shown in Figure 2.7, the posterior probability is unstable with few observations and becomes more stable as the number of observations increases. Single observations can be influential in small samples, and this influence diminishes as the sample size increases. A larger n_{min} automatically reduces the number of interim analyses as well as the Type I errors and requires a smaller correction of the decision threshold, as illustrated in Figure 2.10. However,

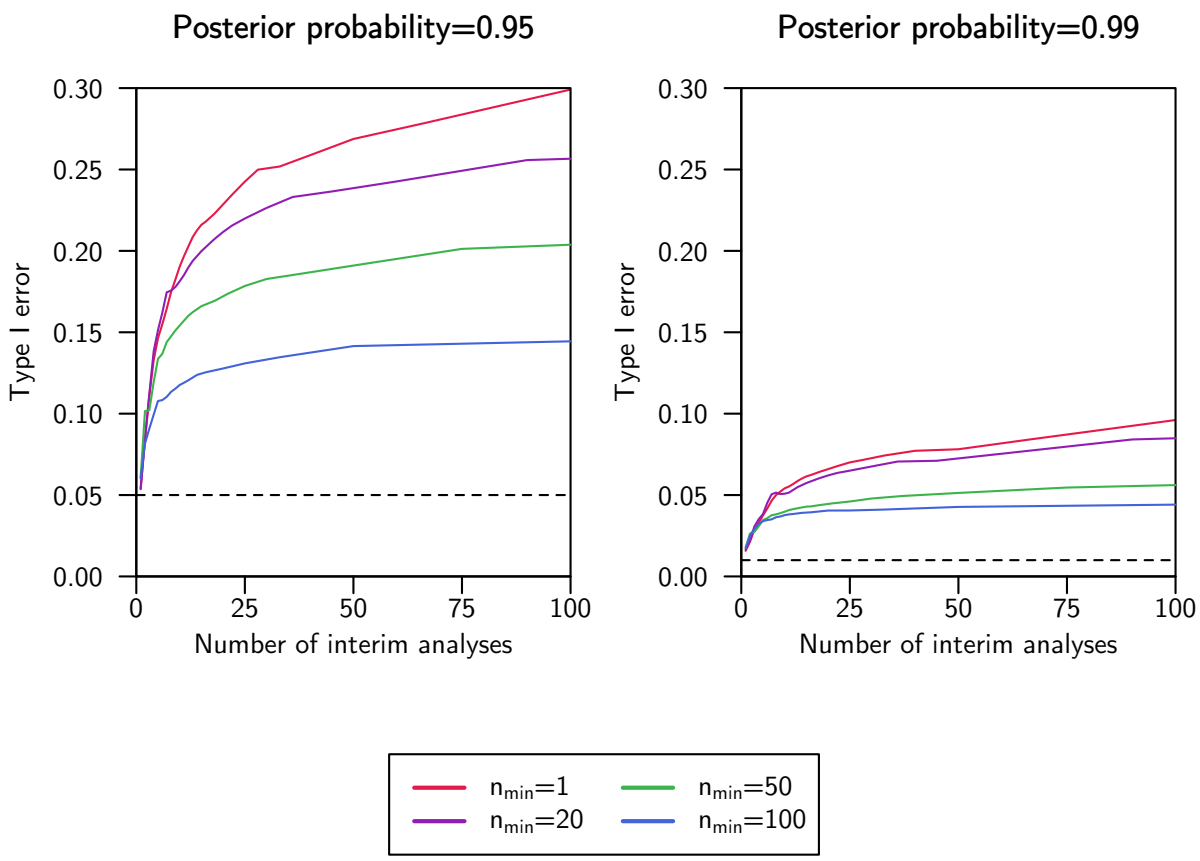


Figure 2.10: The empirical Type I error probability as a function of the number of interim analyses for different n_{min} when the decision threshold is not corrected for the number of interim analyses. Dashed lines indicate the desired thresholds of $\alpha = 0.05$ (posterior probability = 0.95) and $\alpha = 0.01$ (posterior probability = 0.99).

a too large n_{min} limits efficiency: Superiority may have been concluded with a smaller sample and in potential participant recruitment is needlessly extended.

If the first interim analysis did not result in conclusive evidence, the sample size can be increased in several steps. The interim group size of added participants should be chosen with the inconclusive results of the previous analysis in mind, such that the new sample provides a reasonable chance of detecting a treatment difference given the earlier lack of evidence. The number of observations between interim analyses may be the same throughout the trial, or can differ per interim analysis if that would benefit the trial's efficiency. It should be chosen carefully, however, since too small and too large group sizes both reduce efficiency (Jennison & Turnbull, 1999). A too small group size needlessly increases the number of interim analyses, while a too large group size reduces the flexibility to terminate the trial as soon as the decision threshold has been met.

Ideally, the sample size to terminate the trial if the data do not provide sufficient evidence for superiority (n_{max}) equals the sample size that is required to detect the smallest treatment effect of clinical interest (Food and Drug Administration, 2010). In practice, n_{max} will often be limited by the maximum number of available participants and may be smaller than optimal, which has the same consequence as a too small (a priori estimated) sample size: A limited n_{max} restricts the power to detect small treatment differences.

2.6 Concluding remarks

The current chapter presented a Bayesian framework for decision-making with multiple outcomes and illustrated how decisions with two outcomes may help a small sample, when (a) using a decision rule that combines information from two outcomes efficiently, and (b) designing a trial adaptively. Without giving all the mathematical details, I have tried to provide a clear intuition to the approach and software to carry out the analysis.

The proposed approach has several extensions that may accommodate more realistic decisions. First, more than two outcomes can be included, such that researchers might weigh treatment differences on three or more relevant aspects. Increasing the number of outcomes

may further improve efficiency, but more outcomes also increase the complexity of the data analysis.

Second, although equal importance of outcomes was assumed throughout the chapter, unequal importance of outcomes could be incorporated. The Compensatory rule in particular could be adapted easily to, for example, include survival into a decision; an outcome that is in many cases more important than cognitive side effects. However, user-friendly software packages for more outcomes remain to be developed.

Third, the applicability of adaptive designs can be strongly improved with clear guidelines on the concrete choice of design parameters. Optimal design of interim analyses is necessary to do justice to the potential flexibility of adaptive trials.

Chapter 3

Decision-making with multiple correlated binary outcomes in clinical trials

Based on Kavelaars, X., Mulder, J., & Kaptein, M. (2020). Decision-making with multiple correlated binary outcomes in clinical trials. *Statistical Methods in Medical Research*, *29*(11), 3265–3277. https://doi.org/10.1177/0962280220922256

Abstract

Clinical trials often evaluate multiple outcome variables to form a comprehensive picture of the effects of a new treatment. The resulting multidimensional insight contributes to clinically relevant and efficient decision-making about treatment superiority. Common statistical procedures to make these superiority decisions with multiple outcomes have two important shortcomings however: 1) Outcome variables are often modeled individually, and consequently fail to consider the relation between outcomes; and 2) superiority is often defined as a relevant difference on a single, on any, or on all outcomes(s); and lacks a compensatory mechanism that allows large positive effects on one or multiple outcome(s) to outweigh small negative effects on other outcomes. To address these shortcomings, this paper proposes 1) a Bayesian model for the analysis of correlated binary outcomes based on the multivariate Bernoulli distribution; and 2) a flexible decision criterion with a compensatory mechanism that captures the relative importance of the outcomes. A simulation study demonstrates that efficient and unbiased decisions can be made while Type I error rates are properly controlled. The performance of the framework is illustrated for 1) fixed, group sequential, and adaptive designs; and 2) non-informative and informative prior distributions.

3.1 Introduction

Clinical trials often aim to compare the effects of two treatments. To ensure clinical relevance of these comparisons, trials are typically designed to form a comprehensive picture of the treatments by including multiple outcome variables. Collected data about efficacy (e.g., reduction of disease symptoms), safety (e.g., side effects), and other relevant aspects of new treatments are combined into a single, coherent decision regarding treatment superiority. An example of a trial with multiple outcomes is CAR-Study B (Cognitive and Radiation Study B), which investigated an experimental treatment for cancer patients with multiple metastatic brain tumors (Schimmel et al., 2018). Historically, these patients have been treated with radiation of the whole brain (Whole Brain Radiation Therapy; WBRT). This treatment is known to damage healthy brain tissue and to increase the risk of (cognitive) side effects. More recently, local radiation of the individual metastases (stereotactic surgery; SRS) has been proposed as a promising alternative that saves healthy brain tissue and could therefore reduce side effects. The CAR-Study B compared these two treatments based on cognitive functioning, fatigue, and several other outcome variables (Schimmel et al., 2018).

Statistical procedures to arrive at a superiority decision have two components: 1) A statistical model for the collected data; and 2) A decision rule to evaluate the treatment in terms of superiority based on the modeled data. Ideally, the combination of these components forms a decision procedure that satisfies two criteria: Decisions should be clinically relevant and efficient. Clinical relevance ensures that the statistical decision rule corresponds to a meaningful superiority definition, given the clinical context of the treatment. Commonly used decision rules define superiority as one or multiple treatment difference(s) on the most important outcome, on any of the outcomes, or on all of the outcomes (Food and Drug Administration, 2017; Murray et al., 2016; Sozu et al., 2012, 2016). Efficiency refers to achieving acceptable error rates while minimizing the number of patients in the trial. The emphasis on efficiency is motivated by several considerations, such as small patient populations, ethical concerns, limited access to participants, and other

difficulties to enroll a sufficient number of participants (Van de Schoot & Miočević, 2020). In the current paper, we address clinical relevance and efficiency in the context of multiple binary outcomes and propose a framework for statistical decision-making.

In trials with multiple outcomes, it is common to use a univariate modeling procedure for each individual outcome and combine these with one of the aforementioned decision rules (Food and Drug Administration, 2017; Murray et al., 2016). Such decision procedures can be inefficient since they ignore the relationships between outcomes. Incorporating these relations in the modeling procedure is crucial as they directly influence the amount of evidence for a treatment difference as well as the sample size required to achieve satisfactory error rates. A multivariate modeling procedure takes relations between outcomes into account and can therefore be a more efficient and accurate alternative when outcomes are correlated.

Another interesting feature of multivariate models is that they facilitate the use of decision rules that combine multiple outcomes in a flexible way, for example via a compensatory mechanism. Such a mechanism is characterized by the property that beneficial effects are given the opportunity to compensate adverse effects. The flexibility of compensatory decision-making is appealing, since a compensatory mechanism can be naturally extended with impact weights that explicitly take the clinical importance of individual outcome variables into account (Murray et al., 2016). With impact weights, outcome variables of different importance can be combined into a single decision in a straightforward way.

Compensatory rules do not only contribute to clinical relevance, but also have the potential to increase trial efficiency. Effects on individual outcomes may be small (and seemingly unimportant) while the combined treatment effect may be large (and important; O'Brien, 1984; Pocock et al., 1987; Tang et al., 1989), as visualized in Figure 3.1 for fictive data of CAR-Study B. The two displayed bivariate distributions reflect the effects and their uncertainties on cognitive functioning and fatigue for SRS and WBRT. The univariate distributions of both outcomes overlap too much to clearly distinguish the two treatments on individual outcome variables or a combination of them. The bivariate distributions however

clearly distinguish between the two treatments. Consequently, modeling a compensatory treatment effect with equal weights (visualized as the diagonal dashed line) would provide sufficient evidence to consider SRS superior in the presented situation.

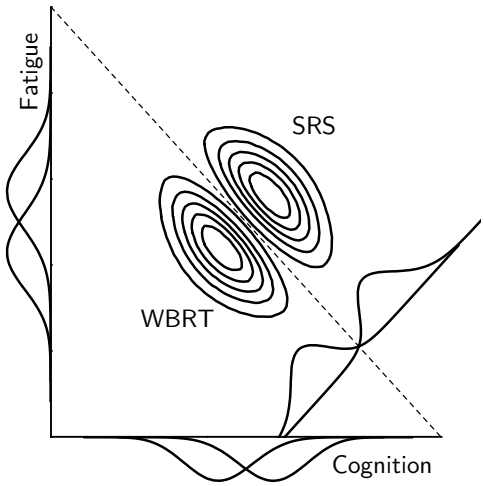


Figure 3.1: Separation of two bivariate distributions (diagonally) versus separation of their univariate distributions (horizontally/vertically) for CAR-Study B. The dashed diagonal line represents a Compensatory decision rule with equal weights. Each distribution reflects the plausibility of the treatment effects on cognitive functioning and fatigue after observing fictive data.

In the current paper, we propose a decision procedure for multivariate decision-making with multiple (correlated) binary outcomes. The procedure consists of two components. First, we model the data with a multivariate Bernoulli distribution, which is a multivariate generalization of the univariate Bernoulli distribution. The model is exact and does not rely on numerical approximations, making it appropriate for small samples. Second, we extend multivariate analysis with a compensatory decision rule to include more comprehensive and flexible definitions of superiority.

The decision procedure is based on a Bayesian multivariate Bernoulli model with a conjugate prior distribution. The motivation for this model is twofold. First, the multivariate Bernoulli model is a natural generalization of the univariate Bernoulli model, which intuitively parametrizes success probabilities per outcome variable. Second, a conjugate prior distribution can greatly facilitate computational procedures for inference. Conjugacy ensures that the form of the posterior distribution is known, making sampling from the posterior distribution straightforward.

Although Bayesian analysis is well-known to allow for inclusion of information external to the trial by means of prior information (Gelman et al., 2013), researchers who wish not to include prior information can obtain results similar to frequentist analysis. The use of a non-informative prior distribution essentially results in a decision based on the likelihood of the data, such that 1) Bayesian and frequentist (point) estimates are equivalent; and 2) the frequentist p-value equals the Bayesian posterior probability of the null hypothesis in one-sided testing (Marsman & Wagenmakers, 2016). Since a (combined) p-value may be difficult to compute for the multivariate Bernoulli model, Bayesian computational procedures can exploit this equivalence and facilitate computations involved in Type I error control (Food and Drug Administration, 2010; Wilson, 2019).

The remainder of the paper is structured as follows. In the next section, we present a multivariate approach to the analysis of multiple binary outcomes. Subsequently, we discuss various decision rules to evaluate treatment differences on multiple outcomes. The framework is evaluated in the *Numerical evaluation* section, and we discuss limitations and extensions in the *Discussion*.

3.2 A model for multivariate analysis of multiple binary outcomes

3.2.1 Notation

We start the introduction of our framework with some notation. The joint response for patient i in treatment j on K outcomes will be denoted by $\mathbf{x}_{j,i} = (x_{j,i,1}, \dots, x_{j,i,K})$, where $i \in \{1, \dots, n_j\}$, and $j \in \{E, C\}$ (i.e., Experimental and Control). The response on outcome k $x_{j,i,k} \in \{0, 1\}$ (0 = failure, 1 = success), such that $\mathbf{x}_{j,i}$ can take on $Q = 2^K$ different combinations $\{1 \dots 11\}, \{1 \dots 10\}, \dots, \{0 \dots 01\}, \{0 \dots 00\}$. The observed frequencies of each possible response combination for treatment j in a dataset of n_j patients are denoted by vector $\mathbf{s}_j$ of length Q. The elements of $\mathbf{s}_j$ add up to n_j, $\sum_{q=1}^{Q} \mathbf{s}_{j,q} = n_j$.

Vector $\boldsymbol{\theta}_j = (\theta_{j,1}, \dots, \theta_{j,K})$ reflects success probabilities of K outcomes for treatment j

in the population. Vector $\boldsymbol{\delta} = (\delta_1, \dots, \delta_K)$ then denotes the treatment differences on K outcomes, where $\delta_k = \theta_{E,k} - \theta_{C,k}$. We use $\boldsymbol{\phi}_j = \phi_{j,1\dots11}, \phi_{j,1\dots10}, \dots, \phi_{j,0\dots01}, \phi_{j,0\dots00}$ to refer to probabilities of joint responses in the population, where $\phi_{j,q}$ denotes the probability of joint response combination $\mathbf{x}_{j,i}$ with configuration q. Vector $\boldsymbol{\phi}_j$ has Q elements, and sums to unity, $\sum_{q=1}^{Q} \boldsymbol{\phi}_{j,q} = 1$. Information about the relation between outcomes k and l is reflected by $\phi_{j,kl}$, which is defined as the sum of those elements of $\boldsymbol{\phi}_j$ that have the kth and l^{th} elements of q equal to 1, e.g., $\phi_{j,11}$ for $K = 2$. Similarly, marginal probability $\theta_{j,k}$ follows from summing all elements of $\boldsymbol{\phi}_j$ with the kth element of q equal to 1. For example, with three outcomes, the success probability of the first outcome is equal to $\theta_{j,1} = \phi_{j,111} + \phi_{j,110} + \phi_{j,101} + \phi_{j,100}$.

3.2.2 Likelihood

The likelihood of joint response $\mathbf{x}_{j,i}$ follows a K-variate Bernoulli distribution (Dai et al., 2013):

$$\begin{aligned} p(\mathbf{x}_{j,i}|\boldsymbol{\phi}_j) &= \text{multivariate Bernoulli}(\mathbf{x}_{j,i}|\boldsymbol{\phi}_j) \\ &= \phi_{j,1\dots11}^{x_{j,1}\times\dots\times x_{j,K}} \phi_{j,1\dots10}^{x_{j,1}\times\dots\times x_{j,K-1}(1-x_{j,K})} \times \dots \times \\ &\quad \phi_{j,0\dots01}^{(1-x_{j,1})\times\dots\times(1-x_{j,K-1})x_{j,K}} \phi_{j,0\dots00}^{(1-x_{j,1}\times\dots\times 1-x_{j,K})}. \end{aligned} \tag{3.1}$$

The multivariate Bernoulli distribution in Equation 3.1 is a specific parametrization of the multinomial distribution. The likelihood of n_j joint responses summarized by cell frequencies in $\mathbf{s}_j$ follows a Q-variate multinomial distribution with parameters $\boldsymbol{\phi}_j$:

$$\begin{aligned} p(\mathbf{s}_j|\boldsymbol{\phi}_j) &= \text{multinomial}(\mathbf{s}_j|\boldsymbol{\phi}_j) \\ &\propto \phi_{j,1\dots11}^{s_{j,1\dots11}} \phi_{j,1\dots10}^{s_{j,1\dots10}} \times \dots \times \phi_{j,0\dots01}^{s_{j,0\dots01}} \phi_{j,0\dots00}^{s_{j,0\dots00}}. \end{aligned} \tag{3.2}$$

Conveniently, the multivariate Bernoulli distribution is consistent under marginalization. That is, marginalizing a $K-$variate Bernoulli distribution with respect to p variables results in a $(K-p)$-variate Bernoulli distribution (Dai et al., 2013). Hence, the univariate Bernoulli distribution is directly related and results from marginalizing $(K-1)$ variables.

The pairwise correlation between variables $x_{j,k}$ and $x_{j,l}$ is reflected by $\rho_{x_{j,k},x_{j,l}}$ (Dai et al.,

2013):

$$\rho_{x_{j,k}x_{j,l}} = \frac{\phi_{j,kl} - \theta_{j,k}\theta_{j,l}}{\sqrt{\theta_{j,k}(1-\theta_{j,k})\theta_{j,l}(1-\theta_{j,l})}}. \tag{3.3}$$

This correlation is over the full range, i.e., $-1 \leq \rho_{x_{j,k},x_{j,l}} \leq 1$ (Olkin & Trikalinos, 2015).

3.2.3 Prior and posterior distribution

A natural choice to model prior information about response probabilities $\boldsymbol{\phi}_j$ is the Dirichlet distribution, since a Dirichlet prior and multinomial likelihood form a conjugate combination. The Q-variate prior Dirichlet distribution has hyperparameters $\boldsymbol{\alpha}_j^0 = (\alpha_{j,11\ldots11}^0, \alpha_{j,11\ldots10}^0, \ldots, \alpha_{j,00\ldots01}^0, \alpha_{j,00\ldots00}^0)$:

$$\begin{aligned} p(\boldsymbol{\phi}_j) &= \text{Dirichlet}(\boldsymbol{\phi}_j|\boldsymbol{\alpha}_j^0) \\ &\propto \phi_{j,1\ldots11}^{\alpha_{j,1\ldots11}^0-1}\phi_{j,1\ldots10}^{\alpha_{j,1\ldots10}^0-1} \times \cdots \times \phi_{j,0\ldots01}^{\alpha_{j,0\ldots01}^0-1}\phi_{j,0\ldots00}^{\alpha_{j,0\ldots00}^0-1}, \end{aligned} \tag{3.4}$$

where each of the prior hyperparameters $\boldsymbol{\alpha}_j^0$ should be larger than zero to ensure a proper prior distribution.

The posterior distribution of $\boldsymbol{\phi}_j$ results from multiplying the likelihood and the prior distribution and follows a Dirichlet distribution with parameters $\boldsymbol{\alpha}_j^n = \boldsymbol{\alpha}_j^0 + \mathbf{s}_j$:

$$\begin{aligned} p(\boldsymbol{\phi}_j|\mathbf{s}_j) =& \text{Dirichlet}(\boldsymbol{\phi}_j|\boldsymbol{\alpha}_j^0 + \mathbf{s}_j) \\ &\propto \phi_{j,1\ldots11}^{s_{j,1\ldots11}}\phi_{j,1\ldots10}^{s_{j,1\ldots10}} \times \cdots \times \phi_{j,0\ldots01}^{s_{j,0\ldots01}}\phi_{j,0\ldots00}^{s_{j,0\ldots00}} \times \\ &\quad \phi_{j,1\ldots11}^{\alpha_{j,1\ldots11}^0-1}\phi_{j,1\ldots10}^{\alpha_{j,1\ldots10}^0-1} \times \cdots \times \phi_{j,0\ldots01}^{\alpha_{j,0\ldots01}^0-1}\phi_{j,0\ldots00}^{\alpha_{j,0\ldots00}^0-1} \\ &\propto \phi_{j,1\ldots11}^{\alpha_{j,1\ldots11}^n-1}\phi_{j,1\ldots10}^{\alpha_{j,1\ldots10}^n-1} \times \cdots \times \phi_{j,0\ldots01}^{\alpha_{j,0\ldots01}^n-1}\phi_{j,0\ldots00}^{\alpha_{j,0\ldots00}^n-1}. \end{aligned} \tag{3.5}$$

Since prior hyperparameters $\boldsymbol{\alpha}_j^0$ impact the posterior distribution of treatment difference $\boldsymbol{\delta}$, specifying them carefully is important. Each of the hyperparameters contains information about one of the observed frequencies $\mathbf{s}_j$ and can be considered a prior frequency that reflects the strength of prior beliefs. Equation 3.5 shows that the influence of prior information depends

on prior frequencies $\boldsymbol{\alpha}_j^0$ relative to observed frequencies $\mathbf{s}_j$. When all elements of $\boldsymbol{\alpha}_j^0$ are set to zero, $\boldsymbol{\alpha}_j^n = \mathbf{s}_j$. This (improper) prior specification results in a posterior mean of $\phi_{j,q}|s_{j,q} = \frac{\alpha_{j,q}^n}{\sum_{p=1}^Q \alpha_{j,p}^n}$, which is equivalent to the frequentist maximum likelihood estimate of $\phi_{j,q} = \frac{s_{j,q}}{\sum_{p=1}^Q s_{j,p}}$. To take advantage of this property with a proper non-informative prior, one could specify hyperparameters slightly larger than zero such that the posterior distribution is essentially completely based on the information in the data (i.e., $\boldsymbol{\alpha}_j^n \approx \mathbf{s}_j$).

To include prior information - when available - in the decision, $\boldsymbol{\alpha}_j^0$ can be set to specific prior frequencies to increase the influence on the decision. These prior frequencies may for example be based on results from related historical trials. We provide more technical details on prior specification in Appendix A *Specification of prior hyperparameters*. There we also highlight the relation between the Dirichlet distribution and the multivariate beta distribution, and demonstrate that the prior and posterior distributions of $\boldsymbol{\theta}_j$ are multivariate beta distributions.

The final superiority decision relies on the posterior distribution of treatment difference $\boldsymbol{\delta}$. Although this distribution does not belong to a known family of distributions, we can approach the distribution of $\boldsymbol{\delta}$ via a two-step transformation of the posterior samples of $\boldsymbol{\phi}_j$. First, a sample of $\boldsymbol{\phi}_j$ is drawn from its known Dirichlet distribution. Next, these draws can be transformed to a sample of $\boldsymbol{\theta}_j$ using the property that joint response frequencies sum to the marginal probabilities. Finally, these samples from the posterior distributions of $\boldsymbol{\theta}_E$ and $\boldsymbol{\theta}_C$ can then be transformed to obtain the posterior distribution of joint treatment difference $\boldsymbol{\delta}$, by subtracting draws of $\boldsymbol{\theta}_C$ from draws of $\boldsymbol{\theta}_E$, i.e., $\boldsymbol{\delta} = \boldsymbol{\theta}_E - \boldsymbol{\theta}_C$. Algorithm 1 in Subsection Implementation of the framework includes pseudocode with the steps required to obtain a sample from the posterior distribution of $\boldsymbol{\delta}$.

3.3 Decision rules for multiple binary outcomes

The current section discusses how the model from the previous section can be used to make treatment superiority decisions. Treatment superiority is defined by the posterior mass in a specific subset of the multivariate parameter space of $\boldsymbol{\delta} = (\delta_1, \dots, \delta_K)$. The complete parameter space will be denoted by $\mathcal{S} \subset (-1, 1)^K$, and the superiority space will be denoted

by $\mathcal{S}_{Sup} \subset S$. Superiority is concluded when a sufficiently large part of the posterior distribution of $\boldsymbol{\delta}$ falls in superiority region $\mathcal{S}_{Sup}$:

$$P(\boldsymbol{\delta} \in \mathcal{S}_{sup} | \mathbf{s}_E, \mathbf{s}_C) > p_{cut} \tag{3.6}$$

where p_{cut} reflects the decision threshold to conclude superiority. The value of this threshold should be chosen to control the Type I error rate α.

3.3.1 Four different decision rules

Different partitions of the parameter space define different superiority criteria to distinguish two treatments. The following decision rules conclude superiority when there is sufficient evidence that:

1. *Single rule:* an a priori specified primary outcome k has a treatment difference larger than zero. The superiority region is denoted by:

$$\mathcal{S}_{Single(k)} = \{\boldsymbol{\delta} | \delta_k > 0\}. \tag{3.7}$$

   Superiority is concluded when

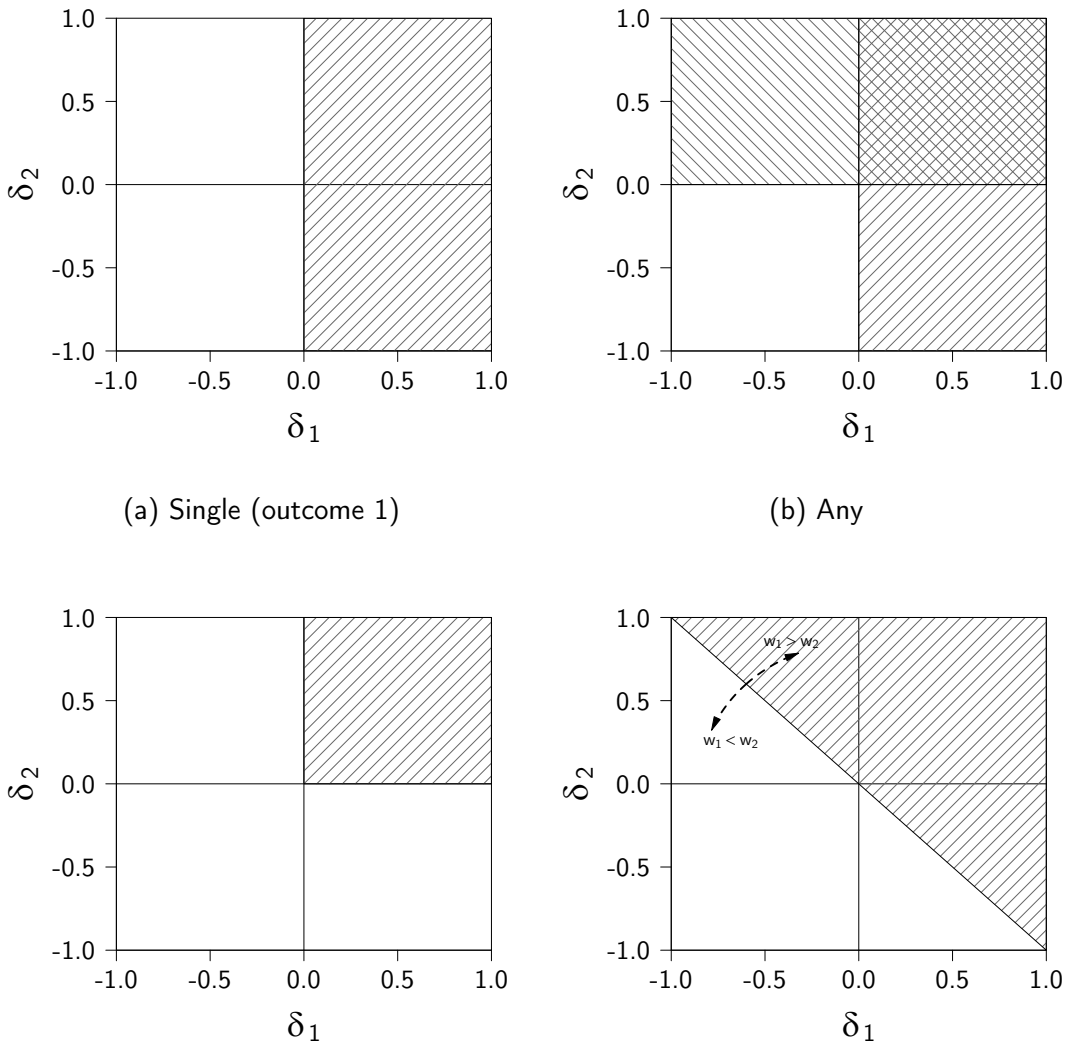
$$P(\boldsymbol{\delta} \in \mathcal{S}_{Single(k)} | \mathbf{s}_E, \mathbf{s}_C) > p_{cut}. \tag{3.8}$$

2. *Any rule:* at least one of the outcomes has a treatment difference larger than zero. The superiority region is a combination of K superiority regions of the Single rule:

$$\mathcal{S}_{Any} = \{\mathcal{S}_{Single_1} \cup \cdots \cup \mathcal{S}_{Single_K}\}.$$

   Superiority is concluded when

$$\max_k P(\boldsymbol{\delta} \in \mathcal{S}_{Single(k)} | \mathbf{s}_E, \mathbf{s}_C) > p_{cut}. \tag{3.9}$$



(a) Single (outcome 1)

(b) Any

(c) All

(d) Compensatory

Figure 3.2: Superiority regions of various decision rules for two outcome variables ($K = 2$). The Any rule is a combination of the two Single rules. The Compensatory rule reflects $\mathbf{w} = (0.5, 0.5)$.

3. *All rule:* all outcomes have a treatment difference larger than zero. Similar to the Any rule, the superiority region is a combination of K superiority regions of the Single rule: The superiority region is denoted by:

$$\mathcal{S}_{All} = \{\mathcal{S}_{Single_1} \cap \cdots \cap \mathcal{S}_{Single_K}\}.$$

Superiority is concluded when

$$\min_k P(\boldsymbol{\delta} \in \mathcal{S}_{Single(k)} | \mathbf{s}_E, \mathbf{s}_C) > p_{cut}. \tag{3.10}$$

Next to facilitating these common decision rules, our framework allows for a Compensatory decision rule:

4. *Compensatory rule:* the weighted sum of treatment differences is larger than zero. The superiority region is denoted by:

$$\mathcal{S}_{Compensatory}(\mathbf{w}) = \{\boldsymbol{\delta} | \sum_{k=1}^{K} w_k \delta_k > 0\} \tag{3.11}$$

where $\mathbf{w} = (w_1, \dots, w_K)$ reflect the weights for outcomes $1, \dots, K$, $0 \leq w_k \leq 1$ and $\sum_{k=1}^{K} w_k = 1$.

Superiority is then concluded when:

$$P(\boldsymbol{\delta} \in \mathcal{S}_{Compensatory}(\mathbf{w}) | \mathbf{s}_E, \mathbf{s}_C) > p_{cut}. \tag{3.12}$$

Figure 3.2 visualizes these four decision rules.

From our discussion of the different decision rules, a number of relationships between them can be identified. First, mathematically the Single rule can be considered a special case of the Compensatory rule with weight $w_k = 1$ for primary outcome k and $w_l = 0$ for all other outcomes. Second, the superiority region of the All rule is a subset of the superiority regions of the other rules, i.e.,

$$\mathcal{S}_{All} \subset \mathcal{S}_{Single}, \mathcal{S}_{Compensatory}, \mathcal{S}_{Any}. \tag{3.13}$$

The Single rule is in turn a subset of the superiority region of the Any rule, such that

$$\mathcal{S}_{Single} \subset \mathcal{S}_{Any}. \tag{3.14}$$

These properties can be observed in Figure 3.2 and translate directly to the amount of evidence provided by data $\mathbf{s}_E$ and $\mathbf{s}_C$. The posterior probability of the All rule is always smallest, while the posterior probability of the Any rule is at least as large as the posterior probability of the Single rule:

$$P(\mathcal{S}_{Any}|\mathbf{s}_E, \mathbf{s}_C) \geq P(\mathcal{S}_{Single}|\mathbf{s}_E, \mathbf{s}_C) > P(\mathcal{S}_{All}|\mathbf{s}_E, \mathbf{s}_C) \tag{3.15}$$
$$P(\mathcal{S}_{Compensatory}|\mathbf{s}_E, \mathbf{s}_C) > P(\mathcal{S}_{All}|\mathbf{s}_E, \mathbf{s}_C).$$

The ordering of the posterior probabilities of different decision rules (Equation 3.15) implies that superiority decisions are most conservative under the All rule and most liberal under the Any rule. In practice, this difference has two consequences. First, to properly control Type I error probabilities for these different decision rules, one needs to set a larger decision threshold p_{cut} for the Any rule than for the All rule. Second, the All rule typically requires the largest sample size to obtain sufficient evidence for a superiority decision.

Additionally, the correlation between treatment differences, ρ_{δ_k,δ_l}, influences the posterior probability to conclude superiority. The correlation influences the overlap with the superiority region, as visualized in Figure 3.3. Consequently, the Single rule is not sensitive to the correlation. A negative correlation requires a smaller sample size than a positive correlation under the Any and Compensatory rules, and vice versa for the All rule.

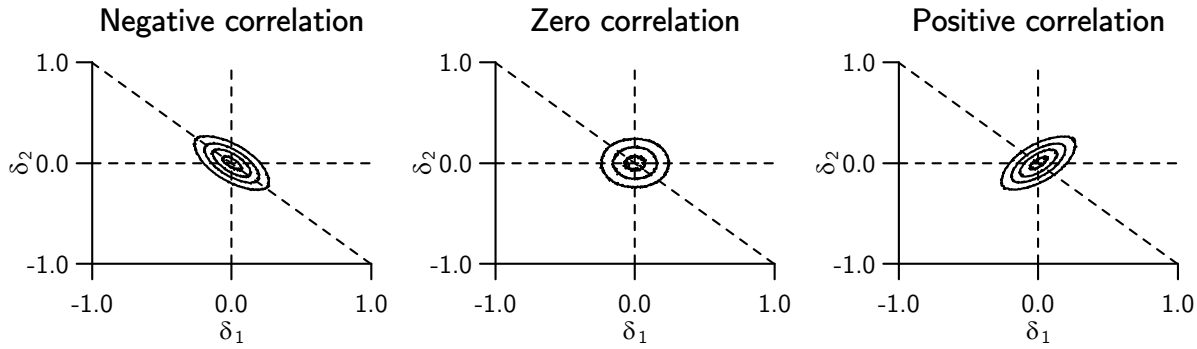


Figure 3.3: Influence of the correlation between two treatment differences on the proportion of overlap between the bivariate distribution of treatment differences $\boldsymbol{\delta}$ and the superiority regions.

3.3.2 Specification of weights of the Compensatory decision rule

To utilize the flexibility of the Compensatory rule, researchers may wish to specify weights $\mathbf{w}$. The current subsection discusses two ways to choose these weights: Specification can be based on the impact of outcome variables or on efficiency of the decision.

Specification of impact weights is guided by substantive considerations to reflect the relative importance of outcomes. When $\mathbf{w} = (\frac{1}{K}, \dots, \frac{1}{K})$, all outcomes are equally important and all success probabilities in $\boldsymbol{\theta}_j$ exert an identical influence on the weighted success probability. Any other specification of $\mathbf{w}$ that satisfies $\sum_{k=1}^{K} w_k = 1$ implies unequal importance of outcomes. To make the implications of importance weight specification more concrete, let us reconsider the two potential side effects of brain cancer treatment in CAR-Study B: cognitive functioning and fatigue (Schimmel et al., 2018). When setting $(w_{cognition}, w_{fatigue}) = (0.50, 0.50)$, both outcomes would be considered equally important and a decrease of (say) 0.10 in fatigue could be compensated by an increase on cognitive functioning of at least 0.10. When $w_{cognition} > 0.50$, cognitive functioning is more influential than fatigue; and vice versa when $w_{cognition} < 0.50$. If $w_{cognition} = 0.75$ and $w_{fatigue} = 0.25$ for example, the treatment difference of cognitive functioning has three times as much impact on the decision as the treatment difference of fatigue.

Efficiency weights are specified with the aim of optimizing the required sample size. As the weights directly affect the amount of evidence for a treatment difference, the efficiency of the Compensatory decision rule can be optimized with values of $\mathbf{w}$ that are a priori expected to maximize the probability of falling in the superiority region. This strategy could be used when efficiency is of major concern, while researchers do not have a strong preference for the substantive priority of specific outcomes. The technical details required to find efficient weights are presented in Appendix B *Specification of efficiency weights*.

3.3.3 Implementation of the framework

The procedure to arrive at a decision using the multivariate analysis procedure proposed in the previous sections is presented in Algorithm 1 for a design with fixed sample size n_j of treatment

j. We present the algorithm for designs with interim analyses in Algorithm 2 in Appendix C *Implementation of the framework in group sequential and adaptive designs*.

Algorithm 1 Decision procedure for a fixed design

1 **Initialize**
  a Choose decision rule
    if Compensatory **then** specify weights $\mathbf{w}$
    if Single **then** specify k
    end if
  for each treatment $j \in \{E, C\}$ **do**
  b Choose prior hyperparameters $\boldsymbol{\alpha}_j^0$
  end for
  c Choose Type I error rate α and power $1 - \beta$
  d Determine decision threshold p_{cut}
    if Any rule **then** $1 - \frac{1}{2}\alpha$
    else $1 - \alpha$
    end if
  e Determine sample size n_j based on anticipated treatment differences $\boldsymbol{\delta}^n$
2 **Collect data and compute evidence**
  for each treatment $j \in \{E, C\}$
  a Collect n_j joint responses $\mathbf{x}_{j,i}$
  b Compute joint response frequencies $\mathbf{s}_j$
  c Compute posterior parameters $\boldsymbol{\alpha}_j^n = \mathbf{s}_j + \boldsymbol{\alpha}_j^0$
  d Sample L posterior draws, $\boldsymbol{\phi}_j^l$, $\boldsymbol{\phi}_j | \boldsymbol{\alpha}_j^n \sim Dirichlet(\boldsymbol{\phi}_j | \boldsymbol{\alpha}_j^n)$
  e Sum draws $\boldsymbol{\phi}_j^l$ to $\boldsymbol{\theta}_j^l$
  end for
  f Transform draws $\boldsymbol{\theta}_j^l$ to $\boldsymbol{\delta}^l$ via $\delta_k^l = \theta_{E,k}^l - \theta_{C,k}^l$
  g Compute posterior probability of treatment superiority $P(\boldsymbol{\delta} \in \mathcal{S}_{Sup} | \mathbf{s}_E, \mathbf{s}_C)$ as the proportion of posterior draws in superiority region $\mathcal{S}_{Sup}$
3 **Make final decision**
  if $P(\boldsymbol{\delta} \in \mathcal{S}_{Sup} | \mathbf{s}_E, \mathbf{s}_C) > P_{cut}$ **then** conclude superiority
  else conclude non-superiority
  end if

3.4 Numerical evaluation

The current section evaluates the performance of the presented multivariate decision framework by means of simulation in the context of two outcomes ($K = 2$). We seek to demonstrate 1) how often the decision procedure results in an (in)correct superiority conclusion to learn

about decision error rates; 2) how many observations are required to conclude superiority with satisfactory error rates to investigate the efficiency of different decision rules, and 3) how well the average estimated treatment difference corresponds to the true treatment difference to examine bias. The current section is structured as follows. We first introduce the simulation conditions, the procedure to compute sample sizes for each of these conditions, and the procedure to generate and evaluate data. We then discuss the results of the simulation.

Table 3.1: Data generating mechanisms (DGM) used in numerical evaluation of the framework.

DGM	δ_1^T	δ_2^T	$\rho^T_{\theta_{j,1},\theta_{j,2}}$	$\theta^T_{E,1}$	$\theta^T_{E,2}$	$\phi^T_{E,11}$	$\theta^T_{C,1}$	$\theta^T_{C,2}$	$\phi^T_{C,11}$
1.1	-0.20	-0.20	-0.30	0.40	0.40	0.09	0.60	0.60	0.29
1.2			0.00			0.16			0.36
1.3			0.30			0.23			0.43
2.1	0.00	0.00	-0.30	0.50	0.50	0.17	0.50	0.50	0.17
2.2			0.00			0.25			0.25
2.3			0.30			0.32			0.32
3.1	0.10	0.10	-0.30	0.55	0.55	0.23	0.45	0.45	0.13
3.2			0.00			0.30			0.20
3.3			0.30			0.38			0.28
4.1	0.20	0.20	-0.30	0.60	0.60	0.29	0.40	0.40	0.09
4.2			0.00			0.36			0.16
4.3			0.30			0.43			0.23
5.1	0.40	0.40	-0.30	0.70	0.70	0.43	0.30	0.30	0.03
5.2			0.00			0.49			0.09
5.3			0.30			0.55			0.15
6.1	0.40	0.00	-0.30	0.70	0.50	0.28	0.30	0.50	0.08
6.2			0.00			0.35			0.15
6.3			0.30			0.42			0.22
7.1	0.20	-0.40	-0.30	0.60	0.30	0.11	0.40	0.70	0.21
7.2			0.00			0.18			0.28
7.3			0.30			0.25			0.35
8.1	0.24	0.08	-0.30	0.62	0.54	0.26	0.38	0.46	0.10
8.2			0.00			0.33			0.17
8.3			0.30			0.41			0.25

Conditions The performance of the framework is examined as a function of the following factors:

1. *Data generating mechanisms:* We generated data of eight treatment difference combinations $\boldsymbol{\delta}^T$ and three correlations between outcomes $\rho_{\theta_{j,1},\theta_{j,2}}$. An overview of these $8 \times 3 = 24$ data generating mechanisms is given in Table 3.1. In the remainder of this section, we refer to these data generating mechanisms with numbered combinations (e.g., 1.2), where the first number reflects treatment difference $\boldsymbol{\delta}^T$ and the second number refers to correlation $\rho^T_{\theta_{j,1},\theta_{j,2}}$.
2. *Decision rules:* The generated data were evaluated with six different decision rules. We used the Single (for outcome $k = 1$), Any, and All rules, as well as three different Compensatory rules: One with equal weights $\mathbf{w} = (0.50, 0.50)$ and two with unequal weights $\mathbf{w} = (0.76, 0.24)$ and $\mathbf{w} = (0.64, 0.36)$. The weight combinations of the latter two Compensatory rules optimize the efficiency of data generating mechanisms with uncorrelated (i.e., 8.2) and correlated (i.e., 8.1) treatment differences respectively, following the procedure in Appendix B *Specification of efficiency weights*. We refer to these three Compensatory rules as Compensatory-Equal (C-E), Compensatory-Unequal Uncorrelated (C-UU) and Compensatory-Unequal Correlated (C-UC) respectively.

Sample size computations To properly control Type I error and power, each of the 24×6 conditions requires a specific sample size. These sample sizes n_j are based on anticipated treatment differences $\boldsymbol{\delta}^n$, that corresponded to the true parameters of each data generating mechanism in Table 3.1 (i.e., $\boldsymbol{\delta}^n = \boldsymbol{\delta}^T$ and $\rho^n_{\theta_{j,1},\theta_{j,2}} = \rho^T_{\theta_{j,1},\theta_{j,2}}$). Procedures to compute sample sizes per treatment group for the different decision rules were the following:

1. For the Single rule, we used a two-proportion $z-$test, where we plugged in the anticipated treatment difference on the first outcome variable (i.e δ^n_1).
2. Following Sozu et al. (2010, 2016) we used multivariate normal approximations of correlated binary outcomes for the All and Any rules.
3. For the Compensatory rule, we used a continuous normal approximation with mean

$\sum_{k=1}^{K} w_k \theta_{j,k}$ and variance $\sum_{k=1}^{K} w_k^2 \sigma_{j,k}^2 + 2 \sum\sum_{k<l} w_k w_l \sigma_{j,kl}$. Here, $\sigma_{j,k}^2 = \theta_{j,k}(1 - \theta_{j,k})$ and $\sigma_{j,kl} = \phi_{j,kl} - \theta_{j,k}\theta_{j,l}$.

The computed sample sizes are presented in Table 3.3. Conditions that should not result in superiority were evaluated at sample size $n_j = 1,000$.

Data generation and evaluation Of each data generating mechanism presented in Table 3.1, we generated 5,000 samples of size $2 \times n_j$. These data were combined with a proper uninformative prior distribution with hyperparameters $\boldsymbol{\alpha}_j^0 = (0.01, \ldots, 0.01)$ to satisfy $\boldsymbol{\alpha}_j^n \approx \mathbf{s}_j$, as discussed in Section A model for multivariate analysis of multiple binary outcomes. We aimed for Type I error rate $\alpha = .05$ and power $1 - \beta = .80$, which corresponds to a decision threshold p_{cut} of $1 - \alpha = 0.95$ (Single, Compensatory, All rules) and $1 - \frac{1}{2}\alpha = 0.975$ (Any rule; Marsman & Wagenmakers, 2016; Sozu et al., 2012, 2016). The generated datasets were evaluated using the procedure in steps 2 and 3 of Algorithm 1.

The proportion of samples that concluded superiority reflects Type I error rates (when false) and power (when correct). We assessed the Type I error rate under the data generating mechanism with the least favorable population values of $\boldsymbol{\delta}^T$ under the null hypothesis in frequentist one-sided significance testing. These are values of $\boldsymbol{\delta}^T$ outside $\mathcal{S}_{Sup}$ that are most difficult to distinguish from values of $\boldsymbol{\delta}^T$ inside $\mathcal{S}_{Sup}$. Adequate Type I error rates for the least favorable treatment differences imply that the Type I error rates of all values of $\boldsymbol{\delta}^T$ outside $\mathcal{S}_{Sup}$ are properly controlled. The least favorable values of $\boldsymbol{\delta}^T$ were reflected by treatment difference 2 for the Single, Any, and Compensatory rules, and treatment difference 6 for the All rule. Bias was computed as the difference between the observed treatment difference at sample size n_j and the true treatment difference $\boldsymbol{\delta}^T$.

3.4.1 Results

The proportion of samples that concluded superiority and the required sample size are presented in Tables 3.2 and 3.3 respectively. Type I error rates were properly controlled around $\alpha = .05$ for each decision rule under its least favorable data generating mechanism. The power was

around .80 in all scenarios with true superiority. Moreover, average treatment differences were estimated without bias (smaller than 0.01 in all conditions).

Given these satisfactory error rates, a comparison of sample sizes provides insight in the efficiency of the approach. We remark here that a comparison of sample sizes is only relevant when the decision rules under consideration have a meaningful definition of superiority. Further, in this discussion of results we primarily focus on the newly introduced Compensatory rule in comparison to the other decision rules. The results demonstrate that the Compensatory rule consistently requires fewer observations than the All rule, and often - in particular when treatment differences are equal (i.e., treatment differences $3 - 5$) - than the Any and the Single rule. Similarly, the Any rule consistently requires fewer observations than the All rule and could be considered an attractive option in terms of sample sizes. Note however that the more lenient Any rule may not result in a meaningful decision for all trials, since the rule would also conclude superiority when the treatment has a small positive treatment effect and large negative treatment effect (i.e., treatment difference 7); A scenario that may not be clinically relevant.

The influence of the relation between outcomes is also apparent: Negative correlations require fewer observations than positive correlations. The variation due to the correlation is considerable: The average sample size almost doubles in scenarios with equal treatment differences (e.g., data generating mechanisms 3.1 vs. 3.3 and 4.1 vs. 4.3).

Comparison of the three different Compensatory rules further highlights the influence of weights $\mathbf{w}$ and illustrates that a Compensatory rule is most efficient when weights have been optimized with respect to the treatment differences and the correlation between them. The Compensatory rule with equal weights (C-E) is most efficient when treatment differences on both outcomes are equally large (treatment differences $3 - 5$), while the Compensatory rule with unequal weights for uncorrelated outcomes (C-UU) is most efficient under data generating mechanism 8.2. The Compensatory rule with unequal weights, optimized for negatively correlated outcomes (C-UC) is most efficient in data generating mechanism 8.1. The Compensatory is less efficient than the Single rule in the scenario with an effect on one

outcome only (treatment difference 6). Effectively, in this situation the Single rule is the Compensatory rule with optimal weights for this specific scenario $\mathbf{w} = (1, 0)$. Utilizing the flexibility of the Compensatory rule to tailor weights to anticipated treatment differences and their correlations thus pays off in terms of efficiency.

Note that in practice it may be difficult to accurately estimate treatment differences and correlations in advance. This uncertainty may result in inaccurate sample size estimates, as demonstrated in Appendix D *Numerical evaluation: Comparison of trial designs*. The simulations in this appendix also show that the approach can be implemented in designs with interim analyses as well, which is particularly useful under uncertainty about anticipated treatment differences. Specifically, we demonstrate that 1) both Type I and Type II error rates increase, while efficiency decreases in a fixed design when the anticipated treatment difference does not correspond to the true treatment difference; and 2) designs with interim analyses could compensate for this uncertainty in terms of error rates and efficiency, albeit at the expense of upward bias.

Further, Appendix E *Numerical evaluation: Comparison of prior specifications* shows how prior information influences the properties of decision-making. Informative priors support efficient decision-making when the prior treatment difference corresponds to the treatment difference in the data. In contrast, evidence is influenced by dissimilarity between prior hyperparameters and data, and may either increase or decrease 1) the required sample size; and 2) the average posterior treatment effect, depending on the nature of the non-correspondence.

Table 3.2: P(Conclude superiority) for different data generating mechanisms (DGM) and decision rules. Bold-faced values indicate the conditions with least favorable values.

DGM	Single	Any	All	C-E	C-UU	C-UC
1.1	0.000	0.000	0.000	0.000	0.000	0.000
1.2	0.000	0.000	0.000	0.000	0.000	0.000
1.3	0.000	0.000	0.000	0.000	0.000	0.000
2.1	**0.051**	**0.048**	0.000	0.049	0.052	0.051
2.2	0.046	0.045	0.003	0.056	0.048	0.054
2.3	0.051	0.045	0.008	**0.049**	**0.049**	**0.049**
3.1	0.810	0.796	0.801	0.807	0.804	0.790
3.2	0.799	0.801	0.804	0.806	0.788	0.791
3.3	0.799	0.807	0.809	0.800	0.797	0.803
4.1	0.794	0.784	0.806	0.811	0.789	0.784
4.2	0.808	0.802	0.814	0.813	0.804	0.803
4.3	0.804	0.801	0.816	0.804	0.796	0.800
5.1	0.807	0.806	0.830	0.881	0.817	0.857
5.2	0.807	0.814	0.838	0.831	0.813	0.813
5.3	0.809	0.847	0.822	0.809	0.798	0.802
6.1	0.811	0.779	0.053	0.824	0.798	0.819
6.2	0.813	0.777	0.045	0.805	0.808	0.820
6.3	0.803	0.758	**0.051**	0.801	0.788	0.803
7.1	0.799	0.789	0.000	0.000	0.863	0.002
7.2	0.804	0.792	0.000	0.000	0.857	0.003
7.3	0.807	0.794	0.000	0.000	0.867	0.005
8.1	0.787	0.782	0.789	0.808	0.804	0.805
8.2	0.777	0.797	0.807	0.804	0.799	0.804
8.3	0.785	0.811	0.807	0.805	0.805	0.806

Table 3.3: Average sample size to correctly conclude superiority for different data generating mechanisms (DGM) and decision rules. Bold-faced values indicate the lowest sample size per data generating mechanism. Conditions with a hyphen should not result in treatment superiority.

DGM	Single	Any	All	C-E	C-UU	C-UC
1.1	-	-	-	-	-	-
1.2	-	-	-	-	-	-
1.3	-	-	-	-	-	-
2.1	-	-	-	-	-	-
2.2	-	-	-	-	-	-
2.3	-	-	-	-	-	-
3.1	307	191	424	**108**	157	119
3.2	307	217	418	**154**	192	162
3.3	307	247	406	**199**	226	206
4.1	75	47	105	**26**	39	29
4.2	75	53	103	**38**	47	40
4.3	75	60	101	**49**	55	50
5.1	17	11	25	**6**	9	7
5.2	17	12	25	**9**	11	**9**
5.3	17	14	24	**11**	12	**11**
6.1	17	21	-	25	**15**	17
6.2	**17**	21	-	36	19	24
6.3	**17**	21	-	47	22	30
7.1	**75**	95	-	-	608	-
7.2	**75**	95	-	-	733	-
7.3	**75**	95	-	-	858	-
8.1	51	56	482	41	38	**36**
8.2	51	60	482	59	**46**	49
8.3	**51**	63	482	76	55	62

3.5 Discussion

The current paper presented a Bayesian framework to efficiently combine multiple binary outcomes into a clinically relevant superiority decision. We highlight two characteristics of the approach.

First, the multivariate Bernoulli model has shown to capture relations properly and support multivariate decision-making. The influence of the correlation between outcomes on the amount of evidence in favor of a specific treatment highlights the urgency to carefully consider these relations in trial design and analysis in practice.

Second, multivariate analysis facilitates comprehensive decision rules such as the Compensatory rule. More specific criteria for superiority can be defined to ensure clinical relevance, while relaxing conditions that are not strictly needed for clinical relevance lowers the sample size required for error control; A fact that researchers may take advantage of in practice where sample size limitations are common (Van de Schoot & Miočević, 2020).

Several other modeling procedures have been proposed for the multivariate analysis of multiple binary outcomes. The majority of these alternatives assume a (latent) normally distributed continuous variable. When these models rely on large sample approximations for decision-making (such as methods presented by Whitehead et al. (2010), Sozu et al. (2010, 2016), and Su et al. (2012); see for an exception Murray et al. (2016)), their applicability is limited, since the validity of z-tests for small samples may be inaccurate. A second class of alternatives uses copula models, which is a flexible approach to model dependencies between multiple univariate marginal distributions. The use of copula structures in discrete data can be challenging however (Panagiotelis et al., 2012). Future research might provide insight in the applicability of copula models for multivariate decision making in clinical trials.

Two additional remarks concerning the number of outcomes should be made. First, the modeling procedure becomes more complex when the number of outcomes increases, since the number of cells increases exponentially. Second, the proposed Compensatory rule has a linear compensatory mechanism. With two outcomes, the outcomes compensate each other directly and the size of a negative effect is maximized by the size of the positive effect. A decision

based on more than two outcomes might have the - potentially undesirable - consequence of compensating a single large negative effect by two or more positive effects. Researchers are encouraged to carefully think about a suitable superiority definition and might consider additional restrictions to the Compensatory rule, such as a maximum size of individual negative effects.

Data and code availability

The R code used to generate results in Section *Numerical evaluation*, Appendix D *Numerical evaluation: Comparison of trial designs*, and Appendix E *Numerical evaluation: Comparison of prior specifications* can be found on https://github.com/XynthiaKavelaars/Decision-making-with-multiple-correlated-binary-outcomes-in-clinical-trials.

Chapter 4

Bayesian multivariate logistic regression for superiority and inferiority decision-making under observed treatment heterogeneity

Based on Kavelaars, X., Mulder, J., & Kaptein, M. (2022b). *Bayesian multivariate logistic regression for superiority and inferiority decision-making under observed treatment heterogeneity.* [Submitted for publication].

Abstract

The effects of treatments may differ between persons with different characteristics. Addressing such treatment heterogeneity is crucial to identify who benefits from a new treatment, but can be complex in the context of multiple correlated outcomes. The current paper presents a novel Bayesian method for superiority and inferiority decision-making in the context of randomized controlled trials with multivariate binary responses and heterogeneous treatment effects. The framework is based on three elements: a) Bayesian multivariate logistic regression analysis with Pólya-Gamma expansion; b) a transformation procedure to transfer obtained regression coefficients to the more intuitive multivariate probability scale (i.e., success probabilities and differences between them); and c) a compatible decision procedure for treatment comparison. Procedures for a priori sample size estimation under a non-informative prior distribution are included. A numerical evaluation demonstrated that decisions based on a priori sample size estimation resulted in anticipated error rates among the trial population as well as subpopulations. Further, average and conditional treatment effect parameters could be estimated unbiasedly when the sample was large enough. Illustration with the International Stroke Trial dataset revealed a trend towards heterogeneous effects among stroke patients: Something that would have remained undetected when analyses were limited to average treatment effects.

4.1 Introduction

The current paper focuses on estimating heterogeneous treatment effects based on covariates in the context of two-arm randomized controlled trials (RCTs) with multiple (correlated) binary outcome variables. Such RCTs are randomized experiments with subjects being assigned at random to either an experimental or a control group, often having the objectives a) to evaluate whether an experimental treatment is superior or inferior to a control condition; b) to inform assignment to eligible subjects in practice (Food and Drug Administration, 2016). Although RCTs are broadly applicable to experimental research in general, we focus on the health domain and refer to psychological and medical interventions in the broad sense when using the word treatment. These interventions include - but are not limited to - behavioral therapies, pharmacological support, and other experimental types of care.

These trials often assess multiple types of (clinical) events (e.g., quitting substance abuse, death), functional measures (e.g., memory decline, ability to walk), and disease symptoms (e.g., fatigue, anxiety; Food and Drug Administration, 2017), which can provide multidimensional insights into the effects of a treatment. Including such comprehensive insights can improve correspondence between statistical and clinical decision-making, since multiple effects of the intervention can be combined and weighed in various ways to provide a single statistical decision regarding superiority or inferiority (e.g., Murray et al., 2016; O'Brien, 1984; Pocock et al., 1987). Whereas performing multiple univariate analyses on individual outcomes is a common strategy, a single multivariate analysis takes correlations into account and can be statistically preferable (Food and Drug Administration, 2017; Murray et al., 2016; Ristl et al., 2018; Senn & Bretz, 2007). Multivariate analysis has the potential to reduce decision errors: Correlations influence the sample sizes required for decision-making with prespecified error rates and provoke under- or overpowerment when falsely omitted (Chow et al., 2017; Kavelaars et al., 2020; Sozu et al., 2010; Xiong et al., 2005).

RCTs often focus on average treatment effects (ATEs) among the study population when comparing interventions (Thall, 2020). Average treatment effects can be sufficiently

insightful when the effects of a treatment are relatively homogeneous over the trial population. However, average effects may give a limited, or even erroneous, impression when the effects of a treatment are heterogeneous and thus interact with characteristics of patients. In that case, treatment effects conditional on a subpopulation contribute to a better understanding of the treatment's potential and are more informative for clinicians advising treatments to patients with specific characteristics. Despite efforts to provide statistical methodology to identify heterogeneous treatment effects (e.g., Jones et al., 2011; Wang et al., 2015; Yang et al., 2021), investigating these effects is not the standard yet: Thall notes that "the great majority of clinical trial designs ignore the possibility of treatment-covariate interactions, and often ignore patient heterogeneity entirely" (Thall, 2020, p.1). This is unfortunate as addressing potential treatment heterogeneity in the evaluation of treatments is crucial to a) identify which patients are likely to benefit from a treatment; and b) optimize treatment results of individual patients via personalized treatment assignment (Goldberger & Buxton, 2013; Hamburg & Collins, 2010; Simon, 2010; Wang et al., 2015). In sum, statistical analysis based on the combination of multiple outcome variables and treatment heterogeneity has the potential to reveal different outcome patterns for different patient profiles, thereby contributing to the personalization of treatment assignment.

An example of a trial with multiple outcomes and potential treatment heterogeneity is the International Stroke Trial (IST; International Stroke Trial Collaborative Group, 1997; Sandercock et al., 2011). Strokes may have far-reaching implications for the quality of life, as they may be recurring and/or lead to long-term impaired (daily) functioning. The IST investigated whether the short-term and long-term perspective of stroke patients can be improved with anti-thrombotic drug therapy. The average treatment differences in the IST were small, so one might conclude that treatment with one of these drugs was marginally effective. However, these overall findings did not show how specific characteristics of patients (e.g., sex or age) and/or disease (e.g., type of stroke or functional status after stroke) potentially interacted with the treatment to produce different perspectives for patients with

different profiles. Average treatment effects do, for example, not reveal whether older patients have better prospects in terms of short-term damage risk and/or long-term recovery potential than younger patients. Clearly, potentially heterogeneous effects such as these would have clinically and psychologically relevant implications and advocate the development of more personalized treatment policies.

Although theoretically relevant in many contemporary RCTs, decision-making under treatment heterogeneity in the multivariate context is considerably more complex compared to the non-heterogeneous and/or univariate setting. Generalizations are subject to assumptions that need to be carefully evaluated in light of the research problem at hand. First, the multivariate setting demands an analysis method that incorporates the correlation between outcome variables (i.e., a multivariate analysis method) to obtain accurate decision error rates (e.g., Kavelaars et al., 2020). For accurate inference regarding conditional treatment effects, the analysis should not only include the overall correlation among the trial population, but should also be flexible enough to deal with correlations that differ over subpopulations. The latter is not evident in existing multivariate analysis methods for binary outcome variables: Some methods impose the marginal correlation structure of the trial population on subpopulations (e.g., multivariate probit models by Chib (1995) or Rossi et al. (2005) and multivariate logit models by Malik and Abraham (1973) and O'Brien and Dunson (2004)). Second, the interpretation of treatment effects can be complex in multivariate non-linear models. Creating insights into so-called marginal effects is strongly recommended in treatment comparison, demanding any multivariate method to return interpretable univariate effects (Food and Drug Administration, 2017; O'Brien & Dunson, 2004). Several existing multivariate models lack insight into marginal distributions (e.g., Malik & Abraham, 1973). Third, multivariate methods may estimate a single regression parameter to capture the relation between a covariate and all outcome variables (e.g., O'Brien & Dunson, 2004; Rossi et al., 2005). The latter assumes that all outcome variables vary identically over the full support of the covariate: An assumption that may be too strict to hold in practice.

As a more flexible alternative to capture the complexity of heterogeneous, multivariate

treatment effects, we build upon an existing Bayesian multivariate Bernoulli framework for superiority decision-making proposed by Kavelaars et al. (2020). The existing procedure consists of three major components: a) a conjugate multivariate Bernoulli model to estimate unknown (regression) parameters; b) a transformation procedure to interpret treatment effects on the (more intuitive) probability scale; and c) a compatible decision procedure to make inferences regarding treatment superiority. The multivariate Bernoulli as an underlying model has advantages over several other approaches, as it relies on a multinomial distribution and has the flexibility to allow univariate effects, correlations between outcomes and multivariate effects to vary with covariates. Although joint response probabilities can provide useful insights, the transformation procedure facilitates the interpretation of treatment comparison: marginal (i.e., univariate) probabilities, multivariate probabilities, and differences between (multivariate) probabilities can be used in inference as well.

The framework is suitable for estimation and inference among the trial population (i.e., ATEs), but does not incorporate patient characteristics to model heterogeneous treatment effects directly. Therefore, we expand the framework with a Bayesian multivariate logistic regression analysis to incorporate potential treatment heterogeneity via the inclusion of covariates, aiming to facilitate treatment comparison among subpopulations and contribute to personalized treatment assignment. The proposed modeling procedure relies on multinomial logistic regression and can model treatment effects and correlations on a subpopulation level and is suitable for estimation and inference among other populations than the trial population. The transformation procedure is essential in this extension, as the model produces multinomial regression coefficients, which have no straightforward interpretation in the context of (multivariate) treatment comparison. Along with the regression model, we include a procedure to compute sample sizes for decision-making with prespecified frequentist error rates.

The paper is organized as follows. In the next section, we introduce the decision framework, including the multivariate logistic regression model to obtain a sample from the multivariate posterior distribution of regression coefficients, a transformation procedure to

find posterior treatment differences, and a decision procedure to draw conclusions regarding treatment superiority and inferiority. The section on capturing heterogeneity explains how the framework can be applied to different patient populations. We evaluate frequentist operating characteristics of the framework via simulation in the numerical evaluation section. Next, we illustrate the methods with data from the International Stroke Trial and conclude the paper with a discussion.

4.2 Decision-framework

4.2.1 Multivariate logistic regression

Response y_i^k is the binary response for subject i on outcome variable $k \in \{1, \dots, K\}$, where $y_i^k \in \{0, 1\}$, $0 =$ failure and $1 =$ success. Vector $\mathbf{y}_i = (y_i^1, \dots, y_i^K)$ is the multivariate (or joint) binary response vector of subject i on K outcomes and has configuration $\mathbf{H}_{q\cdot}$, which is one of the $Q = 2^K$ possible response combinations of length K given in the qth row of matrix $\mathbf{H}$:

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & \dots & 1 & 1 \\ 1 & 1 & \dots & 1 & 0 \\ & & \dots & & \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix} \tag{4.1}$$

The probability of $\mathbf{y}_i$ can be expressed in two meaningful and related ways. First, $\boldsymbol{\theta}_i = (\theta_i^1, \dots, \theta_i^K)$ denotes the vector of K-variate success probabilities on individual outcome $1, \dots, K$, where $\theta_i^k = p(y_i^k = 1)$. Second, $\boldsymbol{\phi}_i = (\phi_i^1, \dots, \phi_i^Q)$ denotes the vector of Q-variate joint response probabilities, where $\phi_i^q = p(\mathbf{y}_i = \mathbf{H}_{q\cdot})$ and sums to unity. The joint response of subject i can be conditioned on covariates in vector $\mathbf{x}_i = (x_{i1}, \dots, x_{iP})$. In this case, the probabilities of response vector $\mathbf{y}_i|\mathbf{x}_i$ are expressed as functions of $\mathbf{x}_i$, namely $\boldsymbol{\phi}_i(\mathbf{x}_i)$ and $\boldsymbol{\theta}_i(\mathbf{x}_i)$.

Joint response probability $\phi_i^q(\mathbf{x}_i)$ maps the dependency of joint response probabilities on

covariates $\mathbf{x}_i$ via a multinomial logistic function:

$$\phi_i^q(\mathbf{x}_i) = \frac{\exp\left[\psi_i^q(\mathbf{x}_i)\right]}{\sum_{r=1}^{Q-1} \exp\left[\psi_i^r(\mathbf{x}_i)\right] + 1} \tag{4.2}$$

for response categories $q = 1, \dots, Q-1$. In Equation 4.2, $\psi_i^q(\mathbf{x}_i)$ reflects the linear predictor of response category q and subject i:

$$\psi_i^q(\mathbf{x}_i) = \beta_0^q + \beta_1^q x_{i1} + \cdots + \beta_P^q x_{iP}. \tag{4.3}$$

Here, x_{ip} can be a treatment indicator, a patient characteristic, or an interaction between these. Vector $\boldsymbol{\beta}^q = (\beta_0^q, \beta_1^q, \dots, \beta_P^q)$ is the vector of regression coefficients of response category q. To ensure identifiability, all regression coefficients of response category Q are fixed at zero, i.e., $\boldsymbol{\beta}^Q = \mathbf{0}$.

The likelihood of response data follows from taking the product over n individual joint response probabilities from Equation 4.2 of Q response categories:

$$l(\mathbf{y}|\boldsymbol{\beta}, \mathbf{x}) = \prod_{i=1}^{n} \prod_{q=1}^{Q-1} \left(\frac{\exp\left[\psi_i^q(\mathbf{x}_i)\right]}{\sum_{r=1}^{Q-1} \exp\left[\psi_i^r(\mathbf{x}_i)\right] + 1} \right)^{I(\mathbf{y}_i = \mathbf{H}_{q\cdot})} \left(\frac{1}{\sum_{r=1}^{Q-1} \exp\left[\psi_i^r(\mathbf{x}_i)\right] + 1} \right)^{I(\mathbf{y}_i = \mathbf{H}_{Q\cdot})}. \tag{4.4}$$

Bayesian analysis is done via the posterior distribution which is given by

$$p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{x}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{x}) p(\boldsymbol{\beta}), \tag{4.5}$$

where $p(\boldsymbol{\beta})$ reflects the prior distribution of the unknown parameters before observing the data. Posterior sampling can be done with a Gibbs sampling algorithm based on a Pólya-Gamma expansion (Polson et al., 2013). Computational details of this procedure can be found in Appendix F.

4.2.2 Transformation to treatment differences

We aim to make the posterior sample of regression coefficients interpretable in terms of a treatment difference, which is defined as the (multivariate) difference between success probabilities of two treatments. To this end, we execute the following multistep procedure with a fictive setup of the IST as running example.

Suppose we are interested in the effect of a combined drug therapy (Heparin plus Aspirin; T_{H+A}) vs. single drug therapy (Aspirin only; T_A) on recurrent stroke on the short-term (y^{strk}) and dependency on the long-term (y^{dep}). There is a total of $Q = 4$ response categories: $\{y^{strk} = 1, y^{dep} = 1\}$, $\{y^{strk} = 1, y^{dep} = 0\}$, $\{y^{strk} = 0, y^{dep} = 1\}$, $\{y^{strk} = 0, y^{dep} = 0\}$, which we refer to as $\{11\}$, $\{10\}$,$\{01\}$, and $\{00\}$ respectively. The treatments are blood thinning agents and may thus interact with the patient's blood pressure. Therefore, we include systolic blood pressure at the time of randomization, resulting in the following model:

$$\psi_i^q(\mathbf{x}_i) = \beta_0^q + \beta_1^q T_i + \beta_2^q bp_i + \beta_2^q bp_i T_i, \tag{4.6}$$

where $\mathbf{x}_i = (T_i, bp_i, bp_i T_i)$. The transformation procedure is then as follows:

1. **Regression coefficients $\boldsymbol{\beta}$ to joint response probabilities $\phi_T(\mathbf{x})$:**

   In the first step, the posterior sample of regression coefficients $\boldsymbol{\beta}$ is transformed into a treatment effect in terms of joint response probabilities $\phi_{Ti}(\mathbf{x}_i)$ for each treatment $T \in \{0, 1\}$. Linear predictor $\psi_i^q(\mathbf{x}_i)$ is then transformed into

$$\phi_i^q(\mathbf{x}_i) = \frac{\exp\left[\psi_i^q(\mathbf{x}_i)\right]}{\sum_{r=1}^{Q-1} \exp\left[\psi_i^r(\mathbf{x}_i)\right] + 1}. \tag{4.2 revisited}$$

   For example, the probability that patient i in the IST does not experience a new stroke

and is dependent after six months can be expressed as:

$$\begin{aligned}\phi^3_{T_i}(\mathbf{x}_i) &= p(\mathbf{y}_i(\mathbf{x}_i) = \{01\}) \\ &= \frac{\exp\left[\psi^3_i(\mathbf{x}_i)\right]}{\sum_{r=1}^{Q-1} \exp\left[\psi^r_i(\mathbf{x}_i)\right] + 1}.\end{aligned} \tag{4.7}$$

Note that we are interested in joint response probability $\boldsymbol{\phi}_T(\mathbf{x})$, which reflects a treatment effect among a population defined by $\mathbf{x}$ and is more general than the joint response probability of an individual patient with covariates $\mathbf{x}_i$. This population can be reflected by an individual patient in some situations, while other cases target the entire study population or a subpopulation of interest. These variations have slightly different computational procedures, which we discuss in more detail in Section 4.3.

2. **Joint response probabilities $\boldsymbol{\phi}_T(\mathbf{x})$ to multivariate success probabilities $\boldsymbol{\theta}_T(\mathbf{x})$:** The next step in the transformation involves the conversion from joint response probabilities $\boldsymbol{\phi}_T(\mathbf{x})$ to multivariate success probabilities of individual outcome variables $\boldsymbol{\theta}_T(\mathbf{x})$. Especially when the number of outcome variables increases, success probabilities are more straightforward in their interpretation than joint response probabilities.

   The relation between both quantities is additive: Success probability θ^k_T on outcome k and treatment T equals the sum of a selection of elements of $\boldsymbol{\phi}_T$, denoted by matrix $\mathbf{U}_k$:

$$\theta^k_T(\mathbf{x}) = \sum_{q=1}^{Q} \phi^q_T(\mathbf{x}) I(\mathbf{H}_{q\cdot} \in \mathbf{U}_k). \tag{4.8}$$

   Selection $\mathbf{U}_k$ consists of the 2^{K-1} rows of $\mathbf{H}$ that have their k^{th} element equal to 1. More concretely, the two outcome variables from the IST are the following combinations, where we drop the dependency on $\mathbf{x}$ for notational simplicity.

$$\mathbf{H} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \ \mathbf{U}_{strk} = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right], \text{ and } \mathbf{U}_{dep} = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right].$$

Hence, the multivariate success probabilities in $\boldsymbol{\theta}_T = (\theta_T^{strk}, \theta_T^{dep})$ consists of univariate success probabilities:

$$\begin{aligned}
\theta_T^{strk} &= p(\mathbf{y}_i(\mathbf{x}_i) = \{11\}) + \ p(\mathbf{y}_i(\mathbf{x}_i) = \{10\}) \\
&= \phi_T^1 + \phi_T^2 \\
\theta_T^{dep} &= p(\mathbf{y}_i(\mathbf{x}_i) = \{11\}) + \ p(\mathbf{y}_i(\mathbf{x}_i) = \{01\}) \\
&= \phi_T^1 + \phi_T^3.
\end{aligned} \tag{4.9}$$

The correlation between these outcome variables is captured in joint response probabilities $\boldsymbol{\phi}_T(\mathbf{x})$ and automatically taken into account in further transformations (Dai et al., 2013; Olkin & Trikalinos, 2015).

3. **Success probabilities $\boldsymbol{\theta}_T(\mathbf{x})$ to treatment differences $\boldsymbol{\delta}(\mathbf{x})$:**

   The treatment difference on outcome k, $\delta^k(\mathbf{x})$, is defined as the difference between the success probabilities of two treatments on outcome k, such that:

$$\delta^k(\mathbf{x}) = \theta_1^k(\mathbf{x}) - \theta_0^k(\mathbf{x}). \tag{4.10}$$

   The K-variate treatment difference is then $\boldsymbol{\delta}(\mathbf{x}) = (\delta^1(\mathbf{x}), \dots, \delta^K(\mathbf{x}))$.

   Multivariate treatment difference $\boldsymbol{\delta} = (\delta^{strk}, \delta^{dep})$ in the IST is a vector of the univariate treatment differences:

$$\begin{aligned}
\delta^{strk} &= \theta_{H+A}^{strk} - \theta_A^{strk} \\
\delta^{dep} &= \theta_{H+A}^{dep} - \theta_A^{dep}.
\end{aligned} \tag{4.11}$$

Applying the three above-mentioned steps to each draw of the posterior sample of $\boldsymbol{\beta}$, results in a posterior sample of multivariate treatment difference $\boldsymbol{\delta}(\mathbf{x})$. This sample provides estimates that can be used for prediction, where various measures of central tendency (e.g., a mean or high posterior density interval) can be used to summarize the sample into a point estimate. Moreover, the sample can be used for statistical inference, as outlined in the next subsection.

4.2.3 Posterior decision-making

Decisions rely on estimated treatment effects and their uncertainties. More formally, multivariate treatment difference $\boldsymbol{\delta}$ has complete parameter space $\mathcal{S} \subset [-1, 1]^K$, which is divided into a rejection region $\mathcal{S}_R$ and a non-rejection region $\mathcal{S}_N$. Rejection region $\mathcal{S}_R$ reflects the part of the parameter space that indicates the treatment difference of interest, while the non-rejection region $\mathcal{S}_N$ refers to the part of the parameter space that would not be considered a (relevant) treatment difference. Rejection regions depend on the type of decision and be composed of multiple subregions if desired (Van Ravenzwaaij et al., 2019). We consider the following three (commonly used) decision types:

1. superiority with region $\mathcal{S}_R \in \mathcal{S}_S$, where the treatment is better;
2. inferiority with region $\mathcal{S}_R \in \mathcal{S}_I$, where the treatment is worse;
3. two-sided with rejection region $\mathcal{S}_R \in \{\mathcal{S}_S, \mathcal{S}_I\}$, where the treatment can be either better or worse.

We would conclude superiority and/or inferiority when the posterior probability that treatment difference $\boldsymbol{\delta}(\mathbf{x})$ lies in the rejection region exceeds a prespecified decision threshold, p_{cut}:

$$p(\boldsymbol{\delta}(\mathbf{x}) \in \mathcal{S}_R | \mathbf{y}) > p_{cut}. \tag{4.12}$$

When the functional form of the posterior distribution is unknown, the rejection probability can be concluded from an MCMC sample of L draws from the posterior distribution of $\boldsymbol{\delta}(\mathbf{x})$. Equation 4.12 is then applied in practice as:

$$\frac{1}{L} \sum_{(l)=1}^{L} I(\boldsymbol{\delta}^{(l)}(\mathbf{x}) \in \mathcal{S}_R | \mathbf{y}) > p_{cut}. \tag{4.13}$$

In a situation with multiple outcome variables, superiority and inferiority can be defined in multiple ways, resulting in different rejection regions (e.g., O'Brien, 1984; Pocock, 1997; Pocock et al., 1987; Prentice, 1997; Tang et al., 1993; Zhao et al., 2007). Although not

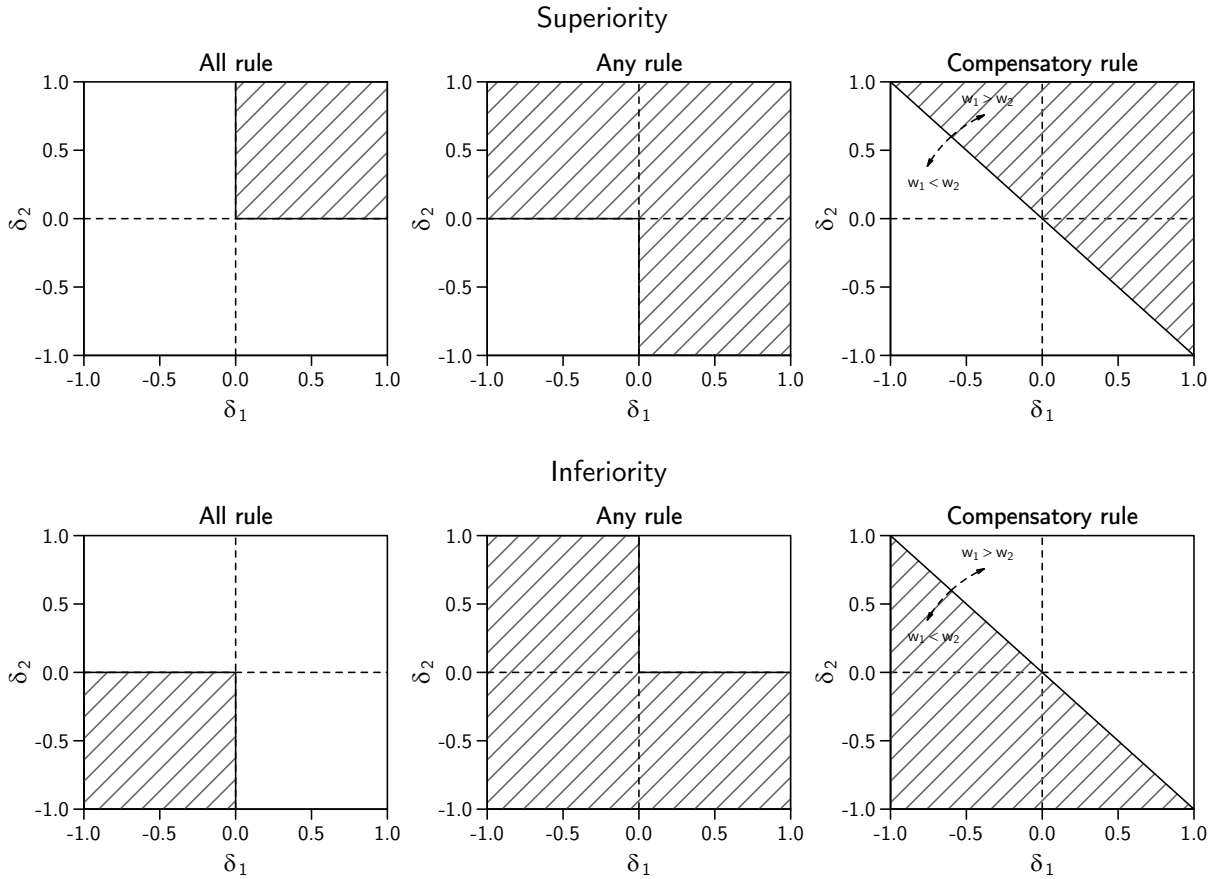


Figure 4.1: Bivariate superiority and inferiority spaces for the All, Any, and Compensatory ($\mathbf{w} = 0.50, 0.50$) rules.

intended as an exhaustive overview, we list three possible rules and graphically present their rejection regions in Figure 4.1. Two of these rules (which we refer to as the "Any" and "All" rules) are presented as part of the regulatory guideline regarding multiple endpoints (Food and Drug Administration, 2017) and have been extensively discussed in literature (e.g., Chuang-Stein et al., 2006; Sozu et al., 2010, 2016; Xiong et al., 2005). The third rule ("Compensatory") is a - relatively unknown - flexible alternative that weighs benefits and risks of treatments by their (clinical) relevance (Kavelaars et al., 2020; Murray et al., 2016).

1. **Any rule:** The Any rule results in superiority or inferiority when the difference between success probabilities is larger or smaller than zero respectively on at least one of the outcome variables (Sozu et al., 2016). The superiority and inferiority spaces are defined

as:

$$\mathcal{S}_S^{Any} = \boldsymbol{\delta}(\mathbf{x})|\max_{1<k<K} \delta^k(\mathbf{x}) > 0 \tag{4.14}$$
$$\mathcal{S}_I^{Any} = \boldsymbol{\delta}(\mathbf{x})|\min_{1<k<K} \delta^k(\mathbf{x}) < 0.$$

2. **All rule:** The All rule results in superiority or inferiority when the difference between success probabilities is larger or smaller than zero respectively on all the outcome variables (Sozu et al., 2010). The superiority and inferiority spaces are defined as:

$$\mathcal{S}_S^{All} = \boldsymbol{\delta}(\mathbf{x})|\min_{1<k<K} \delta^k(\mathbf{x}) > 0 \tag{4.15}$$
$$\mathcal{S}_I^{All} = \boldsymbol{\delta}(\mathbf{x})|\max_{1<k<K} \delta^k(\mathbf{x}) < 0.$$

3. **Compensatory rule:** The Compensatory rule results in superiority or inferiority when the weighted difference between success probabilities is larger or smaller than zero respectively. The superiority and inferiority spaces are defined as:

$$\mathcal{S}_S^{Comp}(\mathbf{w}) = \boldsymbol{\delta}(\mathbf{x})|\delta(\mathbf{w},\mathbf{x}) > 0 \tag{4.16}$$
$$\mathcal{S}_I^{Comp}(\mathbf{w}) = \boldsymbol{\delta}(\mathbf{x})|\delta(\mathbf{w},\mathbf{x}) < 0$$

where $\mathbf{w} = (w^1, \dots, w^K)$ reflect weights of K treatment differences, $\delta(\mathbf{w},\mathbf{x}) = \sum_{k=1}^{K} w^k \delta^k(\mathbf{x})$, $0 \leq w^k \leq 1$ and $\sum_{k=1}^{K} w^k = 1$ (Kavelaars et al., 2020).

4.2.4 Sample size computations

To control decision error rates, methods for a priori sample size estimation are available for variables that follow a multivariate Bernoulli distribution and are eligible for large sample approximation by a (multivariate) normally distributed latent variable (Chow et al., 2017; Sozu et al., 2010, 2016). When combined with a non-informative prior distribution, these procedures have shown to accurately control Type I rate α and Type II error rate β in a

Bayesian multivariate Bernoulli - Dirichlet-model on multivariate response data (Kavelaars et al., 2020). Each of the presented decision rules in Subsection 4.2.3 has an individual procedure to compute sample sizes, as discussed below. These equations provide insight in the required number of observations in absence of prior information and in the influence of the correlation on the sample size. They also allow for verification that correlated outcome variables might result in smaller sample sizes than uncorrelated outcome variables under some conditions detailed in publications by Food and Drug Administration (2017) and Kavelaars et al. (2020). For notational simplicity, we discard the dependence on $\mathbf{x}$ in the remainder of this subsection.

All and Any rules

Sample size computations for the All and Any rules were formulated in Sozu et al. (2010) and Sozu et al. (2016) respectively and rely on the assumption of a multivariate normal latent variable. The power, $1-\beta$, can be expressed in terms of a cumulative K-variate normal distribution $\mathbf{\Psi}_K$ with mean $\mathbf{0}$ and correlation matrix $\mathbf{\Sigma}$ (Sozu et al., 2016):

$$1-\beta = \mathbf{\Psi}_K(c^1, \dots, c^K). \tag{4.17}$$

In Equation 4.17, c^k for outcome k is defined by the decision rule of interest. Further, the off-diagonal elements of $\mathbf{\Sigma}$ denote (estimated) pairwise correlations between outcome variables. For the Any rule,

$$c^k = z_{(1-\frac{\alpha}{K})} - \frac{(\theta_1^k - \theta_0^k)}{\sqrt{\frac{\theta_1^k(1-\theta_1^k)+\theta_0^k(1-\theta_0^k)}{n}}}. \tag{4.18}$$

For the All rule,

$$c^k = -\, z_{(1-\alpha)} + \frac{(\theta_1^k - \theta_0^k)}{\sqrt{\frac{\theta_1^k(1-\theta_1^k)+\theta_0^k(1-\theta_0^k)}{n}}}. \tag{4.19}$$

In Equations 4.18 and 4.19, n is the sample size per treatment and $z_{(.)}$ refers to the selected $(1-\frac{\alpha}{K})$ or $(1-\alpha)$ quantile from the univariate normal distribution.

Since the cumulative multivariate normal distribution does not have a closed-form, the sample size that satisfies targeted decision error rates can be found via the following iterative procedure proposed by Sozu et al. (2010):

1. Plug in estimates of θ_T^k in Equation 4.18 or 4.19.
2. Plug in a starting value for n in Equation 4.18 or 4.19 and calculate the power via Equation 4.17.
3. Repeat step 2 with gradually increasing n until the power exceeds the desired level
4. Select n as the sample size per treatment group

Compensatory rule

Sample sizes for the compensatory rule can be computed using standard methodology for large sample tests with two binomial proportions (Chow et al., 2017, Chapter 4). Plugging in estimates of weighted success probabilities per treatment T, $\theta_T^{\mathbf{w}}$, results in:

$$n = \left[\theta_1^{\mathbf{w}}\left(1-\theta_1^{\mathbf{w}}\right) + \theta_0^{\mathbf{w}}\left(1-\theta_0^{\mathbf{w}}\right)\right]\left[\frac{z_{1-\alpha}+z_{1-\beta}}{\theta_1^{\mathbf{w}}-\theta_0^{\mathbf{w}}}\right]^2, \tag{4.20}$$

where $\theta_T^{\mathbf{w}} = \sum_{k=1}^{K} w^k \theta_T^k$, and $z_{1-\beta}$ is the $(1-\beta)$ quantile of the univariate normal distribution.

4.3 Capturing treatment heterogeneity

In the proposed framework, treatment heterogeneity can be captured by joint response probabilities that reflect conditional treatment effects and thus depend on the characteristics of a subpopulation of interest. We describe two ways to represent subpopulations: by fixed covariate values or by a prespecified interval of the covariate distribution(s). Both representations have their own applications. Specific values of covariates may be relevant when we wish to investigate treatment effects based on individual patients or on patient populations that can be accurately represented by a single number of the covariate (such as a mean or a level of a discrete variable). Intervals of covariate distributions may be sensible

in particular when multiple consecutive covariate values are sufficiently exchangeable to estimate a marginal treatment effect among a population specified by this range. Although such intervals can be specified for discrete covariates as well, their use is particularly reasonable with continuous covariates, as intervals are inherently consistent with the idea of continuity.

We will discuss procedures for fixed values as well as intervals in more detail in the remainder of this subsection. In these discussions, we use a linear predictor $\boldsymbol{\psi}_l^q(\mathbf{x})$ (cf. Equation 4.3) that distinguishes between treatments via a treatment indicator and allows for interaction between the treatment and a covariate. For such a model that includes a single population characteristic x, $\mathbf{x} = (z, T, zT)$ and $\psi_T^q(\mathbf{x})$ is defined as:

$$\psi_T^q(\mathbf{x}) = \beta_0^q + \beta_1^q T + \beta_2^q z + \beta_3^q zT. \tag{4.21}$$

4.3.1 Fixed values of covariate

For a patient population with fixed values of patient covariates, a posterior sample of joint response probabilities $\boldsymbol{\phi}_T(\mathbf{x})$ can be found by plugging in a vector of fixed covariate values $\mathbf{x}$ in linear predictor $\boldsymbol{\psi}_T^{(l)}(\mathbf{x})$. Subsequently applying the multinomial logistic link function in Equation 4.2 to each $\boldsymbol{\psi}_T^{(l)}(\mathbf{x})$ results in joint response probability $\boldsymbol{\phi}_T^{(l)}(\mathbf{x})$ for treatment T. Applying these steps each posterior draw (l) of regression coefficients $\boldsymbol{\beta}^{(l)}$ results in a sample of posterior joint response probabilities. The procedure is presented in Algorithm 3 in Appendix H.

4.3.2 Marginalization over a distribution of covariates

When the population is characterized by a range of covariates, the treatment effect can be marginalized over the interval under consideration, based on available information regarding the distribution of the covariate. A sample of covariate data can be used as input for marginalization. Empirical marginalization involves repeating the fixed values procedure for each subject in the sample to obtain a sample of joint response probabilities for each

posterior draw (l). Averaging the resulting sample of joint response probabilities per treatment results in a marginal joint response probability $\phi_T^{(l)}(\mathbf{x})$ for draw (l). The procedure is presented in Algorithm 4 in the online supplemental materials. Empirical marginalization is computationally efficient for patient populations defined by intervals of more than one continuous covariate. Note however that the procedure is prone to sampling variability in $\mathbf{x}$ and that estimation might depend on the availability of cases with the selected covariate values. Increasing the specificity of subpopulations - often resulting from a higher number of included covariates and/or a limited interval size - will reduce the number of available observations eligible for inclusion[1].

4.4 Numerical evaluation

The current section presents an evaluation of the performance of the proposed multivariate logistic regression procedure. The goal of the evaluation was threefold, and we aimed to demonstrate:

1. how well the obtained regression coefficients and treatment effects correspond to their true values to examine bias;
2. how often the decision procedure results in an (in)correct superiority or inferiority conclusion to learn about decision error rates;
3. how the model performs under a priori sample size estimation to explore the number of required subjects.

4.4.1 Setup

Conditions

The performance of the framework was evaluated in a treatment comparison based on two outcome variables and one covariate. We varied the procedure to compute conditional

[1]If this is the case, (numerical) integration can be an alternative to interpolate the conditional treatment effect distribution of interest.

treatment effects, the effect size, the (sub)population of interest, the procedure to compute the posterior distribution, and the decision rule. Each of these factors will be discussed in the following paragraphs.

Procedure to estimate joint response probabilities We used the two regression-based procedures from Section 4.3 to find the posterior samples of joint response probabilities for two populations of interest defined by:

1. **Fixed covariate values**
2. **Empirical marginalization**

And included a reference approach based on stratification to compare the performance of stratified and regression-based analysis:

3. **Unconditional multivariate Bernoulli - Dirichlet model**
   We used the unconditional multivariate Bernoulli model in (Kavelaars et al., 2020). This model relies on response data and can be used via stratification in the estimation of conditional treatment effects. Samples of treatment-specific joint response probabilities $\boldsymbol{\phi}_T$ could be drawn directly from a posterior Dirichlet distribution with parameters $\boldsymbol{\alpha}_T^n = \boldsymbol{\alpha}^0 + \{\sum_{i=1}^{n} I(T_i = T) I(\mathbf{y}_i = \mathbf{H}_{q\cdot})\}_{q=1}^{Q}$, where $\boldsymbol{\alpha}^0$ is a vector of Q prior hyperparameters.

Effect size We included four treatment differences that varied the heterogeneity of treatment differences:

1. **Conditions 1.1 & 1.2:** A homogeneous treatment effect, with average and conditional treatment differences of zero. This scenario aims to demonstrate the Type I error rate under a least favorable treatment difference for the Any and Compensatory rules in the trial as well as the subpopulation.
2. **Conditions 2.1 & 2.2:** A heterogeneous treatment effect, with an average treatment difference of zero and a conditional treatment effect larger than zero.
3. **Conditions 3.1 & 3.2:** A heterogeneous treatment effect, with average and conditional treatment differences larger than zero. The conditional treatment difference is larger

than the average treatment difference. The effect size is chosen to compare power of different methods, when the sample size should not lead to underpowerment for any of the approaches to the estimation of conditional treatment effects.

4. **Conditions 4.1 & 4.2:** A heterogeneous treatment effect on one of the outcomes with both average and conditional treatment differences larger than zero. The conditional treatment difference is smaller than the average treatment effect. The effect size is chosen such that the expected sample size after stratification of the study sample is smaller than the required sample for evaluation of the conditional treatment effect and aims to investigate the statistical power of regression-based methods when stratification leads to underpowered decisions. Further, this effect size reflects the least favorable treatment difference for a right-sided test of the All rule and should result in a Type I error rate equal to the chosen level of α.

For each of these four effect sizes, we varied the measurement level of the covariate and created a model with a binary covariate and a model with a continuous covariate. This resulted in the eight data generating mechanisms (DGMs) presented in Table 4.1.

Table 4.1: Parameters of average treatment effects (treatment differences and correlations between univariate success probabilities) in the trial and conditional treatment effects in a subpopulation, by data-generating mechanism (DGM).

		Average treatment effect			Conditional treatment effect		
DGM	Covariate	(δ_1, δ_2)	$\delta(\mathbf{w})$	ρ_{θ^1,θ^2}	(δ_1, δ_2)	$\delta(\mathbf{w})$	ρ_{θ^1,θ^2}
1.1	Discrete	(0.000, 0.000)	0.000	-0.160	(0.000, 0.000)	0.000	-0.200
1.2	Continuous	(0.000, 0.000)	0.000	-0.163	(0.000, 0.000)	0.000	-0.207
2.1	Discrete	(0.000, 0.000)	0.000	-0.154	(0.250, 0.150)	0.200	-0.200
2.2	Continuous	(0.000, 0.000)	0.000	-0.157	(0.116, 0.069)	0.092	-0.206
3.1	Discrete	(0.150, 0.050)	0.100	-0.124	(0.400, 0.300)	0.350	-0.200
3.2	Continuous	(0.151, 0.050)	0.101	-0.131	(0.276, 0.169)	0.223	-0.210
4.1	Discrete	(0.400, 0.000)	0.200	-0.194	(0.200, 0.000)	0.100	-0.200
4.2	Continuous	(0.401, 0.000)	0.200	-0.194	(0.323, 0.000)	0.162	-0.205

Patient (sub)population We aimed to assess the treatment difference in two different types of patient populations:

1. **Trial population:**

We assessed the average treatment effect among the trial population. The binary covariate was binomially distributed with a probability of 0.50, while the continuous covariate in the trial population followed a standard normal distribution.

2. **Subpopulation:**

   We assessed the conditional treatment effect (CTE) among patients scoring low on the covariate. The low subpopulation of the binary covariate was described by a value of zero. Note that this subpopulation could not be assigned a range, since subsetting a binary variable inherently results in a single value. Consequently, marginalization reduces to the procedure for fixed covariate values. For the continuous covariate, we specified two different subpopulations. One subpopulation had a value of one standard deviation below the mean, while the other subpopulation was used in the marginalization approaches and defined by a range that entailed all values between the mean and one standard deviation below the mean.

Decision rules and sample size We applied the three decision rules from Subsection 4.4.1:

1. **Any rule**
2. **All rule**
3. **Compensatory rule** with equal weights ($\mathbf{w} = (0.50, 0.50)$)

We computed sample sizes per treatment group via the procedures from Subsection 4.2.4 for conditions with non-zero true average treatment effects. If the true average treatment difference was equal to zero, we used $n = 1,000$ per treatment group. The sample size for the average treatment effect was thus leading for the analysis of both average and conditional treatments. As a result, the power of conditional treatment effects was not targeted at .80, but should exceed this target when the required sample size for a CTE was larger than the sample size for an ATE. Similarly, the power of CTEs with a sample size smaller than the ATE sample size should be lower than .80. The required sample sizes are presented in Table 4.2, where we also included a) the required sample size for the conditional treatment effect in the subpopulation; and b) the sample size after stratification of the trial population. The sample

size after stratification is the expected size in subpopulation analysis of a) response data in the reference approach; and b) covariate data in empirical marginalization.

Table 4.2: Required sample sizes to evaluate the average treatment effect (ATE) and conditional treatment effect (CTE) and expected sample sizes of the subpopulation after stratification (Sub). Bold-faced subsamples are smaller than required for estimation of the CTE.

	Any			All			Compensatory		
DGM	ATE	CTE	Sub	ATE	CTE	Sub	ATE	CTE	Sub
1.1	-	-	1000	-	-	1000	-	-	1000
1.2	-	-	683	-	-	683	-	-	683
2.1	-	45	1000	-	136	1000	-	30	1000
2.2	-	215	683	-	658	683	-	143	683
3.1	154	14	77	1234	34	617	134	9	67
3.2	152	36	52	1219	107	417	131	24	45
4.1	21	93	**11**	-	-	1000	29	122	**15**
4.2	21	33	**8**	-	-	683	29	45	**10**

Procedure

Data generation For each data generating mechanism and each unique (decision-rule specific) sample size, we sampled 1000 datasets. We generated one covariate z and included an interaction between the treatment and the covariate as well, resulting in the following linear predictor ψ_i^q:

$$\psi_i^q(\mathbf{x}_i) = \beta_0^q + \beta_T^q T_i + \beta_1^q z_i + \beta_2^q z_i T_i. \tag{4.22}$$

To generate response data, we first applied the multinomial logistic link function (Equation 4.2) to each true linear predictor $\boldsymbol{\psi}_i(\mathbf{x}_i)$ to obtain joint response probabilities $\boldsymbol{\phi}_i(\mathbf{x}_i)$ for each subject i. Next, we sampled response vector $\mathbf{y}_i|\mathbf{x}_i$ from a multinomial distribution with probabilities $\boldsymbol{\phi}_i(\mathbf{x}_i)$.

Prior distribution For the multivariate logistic regression analysis, we used multivariate normally distributed prior with means $\mathbf{b}^q = \mathbf{0}$ and precision matrix $\mathbf{B}^{0q} = \text{diag}\ (1e^{-2}, \dots, 1e^{-2})$ for all regression coefficients. Prior covariances between regression coefficients were set at zero,

implying that regression coefficients were independent a priori. For the reference approach, we used an improper prior with hyperparameters $\boldsymbol{\alpha}^0 = \mathbf{0}$.

Gibbs sampling The regression coefficients in response categories $1, \ldots, (Q-1)$ were estimated via the Gibbs sampler detailed in the online supplemental materials. We ran two MCMC-chains with $L = 10,000$ iterations plus $1,000$ burn-in iterations. Convergence diagnostics implied that there were no signals of non-convergence when the sample size was large enough. Multivariate Gelman-Rubin convergence diagnostics were below < 1.10 for most of the conditions. We noticed signs of non-convergence (Gelman-Rubin statistic 1.10 to 1.32) in a few datasets generated under mechanisms 4.1 and 4.2 with small sample sizes (i.e., belonging to the Any and Compensatory rules). We generated extra data to replace the datasets with questionable convergence.

Transformation and decision-making We applied the procedures from Subsections 4.2.2 and 4.2.3 to arrive at a decision. In marginalization, we included the selection of subjects that belonged to the subpopulation. We performed a right-sided (superiority) test aiming at a Type I-error rate of $\alpha = .05$. We used a decision threshold $p_{cut} = 1 - \alpha = 0.95$ (Compensatory and All rules) and a for multiple tests corrected $p_{cut} = 1 - \frac{\alpha}{K} = 0.975$ (Any rule) (Kavelaars et al., 2020; Marsman & Wagenmakers, 2016; Sozu et al., 2016).

4.4.2 Results

Bias

Mean estimates of regression coefficients were asymptotically unbiased, implying that bias was negligible ($< .01$) in conditions with a sufficiently large sample. We observed some bias in conditions with smaller samples (DGM 3.1, 3.2, 4.1, and 4.2 under the Any and Compensatory decision rules). Although small-sample bias is a well-documented property of logistic regression in general, we discussed these results in more detail in the online supplemental materials. The bias in regression coefficients was not necessarily problematic

for our actual parameters of interest (success probabilities and differences between them), as transfer to these transformed quantities was not inherent. Even when regression coefficients were slightly biased (DGMs 3.1 and 3.2 under sample sizes of the Any and Compensatory rules), success probabilities and treatment differences could be estimated without bias (absolute bias $< |0.025|$), similar to the conditions without biased regression coefficients. More severe bias of regression coefficients in conditions with smaller sample sizes was not fully corrected in the transformation steps. Treatment effect estimation based on fixed values under DGMs 4.1 and 4.2 resulted in treatment differences with absolute biases up to 0.077 for the Any and Compensatory rules, as shown in Table 4.3. Bias appeared slightly more severe when the covariate was discrete, compared to a continuous covariate. The reference and marginalization approaches could estimate treatment effects without bias, regardless of sample size.

Table 4.3: Comparison of bias in treatment differences by estimation method and decision rule-specific sample size of data generating mechanisms 4.1 and 4.2.

Method	n_{Any} $\boldsymbol{\delta}(\mathbf{x})$	n_{All} $\boldsymbol{\delta}(\mathbf{x})$	$n_{\text{Compensatory}}$ $\delta(\mathbf{w}, \mathbf{x})$
Dgm 4.1 Discrete covariate - Average treatment effect			
Reference	(-0.004, -0.001)	(0.000, 0.000)	0.000
Empirical	(-0.009, -0.004)	(0.000, 0.000)	-0.002
Value	(0.077, -0.026)	(0.001, 0.000)	0.027
Dgm 4.1 Discrete covariate - Conditional treatment effect			
Reference	(-0.002, -0.008)	(-0.001, 0.000)	-0.001
Value	(0.011, -0.002)	(-0.001, 0.000)	0.007
Dgm 4.2 Continuous covariate - Average treatment effect			
Reference	(-0.005, -0.004)	(0.000, 0.000)	-0.002
Empirical	(-0.014, -0.010)	(0.000, 0.000)	-0.007
Value	(0.042, -0.026)	(0.001, 0.000)	0.008
Dgm 4.2 Continuous covariate - Conditional treatment effect			
Reference	(-0.003, -0.008)	(-0.001, 0.000)	-0.005
Empirical	(0.011, -0.013)	(0.000, 0.000)	0.006
Value	(-0.059, 0.005)	(-0.001, 0.000)	-0.010

Decision error rates

Probabilities to conclude superiority of average treatment effects are presented in Table 4.4. Decisions resulted in appropriate Type I error rates around .05 for each of the posterior distribution types under a least favorable scenario of no effect (i.e., DGM 1.1, 1.2, 2.1, 2.2 of Any and Compensatory rules, and 4.1 and 4.2 of the All rule) and the proportions of correct superiority conclusions (i.e., power) were close to the targeted .80. In general, regression-based methods performed comparable to the reference approach. Note that the power of the Compensatory rule in scenario's 4.1 and 4.2 was slightly above .80 in regression-based methods, suggesting that the method was less robust to such small samples compared to the reference approach.

The results of conditional treatment effects in the subpopulations are presented in Table 4.5. Similar to the trial population, Type I error rates were around the targeted .05 under the least favorable scenarios of no effect (DGM 1.1, 1.2 for Any and Compensatory rules) for all estimation methods. The proportion to conclude superiority correctly was above .80 in all scenarios with a sample size exceeding the computed sample size for CTEs (i.e., all DGMs except 4.1 and 4.2). Decisions made with the Any and Compensatory rules in scenarios 4.1 and 4.2 were underpowered due to the use of the ATE sample size, which was smaller than the CTE sample size. A comparison of estimations methods for the continuous covariate revealed that empirical marginalization was generally more powerful than the reference approach. The fixed-values approach could only be compared to the other approaches when the covariate was discrete: In the continuous case, the treatment effect reflected a different (sub)population than empirical marginalization and the reference approach. Here, the reference approach and the fixed value approaches performed similarly in terms of power.

Table 4.4: Proportions of superiority decisions (p) and their standard errors (SE) for ATEs by data-generating mechanism (DGM), estimation method, and decision rule. Bold-faced proportions represent correct rejections (i.e., power).

	Reference		Empirical		Value	
DGM	p	SE	p	SE	p	SE
			Rule = Any			
1.1	0.050	(0.007)	0.058	(0.007)	0.054	(0.007)
1.2	0.044	(0.006)	0.053	(0.007)	0.043	(0.006)
2.1	0.053	(0.007)	0.055	(0.007)	0.052	(0.007)
2.2	0.044	(0.006)	0.049	(0.007)	0.045	(0.007)
3.1	**0.797**	(0.013)	**0.817**	(0.012)	**0.808**	(0.012)
3.2	**0.786**	(0.013)	**0.816**	(0.012)	**0.805**	(0.013)
4.1	**0.770**	(0.013)	**0.815**	(0.012)	**0.842**	(0.012)
4.2	**0.787**	(0.013)	**0.836**	(0.012)	**0.813**	(0.012)
			Rule = All			
1.1	0.001	(0.001)	0.002	(0.001)	0.000	(0.000)
1.2	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)
2.1	0.002	(0.001)	0.002	(0.001)	0.003	(0.002)
2.2	0.003	(0.002)	0.004	(0.002)	0.002	(0.001)
3.1	**0.823**	(0.012)	**0.835**	(0.012)	**0.822**	(0.012)
3.2	**0.788**	(0.013)	**0.799**	(0.013)	**0.813**	(0.012)
4.1	0.048	(0.007)	0.046	(0.007)	0.049	(0.007)
4.2	0.039	(0.006)	0.040	(0.006)	0.041	(0.006)
			Rule = Compensatory			
1.1	0.052	(0.007)	0.056	(0.007)	0.058	(0.007)
1.2	0.045	(0.007)	0.052	(0.007)	0.045	(0.007)
2.1	0.063	(0.008)	0.071	(0.008)	0.055	(0.007)
2.2	0.053	(0.007)	0.065	(0.008)	0.052	(0.007)
3.1	**0.814**	(0.012)	**0.852**	(0.011)	**0.818**	(0.012)
3.2	**0.790**	(0.013)	**0.831**	(0.012)	**0.835**	(0.012)
4.1	**0.819**	(0.012)	**0.842**	(0.012)	**0.865**	(0.011)
4.2	**0.816**	(0.012)	**0.837**	(0.012)	**0.824**	(0.012)

Table 4.5: Proportions of superiority decisions for CTEs (p) and their standard errors (SE) by data-generating mechanism (DGM), estimation method, and decision rule. Bold-faced proportions represent correct rejections (i.e., power).

	Reference		Empirical		Value	
DGM	p	SE	p	SE	p	SE
			Rule = Any			
1.1	0.059	(0.007)			0.064	(0.008)
1.2	0.048	(0.007)	0.060	(0.008)	0.055	(0.007)
2.1	**1.000**	(0.000)			**1.000**	(0.000)
2.2	**1.000**	(0.000)	**1.000**	(0.000)	**1.000**	(0.000)
3.1	**1.000**	(0.000)			**1.000**	(0.000)
3.2	**0.919**	(0.009)	**0.998**	(0.001)	**1.000**	(0.000)
4.1	**0.233**	(0.013)			**0.234**	(0.013)
4.2	**0.355**	(0.015)	**0.542**	(0.016)	**0.175**	(0.012)
			Rule = All			
1.1	0.000	(0.000)			0.000	(0.000)
1.2	0.000	(0.000)	0.001	(0.001)	0.001	(0.001)
2.1	**1.000**	(0.000)			**1.000**	(0.000)
2.2	**0.827**	(0.012)	**0.991**	(0.003)	**1.000**	(0.000)
3.1	**1.000**	(0.000)			**1.000**	(0.000)
3.2	**1.000**	(0.000)	**1.000**	(0.000)	**1.000**	(0.000)
4.1	0.053	(0.007)			0.049	(0.007)
4.2	0.052	(0.007)	0.047	(0.007)	0.049	(0.007)
			Rule = Compensatory			
1.1	0.060	(0.008)			0.057	(0.007)
1.2	0.058	(0.007)	0.053	(0.007)	0.063	(0.008)
2.1	**1.000**	(0.000)			**1.000**	(0.000)
2.2	**1.000**	(0.000)	**1.000**	(0.000)	**1.000**	(0.000)
3.1	**1.000**	(0.000)			**1.000**	(0.000)
3.2	**0.967**	(0.006)	**1.000**	(0.000)	**1.000**	(0.000)
4.1	**0.253**	(0.014)			**0.273**	(0.014)
4.2	**0.380**	(0.015)	**0.589**	(0.016)	**0.231**	(0.013)

4.5 Illustration

We applied the proposed method to a subset of data from the $n = 19,435$ subjects from the International Stroke Trial (International Stroke Trial Collaborative Group, 1997). We selected participants who were alive after six months and were treated with either a combined treatment (Aspirin + medium / high-dose Heparin) or one of the single treatments (Aspirin only), resulting in a sample of $n = 5,657$ participants, of which $n_{H+A} = 1,859$ were in the Heparin + Aspirin group (treatment = 1) and $n_A = 3,798$ subjects were in the Aspirin group (treatment = 0). We fitted the model in Equation 4.6 to compare the effects of the two treatments on a) recurrent stroke within 14 days (0 = no; 1 = yes) and b) dependency after six months (0 = no, 1 = yes) while taking systolic blood pressure of the subjects (Bp) into account.

4.5.1 Method

We applied the two procedures from Subsection 4.3 (fixed values of covariates and empirical marginalization) to assess the multivariate and weighted treatment differences in three different types of patient populations:

1. Average treatment effects in the trial population;
2. Conditional treatment effects in populations defined by a fixed value. Patient populations were defined by six different values of blood pressure, specifically $1, 2,$ and 3 standard deviations below and above the mean.
3. Conditional treatment effects in populations defined by an interval. Patient populations were defined by two different regions of blood pressure: $Bp < -1$ SD (Low), and $Bp > 1$ SD (High).

We specified a diffuse multivariate normally distributed prior with means $\mathbf{b}^q = \mathbf{0}$ and precision matrix $\mathbf{B}^0 = \text{diag}(1e^{-2}, \dots, 1e^{-2})$ for all regression coefficients, except the reference category ($strk = 0, dep = 0$). Prior covariances between regression coefficients were set at zero, implying that regression coefficients were independent a priori. We ran three MCMC-

chains via our proposed Gibbs sampler with $20,000$ iterations plus $10,000$ burn-in iterations. Traceplots showed that chains mixed properly, and the multivariate Gelman-Rubin convergence statistic had a value of 1.000, implying that there were no signals of non-convergence.

We performed two-sided tests for the All, Any, and Compensatory rules. For the Compensatory rule, we assumed that long-term impaired functioning is more important than short-term complications and specified weights $\mathbf{w} = (0.25, 0.75)$ for recurring stroke in 14 days and dependency at 6 months respectively. These weights implied that the long-term outcome was three times more relevant for the decision than the short-term outcome. Since $\boldsymbol{\theta}_T$ reflects failure probabilities rather than success probabilities, the treatment is considered superior when there is sufficient evidence that the treatment difference of interest is *smaller* than zero, while inferiority was concluded when the treatment difference of interest is *larger* than zero. The two-sided test with a targeted Type I-error rate of $\alpha = .05$ was performed with a decision threshold $p_{cut} = 1 - \frac{\alpha}{2} = 0.975$ (Compensatory and All rules) and a for multiple tests corrected $p_{cut} = 1 - \frac{\alpha}{2K} = 0.9875$ (Any rule).

4.5.2 Results

Results are presented in Table 4.6 for different intervals and in Table 4.7 for fixed values of blood pressure. Among the trial population, the regression-based and reference approaches resulted in similar treatment difference estimates and posterior probabilities. Treatment differences were close to zero and each of the decision rules resulted in the conclusion that it does not matter whether Aspirin was administered alone or in combination with Heparin.

These average treatment effects gave a limited impression of the efficacy of Aspirin and Heparin, since a picture of heterogeneous treatment effects emerged when conditional treatment effects among subpopulations were considered separately. As opposed to Aspirin only, the combination of Aspirin and Heparin showed a trend towards higher failure probabilities on both outcome variables for patients with a lower blood pressure, while failure probabilities were generally lower among patients with a higher blood pressure.

A visual comparison of empirical marginalization and stratification of response data (i.e.,

Table 4.6: Average and conditional treatment differences (ATE and CTE respectively) and their posterior probabilities (pp) in the IST data, by range of blood pressure (Bp). Superiority or inferiority was concluded when $>$ or $<$ respectively.

Method	$\boldsymbol{\delta}(Bp)$	pp	Any	All	$\delta(\mathbf{w}, Bp)$	pp	Comp
ATE ($-\infty <$ Bp $< \infty$)			$n_{H+A} = 1859$, $n_A = 3798$				
Reference	(0.005, -0.015)	(0.859, 0.151)	-	-	-0.010	0.182	-
Empirical	(0.004, -0.014)	(0.825, 0.152)	-	-	-0.010	0.178	-
CTE ($-\infty <$ Bp < -1 SD)			$n_{H+A} = 316$, $n_A = 620$				
Reference	(-0.001, 0.066)	(0.459, 0.972)	-	-	0.049	0.970	-
Empirical	(0.012, 0.043)	(0.932, 0.963)	-	-	0.035	0.972	-
CTE ($+1$ SD $<$ Bp $< \infty$)			$n_{H+A} = 290$, $n_A = 646$				
Reference	(-0.009, -0.052)	(0.214, 0.070)	-	-	-0.041	0.063	-
Empirical	(-0.003, -0.081)	(0.330, 0.001)	$>$	-	-0.062	0.001	$>$

the reference approach) resulted in relatively similar estimates and posterior probabilities in the center of the distribution of blood pressure (e.g., between -1 SD and $+1$ SD), but deviated from the regression-based approach in the tails. Point estimates of treatment differences demonstrated a less stable relation between blood pressure and treatment differences after stratification, as shown in Figure 4.2. If the regression approach is flexible enough to properly model the effects over the full support of blood pressure, the different behavior in the tails of the covariate distribution might be explained by the smaller sample size after stratification, as implied by the larger error bars.

Table 4.7: Conditional treatment differences and their posterior probabilities (pp) in the IST data, by range of blood pressure (Bp). Superiority or inferiority was concluded when $>$ or $<$ respectively.

Value	$\boldsymbol{\delta}(Bp)$	pp	Any	All	$\delta(\mathbf{w}, Bp)$	pp	Comp
-3 SD	(0.029, 0.110)	(0.922, 0.994)	$<$	-	0.090	0.996	$<$
-2 SD	(0.017, 0.068)	(0.930, 0.985)	-	-	0.055	0.989	$<$
-1 SD	(0.009, 0.026)	(0.927, 0.908)	-	-	0.022	0.929	-
$+1$ SD	(-0.001, -0.056)	(0.421, 0.002)	$>$	-	-0.042	0.002	$>$
$+2$ SD	(-0.004, -0.097)	(0.294, 0.001)	$>$	-	-0.074	0.001	$>$
$+3$ SD	(-0.007, -0.137)	(0.263, 0.001)	$>$	-	-0.104	0.001	$>$

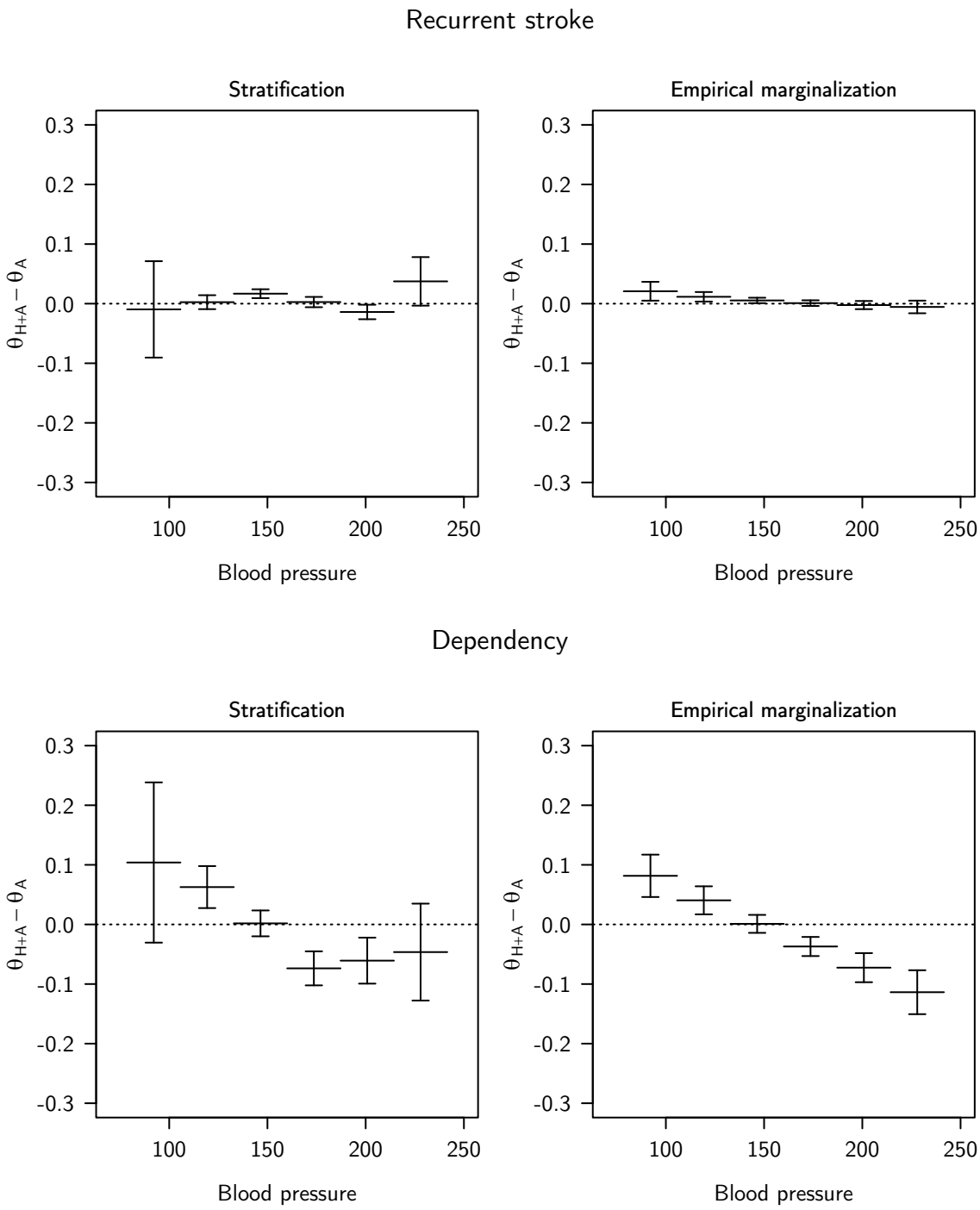


Figure 4.2: Comparison of CTEs and their standard deviations per interval of blood pressure after stratification and empirical marginalization. Each interval reflects one standard deviation.

4.6 Discussion

The current paper proposed a novel multivariate logistic regression framework to identify heterogenous treatment effects on multiple correlated outcome variables. When the sample size was large enough, the proposed regression models were able to reproduce average and conditional treatment differences accurately, and with more robustness against bias than posterior regression coefficients. The model could also make accurate superiority and inferiority decisions among subpopulations, and these decisions were more powerful than those obtained by a stratification approach. Under a priori sample size estimation, anticipated decision error rates were found, when the sample size was not too small. The illustration with the International Stroke Dataset demonstrated how modeling treatment heterogeneity could provide a more in-depth understanding of results beyond average treatment effects.

The model was proposed as an alternative that is flexible enough to model multivariate treatment effects with correlation structures that are free to vary over covariates, supporting accurate decision error rates and a priori sample size computations. This flexibility comes with additional parameters, compared to other multivariate logistic models for correlated binary outcome variables (e.g., Malik & Abraham, 1973; O'Brien & Dunson, 2004) and may result in computational issues when the number of parameters becomes too high. The Gibbs sampling procedure may become unstable when the sample size is too small compared to the number of parameters, although weakly informative priors may be helpful in stabilizing computations (Gelman et al., 2008). Therefore, the model is most suitable for a limited number of outcome variables and covariates.

In practice, researchers are encouraged to consider model assumptions in real data, as highlighted by the illustration with IST data. Additional efforts may be undertaken to verify that the chosen generalized linear model fits the data well enough. If the assumption of linearity on the log-odds scale does not hold, the modeling procedure may benefit from generalization to methods that are more flexible with respect to this assumption, such as (penalized) splines. Again, increased flexibility increases the number of parameters and should be balanced with a)

the general risk of overfitting; and b) computational challenges as outlined above. In a more general sense, the researcher should determine which type of flexibility is most appropriate for the research question and data at hand.

Several directions for future research naturally follow from the current results. First, the procedure theoretically lends itself to out-of-sample prediction to populations within or beyond the covariate range of the trial population. The robustness of the framework in these applications remains to be investigated and may include evaluations of model fit.

Second, research might shed light on further sample size considerations. The presented sample size formulas rely on the size of an estimated treatment effect. Under treatment heterogeneity, average and (multiple) conditional treatment effects have different effect sizes by definition, resulting in different sample sizes and raising the question of which considerations meaningfully guide this choice. Further, in line with our observations, small-sample bias in regression coefficients is a well-documented property of nonlinear regression methods in general (Firth, 1993; Nemes et al., 2009). Although some bias in regression coefficients disappeared during transformation to joint response probabilities, success probabilities, and treatment differences, the mechanism is not yet fully understood. Hence, more light may be shed on circumstances for inheritance of distributional properties in the (non-linear) multinomial logistic transformation to obtain more elaborate insights into the minimum number of observations required for satisfactory model performance. Larger effect sizes (i.e., smaller sample sizes), complexity of the model (i.e., number of parameters), and events per variable are candidate factors to interact in their effects on model performance in small samples (De Jong et al., 2019). There is no short answer to that question, but in practice power among different subpopulations might be balanced with the number of subjects a researcher is willing or able to include in the trial. Therefore, optimum sample sizes in these regression-based decision approaches remain to be investigated more elaborately.

Lastly, causal inference is less straightforward in (stratified) subgroup analysis as conditioning upon covariates might interfere with randomization (European Medicine

Agency, 2019; Food and Drug Administration, 2019). Causal relationships might require additional checking of assumptions and tutorials by Hoogland et al. (2021) and Lipkovich et al. (2016) may be of help.

Acknowledgements

We thank Peter Sandercock on behalf of The International Stroke Trial Collaborative Group for making the data from the International Stroke Trial publicly available.

Data and code availability

The International Stroke Trial data that support the findings of this study are available with the identifier(s) [http://doi.org/10.1186/1745-6215-12-101]. The `R` code used to generate results in the Numerical evaluation and Illustration sections can be found on GitHub https://github.com/XynthiaKavelaars/Bayesian-multivariate-logistic-regression.

Chapter 5

Bayesian multilevel multivariate logistic regression for superiority decision-making under observable treatment heterogeneity

Based on Kavelaars, X., Mulder, J., & Kaptein, M. (2022a). *Bayesian multilevel multivariate logistic regression for superiority decision-making under observable treatment heterogeneity.* [Submitted for publication].

Abstract

In social, medical, and behavioral research we often encounter datasets with a multilevel structure and multiple correlated dependent variables. These data are frequently collected from a study population that distinguishes several subpopulations with different (i.e., heterogeneous) effects of an intervention. Despite the frequent occurrence of such data, methods to analyze them are less common and researchers often resort to either ignoring the multilevel and/or heterogeneous structure, analyzing only a single dependent variable, or a combination of these. These analysis strategies are suboptimal: Ignoring multilevel structures inflates Type I error rates, while neglecting the multivariate or heterogeneous structure masks detailed insights. To analyze such data comprehensively, the current paper presents a novel Bayesian multilevel multivariate logistic regression model. The clustered structure of multilevel data is taken into account, such that posterior inferences can be made with accurate error rates. Further, the model shares information between different subpopulations in the estimation of average and conditional average multivariate treatment effects. To facilitate interpretation, multivariate logistic regression parameters are transformed into posterior success probabilities and differences between them. A numerical evaluation compared our framework to less comprehensive alternatives and highlighted the need to model the multilevel structure: Treatment comparisons based on the multilevel model had targeted Type I error rates, while single-level alternatives resulted in inflated Type I errors. A re-analysis of the Third International Stroke Trial data illustrated how incorporating a multilevel structure, assessing treatment heterogeneity, and combining dependent variables contributed to an in-depth understanding of treatment effects.

5.1 Introduction

In social, medical, and behavioral research we often encounter datasets with a multilevel structure and multiple correlated dependent variables. An example of such a study is the Cognition and Radiation Study B (Schimmel et al., 2022; Schimmel et al., 2018) that investigated whether local brain radiation (stereotactic radiosurgery) preserves cognitive functioning and quality of life better than whole brain radiation in cancer patients with multiple brain metastases. Patients were recruited from multiple hospitals and the treatment was executed in two treatment centers, giving the data a multilevel structure. The authors noted that a) almost half of the reviewed studies were multicenter trials; and b) many studies were designed to assess effectiveness and side effects simultaneously, thus including at least two dependent variables.

Often, these multilevel, multivariate data are collected from a study population that consists of several subpopulations with potentially distinctive (i.e., heterogeneous) effects of an intervention. Examples of such studies are the two International Stroke Trials (IST and IST-3; International Stroke Trial Collaborative Group, 1997; Sandercock et al., 2016; Sandercock et al., 2011; The International Stroke Trial-3 Collaborative Group, 2012), which investigated the effects of antiplatelet and antithrombotic treatments on various (neuro)psychological, functional and psychosocial dependent variables respectively. We discuss the Third International Stroke Trial (IST-3) in more depth as it serves as a running example throughout the paper. The IST-3 investigated the effects of an intravenous thrombolytic treatment on short-term (e.g., recurrent stroke, functional deficits) and long-term (e.g., dependency, depression, pain) indicators of health status among patients who suffered from an acute ischaemic stroke. This trial covered patients from multiple treatment centers in multiple countries and thus clearly has a multilevel structure. Further, the IST-3 data revealed considerable variation in characteristics of patients and disease - such as subtype or severity of stroke, blood pressure, and age - that can be predictive of treatment effects and call for exploration of treatment heterogeneity to gain insight into subpopulation-specific effects (Lindley et al., 2015).

All the trials mentioned so far made treatment comparisons in the context of randomized controlled trials (RCTs): Randomized experiments in which an experimental or a control treatment is randomly assigned and administered to a random sample of patients. RCTs often aim to evaluate whether the experimental treatment is superior or (non-)inferior to the control condition and ultimately guide clinicians in evidence-based assignment of treatments and interventions (Food and Drug Administration, 2016).

Whereas RCTs are considered a golden standard for treatment comparison, their implementation is challenged by a growing demand for personalized treatment (Evans, 2003; Grol & Grimshaw, 2003; Ng et al., 2009; Simon, 2010). That is, clinical practice relies more and more on the idea that different patients react differently to treatments. Treatment prescription is increasingly guided by a trade-off between patient-specific risks and benefits, making the research context for these decisions multivariate and heterogeneous (Murray et al., 2016). While demanding more complex methodology, personalization of treatments can impede the collection of sufficient data for rigorous treatment evaluation. Development of more targeted treatments limits eligibility for participation in trials, thereby making the recruitment of subjects more difficult. As a solution, trials more often span multiple treatment centers or countries. This adds another layer of complexity to the research context: clustered data that often require multilevel analysis.

To meet the methodological demands of these increasingly complex research problems, RCTs ideally provide a) a broad understanding of the treatment's effects on multiple dependent variables; and b) insights into potential dependencies of treatment effects on characteristics of patients; and c) an accurate handling of clustered data structures. In practice, such comprehensive methods are less common, and often researchers resort to either ignoring the multilevel and/or heterogeneous structure, analyzing only a single dependent variable, or a combination of these. Below, we discuss how these three aspects can be implemented in RCT methodology to support research in personalized treatment.

First, many RCTs evaluate more than one dependent variable, often performing univariate analyses (Food and Drug Administration, 2017). As an example, the investigators

of the IST-3 were primarily interested in living independently six months after stroke and secondarily in several other dependent variables, such as recurrent events, adverse reactions to the treatment, and mental health indicators. Analyzing dependent variables independently provides useful insights into treatment effects on each of these dependent variables individually, but discards available information about the relation between them. When the effects on individual dependent variables are complemented with information about their co-occurrences via multivariate analysis, a more detailed picture of treatment effects emerges. Multivariate analysis models relationships between dependent variables and can a) be helpful to detect outcome patterns that would be ignored when dependent variables are considered in isolation; and b) improve the accuracy of sample size computations and error rates in statistical decision-making (Food and Drug Administration, 2017; Kavelaars et al., 2020; Sozu et al., 2016; Su et al., 2012).

Second, incorporating patient and/or disease characteristics in treatment comparison can result in a considerable improvement of the practical value of RCTs. The IST-3 used a sample of diverse patients with different personal and disease characteristics. This variation contains valuable information regarding differences in treatment effects. For example, knowing whether patients with different weights or blood pressures have different chances of a recurrent stroke or independent living has the potential to inform treatment recommendations. When treatments have distinct effects on patients with different characteristics, treatment effects are considered heterogeneous among (sub)populations of patients. In this case, average treatment effects (ATEs) give a global idea of treatment results among the trial population, but have limited value in targeting treatments to specific patients with their individual (disease) characteristics (Hamburg & Collins, 2010; Mirnezami et al., 2012; Schork, 2015). Conditional average treatment effects (CATEs) among specific patient groups provide insight into the variation of treatment effects among the population and help to distinguish patients who ultimately benefit from the treatment from those who do not or may even experience adverse treatment effects. Unfortunately, subgroup-specific treatment comparisons are insufficiently implemented as part of standard trial methodology yet (Thall, 2020). If subgroups are targeted at all, their

effects are often analyzed independently via stratified (or subgroup) analysis. Such a subgroup analysis disregards information from related subgroups and suffers from suboptimal power due to subsetting. Modeling heterogeneity is a more powerful alternative that directly uses the relation between subgroups and allows subgroups to borrow strength from each other (Kaptein, 2014; Kaptein et al., 2015; Kavelaars et al., 2022b).

Third, multilevel data are characterized by observational units that are grouped in clusters. For example, the IST-3 spans multiple treatment centers and multiple countries. Reasons to use multilevel analysis can be both substantive and statistical. From a substantive perspective, multilevel analysis can be useful to explain differences between clusters, while using the information from the entire sample (Gelman & Hill, 2007; Viele et al., 2014). Different trials may - for example - have overlapping but non-identical target populations that can be distinguished by covariate information and may contribute to the understanding of treatment effects. Statistically, differences between clusters should be taken into account for the sake of validity, even if these differences are not of direct interest (Hox et al., 2017; McGlothlin & Viele, 2018; Raudenbush & Bryk, 2001). Clustered data require specific analysis methods that are flexible enough to treat observations from different clusters as more similar to each other than to observations from other clusters. If observations within clusters are indeed more similar, the clustered structure is reflected in variance partitioning: variance between observations within clusters is smaller than variance between observations from different clusters. When clustered observations are treated as independent observations, variance originating from differences between clusters is then erroneously attributed to differences between a manifold of observational units and the unique amount of information is overestimated. As a result, standard errors are overestimated, Type I error rates are inflated, and validity of statistical inference is compromised. The larger the variance between clusters relative to the variance between observational units within clusters, the larger the effect on standard errors. Properly modeling the multilevel structure of clustered data and allowing the parameters to vary over clusters is therefore crucial for accurate statistical decision-making (Hox et al., 2017; Raudenbush & Bryk, 2001).

The current paper presents a Bayesian multilevel multivariate logistic regression (BMMLR) framework to capture the three abovementioned methodological aspects in a comprehensive analysis and decision procedure for treatment comparison. We build upon an existing Bayesian multivariate logistic regression (BMLR) framework for single-level data to analyze multivariate binary data in the presence of treatment heterogeneity and present a multilevel extension to deal with multilevel data. The multilevel aspect adds another layer of complexity, making the analysis a non-trivial endeavor. We discuss the existing BMLR framework first. This framework consists of three coherent elements (Kavelaars et al., 2022b):

1. a multivariate modeling procedure to find unknown regression parameters;
2. a transformation procedure to convert regression parameters to the probability scale to make analysis results more interpretable;
3. a compatible decision procedure to draw conclusions regarding treatment superiority or inferiority with targeted Type I error rates.

The first element, the modeling procedure, assumes multivariate Bernoulli distributed dependent variables and assigns them a multinomial parametrization. A multinomial parametrization is helpful for two reasons, since this a) allows statisticians to draw and build upon existing, established multinomial techniques with tractable (conditional) posterior distributions; and b) has the flexibility to model correlations between dependent variables on the subpopulation level, which contributes to the accuracy of inference under treatment heterogeneity (Dai et al., 2013; Kavelaars et al., 2020, 2022b). Several other multivariate modeling procedures, such as the multivariate probit model by Chib (1995) or multivariate logistic regression models by Malik and Abraham (1973) and O'Brien and Dunson (2004), have a more restrictive correlation structure and are therefore theoretically less suitable to detect treatment heterogeneity with adequate error control. Moreover, the multivariate logistic regression model by Malik and Abraham (1973) does not provide insight into the treatment effects on individual dependent variables. Copula structures have been proposed as promising multivariate alternatives as well, but these models can be difficult to apply to binary dependent variables (Braeken et al., 2007; Nikoloulopoulos & Karlis, 2008;

Panagiotelis et al., 2012). The second element, the transformation procedure, builds upon the close relation between the multinomial and multivariate parametrizations to express results on the scale of (multivariate) success probabilities and differences between them, as a more intuitive alternative to multinomial (log-)odds. The transformed parameters provide understandable insights into the treatment's performance on the trial population (i.e., ATEs) as well as subpopulations of interest (i.e., CATEs). The third element, the decision procedure, conveniently uses the Bayesian nature of the modeling procedure, allowing for inference on the posterior samples of transformed parameters. Decisions can be made in several ways to flexibly combine and weigh multiple dependent variables into a single decision for a population of interest, while taking correlations between dependent variables into account.

The main contribution of the current paper is the extension of the single-level BMLR framework to the multilevel context. The novel Bayesian multilevel multivariate logistic regression (BMMLR) framework provides a multilevel model component and adjusts the transformation and decision procedure accordingly, to make the framework suitable for the multilevel context. The remainder of the paper is structured as follows. Section 5.2 introduces the multilevel multivariate logistic regression model to obtain a sample from the posterior distribution of regression coefficients. Section 5.3 outlines how to transform the obtained regression coefficients to more interpretable treatment effect parameters. Section 5.4 discusses the decision procedure to use the treatment effect parameters for treatment comparison. Section 5.5 demonstrates the performance of the model numerically via simulation and in Section 5.6 the methodology is illustrated with data from the IST-3. The paper concludes with a discussion in Section 5.7.

5.2 BMMLR: Bayesian multilevel multivariate logistic regression

Consider the general case with $K \in \{1, \dots, K\}$ binary dependent variables y_{ji}^k for subject $i \in \{1, \dots, n_j\}$ in cluster $j \in \{1, \dots, J\}$. Outcome y_{ji}^k is Bernoulli distributed with success probability θ_{ji}^k and multivariate vector of K dependent variables, $\mathbf{y}_{ji} = (y_{ji}^1, \dots, y_{ji}^K)$ is multivariate Bernoulli distributed (Dai et al., 2013). The multivariate Bernoulli distribution relies on a hybrid parameterization where a K-variate success probability in $\boldsymbol{\theta}_{ji} = (\theta_{ji}^1, \dots, \theta_{ji}^K)$ is expressed in terms of $Q = 2^K$ multinomial joint response probabilities in $\boldsymbol{\phi}_{ji} = (\phi_{ji}^1, \dots, \phi_{ji}^Q)$ (Dai et al., 2013). The qth joint response probability in $\boldsymbol{\phi}_{ji}$ corresponds to multinomial response combination $\mathbf{h}^q$, which has length K and is given in the qth row of the matrix of joint response combinations denoted by $\mathbf{H}$:

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & \dots & 1 & 1 \\ 1 & 1 & \dots & 1 & 0 \\ & & \dots & & \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix} \tag{5.1}$$

Hence, joint response probability $\phi_{ji}^q = p(\mathbf{y}_{ji} = \mathbf{h}^q)$. Note that joint response probability $\boldsymbol{\phi}_j$ and success probability $\boldsymbol{\theta}_j$ are identical in the univariate situation (i.e., $K = 1$).

5.2.1 Likelihood of the data

The multinomial parametrization of multivariate Bernoulli distributed data allows to model the relation between dependent variables $\mathbf{y}_{ji}$ and one or multiple predictor variables via multinomial logistic regression. Joint response probability ϕ_{ji}^q is then regressed on a vector of P covariates, $\mathbf{x}_{ji} = (x_{ji0}, \dots, x_{ji(P-1)})$. Covariate $x_{ji0} = 1$ is a constant to estimate the intercept and covariate x_{jip} for $p \in \{1, \dots, P-1\}$ can, for example, be a treatment indicator, a patient characteristic, or an interaction between these.

The relation between outcome vector $\mathbf{y}_{ji}$ and covariate vector $\mathbf{x}_{ji}$ is mapped with a multinomial logistic function that expresses the probability of $\mathbf{y}_{ji}$ being in response category

q, conditional on $\mathbf{x}_{ji}$:

$$\begin{aligned}\phi_{ji}^{q} &= p((\mathbf{y}_{ji} = \mathbf{h}^{q})|\mathbf{x}_{ji}) \\ &= \frac{\exp(\psi_{ji}^{q})}{\sum_{r=1}^{Q-1} \exp(\psi_{ji}^{r}) + 1},\end{aligned} \tag{5.2}$$

Here, ψ_{ji}^{q} is a linear predictor:

$$\psi_{ji}^{q} = \mathbf{x}_{ji}^{'}\boldsymbol{\gamma}_{j}^{q} \tag{5.3}$$

In Equation 5.3, regression coefficients for response category q, $\boldsymbol{\gamma}_{j}^{q} = (\gamma_{j0}^{q}, \dots, \gamma_{j(P-1)}^{q})$ are unknown parameters of interest. Regression coefficients of response categories $1, \dots, Q-1$ are estimated, while regression coefficients of response category Q are fixed at zero (i.e., $\boldsymbol{\gamma}_{j}^{Q} = \mathbf{0}$) to ensure identifiability of the model. The entire set of regression coefficients is denoted with $\boldsymbol{\gamma}_{j}$.

Key aspect of multilevel analysis is that regression coefficients $\boldsymbol{\gamma}_{j}^{q}$ vary over clusters. Regression effects in a multilevel model are a linear function in itself:

$$\begin{aligned}\gamma_{jp}^{q} &= \gamma_{p0}^{q} + u_{pj}^{q} \\ \mathbf{u}_{j}^{q} &= (u_{0j}^{q}, \dots, u_{(P-1)j}^{q}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}^{q})\end{aligned} \tag{5.4}$$

Equation 5.4 consists of two elements:

1. Constant γ_{p0}^{q} is the common effect in the population and does not vary over clusters.
2. Error term u_{pj}^{q} covers unexplained variance in γ_{jp}^{q}.

Equation 5.4 can be adjusted to model cluster-specific predictors or cross-level interactions between cluster-level predictors and individual level-predictors. Further, Equation 5.4 can be extended to model mixed effects, which combine regression coefficients that vary over clusters (so-called random effects) and regression coefficients that are identical for all clusters (so-called fixed effects). More information on the specification of more complex linear predictors

can be found in general resources on multilevel models, such as Hox et al. (2017) or Gelman and Hill (2007). In general, it should be noted that each additional random effect increases the number of parameters, affecting computational burden and estimation precision.

5.2.2 Posterior distribution of regression coefficients

Primary goal of BMMLR is estimating the joint posterior distribution of unknown regression coefficients γ_j, their means γ, and their covariance matrices $\mathbf{\Sigma}$. The joint posterior probability of these parameters is given by:

$$p(\gamma_j, \gamma, \mathbf{\Sigma}|\mathbf{y}) \propto p(\mathbf{y}|\gamma_j)p(\gamma_j|\gamma, \mathbf{\Sigma})p(\gamma)p(\mathbf{\Sigma}). \tag{5.5}$$

The posterior probability in Equation 5.5 is proportional to the product of three types of probabilities:

1. The likelihood of the data conditional on cluster-specific regression coefficients, $p(\mathbf{y}|\gamma_j)$, which is the multinomial logistic function given by Equation 5.2;
2. The probability of the cluster-specific regression coefficients γ_j^q conditional on their means γ and covariance matrices $\mathbf{\Sigma}$, $p(\gamma_j|\gamma, \mathbf{\Sigma})$;
3. The prior probabilities of regression coefficient's means γ, $p(\gamma)$, and covariance matrix $\mathbf{\Sigma}$, $p(\mathbf{\Sigma})$, before observing the data.

As the multinomial logistic likelihood (Equation 5.2) does not have a (conditionally) conjugate prior distribution, the functional form of the posterior distribution is unknown and the regression coefficients cannot be sampled directly from the posterior distribution. In Appendix J, we present a Gibbs sampling algorithm based on a Pólya-Gamma auxiliary variable expansion of the likelihood proposed by Polson et al. (2013). The expanded likelihood has a Gaussian form and can be combined with normal prior distributions for regression coefficients γ and an inverse-Wishart distribution on covariance matrix $\mathbf{\Sigma}$. The parameters are known to have conditionally conjugate posterior distributions and allow for direct sampling from their multivariate normal and inverse-Wishart distributions respectively,

resulting in MCMC chains of the joint posterior distribution in Equation 5.5. We also provide a few comments on prior specification for the proposed Gibbs sampling procedure in Appendix J. As an alternative to the proposed Gibbs sampling procedure, sampling from the posterior distribution can be done with standard MCMC methods for non-conjugate prior-likelihood combinations, such as Metropolis-Hastings or Hamiltonian sampling.

5.3 Transformation of posterior regression coefficients to the probability scale

The output of the BMMLR model from Section 5.2 is an MCMC sample of posterior multinomial regression coefficients. These regression coefficients reflect the importance of a predictor on a specific joint response combination and represent - in exponentiated form - the odds compared to reference category Q. While these regression coefficients can be insightful in a truly multinomial research problem, they have no straightforward interpretation in multivariate treatment comparison where marginal effects on individual dependent variables play a central role (Food and Drug Administration, 2017).

Transformation of regression coefficients to the multivariate probability scale provides a convenient solution to gain more intuitive insights into both joint and marginal treatment effects. These transformations rely on the close relationship between multinomial and multivariate parametrizations and can be flexibly obtained for the trial population (i.e., average treatment effects) or for subpopulations (i.e., conditional average treatment effects). They are directly suitable for statistical decision-making regarding treatment comparison.

We use the framework for transformation to the probability scale and decision-making with a posterior sample of multivariate treatment differences introduced in Kavelaars et al. (2020) and Kavelaars et al. (2022b). Technical details of these procedures are presented in Algorithm 5 in Appendix K. We use the remainder of this section to summarize and illustrate the procedure with a toy example from the IST-3-data, where we assume interest in the effect of Alteplase in the experimental condition (T_A) compared to no treatment in the control group

(T_C).

Assume that we re-analyze a part of the IST-3 data using the BMMLR framework and take one of originally presented analyses as a starting point (The International Stroke Trial-3 Collaborative Group, 2012). In the selected analysis, the researchers compared the effects of Alteplase vs. control on their primary outcome, long-term independent living after six months ($Indep6$), among subgroups of patients based on the severity of their initial stroke. In our example, we perform a multivariate analysis of the treatment effects on the primary outcome ($Indep6$) and one of the secondary (short-term) dependent variables: being stroke-free in the first seven days after the initial stroke ($Strk7$). We incorporate severity of the initial stroke as a predictor variable to study heterogeneity, using the grouping criteria from the original trial for the estimation of conditional average treatment effects. We aim to investigate the average treatment effect among the trial population as specified by the original eligibility criteria for inclusion. We are also interested in a potential interaction between the treatment and stroke severity, and investigate the conditional average treatment effects among patients with various severity of stroke. To take the clustered structure of the data into account, we specified a BMMLR mixed-effects model with random slopes for the intercept and the main treatment effect, resulting in the following linear predictor:

$$\begin{aligned}
\psi^q_{ji} &= \gamma^q_{j0} + \gamma^q_{j1} T_{ji} + \beta^q_2 NIHSS_{ji} + \beta^q_3 NIHSS_{ji} T_{ji} \qquad (5.6)\\
\gamma^q_{j0} &= \gamma^q_{00} + u_{0j}\\
\gamma^q_{j1} &= \gamma^q_{10} + u_{1j}.
\end{aligned}$$

In Equation 5.6, $\mathbf{x}_{ji} = (1, T_{ji}, NIHSS_{ji}, NIHSS_{ji} T_{ji})$ with treatment indicator T_{ji} and $NIHSS_{ji}$ being the stroke severity score of subject i in hospital j. The $Q = 4$ resulting joint response categories are $(\{Strk7 = 1, Indep6 = 1\}, \{Strk7 = 1, Indep6 = 0\}, \{Strk7 = 0, Indep6 = 1\}, \{Strk7 = 0, Indep6 = 0\})$, which we refer to as $(\{11\}, \{10\}, \{01\}, \{00\})$.

5.3.1 Transformation to cluster-specific (differences between) probabilities

The main quantity of interest, the multivariate treatment difference, is defined as the difference between multivariate success probabilities of the two treatments:

$$\begin{aligned} \delta_j^{Strk7} &= \theta_{Aj}^{Strk7} - \theta_{Cj}^{Strk7} \\ \delta_j^{Indep6} &= \theta_{Aj}^{Indep6} - \theta_{Cj}^{Indep6} \end{aligned} \tag{5.7}$$

The elements on the right-hand sides of Equation 5.7, marginal success probabilities θ_{Tj}^k, are sums of the multinomial joint response probabilities of all response categories with a success on outcome k:

$$\begin{aligned} \theta_{Tj}^{Strk7} &= I(T_j = T)\big[p((\mathbf{y}_j|T_j) = \{11\}) + p((\mathbf{y}_j|T_j) = \{10\})\big] = \phi_{Tj}^1 + \phi_{Tj}^2 \\ \theta_{Tj}^{Indep6} &= I(T_j = T)\big[p((\mathbf{y}_j|T_j) = \{11\}) + p((\mathbf{y}_j|T_j) = \{01\})\big] = \phi_{Tj}^1 + \phi_{Tj}^3 \end{aligned} \tag{5.8}$$

The multinomial joint response probabilities $\boldsymbol{\phi}_{Tj}$ that form the elements of success probabilities $\boldsymbol{\theta}_{Tj}$ follow from plugging in posterior regression coefficients $\boldsymbol{\gamma}_j$ in the linear predictor (Equation 5.6) and the multinomial logistic link function (Equation 5.2) for prespecified covariates $\mathbf{x}_{ji}$:

$$\phi_{T_{ji}}^q = = \frac{\exp\left(\psi_{Tji}^q\right)}{\sum_{r=1}^{Q-1} \exp\left(\psi_{Tji}^r\right) + 1}. \tag{5.2 revisited}$$

The information in covariate vector $\mathbf{x}_{ji}$ determines the treatment as well as the subpopulation of interest. Subpopulations can be defined as a value, such as a stroke severity score of one standard deviation below or above the mean, that can be plugged in directly into Equations 5.6 and 5.2. When interested in a subpopulation that is defined by an interval, such as the groups of stroke severity in the IST-3, the joint response probability is marginalized over the

specified interval.

Since the model in Section 5.2 resulted in a sample of L posterior draws of each regression coefficient, multivariate treatment differences are computed for each draw (l) separately. The resulting posterior samples can be summarized with standard descriptive methods.

5.3.2 Pooling treatment effects over clusters

As a last step, cluster-specific estimates are pooled into estimates of average or conditional treatment effects among (sub)populations of interest via the following procedure:

$$\boldsymbol{\delta} = \frac{\sum_{j=1}^{J} n_j \boldsymbol{\delta}_j}{\sum_{j=1}^{J} n_j} \tag{5.9}$$

This pooling strategy weighs cluster-specific estimates by cluster size, thereby balancing data with unequal cluster sizes.

5.4 Decision-making based on multivariate treatment effects

The obtained sample of posterior treatment differences can be used for statistical decision-making regarding treatment superiority and inferiority. The multivariate context has multiple options to define superiority and inferiority, leaving much flexibility to combine and prioritize dependent variables in a suitable way. We shortly discuss four different decision rules to give some idea of possibilities, without intending to be exhaustive or complete. The presented rules have different theoretical underpinnings and distinct statistical properties, such as acceptance regions, a priori estimated sample sizes, cutoff values, and error rates. The acceptance regions for superiority decisions of the four presented rules have been graphically presented in Figure 5.1. More details to guide an informed choice for one of these decision rules in practice can

be found in Kavelaars et al. (2020).

Three of these rules originate from guidelines of the Food and Drug Administration (2017). The Food and Drug Administration (FDA) defines superiority as a treatment difference larger than zero on the primary outcome (which we refer to as "Single rule"), on all dependent variables ("All rule") or on any of the dependent variables ("Any rule"). The Single rule reduces the statistical analysis to a univariate problem, using only the treatment difference of independent living after 6 months as a primary outcome (Single rule). The All and Any rules make no distinction in the importance of dependent variables and assume that the short-term and long-term outcomes are either both required for superiority or inferiority (All rule) or are interchangeable (Any rule).

In practice, these rules can oversimplify decision-making. Secondary outcome variables often contribute to treatment evaluation as well, but are given a co-primary status in the All and Any rules or are not formally included in the statistical decision procedure when the Single rule is used (Sozu et al., 2010, 2016). To handle outcomes that differ in relative importance, linear combinations of dependent variables with pre-assigned (importance) weights have been proposed as a flexible alternative (Kavelaars et al., 2020; Murray et al., 2016; O'Brien, 1984; Su et al., 2012; Whitehead et al., 2010). We refer to a linear combination as a Compensatory rule, referring to its inherent mechanism that allows (weighted) positive and negative effects to compensate each other. The Compensatory rule allows the IST-3 data to consider the effects on the long-term much more important than the short-term effect without completely excluding the risk of a recurrent stroke from the final decision. In such a situation, we can assign the primary outcome (*Indep*6) - for example - four times more weight than the secondary outcome (*Strk*7) and consider Alteplase superior to no treatment if a lower chance of dependency is outweighed by a small increase in the risk of a recurrent stroke.

Evidence in favor of the decision rule can be quantified by the proportion of posterior draws of the pooled treatment difference $\boldsymbol{\delta}$ that lie in the decision-rule specific acceptance region, denoted by $\mathcal{S}_R$. A conclusion is reached via comparison to p_{cut}, which is a cutoff value to balance the required amount of evidence with anticipated Type I error rates (Marsman &

Wagenmakers, 2016):

$$p(\boldsymbol{\delta} \in \mathcal{S}_R) > p_{cut}. \tag{5.10}$$

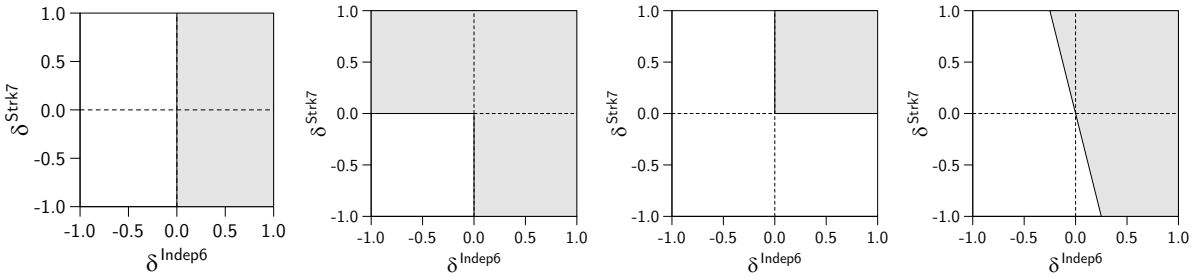
In the multivariate logistic regression model, the probability in Equation 5.10 has no analytical solution. Therefore, decisions are made via the posterior MCMC-sample of L draws. Superiority is concluded when:

$$\frac{1}{L}\sum_{(l)=1}^{L} I(\boldsymbol{\delta}^{(l)} \in \mathcal{S}_R) > p_{cut}. \tag{5.11}$$

Similarly, inferiority is concluded when:

$$\frac{1}{L}\sum_{(l)=1}^{L} I(\boldsymbol{\delta}^{(l)} \in \mathcal{S}_R) < 1 - p_{cut}. \tag{5.12}$$

In Section 5.6, we demonstrate these decision rules with the data from the IST-3 as part of an illustration of the BMMLR framework.



(a) Single (*Indep*6) (b) Any (c) All (d) Compensatory

Figure 5.1: Superiority regions of four decision rules applied to the IST-3. The Compensatory rule has weights $(w^{Strk7}, w^{Indep6}) = (0.20, 0.80)$.

5.5 Numerical evaluation

The current section presents an evaluation of the performance of the proposed BMMLR framework. The goal of the evaluation was twofold, and we aimed to demonstrate:

1. how well the obtained regression coefficients and treatment effects correspond to their true values to examine bias;
2. how often the BMMLR framework results in an (in)correct superiority or inferiority conclusion to learn about decision error rates;

5.5.1 Setup

Model The performance of the multilevel model was evaluated in a treatment comparison based on a two-level model with two dependent variables and one covariate at the subject level. We compared the method to two different (single-level) reference approaches, resulting in the following three modeling procedures:

1. The BMMLR model presented in Section 5.2. We generated response data from a mixed effects model to include random effects while keeping the number of estimated parameters limited. We included an interaction between the treatment and the covariate as well, resulting in the following linear predictor:

$$\begin{aligned} \psi_{ji}^q &= \gamma_{j0}^q + \gamma_{j1}^q T_{ji} + \beta_2^q w_{ji} + \beta_3^q w_{ji} T_{ji} \\ \gamma_{j0}^q &= \gamma_{00}^q + u_{0j} \\ \gamma_{j1}^q &= \gamma_{10}^q + u_{1j}. \end{aligned} \tag{5.13}$$

   In line with previous notation, $\mathbf{x}_{ji} = (1, T_{ji}, w_{ji}, w_{ji} T_{ji})$ in Equation 5.13. Further, $\boldsymbol{\gamma}_j^q = (\gamma_{j0}^q, \gamma_{j1}^q)$ reflects random effects with multivariate normally distributed errors (i.e., $(u_{0j}^q, u_{1j}^q) \sim N(\mathbf{0}, \boldsymbol{\Sigma}^q)$) for the intercept and main effect of the treatment. Regression coefficients $\boldsymbol{\beta}^q = (\beta_2^q, \beta_3^q)$ reflects fixed effects for the covariate and covariate-by-treatment interaction.
2. The single-level Bayesian multivariate logistic regression model (BMLR; Kavelaars et al., 2022b), as a first reference approach. For this model, we use a restricted version of

Equation 5.13 with fixed regression coefficients only:

$$\psi_{ji}^{q} = \beta_0^q + \beta_1^q T_{ji} + \beta_2^q w_{ji} + \beta_3^q w_{ji} T_{ji}, \tag{5.14}$$

MCMC chains were sampled with a simplified version of the Gibbs sampling procedure in Appendix J, that iterates over $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$. The model shares information in the estimation of conditional treatment effects with sufficient power, but does not take the multilevel structure of the data into account.

3. The single-level unconditional Bayesian multivariate Bernoulli analysis (BMB; Kavelaars et al., 2020), as a second reference approach. Bayesian multivariate Bernoulli analysis relies on a conjugate multinomial likelihood and Dirichlet prior. MCMC draws are sampled directly from the posterior Dirichlet distribution with parameters $\sum_{j=1}^{J}\sum_{i=1}^{n_j} I(\mathbf{y}_{ji} = \mathbf{h}^q) + \alpha^{0q}$, where we assigned prior hyperparameters $\boldsymbol{\alpha}^0 = (0.01, 0.01, 0.01, 0.01)$. The approach can estimate homogeneous treatment effects accurately and fast, but cannot deal with multilevel data. Moreover, conditional treatment effects originate from subsampling, which is less powerful than regression due to the isolation from other information.

Effect size We specified a heterogeneous treatment effect, with pooled average treatment differences of zero ($\boldsymbol{\delta} = (0, 0)$, $\delta(\mathbf{w}) = 0$) and pooled conditional treatment differences larger than zero ($\boldsymbol{\delta} = (0.25, 0.15)$, $\delta(\mathbf{w}) = 0.20$). This scenario aimed to demonstrate the Type I error rate among the trial population. It reflects a least favorable treatment difference for the Any and Compensatory rules and should therefore result in the targeted Type I error rate for these rules to be considered accurate. The conditional treatment effect provided insight into the power to conclude superiority among the subpopulation under consideration. Outcome variables were negatively correlated ($\rho_{ATE} = -.157$; $\rho_{CATE} = -.20$). For the BMMLR model, the matrix of random variances, $\boldsymbol{\Sigma}^q$ was specified as:

$$\begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \tag{5.15}$$

for all $q \in 1, \ldots, Q-1$.

Sample size We varied the sample sizes at the cluster and subject level. Since there are no clear guidelines regarding sample size computations in multilevel multivariate logistic regression, we explored performance of the model for different numbers of clusters and different sample sizes within clusters. Specifically, we used number of clusters $J \in \{10, 100\}$ and observations per cluster $n_j \in \{10, 100\}$ for each treatment, resulting in four different sample size combinations.

Procedure

Data generation For each sample size, we sampled 1000 datasets. We assigned n_j participants to each treatment T and generated covariate x from a standard normal distribution. We sampled response vector $\mathbf{y}_{ji}$ from a multinomial distribution with probabilities ϕ_{ji}.

Gibbs sampling Regression coefficients for the BMMLR and BMLR models were estimated via the Gibbs sampling procedure in Appendix J. We ran two MCMC-chains via the Gibbs sampler introduced in Section 5.2 with $L = 50,000$ iterations plus $10,000$ burn-in iterations. This large number of iterations aims to minimize the influence of the potentially high autocorrelations between parameters in multilevel models on the stationary distribution of the parameters. Autocorrelations were highest among random effect parameters γ_j and ranged between 0.107 and 0.781 at lag 1 and reduced to a range of -0.012 – 0.276 at lag 10. Further, following the guidelines in Gelman et al. (2013), we ensured that the multivariate potential scale reduction factor was below 1.10.

For the multilevel model (BMMLR), we specified diffuse priors, which were multivariate normally distributed for regression coefficients and inverse-Wishart distributed for the

covariance matrix:

$$(\beta_2^q, \beta_3^q) \sim N([\begin{smallmatrix}0\\0\end{smallmatrix}], [\begin{smallmatrix}0.1 & 0\\0 & 0.1\end{smallmatrix}]) \tag{5.16}$$
$$(\gamma_{00}^q, \gamma_{10}^q) \sim N([\begin{smallmatrix}0\\0\end{smallmatrix}], [\begin{smallmatrix}0.1 & 0\\0 & 0.1\end{smallmatrix}])$$
$$\boldsymbol{\Sigma}^q \sim \mathcal{W}^{-1}(2, [\begin{smallmatrix}0.1 & 0\\0 & 0.1\end{smallmatrix}]).$$

Regression parameters $\boldsymbol{\beta}^q = \beta_0^q, \dots, \beta_3^q$ in the single-level regression model (BMLR) were - similar to (β_2^q, β_3^q) in the multilevel model - assigned a multivariate normal prior distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ with diagonal entries 0.1 and off-diagonal entries of 0.

Transformation and decision-making We applied the procedures in Algorithm 5 to use the obtained MCMC-chains of posterior regression coefficients for superiority decision-making. We thinned the chains in the transformation procedure with a factor 10 to reduce the computational burden.

We considered two different effects:

1. an average treatment effect for the trial population;
2. a conditional treatment effect for a subpopulation scoring one standard deviation below the mean or lower;

The treatment effects required marginalization over the interval that defined the (sub)population, which we accomplished by averaging over joint response probabilities computed for the empirical sample of data. Cluster-specific treatment effects were weighed by their sample sizes to produce a pooled estimate of the treatment difference.

Decisions were made with a right-sided test for the All, Any, and Compensatory (equal weights, $\mathbf{w} = (0.50, 0.50)$) rules with formal superiority regions:

1. Any rule: $\mathcal{S}_R = \{\boldsymbol{\delta} | \max_{1<k<K} \delta^k > 0\} | \mathbf{y}, \mathbf{x}$ and cut-off value $p_{cut} = 1 - \frac{\alpha}{K}$
2. All Rule: $\mathcal{S}_R = \{\boldsymbol{\delta} | \min_{1<k<K} \delta^k > 0\} | \mathbf{y}, \mathbf{x}$ and cut-off value $p_{cut} = 1 - \alpha$
3. Compensatory rule: $\mathcal{S}_R = \{\boldsymbol{\delta} | \delta(\mathbf{w}) > 0\} | \mathbf{y}, \mathbf{x}$ and cut-off value $p_{cut} = 1 - \alpha$

We computed the probability to conclude superiority (p_{Sup}) as the proportion of posterior treatment differences in the superiority region via Equation 5.10. The targeted Type I-error rate of $\alpha = .05$ corresponded to decision threshold $p_{cut} = 1-\alpha = 0.95$ (Compensatory and All rules) and a for multiple tests corrected threshold $p_{cut} = 1 - \frac{\alpha}{K} = 0.975$ (Any rule) (Kavelaars et al., 2020; Marsman & Wagenmakers, 2016; Sozu et al., 2016).

Software

We conducted our analyses in `R` and made use of several existing packages (R Core Team, 2020). Pólya-Gamma variables were drawn with the `pgdraw` package (Makalic & Schmidt, 2016). Further, we drew variables from the multivariate normal, truncated normal, and Dirichlet distributions with the `MASS`, `msm`, and `MCMCpack` packages respectively (Jackson, 2011; Martin et al., 2011; Venables & Ripley, 2002). MCMC chains were diagnosed with the `coda` and `mcmcse` packages (Flegal et al., 2021; Plummer et al., 2006). We parallelized the simulation procedure with the `foreach` and `doParallel` packages (Microsoft & Weston, 2020a, 2020b) and created LaTeX tables with the `xtable` package (Dahl et al., 2019).

5.5.2 Results

The current subsection presents the results of the simulation study. Presented decision error rates are in Table 5.1.

Bias

Regression coefficients, variance matrices and treatment effects (success probabilities, treatment differences) could be estimated without bias in all sample sizes and data generating mechanisms. The absolute average deviation of mean point estimates from true values was smaller than .01.

Table 5.1: Proportions of superiority decisions (P) and standard errors (SE) by data-generating mechanism, estimation method, and decision rule.

Average treatment effect: $\boldsymbol{\delta} = (0.000, 0.000)$, $\boldsymbol{\delta}(\mathbf{w}) = 0.000$						
	Any		All		Compensatory	
$J = 10$, $n_j = 10$	p	SE	p	SE	p	SE
BMMLR	0.032	(0.006)	0.000	(0.000)	0.042	(0.006)
BMLR	0.055	(0.007)	0.001	(0.001)	0.059	(0.007)
BMB	0.050	(0.007)	0.001	(0.001)	0.046	(0.007)
$J = 100$, $n_j = 10$						
BMMLR	0.053	(0.007)	0.002	(0.001)	0.048	(0.007)
BMLR	0.077	(0.008)	0.003	(0.002)	0.066	(0.008)
BMB	0.069	(0.008)	0.002	(0.001)	0.056	(0.007)
$J = 10$, $n_j = 100$						
BMMLR	0.044	(0.006)	0.000	(0.000)	0.060	(0.008)
BMLR	0.200	(0.013)	0.004	(0.002)	0.125	(0.010)
BMB	0.188	(0.012)	0.003	(0.002)	0.113	(0.010)
$J = 100$, $n_j = 100$						
BMMLR	0.057	(0.007)	0.000	(0.000)	0.054	(0.007)
BMLR	0.252	(0.014)	0.005	(0.002)	0.169	(0.012)
BMB	0.245	(0.014)	0.005	(0.002)	0.159	(0.012)
Conditional treatment effect: $\boldsymbol{\delta} = (0.116, 0.069)$, $\boldsymbol{\delta}(\mathbf{w}) = 0.092$						
	Any		All		Compensatory	
$J = 10$, $n_j = 10$	p	SE	p	SE	p	SE
BMMLR	0.731	(0.014)	0.245	(0.014)	0.920	(0.009)
BMLR	0.397	(0.015)	0.065	(0.008)	0.587	(0.016)
BMB	0.183	(0.012)	0.025	(0.005)	0.294	(0.014)
$J = 100$, $n_j = 10$						
BMMLR	1.000	(0.000)	0.995	(0.002)	1.000	(0.000)
BMLR	1.000	(0.000)	0.868	(0.011)	1.000	(0.000)
BMB	0.933	(0.008)	0.520	(0.016)	0.980	(0.004)
$J = 10$, $n_j = 100$						
BMMLR	1.000	(0.000)	0.949	(0.007)	1.000	(0.000)
BMLR	0.997	(0.002)	0.771	(0.013)	1.000	(0.000)
BMB	0.917	(0.009)	0.445	(0.016)	0.969	(0.005)
$J = 100$, $n_j = 100$						
BMMLR	1.000	(0.000)	1.000	(0.000)	1.000	(0.000)
BMLR	1.000	(0.000)	1.000	(0.000)	1.000	(0.000)
BMB	1.000	(0.000)	1.000	(0.000)	1.000	(0.000)

Decision error rates

Type I error rates The average treatment effect demonstrated that the probability to incorrectly conclude superiority in multilevel regression (BMMLR) was close to the targeted .05 under a least favorable scenario (i.e., Any and Compensatory decision rules). In general, both reference approaches (BMLR and BMB) suffered from inflated Type I error to a similar extent.

The amount of inflation in the (single-level) reference approaches was affected by sample size: A large number of clusters ($J = 100$) and/or a large number of subjects per cluster ($n_j = 100$) had the largest Type I error rates, with the combination $J = 100, n_j = 100$ resulting in the most severe inflation. On the other hand, a small number of clusters and a small number of subjects per cluster ($J = 10, n_j = 10$) resulted in an acceptable Type I error rate for the single-level BMLR model as well, suggesting some robustness against the violation of the assumption of independent observations in the current setup. In general, the number of subjects per cluster appeared more influential on the Type I error rate inflation than the number of clusters, as demonstrated by the two scenarios with an identical total sample size ($J = 10, n_j = 100$ and $J = 100, n_j = 10$): A small number of clusters and a large sample size per cluster resulted in larger Type I error rates than a large number of clusters with a small sample size per cluster. Keeping everything else constant, a larger number of clusters meant more independent units, implying that the assumption of independent observations was violated less severely. In other words, the need for a multilevel model was more prominent when the number of clusters was small. A similar pattern was seen under the All rule, although Type I errors were small in general. This was expected, since a) the All rule is known to be the most conservative of the three introduced rules; and b) the treatment difference was smaller than the least favorable scenario of this decision rule.

Power The conditional treatment effect demonstrated the power to correctly conclude superiority for all three rules. Three results were highlighted. First, the multilevel model (BMMLR) appeared more powerful when the number of clusters was higher as opposed to a

smaller number of clusters. The two conditions with an equal total sample size (e.g., $J = 10, n_j = 100$ and $J = 100, n_j = 10$) showed a .046 difference in power under the All rule in favor of the model with $J = 100$ clusters. The other rules showed the same patterns, but had too high proportions of superiority conclusions to clearly distinguish the sample size conditions: The power in the other conditions equaled or was close to the maximum of 1.000.

Second, the multilevel model (BMMLR) was generally more powerful than the single-level regression model (BMLR) and resulted in more superiority conclusions. Again, differences were best illustrated by the All rule and the condition with small sample sizes for the Any and Compensatory decision rules, as these proportions were well below the maximum. Similar to the Type I error rates, the differences between the proportions of superiority conclusions appeared to be subject to the number of clusters, as demonstrated by a comparison of the two conditions with an identical total sample size under the All rule.

Third, the multivariate Bernoulli model (BMB) has low power overall, despite the underestimation of variance due to falsely assuming independent observations. As a subsampling approach, conditional treatments were fitted on the part of the data that makes up the subpopulation of interest. Especially the $J = 10, n_j = 10$ condition suffered from a small remaining sample size.

5.6 Illustration with IST-3 data

To illustrate the proposed framework with real data, we re-analyzed a subset of data from the Third International Stroke Trial using the BMMLR framework (Sandercock et al., 2016; The International Stroke Trial-3 Collaborative Group, 2012). The included $3,035$ subjects in the IST-3 were recruited from 156 different hospitals in 12 different countries, resulting in multilevel data from patients clustered within hospitals and hospitals clustered within countries. We selected a two-level subset of $1,447$ subjects from 75 hospitals in the United Kingdom with a known health and survival status at six months after the initial stroke and a known or predicted severity score of the initial stroke (NIH Stroke Score; NIHSS) at randomization.

The cluster sizes were skewed and ranged from 1 to 117, with a median cluster size of 7 (SD: 26.66). Of the selected subset of data, $n_A = 716$ subjects were in the Alteplase group (treatment = 1) and $n_C = 731$ subjects were in the control group (treatment = 0). We compared the effects of the two treatments on a) being stroke-free for seven days (0 = no; 1 = yes) and b) long-term independent living at six months (0 = no, 1 = yes), while taking the severity of the initial stroke into account. The NIHSS can range from 0 to 42 with a higher score indicating a more severe stroke. The average stroke severity score in the IST-3 was 13.12 (SD: 6.91) and comparable in both treatment groups.

5.6.1 Method

We fitted our model with random slopes for the intercept and the treatment effect. We sought to compare our multilevel model (BMMLR) to the two single-level models (BMLR and BMB) from the Numerical evaluation Section in treatment comparison of Alteplase and control on dependency after six months (δ^{Indep6}) and recurrent stroke within seven days (δ^{Strk7}). The multilevel model (BMMLR) was fitted with the linear predictor in Equation 5.6 and the linear predictor of the single-level regression model (BMLR) was:

$$\psi_{ji}^{q} = \beta_0^q + \beta_1^q T_{ji} + \beta_2^q NIHSS_{ji} + \beta_3^q NIHSS_{ji} T_{ji} \tag{5.17}$$

We ran two MCMC-chains via the Gibbs samplers and prior distributions specified in the Numerical evaluation. Since the chains of regression coefficients were highly autocorrelated in the multilevel model (lag 10: $\boldsymbol{\beta} : 0.47 - 0.59$; $\boldsymbol{\gamma} : 0.62 - 0.80$, $\boldsymbol{\Sigma} : -0.01 - 0.38$), we sampled a large number of iterations (500, 000) plus 10, 000 burn-in iterations. The multivariate potential scale reduction factor was below 1.01 for all parameters, implying that there were no signals of non-convergence. We thinned MCMC-chains in follow-up posterior transformations with a factor 10 to reduce computational demands, resulting in inference based on $L = 50,000$ draws.

We applied the procedures in Algorithm 5 to the thinned MCMC-chains of posterior regression coefficients to make superiority decisions. We considered (conditional) average

treatment effects among seven different (sub)populations:

1. ATE: average treatment effects for all patients in the trial population;
2. CATE - Low range: conditional average treatment effects for patients with a stroke severity score between 0 and 5;
3. CATE - Mid-Low range: conditional average treatment effects for patients with a stroke severity score between 6 and 14;
4. CATE - Mid-High range: conditional average treatment effects for patients with a stroke severity score between 15 and 24;
5. CATE - High range: conditional average treatment effects for patients with a stroke severity score above 25;
6. CATE - Low value: conditional treatment effects for patients with a stroke severity score of 5.18, corresponding to 1 standard deviation below the mean;
7. CATE - High value: conditional treatment effects for patients with a stroke severity score of 19.03, corresponding to 1 standard deviation above the mean.

The grouping criteria for CATEs of ranges were taken from the original IST-3 paper (The International Stroke Trial-3 Collaborative Group, 2012).

We performed two-sided tests for the All, Any, and Compensatory rules. Similar to the IST-3, we used independent living as the most important outcome in the Compensatory rule and specified weights $\mathbf{w} = (0.20, 0.80)$ for remaining free of strokes and independent living respectively. This specification implied that the long-term outcome had four times more impact on the decision than the short-term outcome. The targeted two-sided Type I-error rate of $\alpha = .05$ corresponded to decision threshold $p_{cut} = 1 - \frac{\alpha}{2} = 0.975$ (Compensatory and All rules) and a for multiple tests corrected threshold $p_{cut} = 1 - \frac{\alpha}{2K} = 0.9875$ (Any rule).

Software In addition to the software packages used in Section 5.5, we used `R` package `haven` to import the dataset (Wickham & Miller, 2021).

Table 5.2: Average (ATE) and conditional average (CATE) treatment effects of the 7 specified (sub)populations of the IST-3, including posterior probabilities (Pop) and superiority ($>$) and inferiority ($<$) conclusions for each decision rule.

	$\boldsymbol{\delta}$	Pop ($\boldsymbol{\delta}$)	Any	All	$\delta(\mathbf{w})$	Pop $\delta(\mathbf{w})$	Comp
ATE		$n_A = 716$, $n_C = 731$					
BMMLR	(-0.114, 0.029)	(0.000, 0.886)	$<$	-	0.000	0.504	-
BMLR	(-0.116, 0.033)	(0.000, 0.941)	$<$	-	0.003	0.572	-
BMB	(-0.117, 0.032)	(0.000, 0.911)	$<$	-	0.003	0.549	-
CATE - Low range		$n_A = 99$, $n_C = 105$					
BMMLR	(-0.078, -0.023)	(0.003, 0.317)	$<$	-	-0.034	0.200	-
BMLR	(-0.081, -0.016)	(0.004, 0.365)	$<$	-	-0.029	0.225	-
BMB	(-0.110, -0.036)	(0.019, 0.318)	-	-	-0.051	0.207	-
CATE - Mid-Low range		$n_A = 327$, $n_C = 334$					
BMMLR	(-0.090, 0.038)	(0.000, 0.884)	$<$	-	0.013	0.679	-
BMLR	(-0.092, 0.044)	(0.000, 0.937)	$<$	-	0.017	0.752	-
BMB	(-0.114, 0.045)	(0.001, 0.853)	$<$	-	0.013	0.642	-
CATE - Mid-High range		$n_A = 237$, $n_C = 252$					
BMMLR	(-0.139, 0.051)	(0.000, 0.992)	$<$ & $>$	-	0.013	0.753	-
BMLR	(-0.141, 0.054)	(0.000, 0.995)	$<$ & $>$	-	0.015	0.783	-
BMB	(-0.118, 0.047)	(0.006, 0.938)	$<$	-	0.014	0.694	-
CATE - High range		$n_A = 53$, $n_C = 40$					
BMMLR	(-0.183, 0.020)	(0.002, 0.980)	$<$	-	-0.021	0.100	-
BMLR	(-0.188, 0.021)	(0.001, 0.982)	$<$	-	-0.021	0.100	-
BMB	(-0.173, 0.019)	(0.069, 0.687)	-	-	-0.019	0.327	-
CATE - Low value							
BMMLR	(-0.078, -0.007)	(0.002, 0.440)	$<$	-	-0.021	0.291	-
BMLR	(-0.080, 0.000)	(0.002, 0.503)	$<$	-	-0.016	0.328	-
CATE - High value							
BMMLR	(-0.140, 0.052)	(0.000, 0.991)	$<$ & $>$	-	0.014	0.751	-
BMLR	(-0.142, 0.055)	(0.000, 0.994)	$<$ & $>$	-	0.015	0.777	-

Note. $\boldsymbol{\delta} = (\delta^{Strk7}, \delta^{Indep6})$, Pop $(\boldsymbol{\delta})$ = (Pop δ^{Strk7}, Pop δ^{Indep6})

5.6.2 Results

Results of different (sub)populations

Table 5.2 shows how different analysis models and different decision rules provide elaborate insights into the effects of Alteplase vs. control on a combination of dependent variables among different (sub)populations. Analysis of the selected data with the BMMLR, BMLR, and BMB models gave the following results.

Average treatment effects The average treatment effect (ATE) among the UK-based part of the trial population showed that the Alteplase group had a lower estimated probability of remaining free of strokes, a higher estimated probability of living independently, and a weighted probability difference close to zero. The three modeling procedures produced similar estimates and unanimously resulted in the conclusions that Alteplase was inferior according to the Any rule due to the effect on being stroke-free, while neither superiority nor inferiority could be concluded from the All or Compensatory rules.

Conditional average treatment effects The four conditional average treatment effects (CATEs) that reflected subpopulations as ranges sketched a more heterogeneous picture than the average treatment effects. Whereas all ranges showed a lower probability of being stroke-free after treatment with Alteplase, these effects varied over subpopulations. Differences between success probabilities of the two treatments appeared to increase with severity of the stroke, such that Alteplase appeared to have the largest negative effect on being stroke-free when the severity of the initial stroke was highest. A more diffuse relation between stroke severity and treatment difference emerged on long-term independent living.

Alteplase resulted in a slightly lower point estimate of the probability of independent living among patients with a Low stroke severity, but resulted in a higher estimated probability of independent living in all categories of more severe strokes. Patients in the Mid-Low and Mid-High ranges of stroke severity had the largest positive effect of Alteplase on independent living. The Low and High stroke severity patients had slightly higher weighted probabilities

after Alteplase compared to the control condition, while patients with a Mid-Low and Mid-High stroke severity had weighted probabilities close to zero.

These non-zero point estimates were not unanimously supported by sufficient evidence to conclude superiority or inferiority. The All and Compensatory rules remained inconclusive for all models among all subpopulations. The BMMLR and BMLR were unanimous in their conclusions for the Any rule: Inferiority was concluded for patients with a Low, Mid-Low and High stroke severity, while both superiority and inferiority were concluded for patients with a Mid-High range stroke severity. The BMB model remained inconclusive in the Low and High ranges and concluded inferiority among patients with a Mid-Low or Mid-High stroke severity, according to the Any rule.

The two conditional average treatment effects (CATEs) that specified subpopulations by values illustrated treatment differences for two hypothetical individual patients. After receiving Alteplase, both types of patients would have a lower probability of remaining free of strokes. Only the patient with a High stroke severity value had a higher probability of long-term independent living. The weighted failure probability difference was slightly below zero for the patient with a Low stroke severity and around zero for the patient with a High stroke severity. Again, the All and Compensatory rules remained inconclusive, whereas the Any rule would result in an inferiority conclusion for the patient with a Low stroke severity and in both inferiority and superiority for the patient with a High stroke severity.

Conclusions and discussion

Several conclusions regarding the BMMLR framework could be drawn from the presented results. First, multilevel analysis did not affect point estimates in the used subset of IST-3 data: BMMLR and BMLR models resulted in similar point estimates of $\boldsymbol{\delta}$ and $\delta(\mathbf{w})$, as expected from the negligible bias in the results of the simulation study. The posterior probabilities of the BMMLR and the BMLR model did not lead to different superiority or inferiority conclusions. These results suggest that the (substantively) clustered nature of this specific subset of data did not correspond to a relevant statistical dependence between observations within clusters. This

might raise the question whether a more restrictive single-level model could be chosen over a more complex multilevel model in absence of substantive reasons to use the latter. In hindsight, we conclude that the restrictions of a single-level model did not notably influence analysis results. Although estimates of random variances could be extracted from the analysis results to provide information about clustering, these estimates are not straightforward to interpret on the parameter scale of interest, namely success probabilities of individual treatments and differences between them. As these parameters are returned on the scale of the linear predictor, their meanings for clustering after (non-)linear transformation to the probability scale were not clear. It would be helpful to have information about clustering beforehand and we concluded that these results call for a proper method to quantify the degree of dependence among observations within clusters prior to the analysis. Such insights could help in clarifying the statistical urgency of a multilevel model and the appropriateness of a single-level model prior to the analysis.

Second, average treatment effects indicated an increased probability of recurrent events and a slightly decreased probability of long-term independent living after receiving the experimental treatment. However, different decision rules led to different conclusions. When the individual treatment effects had to be better on both dependent variables (All rule) or were weighted (Compensatory rule), no superiority or inferiority could be concluded. When any of the dependent variables had to demonstrate a relevant treatment difference (Any rule), both inferiority on recurrent events and superiority on long-term independent living could be concluded. This demonstrated a general potential problem with the Any rule: Contrasting decisions can result from the same analysis. Recall that the Any rule treats all outcome variables as equally important, raising the question of which conclusion to favor for patients in the Mid-High range or with a High value of severity. This problem does not occur with the other rules: The All and Compensatory rules are unambiguous in their conclusions.

Third, conditional (average) treatment effects suggested a trend in heterogeneity on the individual dependent variables that was not reflected by the average treatment effect. These trends were partially supported by superiority and/or inferiority decisions, depending on the

specified decision rule. Even without clear conclusions, conditional treatment effect sizes provided detailed insights: Considering average treatment effects only would have overlooked these trends. Further, the BMB model in the High range demonstrated that subgroup analysis can be a suboptimal approach to estimate conditional average treatment effects, as it can suffer from power loss. The High range subgroup is a relatively small fraction of the total sample size and performing an independent analysis on this group reduces the amount of evidence. This is reflected in the comparison to the BMMLR and BMLR methods: BMB has less extreme posterior probabilities, while treatment effect estimates are similar.

5.7 Discussion

The current paper presented the BMMLR framework as a multilevel extension to the Bayesian multivariate logistic regression (BMLR) analysis framework. The BMMLR framework consisted of three elements:

1. a Bayesian multilevel multivariate logistic regression model;
2. a transformation procedure to interpret results on the (multivariate) probability scale;
3. a statistical decision procedure to draw superiority and inferiority conclusions with targeted frequentist Type I errors

The presented framework accurately handled the multilevel structure of the data in the presence of heterogeneous treatment effects on multiple (correlated) binary dependent variables. A simulation study demonstrated that the proposed model indeed a) estimated average and conditional treatment effects in multilevel data without bias; and b) resulted in statistical decisions with targeted Type I error rates. A multilevel model was clearly superior for clustered data: Naive models that did not take the multilevel structure into account resulted in inflated Type I-error rates. Further, the logistic model promoted information-sharing between clusters and subpopulations, being a more powerful alternative than subgroup analysis to identify heterogeneous treatment effects.

A re-analysis of the IST-3 provided another perspective on the data than the original paper

by The International Stroke Trial-3 Collaborative Group (2012). Detailed insights as well as the varying treatment effects among subpopulations demonstrated the importance of a) a well-considered and specific decision rule; and b) the assessment of treatment heterogeneity. The statistical need for a multilevel model has not clearly become evident for this specific analysis. These results demonstrated that a substantive cluster structure in the data does not necessarily imply a relevant statistical dependence between observations. Gaining insight into the degree of dependence in the dataset is crucial for the choice between alternative models, since that would provide information about the statistical gains of a multilevel model before the data analysis.

Application of the BMMLR framework is not limited to the presented analyses. Theoretically, the model can be adapted to the longitudinal setting, may be used to borrow strength from different trials, or may be extended to data with multiple levels of clustering for example. In practice, such extensions require additional exploration of the (computational) properties of the model, since MCMC sampling procedures appeared sensitive to the amount of autocorrelation and the number of parameters. In a related fashion, carefully choosing which random effects to include is helpful for smooth execution of multilevel analysis. The model has a large number of options regarding specification of the model, giving a lot of flexibility to model cluster effects precisely. Naively including many random effects may not be a good idea: The increasing number of parameters intensifies computations notably and can complicate the translation to a substantially sensible and statistically rigorous model. Similarly, the multinomial setup is most suitable for a limited number of dependent variables. Increasing the number of dependent variables results in a large number of response categories, which may lead to sparsity issues.

Future research might advance the design of the BMMLR framework in multiple ways. First, a priori sample size computation and power analysis have priority in medical research. In line with our findings, larger numbers of clusters are known to be more powerful than larger numbers of subjects within clusters (Snijders, 2005). Expanding and refining knowledge regarding sample sizes in multilevel models aids in strategic experimental design (Moerbeek

et al., 2000, 2001; Raudenbush & Liu, 2000). Additionally, ethical aspects, such as risks and burden of (potentially inferior) treatment, and practical considerations, such as limited access to (large numbers of) subjects, require more in-depth understanding of power and sample sizes. Especially in precision medicine – where treatments are targeted at specific patient populations - numbers of eligible subjects are limited and a priori power analysis helps to manage expectations in terms of duration.

Second, the specification of prior distributions requires consideration. Specification of non-informative priors may not be trivial. The general tendency to choose relatively large variance parameters for normally distributed prior distributions (Gelman et al., 2008), does not necessarily work well with the proposed model. Covering a range far beyond realistic parameter values, can (negatively) affect the efficiency of the sampling procedure and even the resulting posterior distribution. Thus, concrete guidelines for the specification of non-informative priors would be helpful.

Third, pooling of treatment estimates can be done in several other ways than presented. In general, the pooled treatment effect over clusters is a weighted combination of cluster-specific estimates, where the weights aim to balance aspects that influence estimation and are imbalanced over clusters (e.g., cluster size or variance). Whereas we applied a cluster size-based approach, several advanced weighing procedures balance unequal variances within clusters via regularization methods (for overviews, see Gallo, 2000; Jones et al., 1998; Lin, 1999). These weighing methods generally produce shrinkage to the mean a) when group level variance is smaller; and/or b) when sample sizes are smaller (Gelman & Hill, 2007, p. 269). Such weighing procedures have interesting balancing properties but are probably less suitable for trials with clusters of single subjects, such as IST-3. These clusters have no variance, should not be discarded or merged inconsiderately, and call for the exploration of suitable weighing procedures for such data.

Finally, the BMMLR framework and multilevel models for discrete data in general lack a standard way to quantify the degree of clustering and the corresponding statistical need for a multilevel model. Often, the degree of clustering is quantified as the variance between clusters

relative to the variance within clusters, expressed via an intraclass correlation coefficient (ICC). The computation of ICCs in binary data is not straightforward: The variance within clusters - and therefore the ICC - is a function of the predictors in the model and the ICC depends on the prevalence, requiring an alternative approximation to obtain an appropriate estimate of the ICC (Goldstein et al., 2002; Gulliford et al., 2005; Paul et al., 2003; Ridout et al., 1999). We leave the extension of our framework in this direction for future research.

Acknowledgements

We thank Peter Sandercock on behalf of The International Stroke Trial-3 Collaborative Group for making the data from the Third International Stroke Trial publicly available.

Data and code availability

The Third International Stroke Trial data that support the findings of this study are available with the identifiers [https://doi.org/10.1016/S0140-6736(16)30414-7] and [http://doi.org/10.7488/ds/1350]. The `R` code used to generate results in the 5.5 and 5.6 sections can be found on GitHub https://github.com/XynthiaKavelaars/Bayesian-multilevel-multivariate-logistic-regression.

Chapter 6

Discussion

The current dissertation is built upon the idea that the increasing personalization of medicine requires novel research methodology for Randomized Controlled Trials (RCTs) to deal with more comprehensive research questions and complexities in trial data. Statistical methodology can be used to analyze multivariate datasets while, if present, taking a clustered structure into account and/or identifying potential heterogeneous treatment effects. Further, more refined decision strategies would improve alignment of trial conduct and treatment prescription in clinical practice. In this dissertation, we addressed the lack of such methodology and presented a unified multivariate Bayesian methodological framework that captures different combinations of these elements in a flexible manner. This framework focused on efficiently combining information from multiple binary outcome variables and predictor variables in RCTs and consisted of three components:

1. a multivariate modeling procedure that takes the correlation between outcome variables into account;
2. a transformation procedure to interpret model parameters more intuitively on the univariate and multivariate probability scale;
3. a decision procedure to flexibly define decision rules and draw superiority and inferiority conclusions in multivariate treatment comparisons.

Throughout the different chapters, three different multivariate modeling procedures were introduced that allow for the data to be analyzed with various properties, while appropriately controlling Type I error rates. First, the (conjugate) multivariate Bernoulli - Dirichlet model from Chapters 2 and 3 is a powerful and fast multivariate procedure to estimate average treatment effects and is most suitable for application among homogeneous study populations. Compared to more common alternatives, the model used the correlation between outcomes to improve Type II error rates and/or to reduce the sample size.

Second, when (multivariate) treatment effects differ over subpopulations, conditional treatment effects are relevant as well. In this case, the multivariate Bernoulli - Dirichlet model can be used for a multivariate variant of the common (univariate) subgroup analysis. This model performed suboptimally in terms of power as it evaluated treatment effects

among different subpopulations in isolation without using the relation between treatment effects and predictor variables. We presented a more powerful multivariate logistic regression model in Chapter 4 to combine the information from both multiple outcome variables and multiple subpopulations by modeling heterogeneity among multiple outcome variables directly.

Third, to accommodate data with a clustered structure, we developed a multilevel, multivariate logistic regression model in Chapter 5. This multilevel model allows combining the information from multiple outcome variables, multiple subpopulations, and multiple clusters of data in a single model with appropriate Type I error rates. The model provides extensive insights in multivariate treatment effects, is able to detect heterogeneous treatment effects, and takes non-independences of clustered observations into account in the estimation of uncertainty. Throughout the dissertation, we consistently observed how the multivariate nature of the presented models supported the flexible formulation of more refined decision rules, which allowed researchers to reflect upon and choose from a variety of superiority and inferiority definitions and to tailor them to the research question of interest. Chapters 2 and 3 explored statistical characteristics and substantive implications of three commonly used rules and elaborated on an alternative, compensatory decision rule that weighs and balances outcome variables according to their importance - and can be more efficient as a by-effect as well. Application to the (multilevel) multivariate logistic regression models in Chapters 4 and 5 consistently confirmed these findings. A safe conclusion is that choosing one of the default decision rules (Single, All, or Any) is not the best option per se in terms of both their substantive meanings and their statistical properties. The introduced Compensatory rule adds a substantial amount of flexibility compared to these defaults. Nonetheless, the choice of an appropriate decision rule is subject to trial-specific considerations that might demand a different rule compared to the presented ones.

Further, we explored sample size estimation in the multivariate context. The amount of required data is of major importance in medical research and sample size estimation is a delicate problem. Samples need to be large enough for treatments to demonstrate their

effects with sufficient power without being larger than needed. The latter extends the trial and unnecessarily exposes participants to unfavorable treatment conditions. We presented decision rule-specific sample size computations for the different multivariate Bernoulli-Dirichlet model in Chapter 3 and for the multivariate logistic regression model in Chapter 4. These formulas allow for power analysis prior to data collection and help to control Type II error rates. However, a major drawback of a priori sample size computations is their reliance on reasonably precise estimates of treatment effects and correlation coefficients, as demonstrated in Chapters 2 and 3. Any deviation from correct effect size estimates caused inaccuracies in a priori computed sample sizes and affected Type II error rates. In practice, it is difficult to estimate sample sizes accurately since power analysis is subject to multiple uncertain treatment effects together with estimates of the correlation between outcome variables (Rauch & Kieser, 2015). Especially the latter are difficult to estimate, even for experts in the field (O'Hagan et al., 2006; Zondervan-Zwijnenburg et al., 2017). Adaptive designs tolerate these (accumulated) uncertainties much better and are, thus, an attractive alternative to improve statistical power in multivariate analysis, as shown in Chapters 2 and 3. In adaptive designs, interim analyses are performed on incoming data to a) explore whether the trial can be stopped early or should be extended beyond the planned sample size and b) improve decision error rates and efficiency. The improved performance of adaptive designs does not make a priori sample size computations useless: They can serve as input for adaptive trial design and give a rough indication of the required number of subjects.

6.1 Limitations and suggestions for future research

Several aspects of the framework have yet to be critically evaluated. First, the simulation studies have consistently been performed with two outcome variables and—for the multivariate logistic regression models—a limited number of covariates. Larger numbers of variables are common in research practice. For example, the Cognition and Radiation Study B introduced in Chapter 1 reported eight different outcome measures in their protocol, some of which were measured repeatedly over time (Schimmel et al., 2018). Similarly, the Third

International Stroke Trial from Chapter 5 reported 20 clinically relevant characteristics to identify patients with high risks and low benefits from treatment (Lindley et al., 2015; The International Stroke Trial-3 Collaborative Group, 2012). Including several of such variables in analysis can be sensible. The performance with a larger number of variables remains to be investigated as this could lead to sparsity issues and can be expected to challenge the methodology computationally.

Second, efforts can be made to optimize computational aspects of the (multilevel) multivariate logistic regression models. For example, a comparison to other numerical procedures than the Gibbs sampling approach would be insightful. Despite being less tractable than Gibbs sampling algorithms, Metropolis-Hastings and Hamiltonian sampling algorithms can be viable alternatives that deserve exploration.

Third, the specification of prior information in the context of our multivariate framework deserves more attention, especially in the multivariate logistic regression models presented in Chapters 4 and 5. If desired and available, prior information can be included for substantive reasons in all presented models. We provided a few suggestions throughout the chapters but are aware that many applied research problems demand more advanced specifications and more extensive guidance. Research contexts that do not (primarily) intend to inform the analysis via the prior distribution also demand well-considered prior specification, even when default prior specifications are available (Berger, 2006; Depaoli & Van de Schoot, 2017). Particularly in small sample problems, the weight of the prior distribution can be more influential than anticipated on the posterior distribution (Gelman, 2006; Kass & Wasserman, 1996; McNeish, 2016; Smid & Winter, 2020). Chapter 3 showed an example of this effect, where Jeffreys non-informative prior specification differed from another non-informative prior in terms of point estimates and decision error rates in small samples. Computational stability is another relevant factor in prior specification, since prior variance parameters have the potential to (de)stabilize computation (Gelman et al., 2008; Schuurman et al., 2016). In Chapters 4 and 5, we deliberately specified prior covariance matrices with sufficient information to stabilize computations without affecting Type I error rates. A last point worth

elaborating on and investigating is the role of transformation. Seaman et al. (2012) noted that (logit) transformations can change the informativity of the prior unintendedly. In conclusion, naively choosing default (non-informative) prior parameters is not recommended and a better understanding of prior specification in the context of our presented framework would be a valuable addition.

Fourth, robustness checks and model validation approaches should be provided to assess whether the individual components of the framework are the right choices for the research problem at hand. Alternative modeling, transformation, or decision procedures may be more appropriate, as is outlined below. First, the underlying multinomial parametrization of the presented models is flexible in modeling heterogeneous correlations and estimating subgroup correlation coefficients. Several alternative approaches, such as, multivariate logistic and multivariate probit models (Chib, 1995; Malik & Abraham, 1973; O'Brien & Dunson, 2004), impose the marginal correlation structure on heterogeneous treatment effects. Although decision errors are theoretically affected by this restriction, the practical implications for statistical validity are currently not fully clear. If such alternatives appear sufficiently robust regarding heterogeneity in correlations, one of their advantages is that they generally have fewer parameters. Another class of multivariate modeling procedures is formed by copula structures, which can be complex in case of discrete data (Braeken et al., 2007; Chiu & Crump, 2012; Nikoloulopoulos & Karlis, 2008; Panagiotelis et al., 2012). A thorough comparison of modeling procedures can help to better assess their expected fit to a specific dataset and to choose the best option. Second, the choice between the (single-level) multivariate logistic model and the multilevel version of the model partly depends on the presence of clustering of observations (Hox et al., 2017). The degree of clustering is often expressed as an intraclass correlation coefficient, which quantifies the relative variance within and between groups. The computation of an intraclass correlation is not straightforward for binary outcome data, which has led to the construction of several estimators in the past. Differences between their performances can be substantial, especially since the intraclass correlation and the prevalence are related in binary outcome data (Gulliford et al., 2005; Paul

et al., 2003; Ridout et al., 1999). An intraclass correlation coefficient that is suitable for multivariate Bernoulli data would provide insight into the statistical need for a multilevel model.

Finally, adaptive designs have been developed to deal with uncertain estimates of treatment effects and have the potential to become a new standard in the future. They are not without caveats however (Bauer et al., 2016). As debated in three papers by Sanborn and Hills (2014), Rouder (2014), and Sanborn et al. (2014), and as shown in Chapter 3 it is possible to implement adaptive sample size determination in a suboptimal way. Monitoring each incoming observation can be attractive, for example, but naively performing unlimited numbers of interim analyses can lead to serious inflation of Type I error rates (Shi & Yin, 2019). Thus, adaptive designs need further investigation, optimization, and guidance before they are ready for large-scale implementation.

6.2 Concluding remarks

Despite several unanswered questions, the presented framework contributed to RCT methodology for personalized medicine. Taking the multivariate nature of treatment comparison as a starting point, we focused on the analysis of typical data structures and sought to align decision-making in RCTs with the highly personalized setting of clinical practice. Throughout the work, sharing information between outcome variables and subpopulations consistently demonstrated an increase in the value of RCTs in the personalization of medicine by a) borrowing strength from other variables, b) enabling more refined decisions that align research with clinical practice and c) creating more comprehensive insights into heterogeneous treatment effects.

Our work touched upon certain general aspects of decision-making in the clinical context that are open to debate. First, the current dissertation addressed two applications of RCT results: drug admission and treatment prescription in practice. This twofold goal raises practical questions and causes potential friction in the design of the decision procedure. Different stakeholders within and beyond the research process can have different decision

criteria that should be dealt with. Whereas treatment developers and regulatory bodies are primarily concerned with the general safety and efficacy, end-users (e.g., clinicians and patients) can prefer to weigh beneficial effects and side effects differently, in line with professional and personal visions. Especially for treatment prescription, it is worth looking into developing and advancing user-friendly ways to integrate clinical expertise in decision-making. The value of clinical intuition in treatment prescription should not be overlooked or overwhelmed by formal procedures, but can be challenging to incorporate into such a procedure (Lek & Van de Schoot, 2018; O'Hagan et al., 2006; Rietbergen et al., 2014; Zondervan-Zwijnenburg et al., 2017). Involving end users in decision-making raises the questions a) whether they should apply different decision procedures compared to drug developers in the test phase and b) which information should be available to whom and in which form to make the decision procedure suitable for both drug admission and treatment prescription while being sufficiently user-friendly and robust to misinterpretation.

Second, the multivariate logistic regression models in Chapters 4 and 5 allow for interpolation and extrapolation to unobserved values and ranges of covariates. Whether such prediction models provide sufficiently strong evidence to have a place in drug admission is debatable. At the very least, model validation approaches should verify that the linear predictor holds beyond the observed covariate range. This is not only relevant for interpolation and extrapolation: The entire model relies on the assumption of linearity on the log odds. This assumption can be difficult to verify, especially in (un)observed tails of the covariate distribution where data to validate the model are sparse or even absent. It is plausible that predictions—at least—have an intuitive place in treatment prescription even if they do not meet the high standards of drug admission procedures. In absence of observed data and other forms of strong evidence for a specific patient under consideration, prediction models provide an educated guess of the treatment effects for this patient and can serve as guidance to formalize a clinician's decision process. Such a procedure might open doors to a more formal form of shared decision-making for which the clinician and patient jointly prioritize outcomes and apply this jointly weighted decision rule to choose a particular

treatment.

Finally, while the current dissertation focused on more efficient handling of information within the trial, some of the abovementioned discussion points move in the direction of a more integrative approach for which information external to the trial informs design and/or analysis. Although there are sensible reasons to base decisions about drug admission on individual trials, a more integrative approach incorporating external sources of information offers high potential to further support personalized treatments with high-quality evidence. Currently, there are several ways to statistically combine information from multiple trials and borrow strength from external data, such as meta-analytic approaches to combine aggregated trial results, incorporation of prior information from historical trials or experts, or hierarchical modeling approaches to combine raw data (Viele et al., 2014). Proper handling of exchangeability of trials is a prerequisite for information sharing across trials and has recently received attention in the context of master protocols, but it has not been developed into a settled approach yet (see e.g., Hobbs & Landin, 2018; Michael et al., 2020; Neuenschwander et al., 2015). The roles, the potentials, and the limits of current and future methods need to be considered in the light of the demanded statistical rigor to prepare the medical field for principled methods for integration of external information.

Appendix A

Specification of prior hyperparameters

We might facilitate specification of hyperparameters when we consider the prior distribution of joint success probabilities $\boldsymbol{\theta}_j$ rather than the prior distribution of joint response probabilities $\boldsymbol{\phi}_j$. Here we can utilize the facts that 1) the multivariate beta distribution is a transformation of the Dirichlet distribution; and 2) the parameters of the two distributions are identical (Olkin & Trikalinos, 2015).

We present the transformation for $K = 2$, such that $Q = 2^K = 4$. The Dirichlet distribution with hyperparameters $\boldsymbol{\alpha}_j{}^0 = (\alpha^0_{j,11\ldots11}, \alpha^0_{j,11\ldots10}, \ldots, \alpha^0_{j,00\ldots01}, \alpha^0_{j,00\ldots00})$ has the following form:

$$\begin{aligned} p(\boldsymbol{\phi}_j|\boldsymbol{\alpha}^0_j) =& \text{Dirichlet}(\boldsymbol{\phi}_j|\boldsymbol{\alpha}^0_j) \\ \propto& \phi_{j,1\ldots11}^{\alpha^0_{j,1\ldots11}-1}\phi_{j,1\ldots10}^{\alpha^0_{j,1\ldots10}-1} \times \cdots \times \phi_{j,0\ldots01}^{\alpha^0_{j,0\ldots01}-1}\phi_{j,0\ldots00}^{\alpha^0_{j,0\ldots00}-1}. \end{aligned} \tag{A.1}$$

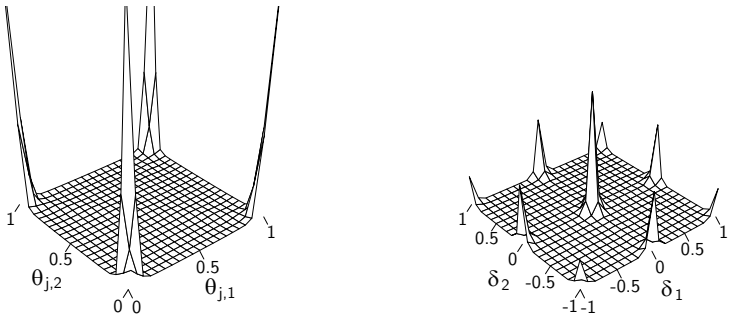
Reparametrizing $\boldsymbol{\phi}_j$ in terms of $\boldsymbol{\theta}_j$ and integrating $\phi_{j,11}$ out transforms the Dirichlet distribution of posterior $\boldsymbol{\phi}_j$ to a multivariate beta posterior distribution of success probabilities $\boldsymbol{\theta}_j$ (Olkin & Trikalinos, 2015):

$$\begin{aligned} p(\boldsymbol{\theta}_j|\boldsymbol{\alpha}^n_j) =& \frac{1}{B(\boldsymbol{\alpha}^n_j)} \int_\Omega \phi_{j,11}^{\alpha^n_{j,11}-1} \times (\theta_{j,1} - \phi_{j,11})^{\alpha^n_{j,10}-1} \times \\ & (\theta_{j,2} - \phi_{j,11})^{\alpha^n_{j,01}-1} \times (1 - \theta_{j,1} - \theta_{j,2} + \phi_{j,11})^{\alpha^n_{j,00}-1} \partial\phi_{j,11}, \end{aligned} \tag{A.2}$$

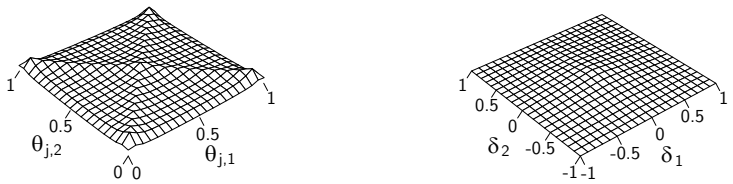
where $\Omega = \phi_{j,11} : \max(0, \theta_{j,1} + \theta_{j,2} - 1) < \phi_{j,11} < \min(\theta_{j,1}, \theta_{j,2})$.

Note that prior $\boldsymbol{\theta}_j$ also follows a multivariate beta distribution. However, the multivariate beta distribution cannot be formally used as a prior distribution in posterior computation, since the distribution is marginalized with respect to the information about the relation between success probabilities in $\phi_{j,kl}$.

Let us further redefine $\boldsymbol{\alpha}^0_j$ as $n^0_j\boldsymbol{\phi}^0_j$ to provide an intuitive specification of prior information. Here n^0_j reflect the amount of prior information and $\boldsymbol{\phi}^0_j$ reflects the prior means of joint response probabilities $\boldsymbol{\phi}_j$. Prior means $\boldsymbol{\phi}^0_j$ relate directly to the prior means of joint success probabilities $\boldsymbol{\theta}^0_j$ and the prior mean treatment difference $\boldsymbol{\delta}^0$, since $\theta^0_{j,k}$ equals the sum of all elements of $\boldsymbol{\phi}^0_j$

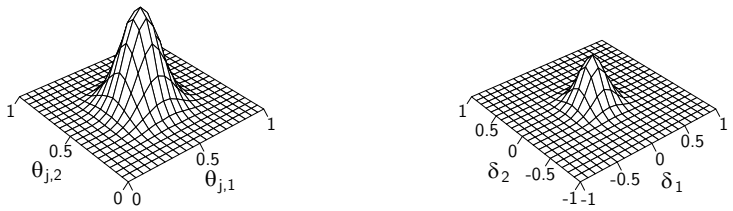


(a) $\boldsymbol{\alpha}_j^0 = (0.01, 0.01, 0.01, 0.01)$; $n_j^0 = 0.04$



(b) $\boldsymbol{\alpha}_j^0 = (0.5, 0.5, 0.5, 0.5)$; $n_j^0 = 2$

(c) $\boldsymbol{\alpha}_j^0 = (1, 1, 1, 1)$; $n_j^0 = 4$



(d) $\boldsymbol{\alpha}_j^0 = (5, 5, 5, 5)$; $n_j^0 = 20$

Figure A.1: Bivariate prior distributions of $\boldsymbol{\theta}_j$ (left) and $\boldsymbol{\delta} = \boldsymbol{\theta}_E - \boldsymbol{\theta}_C$ (right) for various specifications of hyperparameters $\boldsymbol{\alpha}_j^0$ and n_j^0 when $K = 2$. Prior response probability $\boldsymbol{\phi}_j^0 = \frac{1}{4}$.

with the kth element of response combination q equal to 1 and $\delta_k^0 = \theta_{E,k}^0 - \theta_{C,k}^0$. The following paragraph lists the influence of hyperparameters $n_j^0 \boldsymbol{\phi}_j^0$ on the shape of the prior distributions of success probabilities $\boldsymbol{\theta}_j^0$ and treatment differences $\boldsymbol{\delta}$.

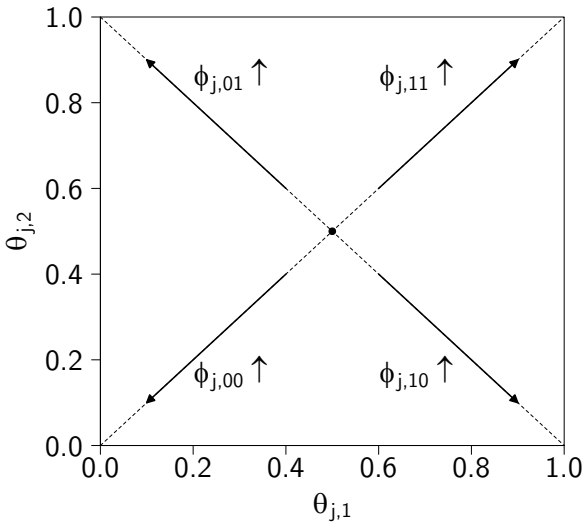


Figure A.2: The influence of $\boldsymbol{\phi}_j$ on the bivariate beta distribution of $\boldsymbol{\theta}_j$ for two outcomes ($K = 2$).

1. The amount of prior information in n_j^0 determines the spread of the prior distribution of treatment j, as visualized in Figure A.1. Large n_j^0 results in a peaked distribution that reflects more prior information, whereas small n_j^0 results in a distribution with heavy tails that conveys little prior information. Parameter n_j^0 can be considered a prior sample size, where each observation has the same influence on the decision as one joint response $\mathbf{x}_{j,i}$.
2. Mean prior success probabilities $\boldsymbol{\theta}_{j,k}^0$ define the center of the prior distribution of success probabilities $\boldsymbol{\theta}_j$, as visualized in Figure A.2. Similarly, mean prior treatment differences $\boldsymbol{\delta}_k^0$ reflect the center of the prior distribution of $\boldsymbol{\delta}$. When $n_E^0 = n_C^0$ and $\boldsymbol{\phi}_E^0 = \boldsymbol{\phi}_C^0$, the prior distribution of the treatment difference $\boldsymbol{\delta}$ is centered around the origin (i.e., $\boldsymbol{\delta}^0 = \mathbf{0}$).
3. The size of $\phi_{j,kl}^0$ relative to $\theta_{j,k}^0 \theta_{j,l}^0$ determines the prior correlation between $\theta_{j,k}$ and $\theta_{j,l}$

(Olkin & Trikalinos, 2015):

$$\rho_{\theta_{j,k}\theta_{j,l}} = \frac{\phi^0_{j,kl} - \theta^0_{j,k}\theta^0_{j,l}}{\sqrt{\theta^0_{j,k}(1-\theta^0_{j,k})\theta^0_{j,l}(1-\theta^0_{j,l})}}. \tag{A.3}$$

As follows from Equation A.3, $\theta_{j,k}$ and $\theta_{j,l}$ are independent a priori if $\phi^0_{j,kl} = \theta^0_{j,k}\theta^0_{j,l}$. When $\phi^0_{j,kl} > \theta^0_{j,k}\theta^0_{j,l}$, parameters are positively correlated, while parameters are negatively correlated when $\phi^0_{j,kl} < \theta^0_{j,k}\theta^0_{j,l}$.

When prior information is known, the introduced properties can be used to make informative choices regarding these parameters. In absence of prior information however, a reasonable prior distribution of success probabilities $\boldsymbol{\theta}_j$ has 1) a small n^0_j, such that the impact on the decision is limited; and 2) $\theta_{j,k} = \frac{1}{2}$, such that successes and failures are equally likely a priori for all treatments j and all outcomes k. Although mathematically straightforward, we remark that estimating prior hyperparameters in practice can be challenging when K is large.

Appendix B

Specification of efficiency weights

Finding weights that maximize efficiency requires maximizing the following function with respect to $\mathbf{w}$:

$$\sum_{\mathbf{s}_E^*,\mathbf{s}_C^*} P(\boldsymbol{\delta} \in \mathcal{S}_{Compensatory}(\mathbf{w})|\mathbf{s}_E^*, \mathbf{s}_C^*) \times P(\mathbf{s}_E^*, \mathbf{s}_C^*|\boldsymbol{\theta}_E^T, \boldsymbol{\theta}_C^T, \boldsymbol{\rho}_{\delta_k,\delta_l}^T) \tag{B.1}$$
$$= P(\boldsymbol{\delta} \in \mathcal{S}_{Compensatory}(\mathbf{w})|\boldsymbol{\theta}_E^T, \boldsymbol{\theta}_C^T, \boldsymbol{\rho}_{\delta_k,\delta_l}^T)$$

where $\mathbf{s}_E^*$ and $\mathbf{s}_C^*$ are the anticipated response frequencies before data collection

$\boldsymbol{\theta}_E^T$ and $\boldsymbol{\theta}_C^T$ are the treatment effects in the population

$\boldsymbol{\rho}_{\delta_k,\delta_l}^T$ is the correlation between δ_k and δ_l in the population.

No analytical solution for Equation B.1 exists. We can however obtain a solution for $\mathbf{w}$ using large sample theory, which dictates that the posterior distribution of $\boldsymbol{\delta}$ can be approximated with a multivariate normal distribution in case of a sufficiently large sample:

$$\boldsymbol{\delta} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{B.2}$$

where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ and

$\boldsymbol{\Sigma}$ has diagonal elements $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_K^2)$ and off-diagonal elements $\boldsymbol{\sigma}_{kl}$.

Consequently, the linear combination $\sum_{k=1}^K w_k\delta_k$ has an approximate normal posterior distribution with mean $\sum_{k=1}^K w_k\mu_k$ and variance $\sum_{k=1}^K w_k^2\sigma_k^2 + 2\sum\sum_{k<l} w_k w_l \sigma_{kl}$. The probability that $\sum_{k=1}^K w_k\delta_k > 0$ then follows from the cumulative normal distribution:

$$P(\sum_{k=1}^K w_k\delta_k > 0) = 1 - \Phi\left(\frac{0 - \sum_{k=1}^K w_k\mu_k}{\sqrt{\sum_{k=1}^K w_k^2\sigma_k^2 + 2\sum\sum_{k<l} w_k w_l \sigma_{kl}}}\right). \tag{B.3}$$

Weights $\mathbf{w}$ that maximize the probability in Equation B.3 result in maximal efficiency. In practice, computing efficient weights is less straightforward since $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown. To facilitate the choice of these parameters for the construction of a normal posterior distribution of $\boldsymbol{\delta}$, we may consider hypothetical datasets of expected joint response frequencies $\mathbf{s}_j^*$ for both treatments j. These frequencies can be used to obtain a sample of $\boldsymbol{\delta}$, that is assumed to follow

a normal distribution when the sample size n_j and the number of draws are sufficiently large. Such a sample provides estimates of $\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$ and $\boldsymbol{\sigma}_{kl}$, that can be plugged in in Equation B.3. We provide an example data configuration for $K = 2$ in Table B.1. This hypothetical dataset would result in $\boldsymbol{\mu} = (0.24, 0.08)$, $\boldsymbol{\sigma}^2 = (0.005, 0.005)$ and $\sigma_{12} = -0.001$, such that optimal weights equal $\mathbf{w} = (0.64, 0.36)$.

Table B.1: Example configuration of anticipated joint response frequencies $\mathbf{s}_E^*$ and $\mathbf{s}_C^*$ for approximation of $\boldsymbol{\mu}$,$\boldsymbol{\sigma}^2$, and σ_{12} for two outcomes.

	$s^*_{E,1} = 1$	$s^*_{E,1} = 0$		$s^*_{C,1} = 1$	$s^*_{C,1} = 0$
$s^*_{E,2} = 1$	262	278	$s^*_{C,2} = 1$	102	358
$s^*_{E,2} = 0$	358	102	$s^*_{C,2} = 0$	278	262

The procedure to find efficient weights simplifies when treatment differences are uncorrelated, i.e., when $\sigma_{kl} = 0$. Maximum evidence is then obtained when weights $\mathbf{w}$ are proportional to treatment difference $\boldsymbol{\delta}$ and $\sigma_k^2 = \sigma_l^2$. For example, when $\boldsymbol{\delta} = (0.30, 0.10)$, weights $\mathbf{w} = (0.75, 0.25)$ are optimal.

Appendix C

Implementation of the framework in group sequential and adaptive designs

The current appendix presents an algorithm with the procedure to arrive at a decision using the multivariate analysis procedure for a group sequential or adaptive design.

Algorithm 2 Decision procedure for a group sequential or adaptive design

1 <u>**Initialize**</u>

a Choose decision rule

if Compensatory **then** specify weights $\mathbf{w}$

if Single **then** specify k

end if

for each treatment $j \in \{E, C\}$ **do**

b Choose prior hyperparameters $\boldsymbol{\alpha}_j^0$

end for

c Choose Type I error rate α and power $1 - \beta$

d Choose number of interim analyses M

e Determine decision threshold p_{cut}

f Determine vector of sample sizes $\mathbf{n}_j^{(\cdot)}$ of length M

if group sequential design **then** $\mathbf{n}_j^{SD} = n_j^{FD} \times \mathbf{n}_{ratio}$, where n_j^{FD} reflects the required sample size for a fixed design and $\mathbf{n}_{ratio}$ reflects M proportions of the final sample size at which to perform interim analyses

else if adaptive design **then** define $\mathbf{n}_j^{AD}$ according to desired monitoring scheme

end if

2 <u>**Perform interim analyses**</u>

$m \leftarrow 0$

repeat

a $m \leftarrow m + 1$

b $n_j \leftarrow m^{th}$ element of $\mathbf{n}_j^{(\cdot)}$

c Collect data and evaluate evidence via Step 2 of Algorithm 1

until $m = M$ **or** $P(\boldsymbol{\delta} \in \mathcal{S}_{Sup} | \mathbf{s}_E, \mathbf{s}_C) > p_{cut}$

3 <u>**Make final decision**</u>

if $P(\boldsymbol{\delta} \in \mathcal{S}_{Sup} | \mathbf{s}_E, \mathbf{s}_C) > p_{cut}$ **then** conclude superiority

else conclude non-superiority

end if

Appendix D

Numerical evaluation: Comparison of trial designs

The Section *Numerical evaluation* (3.4) showed how accurate decision error rates could be obtained with the proposed framework under a fixed design. However, the realization of adequate error rates and efficient decisions depends on the accuracy of sample sizes, and hence on adequate estimates of anticipated treatment differences and correlations. A design based on interim analyses might improve statistical inference under parameter uncertainty, which is especially relevant for the estimation of multiple parameters in multivariate analysis (Berry et al., 2010; Jennison & Turnbull, 1999). Such trials monitor incoming data and terminate data collection as soon as evidence exceeds a prespecified decision threshold. In the current paper, we make a sharp distinction between two of these design types: adaptive and group sequential designs. Here, adaptive designs evaluate the data according to an interim monitoring scheme that does not rely on parameter estimates. Such a monitoring scheme and a decision threshold suffice to start data collection. These designs allow for both early and late termination if the treatment effect appears larger or smaller than anticipated respectively. On the downside, efficiency may be compromised if the number of interim analyses high: A (very) strict decision threshold is then needed to the control Type I error rate under repeated decision-making (Rouder, 2014; Sanborn & Hills, 2014; Shi & Yin, 2019).

In contrast to adaptive designs, group sequential designs rely on anticipated parameters to estimate a maximum sample size in advance, which is often similar to the sample size of a fixed design. Interim analyses are performed at (a limited number of) prespecified proportions of this sample size to allow for early termination. The potential for late termination is limited with such a setup (Jennison & Turnbull, 1999). To control Type I error rates adequately, decision threshold p_{cut} should be adjusted to the number of interim analyses, and may differ per interim analysis (Jennison & Turnbull, 1999; Shi & Yin, 2019). In practice the distinction between group sequential and adaptive designs is less sharp than presented here: Expectations about parameters often (roughly) inform the monitoring scheme in adaptive stopping to limit the number of interim analyses, while conservative parameter estimates allow for late termination of group sequential trials.

In the current appendix, we demonstrate how 1) error rates are influenced by uncertainty

about parameters in a priori sample size estimation; and 2) designs with and without interim analyses perform under this uncertainty. We considered seven different designs:

1. A fixed design with sample size n_j^{FD} computed with three different anticipated treatment differences $\boldsymbol{\delta}^n$:
   (a) True treatment differences ($\boldsymbol{\delta}^n = \boldsymbol{\delta}^T$)
   (b) Overestimated treatment differences ($\boldsymbol{\delta}^n = \boldsymbol{\delta}^T + (0.10, 0.10)$)
   (c) Underestimated treatment differences ($\boldsymbol{\delta}^n = \boldsymbol{\delta}^T - (0.10, 0.10)$)
2. A group sequential design with a maximum of $M = 3$ analyses, evaluated at sample sizes $\mathbf{n}_j^{SD}$ or until superiority is concluded. These sample sizes are computed with three different anticipated treatment differences $\boldsymbol{\delta}^n$:
   (a) True treatment differences ($\boldsymbol{\delta}^n = \boldsymbol{\delta}^T$)
   (b) Overestimated treatment differences ($\boldsymbol{\delta}^n = \boldsymbol{\delta}^T + (0.10, 0.10)$)
   (c) Underestimated treatment differences ($\boldsymbol{\delta}^n = \boldsymbol{\delta}^T - (0.10, 0.10)$)
3. An adaptive design with a maximum of $M = 136$ analyses, evaluated at sample sizes $n_j^{AD} = n_{j,1}, \dots, n_{j,M}$, or until superiority is concluded. The first interim analysis is performed at $n_{j,1}^{AD} = 5$ and monitors every observation until 50 observations have been made. Then the interim group size increases to 5 until $n_{j,M}^{AD} = 500$, such that $\mathbf{n}_j^{AD} = (5, 6, \dots, 49, 50, 55, \dots, 500)$.

Group sequential design We set up a group sequential design with equally spaced interim analyses using the `gsDesign` package (Anderson, 2016). We based computations of interim sample sizes $\mathbf{n}_j^{SD}$ and interim decision threshold p_{cut}^{SD} on the sample size of a fixed design, n_j^{FD}, using the default settings of the `gsDesign()` function for a one-sided test with $\alpha = .05$ and $\beta = .20$.

Adaptive design Decision threshold p_{cut}^{AD} was calibrated to reflect the desired Type I error rate α. The procedure involves repeatedly evaluating a large number of simulated samples from the distributions of least favorable values at sample sizes $\mathbf{n}_j^{AD}$ at different values of p_{cut},

and selecting the decision threshold for which the empirical Type I error rate corresponds to α.

Data generation and evaluation We generated 5,000 samples to compare the seven trial designs for the Compensatory decision rule with equal weights (Comp-E; $\mathbf{w} = (0.50, 0.50)$) and an uninformative prior distribution ($\boldsymbol{\alpha}_j^0 = (0.01, 0.01, 0.01, 0.01)$). We used decision thresholds $p_{cut}^{FD} = 0.95$, $p_{cut}^{SD} = 0.98$, and $p_{cut}^{AD} = 0.9968$, The generated datasets were evaluated using the procedure in Algorithm 2.

D.1 Results

Tables D.1, D.2, and D.3 present the results of the comparison of designs. The performance of fixed and group sequential designs depends on the correspondence between parameter estimates for sample size estimation and the true parameters. When parameters were specified correctly (i.e., $\boldsymbol{\delta}^n = \boldsymbol{\delta}^T$), both designs resulted in a satisfactory Type I error rate and power (Table D.1). The group sequential design was generally more efficient than the fixed design (Table D.2). In this situation, the adaptive design was less efficient than the other designs, since 1) trials are free to continue until superiority has been concluded, resulting in a high (but uncontrolled) power at the expense of a larger sample size; and 2) the increased number of interim analyses in an adaptive design requires a higher decision threshold, which accompanies - on average - a larger average sample size to conclude superiority.

The benefits of interim analyses in terms of decision error rates and efficiency were particularly apparent when anticipated treatment differences did not correspond to the true treatment difference. A group sequential design is mainly advantageous over a fixed design when sample sizes were based on underestimated treatment differences (i.e., $\boldsymbol{\delta}^n < \boldsymbol{\delta}^T$): Probabilities to conclude superiority correctly were well above the planned .80 in both designs, but the group sequential design is more efficient. The adaptive design was especially powerful when anticipated treatment differences were overestimated (i.e., $\boldsymbol{\delta}^n > \boldsymbol{\delta}^T$): Both the fixed and group sequential design had a limited power to conclude superiority.

While the adaptive and group sequential designs outperformed the fixed design in terms of power and efficiency under parameter uncertainty, they do result in upward bias (Table D.3). This effect can be attributed to two different aspects of these designs. First, the effect is most apparent when the sample size is underestimated ($\boldsymbol{\delta}^n > \boldsymbol{\delta}^T$) . Here, the studies that conclude superiority are those with early stops, *because* their effect size at the termination point is larger than the true effect size (Emerson et al., 2007). The effect size is then averaged over a selection of the (upper part of the) sampling distribution of treatment differences. When sample sizes are sufficiently large to allow for timely ($\boldsymbol{\delta}^n = \boldsymbol{\delta}^T$) or late ($\boldsymbol{\delta}^n < \boldsymbol{\delta}^T$) terminations, these high effect sizes are partially compensated by the samples with smaller effect sizes (Schönbrodt et al., 2017).

Second, when naively pooled, average effect sizes from trials that stopped early for efficacy are affected by instability of treatment effects early in data collection (Goodman, 2007; Schou & Marschner, 2013; Senn, 2014; Zhang et al., 2012). With few data points, new observations are quite influential resulting in large variation around the treatment effect in the sample. In contrast, the treatment effect estimate stabilizes as data accumulate (Zhang et al., 2012). Trials that allow for early efficacy stopping exploit the variation of small samples to stop trials at extreme values, but only from the upper tail of the distribution. Extreme treatment effects in the other direction - indicating treatment futility - are given the opportunity to regress to the mean by adding new observations. Pooling these extreme values from early stops with the stabilized values from later stops then results in an overestimated treatment effect. Since our adaptive design has more interim analysis in the beginning of data collection, the adaptive design has more opportunities to include extreme values, resulting in a larger bias compared to the group sequential design.

Table D.1: P(Conclude superiority) for different trial designs (AD = adaptive design, FD = fixed design, SD = group sequential design) and anticipated treatment differences ($\boldsymbol{\delta}^n$) after applying the Compensatory decision rule with equal weights.

DGM	AD	FD			SD		
		$\boldsymbol{\delta}^n = \boldsymbol{\delta}^T$	$\boldsymbol{\delta}^n < \boldsymbol{\delta}^T$	$\boldsymbol{\delta}^n > \boldsymbol{\delta}^T$	$\boldsymbol{\delta}^n = \boldsymbol{\delta}^T$	$\boldsymbol{\delta}^n < \boldsymbol{\delta}^T$	$\boldsymbol{\delta}^n > \boldsymbol{\delta}^T$
1.1	0.001	0.000	0.000	0.000	0.000	0.000	0.000
1.2	0.004	0.000	0.000	0.000	0.000	0.000	0.000
1.3	0.002	0.000	0.000	0.000	0.000	0.000	0.000
2.1	0.047	0.049	0.047	0.052	0.053	0.045	0.051
2.2	0.034	0.056	0.050	0.046	0.046	0.047	0.046
2.3	0.031	0.049	0.046	0.056	0.051	0.054	0.050
3.1	0.990	0.807	1.000	0.366	0.796	1.000	0.389
3.2	0.934	0.806	1.000	0.357	0.795	1.000	0.367
3.3	0.825	0.800	1.000	0.345	0.804	1.000	0.343
4.1	1.000	0.811	1.000	0.545	0.832	1.000	0.615
4.2	1.000	0.813	0.999	0.520	0.810	1.000	0.569
4.3	1.000	0.804	1.000	0.514	0.806	0.999	0.534
5.1	1.000	0.881	0.975	0.837	0.925	0.982	0.899
5.2	1.000	0.831	0.967	0.693	0.871	0.967	0.790
5.3	1.000	0.809	0.958	0.696	0.842	0.969	0.753
6.1	1.000	0.824	1.000	0.552	0.835	1.000	0.633
6.2	1.000	0.805	1.000	0.512	0.819	0.999	0.564
6.3	1.000	0.801	0.999	0.514	0.799	1.000	0.541
7.1	0.007	0.000	0.000	0.000	0.000	0.000	0.000
7.2	0.010	0.000	0.000	0.000	0.000	0.000	0.000
7.3	0.007	0.000	0.000	0.000	0.000	0.000	0.000
8.1	1.000	0.808	1.000	0.494	0.811	1.000	0.538
8.2	1.000	0.804	1.000	0.464	0.808	1.000	0.497
8.3	1.000	0.805	1.000	0.461	0.802	1.000	0.471

Table D.2: Average sample size to correctly conclude superiority for different trial designs (AD = adaptive design, FD = fixed design, SD = group sequential design) and anticipated treatment differences ($\boldsymbol{\delta}^n$) after applying the Compensatory decision rule with equal weights. Data generating mechanisms with a hyphen should not result in treatment superiority.

DGM	AD	FD			SD		
		$\boldsymbol{\delta}^n = \boldsymbol{\delta}^T$	$\boldsymbol{\delta}^n < \boldsymbol{\delta}^T$	$\boldsymbol{\delta}^n > \boldsymbol{\delta}^T$	$\boldsymbol{\delta}^n = \boldsymbol{\delta}^T$	$\boldsymbol{\delta}^n < \boldsymbol{\delta}^T$	$\boldsymbol{\delta}^n > \boldsymbol{\delta}^T$
1.1	-	-	-	-	-	-	-
1.2	-	-	-	-	-	-	-
1.3	-	-	-	-	-	-	-
2.1	-	-	-	-	-	-	-
2.2	-	-	-	-	-	-	-
2.3	-	-	-	-	-	-	-
3.1	159	108	1000	26	91	354	25
3.2	211	154	1000	38	130	403	37
3.3	243	199	1000	49	169	452	48
4.1	39	26	108	11	20	53	9
4.2	60	38	154	16	31	77	15
4.3	80	49	199	21	41	101	20
5.1	9	6	11	4	4	6	3
5.2	14	9	16	5	7	10	4
5.3	18	11	21	7	8	14	6
6.1	37	25	103	11	19	50	9
6.2	57	36	147	15	30	74	14
6.3	76	47	191	20	39	96	19
7.1	-	-	-	-	-	-	-
7.2	-	-	-	-	-	-	-
7.3	-	-	-	-	-	-	-
8.1	62	41	298	15	34	114	13
8.2	94	59	426	22	49	166	21
8.3	123	76	553	28	65	214	27

Table D.3: Average bias for different trial designs (AD = adaptive design, FD = fixed design, SD = group sequential design) and anticipated treatment differences ($\boldsymbol{\delta}^n$) after applying the Compensatory decision rule with equal weights.

DGM	AD	FD			SD		
		$\boldsymbol{\delta}^n = \boldsymbol{\delta}^T$	$\boldsymbol{\delta}^n < \boldsymbol{\delta}^T$	$\boldsymbol{\delta}^n > \boldsymbol{\delta}^T$	$\boldsymbol{\delta}^n = \boldsymbol{\delta}^T$	$\boldsymbol{\delta}^n < \boldsymbol{\delta}^T$	$\boldsymbol{\delta}^n > \boldsymbol{\delta}^T$
1.1	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(-0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)
1.2	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(-0.00, -0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)
1.3	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, -0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)
2.1	(0.01, 0.01)	(-0.00, -0.00)	(0.00, 0.00)	(-0.00, 0.00)	(0.00, -0.00)	(0.00, -0.00)	(-0.00, 0.00)
2.2	(0.01, 0.01)	(0.00, 0.00)	(0.00, -0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(-0.00, 0.00)
2.3	(0.01, 0.01)	(-0.00, -0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, -0.00)	(0.00, 0.00)
3.1	(0.05, 0.05)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.01, 0.01)	(0.00, 0.00)	(0.02, 0.01)
3.2	(0.05, 0.05)	(0.00, 0.00)	(0.00, 0.00)	(0.00, -0.00)	(0.01, 0.01)	(0.00, 0.00)	(0.01, 0.01)
3.3	(0.05, 0.05)	(0.00, 0.00)	(0.00, 0.00)	(-0.00, -0.00)	(0.01, 0.01)	(0.01, 0.01)	(0.01, 0.01)
4.1	(0.07, 0.07)	(0.00, 0.00)	(0.00, 0.00)	(0.00, -0.00)	(0.03, 0.03)	(0.01, 0.01)	(0.04, 0.05)
4.2	(0.07, 0.08)	(0.00, 0.00)	(0.00, 0.00)	(-0.00, -0.00)	(0.03, 0.03)	(0.01, 0.01)	(0.04, 0.04)
4.3	(0.08, 0.08)	(0.00, 0.00)	(0.00, 0.00)	(-0.00, 0.00)	(0.02, 0.02)	(0.01, 0.01)	(0.02, 0.03)
5.1	(0.04, 0.04)	(0.00, -0.01)	(0.00, 0.00)	(-0.01, 0.00)	(0.03, 0.04)	(0.04, 0.04)	(0.06, 0.06)
5.2	(0.07, 0.07)	(-0.01, -0.01)	(0.00, 0.01)	(-0.00, 0.00)	(0.05, 0.04)	(0.04, 0.04)	(0.08, 0.08)
5.3	(0.09, 0.09)	(0.00, -0.01)	(0.00, 0.00)	(0.00, 0.00)	(0.06, 0.06)	(0.04, 0.04)	(0.07, 0.07)
6.1	(0.06, 0.07)	(0.00, 0.00)	(0.00, -0.00)	(-0.01, 0.01)	(0.03, 0.03)	(0.01, 0.02)	(0.04, 0.06)
6.2	(0.07, 0.07)	(0.00, -0.00)	(0.00, -0.00)	(-0.00, -0.00)	(0.02, 0.03)	(0.02, 0.02)	(0.04, 0.03)
6.3	(0.08, 0.08)	(0.00, -0.00)	(0.00, -0.00)	(0.00, -0.00)	(0.02, 0.02)	(0.01, 0.01)	(0.02, 0.03)
7.1	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)
7.2	(0.01, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(-0.00, -0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)
7.3	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(-0.00, -0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)
8.1	(0.07, 0.06)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.02, 0.02)	(0.00, 0.00)	(0.02, 0.03)
8.2	(0.07, 0.07)	(0.00, 0.00)	(0.00, 0.00)	(0.00, -0.00)	(0.02, 0.02)	(0.01, 0.00)	(0.02, 0.02)
8.3	(0.07, 0.07)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.02, 0.02)	(0.00, 0.00)	(0.02, 0.02)

D.2 Discussion

Each of the designs is compatible with the proposed multivariate decision-making framework and has specific advantages. Although fixed designs perform well under accurate sample size estimation, a priori sample size estimation is difficult when multiple parameters are unknown. Sequential or adaptive designs may be beneficial to deal with this parameter uncertainty, albeit at the expense of bias. Whereas adaptive designs deal most flexibly with parameter uncertainty, group sequential designs limit bias more than adaptive designs.

We remark that adaptive designs in particular have their practical challenges. First, updating adaptive designs might require a large logistic effort, which increases with the size

of the study. These designs are therefore easier to implement in small phase I or II studies compared to confirmatory phase II or III studies. Second, we find that the current literature does not offer clear guidance on the specification of adaptive design parameters. Further elaboration on the choice of these parameters would undoubtedly serve trials that stop data collection adaptively.

Appendix E

Numerical evaluation: Comparison of prior specifications

To demonstrate the influence of prior information on the performance of the Compensatory decision rule, we specified six different sets of prior hyperparameters, that are presented in Table E.1. Two of these prior specifications are assumed to be uninformative $(1-2)$. Prior 1 is the non-informative prior that we used in the *Numerical evaluation* Section (3.4) and serves as a reference prior in this comparison. Prior 2 is Jeffreys's prior, which is well-known for its property to remain invariant under transformation of parameters (Yang & Berger, 1996). This is useful since our main interest is typically in the transformed parameters $\boldsymbol{\delta}$ rather than the marginal probabilities $\boldsymbol{\theta}_j$ or the cell probabilities $\boldsymbol{\phi}_j$, on which the treatment-specific prior distributions are specified. Four informative prior specifications (priors $3-6$) include 20 additional observations (i.e., $n_j^0 = 20$). These 20 additional observations show the effects on decisions when the number of prior observations is either higher or lower than sample size n_j. The former occurs in data generating mechanisms with large treatment differences that require sample sizes smaller than 20 (e.g., treatment difference 5), while the latter occurs when treatment differences are small and sample sizes are larger (e.g., treatment difference 3). Prior specifications $3-6$ differ on the correspondence between the prior treatment difference $\boldsymbol{\delta}^0$ and the true treatment difference $\boldsymbol{\delta}^T$ used for data generation. Specifically, we included prior differences that are identical (prior 3), smaller (prior 4), larger (prior 5) or opposite (prior 6) to the true treatment difference. The bivariate beta distributions of prior specifications 1, 2, and 3 for data generating mechanism 2.2 are visually presented in Figure A.1. We ran the introduced procedure for a fixed design (Steps 2 and 3 of Algorithm 1), using sample size n_j for a fixed design estimated based on true treatment differences. These sample sizes were also used in Section *Numerical evaluation* (3.4) and presented in Table 3.3.

Table E.1: Prior specifications used for numerical evaluation. Prior hyperparameters $\boldsymbol{\alpha}_j^0 = n_j^0 \boldsymbol{\phi}_j^0$. True parameters $\boldsymbol{\phi}_E^T$, $\boldsymbol{\phi}_C^T$ and $\boldsymbol{\delta}^T$ can be obtained via the simulation conditions presented in Table 3.1.

Prior	n_j^0	$\boldsymbol{\phi}_E^0$	$\boldsymbol{\phi}_C^0$	$\boldsymbol{\delta}^0$
1	$\frac{1}{25}$	$\frac{1}{4}$	$\frac{1}{4}$	0
2	2	$\frac{1}{4}$	$\frac{1}{4}$	0
3	20	$\boldsymbol{\phi}_E^T$	$\boldsymbol{\phi}_C^T$	$\boldsymbol{\delta}^T$
4	20	$\boldsymbol{\phi}_E^T - 0.05$	$\boldsymbol{\phi}_C^T + 0.05$	$\boldsymbol{\delta}^T - 0.10$
5	20	$\boldsymbol{\phi}_E^T + 0.05$	$\boldsymbol{\phi}_C^T - 0.05$	$\boldsymbol{\delta}^T + 0.10$
6	20	$\boldsymbol{\phi}_C^T$	$\boldsymbol{\phi}_E^T$	$-\boldsymbol{\delta}^T$

E.1 Results

The two uninformative priors do not noticeably influence the probability to conclude superiority (Table E.2) or the average treatment effect (Table E.3) in the majority of data generating mechanisms. An exception is a large treatment difference (5.1 – 5.3) where Jeffreys's prior (i.e., prior 2) lowered power and biased the treatment effect downwards. Here, $\mathbf{s}_j$ is too small to satisfy $\boldsymbol{\alpha}_j^n \approx \mathbf{s}_j$. A smaller prior sample size n_j^0 (prior 1) resulted in an unbiased estimate of the average treatment difference.

An informative prior distribution influences the probability to conclude superiority as well as the average treatment effect, depending on prior treatment difference $\boldsymbol{\delta}^0$. Prior information improves decision-making when the prior treatment effect equals the true treatment effect (i.e., $\boldsymbol{\delta}^0 = \boldsymbol{\delta}^T$; prior 3). This situation increases power, without influencing Type I error or the average posterior treatment effect

In contrast, prior information affects the decision when prior and true treatment effects do not correspond. When the prior treatment effect is less strong than the true treatment effect (i.e., $\boldsymbol{\delta}^0 < \boldsymbol{\delta}^T$, prior 4), the Type I error as well as the probability to conclude superiority correctly are lowered, although the effect is masked in treatment differences 4 – 6 by the (relatively large) number of prior observations n_j^0. Moreover, the average posterior treatment effect is lower than the treatment effect of the data $\boldsymbol{\delta}^T$ especially when the treatment difference is large and the required sample size is low (5.1 – 5.3).

When the prior treatment effect is stronger than the true treatment effect (i.e., $\boldsymbol{\delta}^0 > \boldsymbol{\delta}^T$, prior 5), the Type I error and the probability to conclude superiority are above the planned .05 and .80. Moreover, the average posterior treatment effect exceeds the treatment effect of the data $\boldsymbol{\delta}^T$, in particular when the treatment effect is large (5.1 – 5.3). An opposite prior treatment effect (i.e., $\boldsymbol{\delta}^0 = -\boldsymbol{\delta}^T$, prior 6) results in a lower probability to conclude superiority as well as an average posterior treatment effect that differs from true treatment difference $\boldsymbol{\delta}^T$.

In general, the effect of prior information is strongest in 5.1 – 5.3 and 8.1 – 8.3, where prior sample size n_j^0 is relatively large compared to sample size n_j, resulting in a larger influence of $\boldsymbol{\alpha}_j^0$ on $\boldsymbol{\alpha}_j^n$. Note that in practice, a difference between results with and without prior information signals a conflict between the informative prior and the data, and does not necessarily reflect an invalid decision.

Table E.2: P(Conclude superiority) for six different prior specifications (see Table E.1) after applying the Compensatory decision rule with equal weights.

DGM	1	2	3	4	5	6
1.1	0.000	0.000	0.000	0.000	0.000	0.000
1.2	0.000	0.000	0.000	0.000	0.000	0.000
1.3	0.000	0.000	0.000	0.000	0.000	0.000
2.1	0.049	0.055	0.049	0.037	0.067	0.049
2.2	0.056	0.050	0.049	0.038	0.054	0.055
2.3	0.049	0.048	0.045	0.035	0.064	0.053
3.1	0.807	0.798	0.872	0.753	0.951	0.606
3.2	0.806	0.803	0.855	0.775	0.911	0.674
3.3	0.800	0.793	0.841	0.782	0.895	0.688
4.1	0.811	0.791	0.987	0.905	0.999	0.043
4.2	0.813	0.794	0.967	0.867	0.990	0.178
4.3	0.804	0.802	0.939	0.854	0.979	0.313
5.1	0.881	0.704	1.000	1.000	1.000	0.000
5.2	0.831	0.787	1.000	1.000	1.000	0.000
5.3	0.809	0.761	1.000	0.998	1.000	0.000
6.1	0.824	0.789	0.990	0.900	0.999	0.030
6.2	0.805	0.797	0.965	0.874	0.991	0.145
6.3	0.801	0.792	0.946	0.856	0.984	0.282
7.1	0.000	0.000	0.000	0.000	0.000	0.000
7.2	0.000	0.000	0.000	0.000	0.000	0.000
7.3	0.000	0.000	0.000	0.000	0.000	0.000
8.1	0.808	0.792	0.957	0.837	0.992	0.224
8.2	0.804	0.799	0.925	0.823	0.975	0.387
8.3	0.805	0.793	0.896	0.811	0.955	0.480

Table E.3: Average bias for six different prior specifications (see Table E.1) after applying the Compensatory decision rule with equal weights.

DGM	1	2	3	4	5	6
1.1	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.01, 0.01)
1.2	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.01, 0.01)
1.3	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.01, 0.01)
2.1	(-0.00, -0.00)	(0.00, 0.00)	(-0.00, 0.00)	(-0.00, -0.00)	(0.00, 0.00)	(0.00, -0.00)
2.2	(0.00, 0.00)	(-0.00, -0.00)	(-0.00, -0.00)	(-0.00, -0.00)	(0.00, 0.00)	(0.00, -0.00)
2.3	(-0.00, -0.00)	(0.00, -0.00)	(-0.00, -0.00)	(-0.00, -0.00)	(0.00, 0.00)	(-0.00, -0.00)
3.1	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(-0.02, -0.02)	(0.01, 0.02)	(-0.03, -0.03)
3.2	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(-0.01, -0.01)	(0.01, 0.01)	(-0.02, -0.02)
3.3	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(-0.01, -0.01)	(0.01, 0.01)	(-0.02, -0.02)
4.1	(0.00, 0.00)	(-0.01, -0.02)	(0.00, 0.00)	(-0.04, -0.04)	(0.04, 0.04)	(-0.17, -0.17)
4.2	(0.00, 0.00)	(-0.01, -0.01)	(0.00, 0.00)	(-0.03, -0.04)	(0.04, 0.03)	(-0.14, -0.14)
4.3	(0.00, 0.00)	(-0.01, -0.01)	(0.00, 0.00)	(-0.03, -0.03)	(0.03, 0.03)	(-0.11, -0.12)
5.1	(0.00, -0.01)	(-0.10, -0.10)	(0.00, 0.00)	(-0.08, -0.08)	(0.08, 0.08)	(-0.62, -0.62)
5.2	(-0.01, -0.01)	(-0.07, -0.07)	(0.00, 0.00)	(-0.07, -0.07)	(0.07, 0.07)	(-0.55, -0.55)
5.3	(0.00, -0.01)	(-0.06, -0.06)	(0.00, 0.00)	(-0.06, -0.06)	(0.06, 0.07)	(-0.52, -0.52)
6.1	(0.00, 0.00)	(-0.03, 0.00)	(0.00, -0.00)	(-0.05, -0.05)	(0.04, 0.04)	(-0.36, 0.00)
6.2	(0.00, -0.00)	(-0.02, 0.00)	(0.00, -0.00)	(-0.04, -0.03)	(0.03, 0.04)	(-0.29, -0.00)
6.3	(0.00, -0.00)	(-0.02, -0.00)	(0.00, 0.00)	(-0.03, -0.03)	(0.03, 0.03)	(-0.24, -0.00)
7.1	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(-0.01, 0.02)
7.2	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(-0.01, 0.02)
7.3	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(-0.01, 0.02)
8.1	(0.00, 0.00)	(-0.01, 0.00)	(0.00, 0.00)	(-0.03, -0.03)	(0.03, 0.03)	(-0.16, -0.05)
8.2	(0.00, 0.00)	(-0.01, 0.00)	(0.00, 0.00)	(-0.03, -0.02)	(0.03, 0.03)	(-0.12, -0.04)
8.3	(0.00, 0.00)	(-0.01, 0.00)	(0.00, 0.00)	(-0.02, -0.02)	(0.02, 0.02)	(-0.10, -0.03)

Appendix F

Details of posterior computation

The current section describes the Gibbs sampling procedure used to obtain parameters. To simplify notations, we omit the dependence on $\mathbf{x}$ in denoting functions that rely on covariates $(\boldsymbol{\phi}, \boldsymbol{\theta})$.

Starting from the likelihood of individual K-variate response $\mathbf{y}_i$ (Equation 4.2), the likelihood of n K-variate responses follows from taking the product over n individual joint response probabilities in Q response categories:

$$l(\mathbf{y}|\boldsymbol{\beta}, \mathbf{x}) = \prod_{i=1}^{n} \prod_{q=1}^{Q-1} \left(\frac{\exp\left[\psi_i^q\right]}{\sum_{r=1}^{Q-1} \exp\left[\psi_i^r\right] + 1} \right)^{I(\mathbf{y}_i = q)} \left(\frac{1}{\sum_{r=1}^{Q-1} \exp\left[\psi_i^r\right] + 1} \right)^{I(\mathbf{y}_i = Q)} . \tag{F.1}$$

Following Polson et al. (2013), we introduce the Pólya-gamma variable by rewriting the multivariate likelihood in Equation F.1 as a series of binomial likelihoods. The likelihood of $\mathbf{y}$ conditional on the parameters of the qth response category, $\boldsymbol{\beta}^q$, then equals:

$$l(\mathbf{y}|\boldsymbol{\beta}^q, \boldsymbol{\beta}^{-q}) = \prod_{i=1}^{n} \left(\frac{\exp\left[\eta_i^q\right])}{\exp\left[\eta_i^q\right] + 1} \right)^{I(\mathbf{y}_i = q)} \left(\frac{1}{\exp\left[\eta_i^q\right] + 1} \right)^{1 - I(\mathbf{y}_i = q)} \tag{F.2}$$

where $-q$ refers to all rows in $\mathbf{H}$ not having index q and $\eta_i^q = \psi_i^q - \ln\left(\sum_{m \neq \mathbf{H}_{q\cdot}} \exp\left[\psi_i^m\right] \right)$.

The Pólya-Gamma transformation to a Gaussian distribution relies on the following equality (Polson et al., 2013):

$$\frac{\exp\left[\eta_i^q\right]}{\exp\left[\eta_i^q\right] + 1} = 2\exp\left[(y_i - \frac{1}{2})\eta_i^q \right] \int_0^\infty \exp\left[\frac{-\omega_i \eta_i^{q2}}{2} \right] p(\omega_i^q) d\omega_i^q \tag{F.3}$$

where ω_i^q has a Pólya-Gamma distribution, i.e., $p(\omega_i^q) \sim PG(1, \psi_i^q)$.

If we use the equality in Equation F.3, the binomial likelihood in Equation F.2 can be transformed to a multivariate Gaussian likelihood by including an auxiliary Pólya-Gamma

variable ω_i^q (Polson et al., 2013):

$$
\begin{aligned}
l(\mathbf{y}|\boldsymbol{\beta}^q, \boldsymbol{\beta}^{-q}) &= \prod_{i=1}^{n} \frac{\exp\left[\eta_i^q\right]}{\exp\left[\eta_i^q\right]+1} \\
&= \prod_{i=1}^{n} 2\exp\left[(y_i - \frac{1}{2})\eta_i^q\right] \int_0^\infty \exp\left[\frac{-\omega_i^q \eta_i^{q2}}{2}\right] p(\omega_i^q) d\omega_i^q \\
&= \prod_{i=1}^{n} \exp\left[\kappa_i^q \omega_i^q \eta_i^q - \frac{1}{2}(\eta_i^q)^2 \omega_i^q\right] PG(\omega_i^q|1,0) \\
&\propto \exp\left[\frac{1}{2}(2\boldsymbol{\kappa}^q \boldsymbol{\omega}^q \boldsymbol{\eta}^q - \boldsymbol{\omega}^q (\boldsymbol{\eta}^q)^2\right] \\
&\propto \exp\left[-\frac{1}{2}(\boldsymbol{\kappa}^q - \boldsymbol{\eta}^q)^T \boldsymbol{\Omega}^q (\boldsymbol{\kappa}^q - \boldsymbol{\eta}^q)\right] \\
&= \exp\left[-\frac{1}{2}(\boldsymbol{\kappa}^q - \mathbf{X}\boldsymbol{\beta}^q + \ln[\sum_{m \neq q} \exp(\mathbf{X}\boldsymbol{\beta}^m)])^T \boldsymbol{\Omega}^q \right. \\
&\qquad \left. (\boldsymbol{\kappa}^q - \mathbf{X}\boldsymbol{\beta}^q + \ln[\sum_{m \neq q} \exp\left[\mathbf{X}\boldsymbol{\beta}^m\right]]\right],
\end{aligned}
\tag{F.4}
$$

where $\kappa_i^q = \frac{l(y_i = \mathbf{H}_q \ldots) - \frac{1}{2}}{\omega_i^q}$, $\boldsymbol{\kappa}^q = (\kappa_1^q, \ldots, \kappa_n^q)$, $\boldsymbol{\omega}^q = (\omega_1^q, \ldots, \omega_n^q)$, and $\boldsymbol{\Omega}^q = \text{diag}(\boldsymbol{\omega}^q)$.

F.1 Prior distribution

The Gaussian likelihood in Equation F.4 is conditionally conjugate with a normal prior distribution on regression coefficients $\boldsymbol{\beta}^q$:

$$
\boldsymbol{\beta}^q \sim N(\mathbf{b}^q, \mathbf{B}^{0q}) \tag{F.5}
$$

where $\mathbf{b}^q$ is the vector of prior means of regression coefficient vector $\boldsymbol{\beta}^q$ and $\mathbf{B}^{0q}$ is a $P \times P$ symmetric square matrix reflecting the prior precision of regression coefficients $\boldsymbol{\beta}^q$. A researcher who is willing to include prior information regarding treatment effects into the analysis, has several options to specify prior hyperparameters for a normally distributed prior that is compatible with the Gibbs sampling procedure (e.g., Ibrahim & Chen, 2000; Sullivan & Greenland, 2012). We discuss the specification of informative prior means $\mathbf{b}^q$ in terms of joint response probabilities $\boldsymbol{\phi}$ in the next Appendix.

F.2 Posterior distribution

Bayesian statistical inference is done via the posterior distribution which is given by:

$$p(\boldsymbol{\beta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{x})p(\boldsymbol{\beta}), \tag{F.6}$$

The combination of a Pólya-Gamma transformed Gaussian likelihood (Equation F.4) and a normal prior distribution (Equation F.5) respectively is proportional to a normally distributed posterior distribution, conditionally on Pólya-Gamma variables in $\boldsymbol{\omega}^q$ (Polson et al., 2013):

$$\begin{aligned} p(\boldsymbol{\beta}^q|\mathbf{Y}, \boldsymbol{\Omega}^q) \propto & p(\mathbf{y}|\boldsymbol{\beta}^q, \boldsymbol{\omega}^q)p(\boldsymbol{\beta}^q) \\ \propto & \exp\left[-\frac{1}{2}(\boldsymbol{\kappa}^q - \mathbf{X}\boldsymbol{\beta}^q + \ln[\sum_{m\neq q}\exp[\mathbf{X}\boldsymbol{\beta}^m]])^T\boldsymbol{\Omega}^q \right. \\ & \left. (\boldsymbol{\kappa}^q - \mathbf{X}\boldsymbol{\beta}^q + \ln[\sum_{m\neq q}\exp[\mathbf{X}\boldsymbol{\beta}^m]])\right] \times \\ & \exp\left[-\frac{1}{2}(\boldsymbol{\beta}^q - \mathbf{b}^q)^T(\mathbf{B}^q)^{-1}(\boldsymbol{\beta}^q - \mathbf{b}^q)\right] \\ \propto & N\left(\mathbf{V}^q(\mathbf{X}^T\boldsymbol{\Omega}^q(\boldsymbol{\kappa}^q + \ln[\sum_{m\neq q}\exp[\mathbf{X}\boldsymbol{\beta}^m]]) + (\mathbf{B}^q)^{-1}\mathbf{b}^q), \mathbf{V}^q\right) \end{aligned} \tag{F.7}$$

where $\mathbf{V}^q = (\mathbf{X}^T\boldsymbol{\Omega}^q\mathbf{X} + (\mathbf{B}^q)^{-1})^{-1}$. Similarly, subject-specific variable ω_i^q follows a Pólya-Gamma distribution that depends on regression coefficients $\boldsymbol{\beta}^q$ via linear predictor ψ_i^q.

Updating these two conditional distributions via a Gibbs sampling procedure results in a sample from the posterior distribution of $\boldsymbol{\beta}$. Specifically, the sampling procedure involves iterating L times over the following two steps for $q = 1, \dots, Q-1$, while keeping $\boldsymbol{\beta}^Q$ fixed at zero:

1. Draw a vector of $P+1$ regression coefficients $\boldsymbol{\beta}^q|\boldsymbol{\omega}^q$ from a multivariate normal distribution with mean vector $\mathbf{m}^q$ and precision matrix $\mathbf{V}^q$.

$$\boldsymbol{\beta}^q|\boldsymbol{\omega}^q \sim N(\mathbf{m}^q, \mathbf{V}^q) \tag{F.8}$$

$$\text{where} \quad [\mathbf{V}^q]^{-1} = \mathbf{X}\boldsymbol{\Omega}^q\mathbf{X} + [\mathbf{V}^{0q}]^{-1}$$

$$\mathbf{m}^q = \mathbf{V}^q(\mathbf{X}(\boldsymbol{\kappa}^q + \boldsymbol{\Omega}^q\mathbf{c}) + [\mathbf{V}^{0q}]^{-1}\mathbf{m}^{0q})$$

$$\mathbf{c} = \left\{ \ln \left(\sum_{m \neq q} \exp\left[\psi_i^m\right] \right)_{i=1}^{n} \right\}.$$

2. Sample $\boldsymbol{\omega}^q|\boldsymbol{\beta}^q$ as a vector of n draws $\omega_i^q|\boldsymbol{\beta}^q$ from a Pólya-Gamma distribution:

$$\omega_i^q|\boldsymbol{\beta}^q \sim PG(1, \psi_i^q - \ln \sum_{m \neq q} \exp\left[\psi_i^m\right]). \tag{F.9}$$

The Gibbs sampling procedure results in a sample of L sets of regression coefficients from the posterior distribution of $\boldsymbol{\beta}$.

Appendix G

Specification of prior means of regression coefficients

In the current Section, we introduce a procedure to determine prior means, based on beliefs regarding success probabilities and correlations between them. We outline the procedure for two outcome variables and a linear predictor ψ with one covariate and an interaction between the treatment and the covariate:

$$\psi_T^q = \beta_0^q + \beta_1^q T + \beta_2^q x + \beta_3^q x \times T \tag{G.1}$$

First, choose x_{L} and x_{H} as low and high values of covariate x respectively. Next, specify success probabilities and correlations $\boldsymbol{\theta}_T(x^L)$, $\rho_T(x^L)$, $\boldsymbol{\theta}_T(x^H)$, and $\rho_T(x^H)$ for each treatment T that accompany the low and high values of covariates respectively. These success probabilities $\boldsymbol{\theta}_T(x^{\cdot})$ and correlations $\rho_T(x^{\cdot})$ can be transformed to joint response probabilities $\boldsymbol{\phi}_T(x^{\cdot})$ via the following set of equations:

$$\begin{aligned}
\phi_T^{11}(x^{\cdot}) &= \rho_T(x^{\cdot})\sqrt{\theta_T^1(x^{\cdot})\left[1-\theta_T^1(x^{\cdot})\right]\theta_T^2(x^{\cdot})\left[1-\theta_T^2(x^{\cdot})\right]} + \theta_T^1(x^{\cdot})\theta_T^2(x^{\cdot}) \\
\phi_T^{10}(x^{\cdot}) &= \theta_T^1(x^{\cdot}) - \phi_T^{11}(x^{\cdot}) \\
\phi_T^{01}(x^{\cdot}) &= \theta_T^2(x^{\cdot}) - \phi_T^{11}(x^{\cdot}) \\
\phi_T^{00}(x^{\cdot}) &= 1 - \theta_T^1(x^{\cdot}) - \theta_T^2(x^{\cdot}) + \phi_T^{11}(x^{\cdot})
\end{aligned} \tag{G.2}$$

For each response category q, joint responses $\phi_T^{q\cdot}$ can be transformed to linear predictor $\psi_T^{q\cdot}$ using the multinomial logistic link function in Equation 4.2.

Solving these linear predictors for $\boldsymbol{\beta}^q$ results in the following definitions of the elements in $\boldsymbol{\beta}^q$:

$$\begin{aligned}
\beta_0^q &= \frac{x^{\mathrm{H}}\psi_0^q(x^L) - x^{\mathrm{L}}\psi_0^q(x^H)}{x^{\mathrm{H}} - x^{\mathrm{L}}} \\
\beta_1^q &= \frac{x^{\mathrm{H}}\left[\psi_1^q(x^L) - \psi_0^q(x^L)\right] + x^{\mathrm{L}}\left[\psi_0^q(x^H) - \psi_1^q(x^H)\right]}{x^{\mathrm{H}} - x^{\mathrm{L}}} \\
\beta_2^q &= \frac{\psi_0^q(x^H) - \psi_0^q(x^L)}{x^{\mathrm{H}} - x^{\mathrm{L}}} \\
\beta_3^q &= \frac{\psi_1^q(x^H) - \psi_0^q(x^H) - \psi_1^q(x^L) + \psi_0^q(x^L)}{x^{\mathrm{H}} - x^{\mathrm{L}}}
\end{aligned} \tag{G.3}$$

For example, if we would believe that treatment have the following parameters:

$$\boldsymbol{\theta}_1^L = (0.60, 0.70), \rho_1^L = -0.30$$

$$\boldsymbol{\theta}_1^H = (0.40, 0.30), \rho_1^H = -0.30$$

$$\boldsymbol{\theta}_0^L = (0.40, 0.30), \rho_0^L = -0.30$$

$$\boldsymbol{\theta}_0^H = (0.60, 0.70), \rho_0^H = -0.30,$$

then the regression coefficients would be as presented in Table G.1.

Table G.1: Example of means of the prior distribution of regression coefficients

	$q = 1$	$q = 2$	$q = 3$	$q = 4$
β_0^q	-0.000	0.766	0.766	0.000
β_1^q	0.000	0.000	0.000	0.000
β_2^q	1.902	0.781	1.121	0.000
β_3^q	-3.804	-1.562	-2.241	0.000

Appendix H

Procedures for estimation and inference over a specified (sub)population

Algorithm 3 Transformation of posterior regression coefficients to posterior joint response probabilities based on fixed covariate values.

Define $\mathbf{x} = x_2, \dots, x_P$ as a vector of covariate values of interest
Let $\boldsymbol{\beta}^Q = (0, \dots, 0)$
for draw $(l) \leftarrow 1 : L$ **do**
  for treatment $T \leftarrow 0 : 1$ **do**
    for joint response $q \leftarrow 1 : Q$ **do**
      Compute $\psi_T^{q(l)} = \beta_0^{q(l)} + \beta_1^{q(l)} T + \beta_2^{q(l)} x + \beta_3^{q(l)} x \times T$
      Compute $\phi_T^{q(l)} = \frac{\exp\left[\psi_T^{q(l)}\right]}{\sum_{r=1}^{Q-1} \exp\left[\psi_T^{r(l)}\right] + 1}$
    end for
  end for
end for

Algorithm 4 Transformation of posterior regression coefficients to posterior joint response probabilities based on empirical marginalization.

Let $\boldsymbol{\beta}^Q = (0, \dots, 0)$
for draw $(l) \leftarrow 1 : L$ **do**
  for subject $i \leftarrow 1 : n$ **do**
    for joint response $q \leftarrow 1 : Q$ **do**
      Compute $\psi_i^{q(l)} = \beta_1^{q(l)} T_i + \beta_2^{q(l)} x_i + \beta_3^{q(l)} x_i \times T_i$
      Compute $\phi_i^{q(l)} = \frac{\exp\left[\psi_i^{q(l)}\right]}{\sum_{r=1}^{Q-1} \exp\left[\psi_i^{r(l)}\right] + 1}$
      for $T \leftarrow 0 : 1$ **do**
        Compute $\phi_T^{q(l)}(\mathbf{x}) = \frac{1}{\sum_{i=1}^{n} I(T_i = T)} \phi_i^{q(l)} I(T_i = T)$
      end for
    end for
  end for
end for

Appendix I

Observed bias in regression coefficients

The simulation study in Section 4.4 showed that mean estimates of regression coefficients were asymptotically unbiased. Bias was negligible ($< .01$) in conditions with a sufficiently large sample, while we observed some bias in conditions with smaller samples (DGM 3.1, 3.2, 4.1, and 4.2 under the Any and Compensatory decision rules). Of these conditions, bias was most prominent in data generating mechanisms 4.1 and 4.2 under the sample sizes used for the Any ($n = 21$) and Compensatory ($n = 29$) rules. The histograms of median regression coefficient for one of these conditions (DGM 4.2, Compensatory rule) are shown in Figure I.1, revealing that some regression coefficients were skewed in the extreme direction.

The bias in regression coefficients is a well-documented property of the (non-linear) logistic transformation (e.g., Firth, 1993). When bias was mild, the multinomial logistic transformation needed to obtain joint responses (Equation 4.2) appeared to normalize the skewed posterior samples of regression coefficients. More severe bias in conditions with smaller sample sizes was not fully corrected in the transformation steps. Treatment effect estimation based on fixed values under DGMs 4.1 and 4.2 resulted in treatment differences with absolute biases up to 0.077 for the Any and Compensatory rules, as shown in Table 4.3. Bias appeared slightly more severe when the covariate was discrete, compared to a continuous covariate. The reference and marginalization approaches could estimate treatment effects without bias, regardless of sample size.

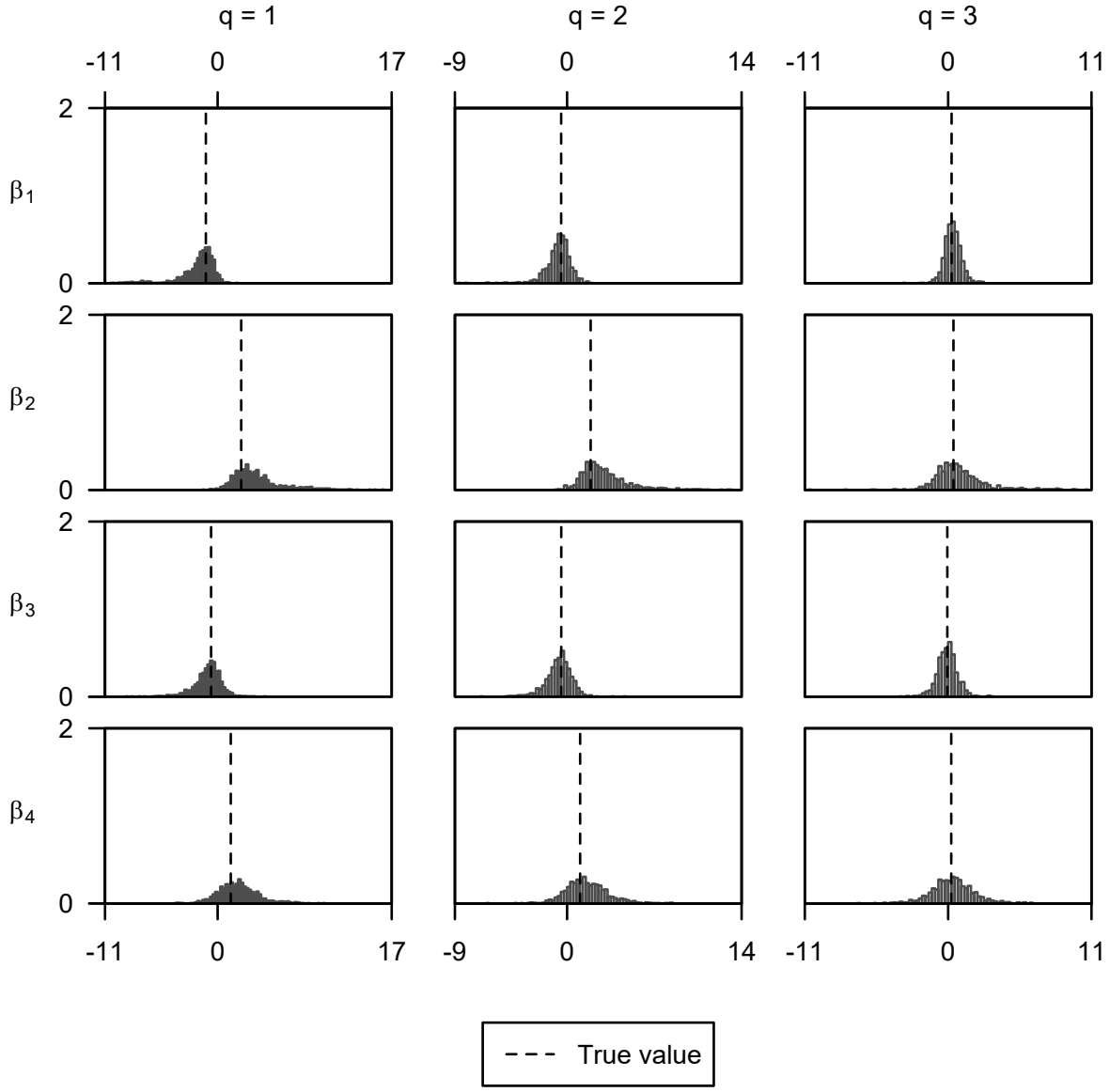


Figure I.1: Histograms of median regression coefficients fitted for application of the Compensatory rule under DGM 4.2.

Appendix J

Gibbs sampling procedure based on Pólya-Gamma expansion

J.1 Random effects model

Bayesian analysis relies on the posterior distribution of regression coefficients, which is proportional to the likelihood of the data and the prior distribution:

$$p(\gamma_j, \gamma, \mathbf{\Sigma}|\mathbf{y}) \propto p(\mathbf{y}|\gamma_j)p(\gamma_j|\gamma, \mathbf{\Sigma})p(\gamma)p(\mathbf{\Sigma}). \tag{J.1}$$

The multinomial logistic likelihood (Equation 5.2) can be expanded with a Pólya-Gamma auxiliary variable to suit a Gibbs sampling procedure. This expansion relies on the following equality (Polson et al., 2013):

$$\begin{aligned} p((\mathbf{y}_j = \mathbf{h}^q)|\gamma_j^q, \gamma_j^{-q}, \boldsymbol{\omega}_j^q) &= \frac{\exp\left(\mathbf{x}_{ji}\gamma_j^q\right)}{\sum_{r=1}^{Q-1} \exp\left(\mathbf{x}_{ji}\gamma_j^r\right) + 1}, \\ &\propto \exp\left[-\frac{1}{2}(\boldsymbol{\kappa}_j^q - \boldsymbol{\eta}_j^q)^T \boldsymbol{\Omega}_j^q (\boldsymbol{\kappa}_j^q - \boldsymbol{\eta}_j^q)\right], \end{aligned} \tag{J.2}$$

where $\mathbf{X}_j$ is a matrix filled with n_j rows of covariate vectors $\mathbf{x}_{ji}$ and $\boldsymbol{\eta}_j^q = \mathbf{X}_j\gamma_j^q - \ln[\sum_{m \neq q} \exp(\mathbf{X}_j\gamma_j^m)]$, $\boldsymbol{\kappa}_j^q = \frac{I(\mathbf{y}_j = \mathbf{h}^q) - \frac{1}{2}}{\boldsymbol{\omega}_j^q}$.

Equation J.2 can be recognized as the kernel of a multivariate Gaussian likelihood of working variable $\boldsymbol{\kappa}_j^q$ (Polson et al., 2013):

$$\boldsymbol{\kappa}_j^q \sim N\left(\boldsymbol{\eta}_j^q, \{\boldsymbol{\Omega}_j^q\}^{-1}\right) \tag{J.3}$$

Here, $\boldsymbol{\Omega}_j^q$ reflects the diagonal matrix of Pólya-Gamma distributed variables $\boldsymbol{\omega}_j^q = (\omega_{j1}^q, \dots, \omega_{jn_j}^q)$. A Gibbs sampler can be constructed when the likelihood in Equation J.3 is combined with multivariate normal prior distributions on random regression coefficients $\gamma_j^q|\gamma^q, \mathbf{\Sigma}^q$ and mean random regression coefficients γ^q, and an inverse-Wishart prior

distribution on covariance matrix $\mathbf{\Sigma}^q$:

$$\gamma_j^q \sim N(\gamma^q, \mathbf{\Sigma}^q) \tag{J.4}$$
$$\gamma^q \sim N(\mathbf{g}^q, \mathbf{G}^q)$$
$$\mathbf{\Sigma}^q \sim \mathcal{W}^{-1}(j^0, \mathbf{S}^q)$$

The resulting Gibbs sampler consists of the following steps:

1. Sample mean regression coefficients:

$$\gamma^{q(l)} \sim N\left(\mathbf{V}_\gamma^q(\{\mathbf{\Sigma}^{q(l-1)}\}^{-1}\sum_{j=1}^{J}\gamma_j^{q(l-1)} + \mathbf{G}^q\mathbf{g}^q), \mathbf{V}_\gamma^q\right)$$

   with prior mean vector $\mathbf{g}^q$, prior precision matrix $\mathbf{G}^q$ and posterior variance matrix $\mathbf{V}_\gamma = (J\{\mathbf{\Sigma}^{q(l-1)}\}^{-1} + \mathbf{G}^q)^{-1}$.

2. Sample covariance matrices of regression coefficients:

$$\mathbf{\Sigma}^{q(l)} \sim \mathcal{W}^{-1}\left(j^0 + J, \mathbf{S}^q + \sum_{j=1}^{J}\left(\gamma_j^{q(l-1)} - \gamma^{q(l)}\right)\left(\gamma_j^{q(l-1)} - \gamma^{q(l)}\right)^T\right)$$

   with prior hyperparameters $j^0 \geq P$ and $\mathbf{S}^q$.

3. For each j, sample random regression coefficients:

$$\gamma_j^{q(l)} \sim N\left(\mathbf{V}_{\gamma_j^q}^q(\mathbf{X}_j\mathbf{\Omega}_j^{q(l-1)}(\boldsymbol{\kappa}_j^{q(l-1)} + \ln[\sum_{m\neq q}\exp(\mathbf{X}_j\gamma_j^{m(l)})]) + \{\mathbf{\Sigma}^{q(l)}\}^{-1}\gamma^{q(l)}), \mathbf{V}_{\gamma_j}^q\right)$$

   with prior mean vector $\gamma^{q(l)}$, prior precision matrix $\mathbf{\Sigma}^{q(l)}$, posterior variance matrix $\mathbf{V}_{\gamma_j}^q = ({\mathbf{X}_j}^T\mathbf{\Omega}_j^{q(l-1)}\mathbf{X}_j + \{\mathbf{\Sigma}^{q(l)}\}^{-1})^{-1}$, and diagonal matrix of Pólya-Gamma variables $\mathbf{\Omega}_j^{q(l-1)} = \text{diag}(\omega_{j1}^{q(l-1)}, \dots, \omega_{jn_j}^{q(l-1)})$.

4. For each j and i, sample Pólya-Gamma variables:

$$\omega_{ji}^{q(l)} \sim PG(1, \eta_{ji}^{q(l)})$$

The remainder of this section shows the derivations of the full conditional distributions.

J.1.1 Deriving the likelihood function

The following equality forms the basis to rewrite the multinomial likelihood in Equation 5.2 as a Gaussian likelihood (Polson et al., 2013):

$$p((\mathbf{y}_j = \mathbf{h}^q)|\gamma_j, \boldsymbol{\omega}_j^q, \mathbf{x}_j) = \frac{\exp\left(\mathbf{x}_{ji}\gamma_j^q\right)}{\sum_{r=1}^{Q-1} \exp\left(\mathbf{x}_{ji}\gamma_j^r\right) + 1}, \tag{J.5}$$
$$= \prod_{i=1}^{n_j} 2 \exp\left[\kappa_{ji}^q \omega_{ji}^q \eta_{ji}^q\right] \int_0^\infty \exp\left[\frac{-\omega_{ji}^q (\eta_{ji}^q)^2}{2}\right] p(\omega_{ji}^q) d\omega_{ji}^q$$

where $\omega_{ji}^q \sim PG(1, \eta_{ji}^q)$ is a Pólya-Gamma distributed variable,
where $\eta_{ji}^q = \mathbf{x}_{ji}\gamma_j^q - \ln\left[\sum_{m \neq q} \exp(\mathbf{x}_{ji}\gamma_j^m)\right]$,
and where working variable $\boldsymbol{\kappa}_j^q = \frac{I(\mathbf{y}_j = \mathbf{h}^q) - \frac{1}{2}}{\boldsymbol{\omega}_j^q}$.

Further algebraic transformation results in the kernel of a Gaussian likelihood:

$$p((\mathbf{y}_j = \mathbf{h}^q)|.) = \prod_{i=1}^{n_j} 2 \exp\left[\kappa_{ji}^q \omega_{ji}^q \eta_{ji}^q\right] \int_0^\infty \exp\left[\frac{-\omega_{ji}^q (\eta_{ji}^q)^2}{2}\right] p(\omega_{ji}^q) d\omega_{ji}^q \tag{J.6}$$
$$\propto \exp\left[\frac{1}{2}(\boldsymbol{\kappa}_j^q \boldsymbol{\omega}_j^q \boldsymbol{\eta}_j^q - \boldsymbol{\omega}_j^q (\boldsymbol{\eta}_j^q)^2\right]$$
$$\propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\kappa}_j^q - \boldsymbol{\eta}_j^q\right)^T \boldsymbol{\Omega}_j^q \left(\boldsymbol{\kappa}_j^q - \boldsymbol{\eta}_j^q\right)\right],$$

Hence, working variable $\boldsymbol{\kappa}_j^q$ is multivariate normally distributed:

$$\boldsymbol{\kappa}_j^q \sim N\left(\boldsymbol{\eta}_j^q, \{\boldsymbol{\Omega}_j^q\}^{-1}\right). \tag{J.7}$$

J.1.2 Deriving conditional posterior distributions

Random regression coefficients γ_j^q

Using the likelihood in Equation J.7 and prior distribution $\gamma_j^q \sim N(\gamma^q, \{\mathbf{\Sigma}^q\})$, the conditional posterior distribution of random regression coefficients γ_j^q is also a multivariate normal distribution:

$$
\begin{aligned}
p(\gamma_j^q|.) \propto & p(\mathbf{y}_j|\gamma_j^q, \gamma_j^{-q}, \boldsymbol{\omega}_j^q, \mathbf{x})p(\gamma_j^q) \qquad \text{(J.8)}\\
\propto & \exp\left[-\frac{1}{2}\left(\boldsymbol{\kappa}_j^q - \boldsymbol{\eta}_j^q\right)^T \mathbf{\Omega}_j^q \left(\boldsymbol{\kappa}_j^q - \boldsymbol{\eta}_j^q\right)\right] \times \\
& \exp\left[-\frac{1}{2}\left(\gamma_j^q - \gamma^q\right)^T \{\mathbf{\Sigma}^q\}^{-1}\left(\gamma_j^q - \gamma^q\right)\right] \\
\propto & \exp\Bigg[-\frac{1}{2}\Big(\{\gamma_j^q\}^T(\{\mathbf{X}_j\}^T\mathbf{\Omega}_j^q\mathbf{X}_j + \{\mathbf{\Sigma}^q\}^{-1})\gamma_j^q - 2\{\gamma_j^q\}^T \\
& \qquad (\{\mathbf{X}_j\}^T\mathbf{\Omega}_j^q(\boldsymbol{\kappa}_j^q + \ln[\sum_{m\neq q}\exp(\mathbf{X}_j\gamma_j^m)]) + \{\mathbf{\Sigma}^q\}^{-1}\gamma^q)\Big)\Bigg] \\
\propto & \exp\Bigg[-\frac{1}{2}\left(\gamma_j^q - \mathbf{V}_{\gamma_j}^q(\{\mathbf{X}_j\}^T\mathbf{\Omega}_j^q(\boldsymbol{\kappa}_j^q + \ln[\sum_{m\neq q}\exp(\mathbf{X}_j\gamma_j^m)]) + \{\mathbf{\Sigma}^q\}^{-1}\gamma^q)\right)^T \\
& \qquad \{\mathbf{V}_{\gamma_j}^q\}^{-1}\left(\gamma_j^q - \mathbf{V}_{\gamma_j}^q(\{\mathbf{X}_j\}^T\mathbf{\Omega}_j^q(\boldsymbol{\kappa}_j^q + \ln[\sum_{m\neq q}\exp(\mathbf{X}_j\gamma_j^m)]) + \{\mathbf{\Sigma}^q\}^{-1}\gamma^q)\right)\Bigg] \\
\sim & N\left(\mathbf{V}_{\gamma_j}^q(\mathbf{X}_j\mathbf{\Omega}_j^q(\boldsymbol{\kappa}_j^q + \ln[\sum_{m\neq q}\exp(\mathbf{X}_j\gamma_j^m)]) + \{\mathbf{\Sigma}^q\}^{-1}\gamma^q), \mathbf{V}_{\gamma_j}^q\right)
\end{aligned}
$$

with prior mean vector γ^q, prior variance matrix $\mathbf{\Sigma}^q$ and posterior variance matrix $\mathbf{V}_{\gamma_j}^q = ({\mathbf{X}_j}^T\mathbf{\Omega}_j^q\mathbf{X}_j + \{\mathbf{\Sigma}^q\}^{-1})^{-1}$.

Random mean γ^q

When the posterior distribution of γ_j^q (Equation J.8) is included as a likelihood and combined with a $N(\mathbf{g}^q, \{\mathbf{G}^q\}^{-1})$ prior distribution, the conditional posterior distribution of random mean

γ^q is another multivariate normal distribution:

$$
\begin{aligned}
p(\gamma^q|.) \propto & \prod_{j=1}^{J} p(\gamma_j^q|\gamma^q, \mathbf{\Sigma}^q)p(\gamma^q) \qquad \text{(J.9)} \\
\propto & \prod_{j=1}^{J} \exp\left[-\frac{1}{2}(\gamma_j^q - \gamma^q)^T\{\mathbf{\Sigma}^q\}^{-1}(\gamma_j^q - \gamma^q)\right] \times \exp\left[-\frac{1}{2}(\gamma^q - \mathbf{g}^q)^T\mathbf{G}^q(\gamma^q - \mathbf{g}^q)\right] \\
\propto & \exp\left[-\frac{1}{2}(\{\gamma^q\}^T\left(J\{\mathbf{\Sigma}^q\}^{-1}\right)\gamma^q) - 2\{\gamma^q\}^T\left(\{\mathbf{\Sigma}^q\}^{-1}\sum_{j=1}^{J}\gamma_j^q\right)\right]\times \\
& \exp\left[-\frac{1}{2}\{\gamma^q\}^T\mathbf{G}^q\gamma^q - 2\{\gamma^q\}^T\mathbf{G}^q\mathbf{g}^q\right] \\
\propto & \exp\left[-\frac{1}{2}\{\gamma^q\}^T\left(J\{\mathbf{\Sigma}^q\}^{-1} + \mathbf{G}^q\right)\gamma^q - 2\{\gamma^q\}^T\left(\{\mathbf{\Sigma}^q\}^{-1}\sum_{j=1}^{J}\gamma_j^q + \mathbf{G}^q\mathbf{g}^q\right)\right] \\
\propto & \exp\left[-\frac{1}{2}\left(\gamma^q - \mathbf{V}_\gamma^q\left(\{\mathbf{\Sigma}^q\}^{-1}\sum_{j=1}^{J}\gamma_j^q + \mathbf{G}^q\mathbf{g}^q\right)\right)^T\{\mathbf{V}_\gamma^q\}^{-1}\right. \\
& \qquad \left.\left(\gamma^q - \mathbf{V}_\gamma^q\left(\{\mathbf{\Sigma}^q\}^{-1}\sum_{j=1}^{J}\gamma_j^q + \mathbf{G}^q\mathbf{g}^q\right)\right)\right] \\
\sim & N\left(\mathbf{V}_\gamma^q\left(\{\mathbf{\Sigma}^q\}^{-1}\sum_{j=1}^{J}\gamma_j^q + \mathbf{G}^q\mathbf{g}^q\right), \mathbf{V}_\gamma^q\right),
\end{aligned}
$$

with prior mean vector $\mathbf{g}^q$, prior precision matrix $\mathbf{G}^q$, and posterior variance matrix $\mathbf{V}_\gamma = (J\{\mathbf{\Sigma}^q\}^{-1} + \mathbf{G}^q)^{-1}$.

Random variance $\mathbf{\Sigma}^q$

When the posterior distribution of γ_j^q (Equation J.8) is included as a likelihood and combined with an inverse Wishart $\mathcal{W}^{-1}(j^0, \mathbf{S}^q)$ prior, the conditional posterior distribution of random

variance $\mathbf{\Sigma}^q$ is proportional to an inverse Wishart distribution:

$$
\begin{aligned}
p(\mathbf{\Sigma}^q|.) &\propto p(\gamma_j^q|\gamma^q, \mathbf{\Sigma}^q)p\{\mathbf{\Sigma}^q\} \qquad \text{(J.10)}\\
&\propto \prod_{j=1}^{J} |\mathbf{\Sigma}^q|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(\gamma_j^q - \gamma^q)^T\{\mathbf{\Sigma}^q\}^{-1}(\gamma_j^q - \gamma^q)\right] \times \\
&\quad |\mathbf{\Sigma}^q|^{\frac{1}{2}(j^0+p+1)} \exp\left[-\frac{1}{2}tr(\mathbf{S}^q\{\mathbf{\Sigma}^q\}^{-1})\right] \\
&\propto |\mathbf{\Sigma}^q|^{-\frac{1}{2}(j^0+J+P+1)} \exp\left[-\frac{1}{2}tr\left(\left(\mathbf{S}^q + \sum_{j=1}^{J}(\gamma_j^q - \gamma^q)(\gamma_j^q - \gamma^q)^T\right)\{\mathbf{\Sigma}^q\}^{-1}\right)\right] \\
&\sim \mathcal{W}^{-1}\left(j^0 + J, \mathbf{S}^q + \sum_{j=1}^{J}(\gamma_j^q - \gamma^q)(\gamma_j^q - \gamma^q)^T\right).
\end{aligned}
$$

J.2 Mixed effects model

A mixed effect model is defined as follows:

$$\phi_{ji}^q = f(\mathbf{x}_{ji}^F \boldsymbol{\beta}^q + \mathbf{x}_{ji}^R \gamma_j^q) \qquad \text{(J.11)}$$

where $\mathbf{x}_{ji}^F$ and $\mathbf{x}_{ji}^R$ are vectors of fixed and random covariates respectively. Vectors $\boldsymbol{\beta}^q$ and γ_j^q reflect the accompanying fixed and random regression coefficients. Function f refers to the multinomial logistic likelihood function.

The multivariate normal distribution of working variable $\boldsymbol{\kappa}_j^q$ then has the following form:

$$\boldsymbol{\kappa}_j^q \sim N\left(\boldsymbol{\eta}_j^q, \{\mathbf{\Omega}_j^q\}^{-1}\right). \qquad \text{(J.12)}$$

Here, $\boldsymbol{\eta}_j^q = \mathbf{X}_j^F \boldsymbol{\beta}^q + \mathbf{X}_j^R \gamma_j^q - \ln[\sum_{m \neq q} \exp(\mathbf{X}_j^F \boldsymbol{\beta}^m + \mathbf{X}_j^R \gamma_j^m)]$. The likelihood in Equation J.12 can be combined with the prior distributions in Equation J.4, complemented with a multivariate normally distributed prior on $\boldsymbol{\beta}^q$:

$$\boldsymbol{\beta}^q \sim N(\mathbf{b}^q, \mathbf{B}^q) \qquad \text{(J.13)}$$

The Gibbs sampling algorithm in list J.1 is extended with a distinct step for the fixed regression coefficients:

1. Sample fixed regression coefficients:

$$\boldsymbol{\beta}^{q(l)} \sim N\left(\mathbf{V}_{\boldsymbol{\beta}}^{q}(\sum_{j=1}^{J}\mathbf{X}_j^{F\,T}\boldsymbol{\Omega}_j^{q(l-1)}(\boldsymbol{\kappa}_j^{q(l-1)} - \mathbf{X}_j^{R}\boldsymbol{\gamma}_j^{q(l-1)} + \ln[\sum_{m\neq q}\exp(\mathbf{X}_j^{F}\boldsymbol{\beta}^{m(l)} + \mathbf{X}_j^{R}\boldsymbol{\gamma}_j^{m(l-1)})]) + \mathbf{B}^{q}\mathbf{b}^{q}), \mathbf{V}_{\boldsymbol{\beta}}^{q}\right)$$

   with prior mean vector $\mathbf{b}^q$, prior precision matrix $\mathbf{B}^q$ and posterior variance matrix $\mathbf{V}_{\boldsymbol{\beta}}^{q} = (\sum_{j=1}^{J}\mathbf{X}_j^{F\,T}\boldsymbol{\Omega}_j^{q(l-1)}\mathbf{X}_j^{F} + \mathbf{B}^{q})^{-1}$.

2. Sample mean random regression coefficients:

$$\boldsymbol{\gamma}^{q(l)} \sim N\left(\mathbf{V}_{\boldsymbol{\gamma}}^{q}(\{\boldsymbol{\Sigma}^{q(l-1)}\}^{-1}\sum_{j=1}^{J}\boldsymbol{\gamma}_j^{q(l-1)} + \mathbf{G}^{q}\mathbf{g}^{q}), \mathbf{V}_{\boldsymbol{\gamma}}^{q}\right)$$

   with prior mean vector $\mathbf{g}^q$, prior precision matrix $\mathbf{G}^q$ and posterior variance matrix $\mathbf{V}_{\boldsymbol{\gamma}} = (J\{\boldsymbol{\Sigma}^{q(l-1)}\}^{-1} + \mathbf{G}^{q})^{-1}$.

3. Sample covariance matrices of random regression coefficients:

$$\boldsymbol{\Sigma}^{q(l)} \sim \mathcal{W}^{-1}\left(j^0 + J, \boldsymbol{\Sigma}^0 + \sum_{j=1}^{J}\left(\boldsymbol{\gamma}_j^{q(l-1)} - \boldsymbol{\gamma}^{q(l)}\right)\left(\boldsymbol{\gamma}_j^{q(l-1)} - \boldsymbol{\gamma}^{q(l)}\right)^{T}\right)$$

   with prior hyperparameters $j^0 \geq P^R$ and $\boldsymbol{\Sigma}^0$.

4. For each j, sample random regression coefficients:

$$\boldsymbol{\gamma}_j^{q(l)} \sim N\left(\mathbf{V}_{\boldsymbol{\gamma}_j^q}^{q}(\mathbf{X}_j^{R}\boldsymbol{\Omega}_j^{q(l-1)}(\boldsymbol{\kappa}_j^{q(l-1)} - \mathbf{X}_j^{F}\boldsymbol{\beta}^{q(l)} + \ln[\sum_{m\neq q}\exp(\mathbf{X}_j^{F}\boldsymbol{\beta}^{m(l)} + \mathbf{X}_j^{R}\boldsymbol{\gamma}_j^{m(l)})]) + \{\boldsymbol{\Sigma}^{q}\}^{-1}\boldsymbol{\gamma}^{q}), \mathbf{V}_{\boldsymbol{\gamma}_j}^{q}\right)$$

   with prior mean vector $\boldsymbol{\gamma}^{q(l)}$, prior precision matrix $\boldsymbol{\Sigma}^{q(l)}$ and posterior variance matrix $\mathbf{V}_{\boldsymbol{\gamma}_j}^{q} = (\mathbf{X}_j^{R\,T}\boldsymbol{\Omega}_j^{q(l-1)}\mathbf{X}_j^{R} + \{\boldsymbol{\Sigma}^{q(l)}\}^{-1})^{-1}$.

5. For each j and i, sample Pólya-Gamma variables:

$$\omega_{ji}^{q(l)} \sim PG(1, \eta_{ji}^{q(l)})$$

J.3 A note on prior specification

J.3.1 Regression parameters

In the Gibbs sampling framework, regression coefficients are normally distributed with a mean and covariance matrix. We shortly discuss the role of these parameters below. The covariance matrix defines the spread of the distribution and therefore has a substantial influence on informativity: Small variance parameters increase prior information. When non-informativity is preferable, large variance parameters are not the simple answer, as they may destabilize computations in Bayesian logistic regression analysis (Gelman et al., 2008). Jeffreys prior could be an option, but sufficiently stable computation is not guaranteed (Gelman et al., 2008; Poirier, 1994). The challenge is therefore to specify prior variance parameters that are both sufficiently small to support stable analysis and to give a realistic support of the parameter and at the same time sufficiently large to be considered vague.

The mean hyperparameters define the center of the distribution and become increasingly influential on the posterior distribution when the variance of the distribution is small. The relevance of adequate mean hyperparameters therefore increases with the informativity of the analysis. It should be noted that prior information of mean regression coefficients is not always available in the required parametrization. Researchers may be more likely to have information available in terms of (success) probabilities rather than logistic regression parameters. Kavelaars et al. (2022b) propose an approach to compute mean hyperparameters for the context of treatment comparison in the presence of a single patient characteristics, based on expected joint response probabilities.

J.3.2 Covariance matrices

The covariance matrix follows an inverse-Wishart distribution with parameters. Specifying a non-informative prior on covariance matrices and variance parameters in general is not straightforward (Gelman, 2006; Schuurman et al., 2016). The informativity of the inverse-Wishart distribution is sensitive to the size of variance parameters: small variances make inverse-Wishart distributions more informative. Naively specifying standard prior hyperparameters without consideration of prior information or data at hand may result in an undesirably large prior influence. Weakly informative (data-based) prior specification may be superior, if not essential for computational stability (Gelman, 2006).

Appendix K

Procedure for transformation to the probability scale and decision-making

Algorithm 5 Procedure for statistical decision-making with posterior regression coefficients

Step 1. Transform regression coefficients to treatment differences
Let $\gamma_j^Q = (0, \dots, 0)$ and $\mathbf{x} = (1, T_j, w_j, \dots)$
for draw $(l) \leftarrow 1 : L$ **do**
  for cluster $j \leftarrow 1 : J$ **do**
    Compute joint response probabilities
    for treatment $T \leftarrow 0 : 1$ **do**
      for joint response category $q \leftarrow 1 : Q$ **do**
        if Population of interest defined by a range of values of w **then**

$$\text{Compute } \phi_{Tj}^{q(l)} = \int_w \frac{\exp\left[\mathbf{x}_j' \gamma_j^{q(l)}\right]}{\sum_{r=1}^{Q-1} \exp\left[\mathbf{x}_j' \gamma_j^{r(l)}\right] + 1} dw$$

        end if
        if Population of interest defined by a fixed value of w **then**

$$\text{Compute } \phi_{Tj}^{q(l)} = \frac{\exp\left[\mathbf{x}_j' \gamma_j^{q(l)}\right]}{\sum_{r=1}^{Q-1} \exp\left[\mathbf{x}_j' \gamma_j^{r(l)}\right] + 1}$$

        end if
      end for
      Compute multivariate success probabilities
      for outcome $k \leftarrow 1 : K$ **do**

$$\text{Compute } \theta_{Tj}^{q(l)} = \sum_{q=1}^{Q} \phi_{Tj}^{q(l)} I(\mathbf{h}^q \in \mathbf{U}_k)$$

        Compute multivariate treatment difference
        Compute $\delta_j^{k(l)} = \theta_{1j}^{k(l)} - \theta_{0j}^{k(l)}$
      end for
    end for
  end for
  for outcome $k \leftarrow 1 : K$ **do**

$$\text{Pool } \delta^{k(l)} = \sum_{j=1}^{J} \frac{n_j}{\sum_{j=1}^{J} n_j} \delta_j^{k(l)}$$

  end for
end for
Step 2. Make superiority decision
Define superiority region $\mathcal{S}_R$
Draw conclusion
if $\frac{1}{L} \sum_{(l)=1}^{L} I(\boldsymbol{\delta}^{(l)} \in \mathcal{S}_R) > p_{cut}$ **then** Conclude superiority
else Conclude non-superiority
end if

Bibliography

Anderson, K. M. (2016). *gsDesign: An R package for designing group sequential clinical trials, version 3.0 manual*.

Bauer, P., Bretz, F., Dragalin, V., König, F., & Wassmer, G. (2016). Twenty-five years of confirmatory adaptive designs: Opportunities and pitfalls. *Statistics in Medicine*, *35*(3), 325–347. https://doi.org/10.1002/sim.6472

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*(3). https://doi.org/10.1214/06-ba115

Berger, J. O. (2010). *Statistical decision theory and Bayesian analysis*. Springer New York.

Berger, J. O., Bernardo, J. M., & Sun, D. (2015). Overall objective priors. *Bayesian Analysis*, *10*(1), 189–221. https://doi.org/10.1214/14-BA915

Berry, S. M., Carlin, B. P., Lee, J. J., & Muller, P. (2010). *Bayesian adaptive methods for clinical trials*. CRC press.

Biswas, S., Liu, D. D., Lee, J. J., & Berry, D. A. (2009). Bayesian clinical trials at the University of Texas MD Anderson cancer center. *Clinical Trials*, *6*(3), 205–216. https://doi.org/10.1177/1740774519871471

Braeken, J., Tuerlinckx, F., & De Boeck, P. (2007). Copula functions for residual dependency. *Psychometrika*, *72*(3), 393. https://doi.org/10.1007/s11336-007-9005-4

Brown, P. D., Jaeckle, K., Ballman, K. V., Farace, E., Cerhan, J. H., Anderson, S. K., Carrero, X. W., Barker, F. G., Deming, R., Burri, S. H., Ménard, C., Chung, C., Stieber, V. W., Pollock, B. E., Galanis, E., Buckner, J. C., & Asher, A. L. (2016). Effect of radiosurgery alone vs radiosurgery with whole brain radiation therapy on cognitive function in patients with 1 to 3 brain metastases. *JAMA*, *316*(4), 401. https://doi.org/10.1001/jama.2016.9839

Chang, E. L., Wefel, J. S., Hess, K. R., Allen, P. K., Lang, F. F., Kornguth, D. G., Arbuckle, R. B., Swint, J. M., Shiu, A. S., Maor, M. H., & Meyers, C. A. (2009). Neurocognition in patients with brain metastases treated with radiosurgery or radiosurgery plus whole-brain irradiation: A randomised controlled trial. *The Lancet Oncology*, *10*(11), 1037–1044. https://doi.org/10.1016/s1470-2045(09)70263-3

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*(432), 1313–1321. https://doi.org/10.1080/01621459.1995.10476635

Chiu, W. A., & Crump, K. S. (2012). Using copulas to introduce dependence in dose-response modeling of multiple binary endpoints. *Journal of Agricultural, Biological, and Environmental Statistics*, *17*(1), 107–127. https://doi.org/10.1007/s13253-011-0078-2

Chow, S.-C., Shao, J., Wang, H., & Lokhnygina, Y. (2017). *Sample size calculations in clinical research* (3rd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9781315183084

Chuang-Stein, C., Stryszak, P., Dmitrienko, A., & Offen, W. (2006). Challenge of multiple co-primary endpoints: A new approach. *Statistics in Medicine*, *26*(6), 1181–1192. https://doi.org/10.1002/sim.2604

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., & Swinton, J. (2019). *xtable: Export tables to LaTeX or HTML* [R package version 1.8-4]. https://CRAN.R-project.org/package=xtable

Dai, B., Ding, S., & Wahba, G. (2013). Multivariate Bernoulli distribution. *Bernoulli*, *19*(4), 1465–1483. https://doi.org/10.3150/12-BEJSP10

De Jong, V. M. T., Eijkemans, M. J. C., Calster, B., Timmerman, D., Moons, K. G. M., Steyerberg, E. W., & van Smeden, M. (2019). Sample size considerations and predictive performance of multinomial logistic prediction models. *Statistics in Medicine*, *38*(9), 1601–1619. https://doi.org/10.1002/sim.8063

Depaoli, S., & Van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods*, *22*(2), 240–261. https://doi.org/10.1037/met0000065

Eisuke, A., & Aoyama, H. (2011). The role of whole brain radiation therapy for the management of brain metastases in the era of stereotactic radiosurgery. *Current Oncology Reports*, *14*(1), 79–84. https://doi.org/10.1007/s11912-011-0201-0

Ellis, T. L., Neal, M. T., & Chan, M. D. (2012). The role of surgery, radiosurgery and whole brain radiation therapy in the management of patients with metastatic brain tumors.

International Journal of Surgical Oncology, *2012*, 1–10. https://doi.org/10.1155/2012/952345

Emerson, S. S., Kittelson, J. M., & Gillen, D. L. (2007). Frequentist evaluation of group sequential clinical trial designs. *Statistics in Medicine*, *26*(28), 5047–5080. https://doi.org/10.1002/sim.2901

European Medicine Agency. (2019). *Guideline on the investigation of subgroups in confirmatory clinical trials*.

Evans, D. (2003). Hierarchy of evidence: A framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing*, *12*(1), 77–84. https://doi.org/10.1046/j.1365-2702.2003.00662.x

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38. https://doi.org/10.1093/biomet/80.1.27

Flegal, J. M., Hughes, J., Vats, D., Dai, N., Gupta, K., & Maji, U. (2021). *Mcmcse: Monte Carlo standard errors for MCMC*.

Food and Drug Administration. (2010). *Guidance for industry adaptive design clinical trials for drugs and biologics. 2010*. Center for Biologics Evaluation; Research (CBER).

Food and Drug Administration. (2016). *Non-inferiority clinical trials to establish effectiveness: Guidance for industry*. Center for Biologics Evaluation; Research (CBER).

Food and Drug Administration. (2017). *Multiple endpoints in clinical trials guidance for industry.* Center for Biologics Evaluation; Research (CBER).

Food and Drug Administration. (2019). *Enrichment strategies for clinical trials to support determination of effectiveness of human drugs and biological products: Guidance for industry.* Center for Biologics Evaluation; Research (CBER).

Gallo, P. P. (2000). Center-weighting issues in multicenter clinical trials. *Journal of Biopharmaceutical Statistics*, *10*(2), 145–163. https://doi.org/10.1081/bip-100101019

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman; Hall/CRC.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383. https://doi.org/10.1214/08-aoas191

Goldberger, J. J., & Buxton, A. E. (2013). Personalized medicine vs guideline-based medicine. *JAMA*, *309*(24), 2559. https://doi.org/10.1001/jama.2013.6629

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, *1*(4), 223–231. https://doi.org/10.1207/s15328031us0104_02

Goodman, S. N. (2007). Stopping at nothing? Some dilemmas of data monitoring in clinical trials. *Annals of internal medicine*, *146*(12), 882–887. https://doi.org/10.7326/0003-4819-146-12-200706190-00010

Grol, R., & Grimshaw, J. (2003). From best evidence to best practice: Effective implementation of change in patients' care. *The Lancet*, *362*(9391), 1225–1230. https://doi.org/10.1016/s0140-6736(03)14546-1

Gulliford, M., Adams, G., Ukoumunne, O., Latinovic, R., Chinn, S., & Campbell, M. (2005). Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *Journal of Clinical Epidemiology*, *58*(3), 246–251. https://doi.org/10.1016/j.jclinepi.2004.08.012

Habets, E. J., Dirven, L., Wiggenraad, R. G., Kanter, A. V.-d., à Nijeholt, G. J. L., Zwinkels, H., Klein, M., & Taphoorn, M. J. (2015). Neurocognitive functioning and health-related quality of life in patients treated with stereotactic radiotherapy for

brain metastases: A prospective study. *Neuro-Oncology*, *18*(3), 435–444. https://doi.org/10.1093/neuonc/nov186

Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine*, *363*(4), 301–304. https://doi.org/10.1056/nejmp1006304

Harbour, R., & Miller, J. (2001). A new system for grading recommendations in evidence based guidelines. *BMJ*, *323*(7308), 334–336. https://doi.org/10.1136/bmj.323.7308.334

Hobbs, B. P., & Landin, R. (2018). Bayesian basket trial design with exchangeability monitoring. *Statistics in Medicine*, *37*(25), 3557–3572. https://doi.org/10.1002/sim.7893

Hoogland, J., IntHout, J., Belias, M., Rovers, M. M., Riley, R. D., Jr, F. E. H., Moons, K. G. M., Debray, T. P. A., & Reitsma, J. B. (2021). A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. *Statistics in Medicine*, *40*(26), 5961–5981. https://doi.org/10.1002/sim.9154

Hox, J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis*. Taylor & Francis Ltd.

Ibrahim, J. G., & Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, *15*(1), 46–60. https://doi.org/10.1214/ss/1009212673

International Stroke Trial Collaborative Group. (1997). The International Stroke Trial (IST): A randomised trial of Aspirin, subcutaneous Heparin, both, or neither among 19435 patients with acute ischaemic stroke. *The Lancet*, *349*(9065), 1569–1581. https://doi.org/10.1016/s0140-6736(97)04011-7

Jackson, C. H. (2011). Multi-state models for panel data: The msm package for R. *Journal of Statistical Software*, *38*(8), 1–29. http://www.jstatsoft.org/v38/i08/

Jennison, C., & Turnbull, B. W. (1999). *Group sequential methods with applications to clinical trials*. Chapman; Hall/CRC. https://doi.org/10.1201/9780367805326

Jones, B., D., T., Wang, J., & Lewis, J. A. (1998). A comparison of various estimators of a treatment difference for a multi-centre clinical trial. *Statistics in Medicine*, *17*(15-

16), 1767–1777. https://doi.org/{10.1002/(SICI)1097-0258(19980815/30)17:15/16<1767::AID-SIM978>3.0.CO;2-H}

Jones, H. E., Ohlssen, D. I., Neuenschwander, B., Racine, A., & Branson, M. (2011). Bayesian models for subgroup analysis in clinical trials. *Clinical Trials*, *8*(2), 129–143. https://doi.org/10.1177/1740774510396933

Kaptein, M. (2014). The use of Thompson sampling to increase estimation precision. *Behavior Research Methods*, *47*(2), 409–423. https://doi.org/10.3758/s13428-014-0480-0

Kaptein, M., Markopoulos, P., de Ruyter, B., & Aarts, E. (2015). Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies*, *77*, 38–51. https://doi.org/10.1016/j.ijhcs.2015.01.004

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*(435), 1343–1370. https://doi.org/10.1080/01621459.1996.10477003

Kavelaars, X. (2020). Going multivariate in clinical trial studies: A Bayesian framework for multiple binary outcomes. In R. Van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners*. Routledge. https://doi.org/10.4324/9780429273872

Kavelaars, X., Mulder, J., & Kaptein, M. (2020). Decision-making with multiple correlated binary outcomes in clinical trials. *Statistical Methods in Medical Research*, *29*(11), 3265–3277. https://doi.org/10.1177/0962280220922256

Kavelaars, X., Mulder, J., & Kaptein, M. (2022a). *Bayesian multilevel multivariate logistic regression for superiority decision-making under observable treatment heterogeneity.* [Submitted for publication].

Kavelaars, X., Mulder, J., & Kaptein, M. (2022b). *Bayesian multivariate logistic regression for superiority and inferiority decision-making under observed treatment heterogeneity.* [Submitted for publication].

Khalsa, S. S. S., Chinn, M., Krucoff, M., & Sherman, J. H. (2013). The role of stereotactic radiosurgery for multiple brain metastases in stable systemic disease: A review of the literature. *Acta Neurochirurgica*, *155*(7), 1321–1328. https://doi.org/10.1007/s00701-013-1701-5

Lek, K., & Van de Schoot, R. (2018). Development and evaluation of a digital expert elicitation method aimed at fostering elementary school teachers' diagnostic competence. *Frontiers in Education*, *3*. https://doi.org/10.3389/feduc.2018.00082

Leon-Novelo, L. G., Bekele, B. N., Müller, P., Quintana, F., & Wathen, K. (2012). Borrowing strength with nonexchangeable priors over subpopulations. *Biometrics*, *68*(2), 550–558. https://doi.org/10.1111/j.1541-0420.2011.01693.x

Lin, Z. (1999). An issue of statistical analysis in controlled multi-centre studies: How shall we weight the centres? *Statistics in Medicine*, *18*(4), 365–373. https://doi.org/10.1002/(sici)1097-0258(19990228)18:4<365::aid-sim46>3.0.co;2-2

Lindley, R. I., Wardlaw, J. M., Whiteley, W. N., Cohen, G., Blackwell, L., Murray, G. D., Sandercock, P. A., Baigent, C., Chadwick, D., Tyrrell, P., Lowe, G., Dennis, M., Innes, K., Goodare, H., Farrall, A., von Kummer, R., Cala, L., von Heijne, A., Morris, Z., ... Isaakson, E. (2015). Alteplase for acute ischemic stroke. *Stroke*, *46*(3), 746–756. https://doi.org/10.1161/strokeaha.114.006573

Lipkovich, I., Dmitrienko, A., & D'Agostino, R. B. (2016). Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, *36*(1), 136–196. https://doi.org/10.1002/sim.7064

Makalic, E., & Schmidt, D. (2016). High-dimensional Bayesian regularised regression with the bayesreg package. *arXiv*. https://doi.org/10.48550/ARXIV.1611.06649

Malik, H. J., & Abraham, B. (1973). Multivariate logistic distributions. *The Annals of Statistics*, *1*(3), 588–590. https://doi.org/10.1214/aos/1176342430

Marsman, M., & Wagenmakers, E.-J. (2016). Three insights from a Bayesian interpretation of the one-sided p value. *Educational and Psychological Measurement*, *77*(3), 529–539. https://doi.org/10.1177/0013164416669201

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software*, *42*(9), 22. https://doi.org/10.18637/jss.v042.i09

McGlothlin, A. E., & Viele, K. (2018). Bayesian hierarchical models. *JAMA*, *320*(22), 2365. https://doi.org/10.1001/jama.2018.17977

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 750–773. https://doi.org/10.1080/10705511.2016.1186549

Michael, J. K., Chen, N., Alexander, M. K., Jiang, X., H., A. X., & Brian, P. H. (2020). Analyzing basket trials under multisource exchangeability assumptions. *The R Journal*, *12*(2), 342. https://doi.org/10.32614/rj-2021-020

Microsoft & Weston, S. (2020a). *doParallel: Foreach parallel adaptor for the 'parallel' package* [R package version 1.0.16].

Microsoft & Weston, S. (2020b). *foreach: Provides foreach looping construct* [R package version 1.5.1].

Miočević, M., Levy, R., & Savord, A. (2020a). The role of exchangeability in sequential updating of findings from small studies and the challenges of identifying exchangeable data sets. In R. Van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners*. Routledge. https://doi.org/10.4324/9780429273872

Miočević, M., Levy, R., & Van de Schoot, R. (2020b). Introduction to Bayesian statistics. In R. Van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners*. Routledge. https://doi.org/10.4324/9780429273872

Mirnezami, R., Nicholson, J., & Darzi, A. (2012). Preparing for precision medicine. *New England Journal of Medicine*, *366*(6), 489–491. https://doi.org/10.1056/nejmp1114866

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2001). Optimal experimental designs for multilevel logistic models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *50*(1), 17–30. https://doi.org/10.1111/1467-9884.00257

Moerbeek, M., van Breukelen, G. J. P., & Berger, M. P. F. (2000). Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics*, *25*(3), 271. https://doi.org/10.2307/1165206

Murray, T. A., Thall, P. F., & Yuan, Y. (2016). Utility-based designs for randomized comparative trials with categorical outcomes. *Statistics in Medicine*, *35*(24), 4285–4305. https://doi.org/10.1002/sim.6989

Nemes, S., Jonasson, J. M., Genell, A., & Steineck, G. (2009). Bias in odds ratios by logistic regression modelling and sample size. *BMC Medical Research Methodology*, *9*(1). https://doi.org/10.1186/1471-2288-9-56

Neuenschwander, B., Wandel, S., Roychoudhury, S., & Bailey, S. (2015). Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharmaceutical Statistics*, *15*(2), 123–134. https://doi.org/10.1002/pst.1730

Ng, P. C., Murray, S. S., Levy, S., & Venter, J. C. (2009). An agenda for personalized medicine. *Nature*, *461*(7265), 724–726. https://doi.org/10.1038/461724a

Nikoloulopoulos, A. K., & Karlis, D. (2008). Multivariate logit copula model with an application to dental data. *Statistics in Medicine*, *27*(30), 6393–6406. https://doi.org/10.1002/sim.3449

Niranjan, A., Lunsford, L. D., & Emerick, R. L. (2012). Stereotactic radiosurgery for patients with metastatic brain tumors: Development of a consensus radiosurgery guideline recommendation. In D. Kim & L. Lunsford (Eds.), *Current and future management of brain metastasis.* (pp. 123–138). Karger. https://doi.org/10.1159/000331185

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, 1079–1087. https://doi.org/10.2307/2531158

O'Brien, S. M., & Dunson, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics*, *60*(3), 739–746. https://doi.org/10.1111/j.0006-341x.2004.00224.x

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., & Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. Wiley. https://doi.org/10.1002/0470033312

Oliveira, R., & Teixeira-Pinto, A. (2015). Analyzing multiple outcomes: Is it really worth the use of multivariate linear regression? *Journal of Biometrics & Biostatistics*, *06*(04). https://doi.org/10.4172/2155-6180.1000256

Olkin, I., & Trikalinos, T. A. (2015). Constructions for a bivariate beta distribution. *Statistics & Probability Letters*, *96*, 54–60. https://doi.org/10.1016/j.spl.2014.09.013

Panagiotelis, A., Czado, C., & Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, *107*(499), 1063–1072. https://doi.org/10.1080/01621459.2012.682850

Paul, S., Saha, K., & Balasooriya, U. (2003). An empirical investigation of different operating characteristics of several estimators of the intraclass correlation in the analysis of binary data. *Journal of Statistical Computation and Simulation*, *73*(7), 507–523. https://doi.org/10.1080/0094965021000050883

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). Coda: Convergence diagnosis and output analysis for MCMC. *R News*, *6*(1), 7–11.

Pocock, S. J. (1997). Clinical trials with multiple outcomes: A statistical perspective on their design, analysis, and interpretation. *Controlled clinical trials*, *18*(6), 530–545. https://doi.org/10.1016/S0197-2456(97)00008-1

Pocock, S. J., Geller, N. L., & Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics*, 487–498. https://doi.org/10.2307/2531989

Poirier, D. (1994). Jeffreys prior for logit models. *Journal of Econometrics*, *63*(2), 327–339. https://doi.org/10.1016/0304-4076(93)01556-2

Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, *108*(504), 1339–1349. https://doi.org/10.1080/01621459.2013.829001

Prentice, R. L. (1997). Discussion: On the role and analysis of secondary outcomes in clinical trials. *Controlled Clinical Trials*, *18*(6), 561–567. https://doi.org/10.1016/s0197-2456(96)00105-5

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rauch, G., & Kieser, M. (2015). Adaptive designs for clinical trials with multiple endpoints. *Clinical Investigation*, *5*(5), 433–435. https://doi.org/10.4155/cli.14.138

Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical linear models: Applications and data analysis methods*. SAGE.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199–213. https://doi.org/10.1037/1082-989x.5.2.199

Renfro, L. A., & Sargent, D. (2017). Statistical controversies in clinical research: Basket trials, umbrella trials, and other master protocols: A review and examples. *Annals of Oncology*, *28*(1), 34–43. https://doi.org/10.1093/annonc/mdw413

Renfro, L. A., & Mandrekar, S. J. (2017). Definitions and statistical properties of master protocols for personalized medicine in oncology. *Journal of Biopharmaceutical Statistics*, *28*(2), 217–228. https://doi.org/10.1080/10543406.2017.1372778

Ridout, M. S., Demetrio, C. G. B., & Firth, D. (1999). Estimating intraclass correlation for binary data. *Biometrics*, *55*(1), 137–148. https://doi.org/10.1111/j.0006-341x.1999.00137.x

Rietbergen, C., Groenwold, R. H. H., Hoijtink, H. J. A., Moons, K. G. M., & Klugkist, I. (2014). Expert elicitation of study weights for Bayesian analysis and meta-analysis. *Journal of Mixed Methods Research*, *10*(2), 168–181. https://doi.org/10.1177/1558689814553850

Ristl, R., Urach, S., Rosenkranz, G., & Posch, M. (2018). Methods for the analysis of multiple endpoints in small populations: A review. *Journal of Biopharmaceutical Statistics*, *29*(1), 1–29. https://doi.org/10.1080/10543406.2018.1489402

Rossi, P. E., Allenby, G. M., & McCulloch, R. (2005). *Bayesian statistics and marketing*. Wiley.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic bulletin & review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, *21*(2), 283–300. https://doi.org/10.3758/s13423-013-0518-9

Sanborn, A. N., Hills, T. T., Dougherty, M. R., Thomas, R. P., Erica, C. Y., & Sprenger, A. M. (2014). Reply to Rouder (2014): Good frequentist properties raise confidence. *Psychonomic Bulletin & Review*, *21*(2), 309–311. https://doi.org/10.3758/s13423-014-0607-4

Sandercock, P. A. G., Wardlaw, J., Lindley, R., Cohen, G., & Whiteley, W. (2016). The third International Stroke Trial (IST-3), 2000-2015. [dataset]. https://doi.org/10.7488/DS/1350

Sandercock, P. A., Niewada, M., & Członkowska, A. (2011). The International Stroke Trial database. *Trials*, *12*(1). https://doi.org/10.1186/1745-6215-12-101

Schimmel, W. C. M., Verhaak, E., Hanssens, P. E. J., Kavelaars, X. M., Mulder, J., Kaptein, M. C., Sitskoorn, M. M., & Gehring, K. (2022). P01.06.b interim results from CAR-Study B: An ongoing randomized trial on the effect of SRS or WBRT on cognitive performance in patients with 11-20 brain metastases. *Neuro-Oncology*, *24*(Suppl._2), ii24–ii24. https://doi.org/10.1093/neuonc/noac174.078

Schimmel, W. C. M., Verhaak, E., Hanssens, P. E. J., Gehring, K., & Sitskoorn, M. M. (2018). A randomised trial to compare cognitive outcome after Gamma Knife radiosurgery versus whole brain radiation therapy in patients with multiple brain metastases: Research protocol CAR-study B. *BMC Cancer*, *18*(1), 218. https://doi.org/10.1186/s12885-018-4106-2

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322. https://doi.org/10.1037/met0000061

Schork, N. J. (2015). Personalized medicine: Time for one-person trials. *Nature*, *520*(7549), 609–611. https://doi.org/10.1038/520609a

Schou, I. M., & Marschner, I. C. (2013). Meta-analysis of clinical trials with early stopping: An investigation of potential bias. *Statistics in Medicine*, *32*(28), 4859–4874. https://doi.org/10.1002/sim.5893

Schuurman, N. K., Grasman, R. P. P. P., & Hamaker, E. L. (2016). A comparison of inverse-Wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research*, *51*(2-3), 185–206. https://doi.org/10.1080/00273171.2015.1065398

Seaman, J. W., Seaman, J. W., & Stamey, J. D. (2012). Hidden dangers of specifying noninformative priors. *The American Statistician*, *66*(2), 77–84. https://doi.org/10.1080/00031305.2012.695938

Senn, S. (2014). A note regarding meta-analysis of sequential trials with stopping for efficacy. *Pharmaceutical Statistics*, *13*(6), 371–375. https://doi.org/10.1002/pst.1639

Senn, S., & Bretz, F. (2007). Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics*, *6*(3), 161–170. https://doi.org/10.1002/pst.301

Shi, H., & Yin, G. (2019). Control of Type I error rates in Bayesian sequential designs. *Bayesian Analysis*, *14*(2), 399–425. https://doi.org/10.1214/18-BA1109

Simon, R. (2010). Clinical trials for predictive medicine: New challenges and paradigms. *Clinical Trials*, *7*(5), 516–524. https://doi.org/10.1177/1740774510366454

Smid, S. C., & Winter, S. D. (2020). Dangers of the defaults: A tutorial on the impact of default priors when using Bayesian SEM with small samples. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.611963

Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In B. Everitt & D. HowelL (Eds.), *Encyclopedia of statistics in behavioral science*. Wiley. https://doi.org/10.1002/0470013192.bsa492

Sozu, T., Sugimoto, T., & Hamasaki, T. (2010). Sample size determination in clinical trials with multiple co-primary binary endpoints. *Statistics in Medicine*, *29*, 2169–2179. https://doi.org/10.1002/sim.3972

Sozu, T., Sugimoto, T., & Hamasaki, T. (2012). Sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables. *Biometrical Journal*, *54*, 716–729. https://doi.org/10.1002/bimj.201100221

Sozu, T., Sugimoto, T., & Hamasaki, T. (2016). Reducing unnecessary measurements in clinical trials with multiple primary endpoints. *Journal of Biopharmaceutical Statistics*, *26*(4), 631–643. https://doi.org/10.1080/10543406.2015.1052497

Su, T.-L., Glimm, E., Whitehead, J., & Branson, M. (2012). An evaluation of methods for testing hypotheses relating to two endpoints in a single clinical trial. *Pharmaceutical Statistics*, *11*(2), 107–117. https://doi.org/10.1002/pst.504

Sullivan, S. G., & Greenland, S. (2012). Bayesian regression in SAS software. *International Journal of Epidemiology*, *42*(1), 308–317. https://doi.org/10.1093/ije/dys213

Tang, D.-I., Geller, N. L., & Pocock, S. J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics*, *49*(1), 23. https://doi.org/10.2307/2532599

Tang, D.-I., Gnecco, C., & Geller, N. L. (1989). Design of group sequential clinical trials with multiple endpoints. *Journal of the American Statistical Association*, *84*(407), 775–779. https://doi.org/10.1111/j.0006-341X.2004.00146.x

Thall, P. F. (2020). Bayesian cancer clinical trial designs with subgroup-specific decisions. *Contemporary Clinical Trials*, *90*, 105860. https://doi.org/10.1016/j.cct.2019.105860

The International Stroke Trial-3 Collaborative Group. (2012). The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6 h of acute ischaemic stroke (the third International Stroke Trial [IST-3]): A randomised controlled trial. *The Lancet*, *379*(9834), 2352–2363. https://doi.org/10.1016/s0140-6736(12)60768-5

Thorlund, K., Haggstrom, J., Park, J. J., & Mills, E. J. (2018). Key design considerations for adaptive clinical trials: A primer for clinicians. *BMJ*, *360*, k698. https://doi.org/10.1136/bmj.k698

Van de Schoot, R., & Miočević, M. (2020). *Small sample size solutions: A guide for applied researchers and practitioners*. Routledge.

Van de Schoot, R., Veen, D., Smeets, L., Winter, S. D., & Depaoli, S. (2020). A tutorial on using the wambs checklist to avoid the misuse of Bayesian statistics. In R. Van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners*. Routledge. https://doi.org/10.4324/9780429273872

Van Ravenzwaaij, D., Monden, R., Tendeiro, J. N., & Ioannidis, J. P. A. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC Medical Research Methodology*, *19*(1). https://doi.org/10.1186/s12874-019-0699-7

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer. http://www.stats.ox.ac.uk/pub/MASS4/

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., Kinnersley, N., Lindborg, S., et al. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, *13*(1), 41–54. https://doi.org/10.1002/pst.1589

Wang, M., Spiegelman, D., Kuchiba, A., Lochhead, P., Kim, S., Chan, A. T., Poole, E. M., Tamimi, R., Tworoger, S. S., Giovannucci, E., Rosner, B., & Ogino, S. (2015). Statistical methods for studying disease subtype heterogeneity. *Statistics in Medicine*, *35*(5), 782–800. https://doi.org/10.1002/sim.6793

Whitehead, J., Branson, M., & Todd, S. (2010). A combined score test for binary and ordinal endpoints from clinical trials. *Statistics in Medicine*, *29*(5), 521–532. https://doi.org/10.1002/sim.3822

Wickham, H., & Miller, E. (2021). *haven: Import and export 'SPSS', 'Stata' and 'SAS' files* [R package version 2.4.3].

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, *116*(4), 1195–1200. https://doi.org/10.1073/pnas.1814092116

Woodcock, J., & LaVange, L. M. (2017). Master protocols to study multiple therapies, multiple diseases, or both. *New England Journal of Medicine*, *377*(1), 62–70. https://doi.org/10.1056/NEJMra1510062

Xiong, C., Yu, K., Gao, F., Yan, Y., & Zhang, Z. (2005). Power and sample size for clinical trials when efficacy is required in multiple endpoints: Application to an Alzheimer's treatment trial. *Clinical Trials*, *2*(5), 387–393. https://doi.org/10.1191/1740774505cn112oa

Yamamoto, M., Serizawa, T., Shuto, T., Akabane, A., Higuchi, Y., Kawagishi, J., Yamanaka, K., Sato, Y., Jokura, H., Yomo, S., Nagano, O., Kenai, H., Moriki, A., Suzuki, S., Kida, Y., Iwai, Y., Hayashi, M., Onishi, H., Gondo, M., ... Tsuchiya, K. (2014). Stereotactic radiosurgery for patients with multiple brain metastases (JLGK0901): A multi-institutional prospective observational study. *The Lancet Oncology*, *15*(4), 387–395. https://doi.org/10.1016/s1470-2045(14)70061-0

Yang, R., & Berger, J. O. (1996). *A catalog of noninformative priors*. Institute of Statistics; Decision Sciences, Duke University.

Yang, S., Li, F., Thomas, L. E., & Li, F. (2021). Covariate adjustment in subgroup analyses of randomized clinical trials: A propensity score approach. *Clinical Trials*, *18*(5), 570–581. https://doi.org/10.1177/17407745211028588

Zhang, J. J., Blumenthal, G. M., He, K., Tang, S., Cortazar, P., & Sridhara, R. (2012). Overestimation of the effect size in group sequential trials. *Clinical Cancer Research*, *18*(18), 4872–4876. https://doi.org/10.1158/1078-0432.CCR-11-3118

Zhao, Y., Grambsch, P. M., & Neaton, J. D. (2007). A decision rule for sequential monitoring of clinical trials with a primary and supportive outcome. *Clinical Trials*, *4*(2), 140–153. https://doi.org/10.1177/1740774507076936

Zondervan-Zwijnenburg, M., Van de Schoot-Hubeek, W., Lek, K., Hoijtink, H., & Van de Schoot, R. (2017). Application and evaluation of an expert judgment elicitation procedure for correlations. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00090

Summary

In medical research, Randomized Controlled Trials (RCTs) are considered the gold standard by which we evaluate the effects of new treatments, therapies, and interventions. While being a robust standard with favorable properties, one of the challenges to RCT methodology is the personalization of medicine: The ideas that patients with different characteristics respond differently to treatments and that we can prescribe better treatments to patients if we take these characteristics into account. Personalization demands RCTs to answer new, more complex research questions and to provide additional information that clinicians need to support treatment prescriptions to individual patients with adequate evidence. At the same time, personalization potentially affects characteristics of datasets that are relevant for the choice of analysis techniques.

These two developments require novel methods a) to create more extensive overviews of treatment effects among a range of diverse patient populations; and/or b) to reduce the required number of participants without compromising decision error rates. Thus, sharing information between outcome variables and subpopulations can greatly improve the value of RCTs in personalized medicine, since it a) borrows strength from other variables to improve the efficiency of clinical trial methodology; b) enables more refined decisions thereby facilitating alignment of trial conduct and clinical decision-making; and c) creates more comprehensive insights into the way treatment effects vary over related, but different subpopulations.

In the current dissertation, we implemented the idea of information-sharing in a Bayesian multivariate framework for RCT data with multiple correlated binary outcome variables. Central to the framework are three components: a multivariate analysis model for multiple binary outcome variables to benefit from the correlation between outcome variables; a transformation procedure to make the resulting model parameters interpretable in terms of (multivariate) success probabilities and differences between them; and a decision procedure to make treatment comparisons and draw conclusions regarding superiority and inferiority with prespecified frequentist error rates. Together, these three components form a comprehensive framework for statistical analysis and decision-making with multiple (correlated) binary outcome variables.

Throughout the dissertation we presented and evaluated three increasingly complex variations of the modeling element of the framework. First, we presented a conjugate Bayesian analysis technique based on a multivariate Bernoulli model to analyze multiple binary outcome variables and the relation between them simultaneously. Second, we presented a multivariate logistic regression model to also include the relation with observed covariates in the analysis, to enable decision-making for (groups of) patients with specific characteristics. Finally, we extended the presented multivariate logistic regression model to the multilevel context to accommodate clustered data as well.

Acknowledgements

Of course, I would not have been able to write this dissertation without the help of many other people. I am grateful to everyone who contributed to the research or writing process directly or indirectly. Here, I would like to mention a few persons in particular.

First and foremost, I would like to thank my supervisors, Maurits and Joris. Thank you for giving me the chance to start this project, for supporting me throughout this journey, for investing time and effort in me, and for answering all my questions. I learned a lot from your knowledge and advice. I appreciated the increasing amount of independence I was given over the course of time. Maurits, thank you for your punctuality and your widely applicable writing tips. Joris, thank you for your inspiring suggestions and your calm approach to supervision.

I would also like to thank the members of my PhD committee. Thank you for your time to read, review, and discuss my manuscript, as well as for attending my defense. Prof. Klugkist and dr. Van Smeden did not only serve on my committee, but also provided me with valuable lessons on methodology and statistics during my time in the M&S Research Master program at Utrecht University. Their teachings laid a solid foundation for my dissertation, and I am grateful for the rigorous preparation for academia that I received from them and other teachers.

Further, I want to thank the Dutch Research Council (Nederlandse Organisatie voor Wetenschappelijk Onderzoek; NWO) for funding this project. The grant application owes its success to the valuable input and feedback received from the colleagues in MTO who assisted me in obtaining it. Also, the project cannot be seen apart from the research team of Car Study-B: Wietske, Eline, Karin, Margriet, and Patrick. Thank you for inspiring my PhD project and for providing the opportunity to be involved in the trial.

To my (former) colleagues from the MTO department: thank you for all the inspiring chats, talks, discussions, and other encounters that were fun, educational, or both. Anne-Marie and Marieke, our (former) secretaries, thank you for your help with whatever question I had. Esther and Chris, going against the grain can appear almost inevitable when trying to live up

to values and principles. Thank you for your share in paving the road and for supporting me in navigating through academia.

To the members of the Computational Personalization lab: I appreciated all your valuable insights, elaborate feedback, and interesting discussions. As much as I learned from your feedback on my own work, I also learned a lot from your research projects and processes.

A special thanks to my paranymphs, Ylva and Lingjie. Thank you for your support throughout the years, and especially during my defense preparation. Our enjoyable (online and offline) meetings are memorable, as are our lively discussions. I hope the next round will come soon!

Apart from the support I received from work, there were also several individuals from my personal circle who made valuable contributions. Bernadette, thank you for the insights you shared, the seeds you planted, and the practical suggestions you provided to help me navigate through the early years of this trajectory. You could not have foreseen how much I needed many of these lessons in later years. Ruth, thank you for your great flexibility, your almost unlimited patience, and your genuine curiosity. Lita, thank you for your care and your knowledge-sharing.

Manon and Razma, thank you for maintaining our friendship ever since high school and supporting me through my PhD as well. Whether it was to blow off steam, to detach from work, or to just engage in our (more or less frequent) drinks, walks, and talks: I all enjoyed them and hope for many more to come!

Lynn and Timo, thank you for all the fun and the serious moments we shared together. You have no idea how much I learned from the two of you.

Grandma, thank you for the wise lessons you taught me and for the times you looked after me when I was young. At the time of writing, I am not sure whether you will be able to attend my defense. Your presence would be a gift to me!

Veruschka, thank you for being readily available to assist me whenever I asked for help. A special thanks for taking care of my furry friends from time to time. They could not have wished for a better petsitter.

To my parents: thank you for the many opportunities you gave me to develop myself. Throughout my life, you provided me with a lot of support and with many wise lessons that directly or indirectly made their contributions to this dissertation. These are things that I will never take for granted.

Last, but definitely not least, Alexe. Thank you for putting your creativity and effort into the cover design. You can be proud of the way you translated your drawing skills into such a personalized cover. While the list of other things that I'm grateful to you for is quite long, I'll keep it short here. Just a big thank you, for everything.