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

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

Xynthia Kavelaars[a]*, Joris Mulder[a], and Maurits Kaptein [b]

[a] Department of Methodology and Statistics, Tilburg University, The Netherlands

[b] Jheronimus Academy of Data Science, The Netherlands

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 tranfer 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.

Keywords: Bayesian multivariate logistic regression, treatment heterogeneity, multiple outcome variables, Bayesian analysis, Pólya-Gamma, subgroup analysis

*Contact: X.M. Kavelaars x.m.kavelaars@tilburguniversity.edu

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 to refer to psychological and medical interventions in the broad sense with 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 weighted in various ways to provide a single statistical decision regarding superiority or inferiority (e.g. Pocock et al., 1987; O'Brien, 1984; Murray et al., 2016). 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 (Senn and Bretz, 2007; Ristl et al., 2018; Food and Drug Administration, 2017; Murray et al., 2016). 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. Wang et al., 2015; Yang et al., 2021; Jones et al., 2011), 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 and Buxton, 2013; Hamburg and Collins, 2010; Wang et al., 2015; Simon, 2010). 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; Sandercock et al., 2011; International Stroke Trial Collaborative Group, 1997). 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 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 in 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 and Dunson, 2004). Several existing multivariate models lack insight into marginal distributions (e.g Malik and 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 and 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 faciliate treatment comparison among subpopulations and contribute to personalized treatment assignment. The proposed modelling 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.

2 Decision-framework

2.1 Multivariate logistic regression

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

$$\boldsymbol{H} = \begin{bmatrix} 1 & 1 & \ldots & 1 & 1 \\ 1 & 1 & \ldots & 1 & 0 \\ & & \ldots & & \\ 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & \ldots & 0 & 0 \end{bmatrix} \tag{1}$$

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

Joint response probability $\phi_i^q(\boldsymbol{x}_i)$ maps the dependency of joint response probabilities on covariates $\boldsymbol{x}_i$ via a multinomial logistic function:

$$\phi_i^q(\boldsymbol{x}_i) = \frac{\exp\left[\psi_i^q(\boldsymbol{x}_i)\right]}{\sum_{r=1}^{Q-1} \exp\left[\psi_i^r(\boldsymbol{x}_i)\right] + 1} \tag{2}$$

for response categories $q = 1, \ldots, Q-1$. In Equation 2, $\psi_i^q(\boldsymbol{x}_i)$ reflects the linear predictor of response

category q and subject i:

$$\psi_i^q(\boldsymbol{x}_i) = \beta_0^q + \beta_1^q x_{i1} + \cdots + \beta_P^q x_{iP}. \tag{3}$$

Here, x_{ip} can be a treatment indicator, a patient characteristic, or an interaction between these. Vector $\boldsymbol{\beta}^q = (\beta_0^q, \beta_1^q, \ldots, \beta_P^q)$ 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 2 of Q response categories:

$$l(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{x}) = \prod_{i=1}^{n} \prod_{q=1}^{Q-1} \left(\frac{\exp\left[\psi_i^q(\boldsymbol{x}_i)\right]}{\sum_{r=1}^{Q-1} \exp\left[\psi_i^r(\boldsymbol{x}_i)\right] + 1} \right)^{I(\boldsymbol{y}_i = \boldsymbol{H}_{q\cdot})} \left(\frac{1}{\sum_{r=1}^{Q-1} \exp\left[\psi_i^r(\boldsymbol{x}_i)\right] + 1} \right)^{I(\boldsymbol{y}_i = \boldsymbol{H}_{Q\cdot})}, \tag{4}$$

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

$$p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{x}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{x}) p(\boldsymbol{\beta}), \tag{5}$$

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 Polya-Gamma expansion (Polson et al., 2013). Computational details of this procedure can be found in Appendix A.

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 trial as running example.

Suppose we are interested in the effect of a combined drug therapy (Heparin plus Asparin; 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:

$$\psi_i^q(\boldsymbol{x}_i) = \beta_0^q + \beta_1^q T_i + \beta_2^q bp_i + \beta_2^q bp_i T_i, \tag{6}$$

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

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

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

$$\phi_i^q(\boldsymbol{x}_i) = \frac{\exp\left[\psi_i^q(\boldsymbol{x}_i)\right]}{\sum_{r=1}^{Q-1} \exp\left[\psi_i^r(\boldsymbol{x}_i)\right] + 1} \tag{2 revisited}$$

   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:

$$\begin{aligned} \phi_{T_i}^3(\boldsymbol{x}_i) &= p(\boldsymbol{y}_i(\boldsymbol{x}_i) = \{01\}) \\ &= \frac{\exp\left[\psi_i^3(\boldsymbol{x}_i)\right]}{\sum_{r=1}^{Q-1} \exp\left[\psi_i^r(\boldsymbol{x}_i)\right] + 1}. \end{aligned} \tag{7}$$

   Note that we are interested in joint response probability $\boldsymbol{\phi}_T(\boldsymbol{x})$, which reflects a treatment effect among a population defined by $\boldsymbol{x}$ and is more general than the joint response probability of an individual patient with covariates $\boldsymbol{x}_i$. 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 3.

2. **Joint response probabilities $\boldsymbol{\phi}_T(\boldsymbol{x})$ to multivariate success probabilities $\boldsymbol{\theta}_T(\boldsymbol{x})$:**

   The next step in the transformation involves the conversion from joint response probabilities $\boldsymbol{\phi}_T(\boldsymbol{x})$ to multivariate success probabilities of individual outcome variables $\boldsymbol{\theta}_T(\boldsymbol{x})$. 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_T^k on outcome k and treatment

T equals the sum of a selection of elements of $\boldsymbol{\phi}_T$, denoted by matrix $\boldsymbol{U}_k$:

$$\theta_T^k(\boldsymbol{x}) = \sum_{q=1}^{Q} \phi_T^q(\boldsymbol{x}) I(\boldsymbol{H}_{q\cdot} \in \boldsymbol{U}_k). \tag{8}$$

Selection $\boldsymbol{U}_k$ consists of the 2^{K-1} rows of $\boldsymbol{H}$ 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 $\boldsymbol{x}$ for notational simplicity.

$$\boldsymbol{H} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \boldsymbol{U}_{strk} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \text{ and } \boldsymbol{U}_{dep} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

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

$$\begin{aligned} \theta_T^{strk} &= p(\boldsymbol{y}_i(\boldsymbol{x}_i) = \{11\}) + \ p(\boldsymbol{y}_i(\boldsymbol{x}_i) = \{10\}) \\ &= \phi_T^1 + \phi_T^2 \\ \theta_T^{dep} &= p(\boldsymbol{y}_i(\boldsymbol{x}_i) = \{11\}) + \ p(\boldsymbol{y}_i(\boldsymbol{x}_i) = \{01\}) \\ &= \phi_T^1 + \phi_T^3. \end{aligned} \tag{9}$$

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

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

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

$$\delta^k(\boldsymbol{x}) = \theta_1^k(\boldsymbol{x}) - \theta_0^k(\boldsymbol{x}). \tag{10}$$

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

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

differences:

$$\delta^{strk} = \theta^{strk}_{H+A} - \theta^{strk}_{A}$$
$$\delta^{dep} = \theta^{dep}_{H+A} - \theta^{dep}_{A}. \tag{11}$$

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 $\boldsymbol{\delta}(\boldsymbol{x})$. 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.

2.3 Posterior decision-making

Decisions rely on estimated treatment effects and their uncertainties. More formally, multivariate treatment difference $\boldsymbol{\delta}$ has complete parameter spaces $\mathcal{S} \subset [-1, 1]^K$, which is divided into a rejection region $\mathcal{S}_R$ and an 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 $\boldsymbol{\delta}(\boldsymbol{x})$ lies in the rejection region exceeds a prespecified decision threshold, p_{cut}:

$$p(\boldsymbol{\delta}(\boldsymbol{x}) \in \mathcal{S}_R | \boldsymbol{y}) > p_{cut}. \tag{12}$$

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 $\boldsymbol{\delta}(\boldsymbol{x})$. Equation 12 is then applied in

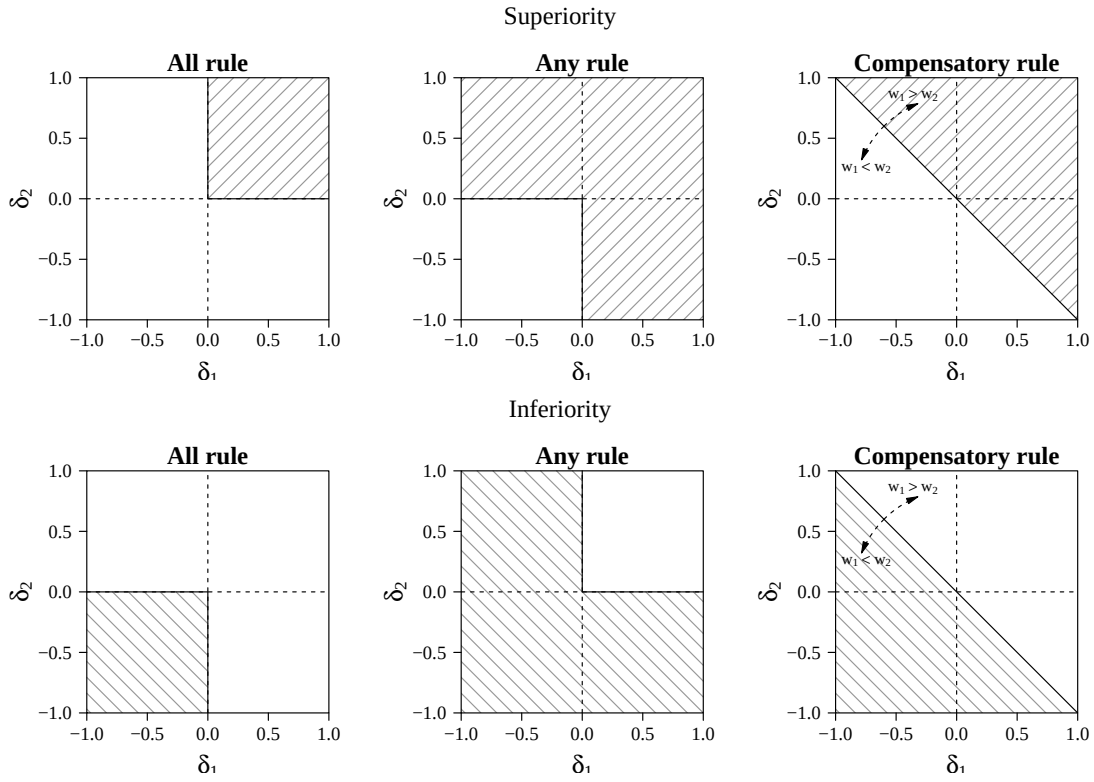

Figure 1

practice as:

$$\frac{1}{L}\sum_{(l)=1}^{L}\boldsymbol{\delta}^{(l)}(\boldsymbol{x}) \in \mathcal{S}_R|\boldsymbol{y} > p_{cut}. \tag{13}$$

In a situation with multiple outcome variables, superiority and inferiority can be defined in multiple ways, resulting in different rejection regions (e.g Pocock et al., 1987; Pocock, 1997; O'Brien, 1984; Prentice, 1997; Tang et al., 1993; Zhao et al., 2007). Although not intended as an exhaustive overview, we list three possible rules and graphically present their rejection regions in Figure 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 (Murray et al., 2016; Kavelaars et al., 2020).

1. **Any rule:** The Any rule results in superiority or inferiority when the difference between success prob-

abilities 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:

$$\mathcal{S}_S^{Any} = \boldsymbol{\delta}(\boldsymbol{x})|\max_{1<k<K} \delta^k(\boldsymbol{x}) > 0 \tag{14}$$
$$\mathcal{S}_I^{Any} = \boldsymbol{\delta}(\boldsymbol{x})|\min_{1<k<K} \delta^k(\boldsymbol{x}) < 0.$$

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 of the outcome variables (Sozu et al., 2010). The superiority and inferiority spaces are defined as:

$$\mathcal{S}_S^{All} = \boldsymbol{\delta}(\boldsymbol{x})|\min_{1<k<K} \delta^k(\boldsymbol{x}) > 0 \tag{15}$$
$$\mathcal{S}_I^{All} = \boldsymbol{\delta}(\boldsymbol{x})|\max_{1<k<K} \delta^k(\boldsymbol{x}) < 0.$$

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:

$$\mathcal{S}_S^{Comp}(\boldsymbol{w}) = \boldsymbol{\delta}(\boldsymbol{x})|\delta(\boldsymbol{w},\boldsymbol{x}) > 0 \tag{16}$$
$$\mathcal{S}_I^{Comp}(\boldsymbol{w}) = \boldsymbol{\delta}(\boldsymbol{x})|\delta(\boldsymbol{w},\boldsymbol{x}) < 0$$

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

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 (Sozu et al., 2016, 2010; Chow et al., 2017). 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 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 Food and Drug Administration (2017) and Kavelaars et al. (2020). For notational simplicity, we discard the dependence on $\boldsymbol{x}$ in the remainder of this subsection.

2.4.1 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 $\boldsymbol{\Psi}_K$ with mean $\mathbf{0}$ and correlation matrix $\boldsymbol{\Sigma}$ (Sozu et al., 2016):

$$1-\beta = \boldsymbol{\Psi}_K(c^1, \ldots, c^K). \tag{17}$$

In Equation 17, c^k for outcome k is defined by the decision rule of interest. Further, the off-diagonal elements of $\boldsymbol{\Sigma}$ denote (estimated) pairwise correlations between outcome variables.
For the Any rule,

$$c^k = z_{(1-\frac{\alpha}{K})} - \frac{(\theta_1^k - \theta_0^k)}{\sqrt{\frac{\theta_1^k(1-\theta_1^k)+\theta_0^k(1-\theta_0^k)}{n}}}. \tag{18}$$

For the All rule,

$$c^k = -z_{(1-\alpha)} + \frac{(\theta_1^k - \theta_0^k)}{\sqrt{\frac{\theta_1^k(1-\theta_1^k)+\theta_0^k(1-\theta_0^k)}{n}}}. \tag{19}$$

In Equations 18 and 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. (Sozu et al., 2010):

1. Plug in estimates of θ_T^k in Equation 18 or 19.

2. Plug in a starting value for n in Equation 18 or 19 and calculate the power via Equation 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

2.4.2 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, $\theta_T^{\boldsymbol{w}}$, results in:

$$n = \left[\theta_1^{\boldsymbol{w}}\left(1-\theta_1^{\boldsymbol{w}}\right) + \theta_0^{\boldsymbol{w}}\left(1-\theta_0^{\boldsymbol{w}}\right)\right]\left[\frac{z_{1-\alpha}+z_{1-\beta}}{\theta_1^{\boldsymbol{w}}-\theta_0^{\boldsymbol{w}}}\right]^2, \tag{20}$$

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

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 $\boldsymbol{\psi}_i^q(\boldsymbol{x})$ (cf. Equation 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, $\boldsymbol{x} = (z, T, zT)$ and $\psi_T^q(\boldsymbol{x})$ is defined as:

$$\psi_T^q(\boldsymbol{x}) = \beta_0^q + \beta_1^q T + \beta_2^q z + \beta_3^q zT. \tag{21}$$

3.1 Fixed values of covariate

For a patient population with fixed values of patient covariates, a posterior sample of joint response probabilities $\boldsymbol{\phi}_T(\boldsymbol{x})$ can be found by plugging in a vector of fixed covariate values $\boldsymbol{x}$ in linear predictor $\boldsymbol{\psi}_T^{(l)}(\boldsymbol{x})$. Subsequently applying the multinomial logistic link function in Equation 2 to each $\boldsymbol{\psi}_T^{(l)}(\boldsymbol{x})$ results in joint response probability $\boldsymbol{\phi}_T^{(l)}(\boldsymbol{x})$ 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 1 in Appendix C.

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 $\boldsymbol{\phi}_T^{(l)}(\boldsymbol{x})$ for draw (l). The procedure is presented in Algorithm 2 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 $\boldsymbol{x}$ 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 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] If this is the case, (numerical) integration can be an alternative to interpolate the conditional treatment effect distribution of interest.

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.

4.1 Setup

4.1.1 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 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 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 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 $\boldsymbol{\alpha}_T^n = \boldsymbol{\alpha}^0 + \{\sum_{i=1}^{n} I(T_i = T) I(\boldsymbol{y}_i = \boldsymbol{H}_{q\cdot})\}_{q=1}^{Q}$, 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 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 1.

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.

Table 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)	$\delta(\boldsymbol{w})$	ρ_{θ^1,θ^2}	(δ_1, δ_2)	$\delta(\boldsymbol{w})$	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

2. **Subpopulation:**

   We assessed the conditional treatment effect 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. As a consequence, 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 4.1.2:

1. **Any rule**

2. **All rule**

3. **Compensatory rule** with equal weights ($\boldsymbol{w} = (0.50, 0.50)$)

We computed sample sizes per treatment group via the procedures from Subsection 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 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 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**

4.1.2 Procedure

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

$$\psi_i^q(x_i) = \beta_0^q + \beta_T^q T_i + \beta_1^q z_i + \beta_2^q z_i T_i. \tag{22}$$

To generate response data, we first applied the multinomial logistic link function (Equation 2) to each true linear predictor $\boldsymbol{\psi}_i(x_i)$ to obtain joint response probabilities $\boldsymbol{\phi}_i(x_i)$ for each subject i. Next, we sampled response vector $\boldsymbol{y}_i|\boldsymbol{x}_i$ from a multinomial distribution with probabilities $\boldsymbol{\phi}_i(x_i)$.

Prior distribution For the multivariate logistic regression analysis, we used multivariate normally distributed prior with means $\boldsymbol{b}^q = \mathbf{0}$ and precision matrix $\boldsymbol{B}^{0q} = \text{diag}\ (1e^{-2}, \ldots, 1e^{-2})$ 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 $1,000$ burnin 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 2.2 and 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) (Marsman and Wagenmakers, 2016; Kavelaars et al., 2020; Sozu et al., 2016).

4.2 Results

4.2.1 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 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 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} $\boldsymbol{\delta}(\boldsymbol{x})$	n_{All} $\boldsymbol{\delta}(\boldsymbol{x})$	$n_{\text{Compensatory}}$ $\delta(\boldsymbol{w}, \boldsymbol{x})$
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

4.2.2 Decision error rates

Probabilities to conclude superiority of average treatment effects are presented in Table 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 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

Table 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 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)

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.

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 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.

5.1 Method

We applied the two procedures from Subsection 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 $\boldsymbol{b}^q = \mathbf{0}$ and precision matrix $\boldsymbol{B}^0 = \text{diag}(1e^{-2}, \ldots, 1e^{-2})$ 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 $10,000$ burnin 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 $\boldsymbol{w} = (0.25, 0.75)$ for recurring stroke in 14 days and dependency at 6 months respectively. These weights implied that the longterm outcome was three times more relevant for the decision than the shortterm 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).

Table 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	$\delta(\boldsymbol{w}, Bp)$	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	$>$

Table 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	$\delta(\boldsymbol{w}, Bp)$	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	$>$

5.2 Results

Results are presented in Table 6 for different intervals and in Table 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. 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 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.

6 Discussion

The current paper proposed a novel multivariate logistic regression framework to identify heteregenous 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 flexiblity comes with additional parameters, compared to other multivariate logistic models for correlated binary outcome variables (e.g. Malik and Abraham, 1973; O'Brien and 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 modelling procedure may benefit from generalization to methods that are more flexible with respect to

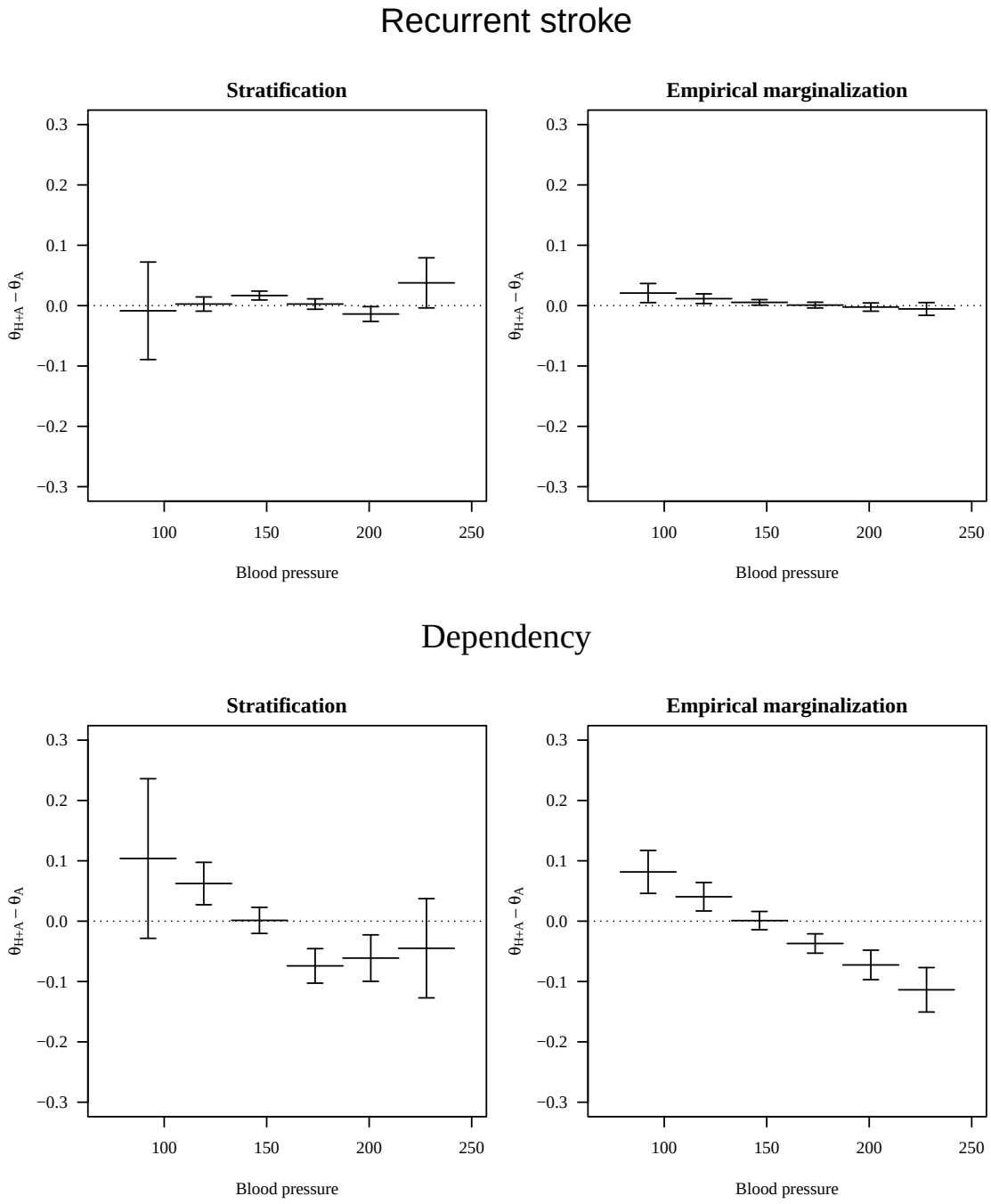
Recurrent stroke
Stratification
Empirical marginalization
$\theta_{H+A} - \theta_A$
Blood pressure
Dependency
Stratification
Empirical marginalization
$\theta_{H+A} - \theta_A$
Blood pressure

Figure 2

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 determines 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 for 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 relies 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 which considerations meanfully 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 in 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 (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 three anonymous reviewers for their valuable feedback on an earlier draft. Also, we thank The International Stroke Trial Collaborative Group for making the data from the second International Stroke Trial publicly available.

Funding

The current work was supported by the Dutch Research Council (NWO) [no. 406.18.505]. The second International Stroke Trial was principally funded by the UK Medical Research Council, the UK Stroke Association, and the European Union BIOMED-1 program. Limited support for collaborators' meetings and travel was provided by Eli Lilly, Sterling Winthrop (now Bayer USA), Sanofi, and Bayer UK. Follow-up in Australia was supported by a grant from the National Heart Foundation and in Canada by a Nova Scotia Heart and Stroke Foundation grant. Czech Republic IST was supported by a grant from the IGA Ministry of Health. India IST was supported by the McMaster INCLEN program and the All India Institute of Medical Sciences. The IST in New Zealand was funded by the Julius Brendel Trust and the Lottery Grants Board. In Norway, the IST was supported by the Norwegian Council on Cardiovascular Disease and Nycomed (for insurance).

Declaration of interest

The Authors declare that there is no conflict of interest.

Data 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-logis

References

Chen, M.-H. and Ibrahim, J. G. (2000). Power prior distributions for regression models. *Statistical Science*, 15(1):46–60.

Chib, S. (1995). Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321.

Chow, S.-C., Shao, J., Wang, H., and Lokhnygina, Y. (2017). *Sample Size Calculations in Clinical Research: Third edition*. Chapman and Hall/CRC.

Chuang-Stein, C., Stryszak, P., Dmitrienko, A., and Offen, W. (2006). Challenge of multiple co-primary endpoints: a new approach. *Statistics in Medicine*, 26(6):1181–1192.

Dai, B., Ding, S., Wahba, G., et al. (2013). Multivariate bernoulli distribution. *Bernoulli*, 19(4):1465–1483.

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

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38.

Food and Drug Administration (2016). *Non-Inferiority Clinical Trials to Establish Effectiveness: Guidance for Industry*.

Food and Drug Administration (2017). *Multiple Endpoints in Clinical Trials Guidance for Industry*. Center for Biologics Evaluation and 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 and Research (CBER).

Gelman, A., Jakulin, A., Pittau, M. G., and 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.

Goldberger, J. J. and Buxton, A. E. (2013). Personalized medicine vs guideline-based medicine. *JAMA*, 309(24):2559.

Hamburg, M. A. and Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304.

Hoogland, J., IntHout, J., Belias, M., Rovers, M. M., Riley, R. D., Jr, F. E. H., Moons, K. G. M., Debray, T. P. A., and 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.

International Stroke Trial Collaborative Group (1997). The international stroke trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke. *The Lancet*, 349(9065):1569–1581.

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

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

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

Lipkovich, I., Dmitrienko, A., and B., R. (2016). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, 36(1):136–196.

Malik, H. J. and Abraham, B. (1973). Multivariate logistic distributions. *The Annals of Statistics*, 1(3):588–590.

Marsman, M. and Wagenmakers, E.-J. (2016). Three insights from a bayesian interpretation of the one-sided p value. *Educational and Psychological Measurement*, 77(3):529–539.

Murray, T. A., Thall, P. F., and Yuan, Y. (2016). Utility-based designs for randomized comparative trials with categorical outcomes. *Statistics in medicine*, 35(24):4285–4305.

Nemes, S., Jonasson, J. M., Genell, A., and Steineck, G. (2009). Bias in odds ratios by logistic regression modelling and sample size. *BMC Medical Research Methodology*, 9(1).

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, pages 1079–1087.

O'Brien, S. M. and Dunson, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics*, 60(3):739–746.

Olkin, I. and Trikalinos, T. A. (2015). Constructions for a bivariate beta distribution. *Statistics & Probability Letters*, 96:54 – 60.

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.

Pocock, S. J., Geller, N. L., and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics*, 43(3):487.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349.

Prentice, R. L. (1997). Discussion: On the role and analysis of secondary outcomes in clinical trials. *Controlled Clinical Trials*, 18(6):561–567.

Ristl, R., Urach, S., Rosenkranz, G., and Posch, M. (2018). Methods for the analysis of multiple endpoints in small populations: A review. *Journal of Biopharmaceutical Statistics*, 29(1):1–29.

Rossi, P. E., Allenby, G. M., and McCulloch, R. (2005). *Bayesian statistics and marketing*. John Wiley & Sons.

Sandercock, P. A., , Niewada, M., and Członkowska, A. (2011). The international stroke trial database. *Trials*, 12(1).

Senn, S. and Bretz, F. (2007). Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics*, 6(3):161–170.

Simon, R. (2010). Clinical trials for predictive medicine: new challenges and paradigms. *Clinical Trials*, 7(5):516–524.

Sozu, T., Sugimoto, T., and Hamasaki, T. (2010). Sample size determination in clinical trials with multiple co-primary binary endpoints. *Statistics in medicine*, 29:2169–2179.

Sozu, T., Sugimoto, T., and Hamasaki, T. (2016). Reducing unnecessary measurements in clinical trials with multiple primary endpoints. *Journal of biopharmaceutical statistics*, 26(4):631–643.

Sullivan, S. G. and Greenland, S. (2012). Bayesian regression in SAS software. *International Journal of Epidemiology*, 42(1):308–317.

Tang, D.-I., Geller, N. L., and Pocock, S. J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics*, 49(1):23.

Thall, P. F. (2020). Bayesian cancer clinical trial designs with subgroup-specific decisions. *Contemporary Clinical Trials*, 90:105860.

van Ravenzwaaij, D., Monden, R., Tendeiro, J. N., and Ioannidis, J. P. A. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC Medical Research Methodology*, 19(1).

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., and Ogino, S. (2015). Statistical methods for studying disease subtype heterogeneity. *Statistics in Medicine*, 35(5):782–800.

Xiong, C., Yu, K., Gao, F., Yan, Y., and 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.

Yang, S., Li, F., Thomas, L. E., and Li, F. (2021). Covariate adjustment in subgroup analyses of randomized clinical trials: A propensity score approach. *Clinical Trials*, 18(5):570–581.

Zhao, Y., Grambsch, P. M., and 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.

A 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 (e.g. ϕ, θ).

Starting from the likelihood of individual K-variate response $\mathbf{y}_i$ (Equation 2), the likelihood of n K-variate responses follows from taking the product over n individual joint response probabilities in Q response categories:

$$l(\mathbf{y}|\beta,\mathbf{x}) = \prod_{i=1}^{n}\prod_{q=1}^{Q-1}\left(\frac{\exp\left[\psi_i^q\right]}{\sum_{r=1}^{Q-1}\exp\left[\psi_i^r\right]+1}\right)^{I(\mathbf{y}_i=q)}\left(\frac{1}{\sum_{r=1}^{Q-1}\exp\left[\psi_i^r\right]+1}\right)^{I(\mathbf{y}_i=Q)}. \tag{23}$$

Following Polson et al. (Polson et al., 2013), we introduce the Pólya-gamma variable by rewriting the multivariate likelihood in Equation 23 as a series of binomial likelihoods. The likelihood of $\mathbf{y}$ conditional on the parameters of the q^{th} response category, β^q, then equals:

$$l(\mathbf{y}|\beta^q,\beta^{-q}) = \prod_{i=1}^{n}\left(\frac{\exp\left[\eta_i^q\right])}{\exp\left[\eta_i^q\right]+1}\right)^{I(\mathbf{y}_i=q)}\left(\frac{1}{\exp\left[\eta_i^q\right]+1}\right)^{1-I(\mathbf{y}_i=q)} \tag{24}$$

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

$$\frac{\exp\left[\eta_i^q\right]}{\exp\left[\eta_i^q\right]+1} = 2\exp\left[(y_i-\frac{1}{2})\eta_i^q\right]\int_0^\infty \exp\left[\frac{-\omega_i\eta_i^{q\,2}}{2}\right]p(\omega_i^q)d\omega_i^q \tag{25}$$

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

If we use the equality in Equation 25, the binomial likelihood in Equation 24 can be transformed to a multivariate Gaussian likelihood by including an auxiliary Pólya-Gamma variable ω_i^q (Polson et al., 2013):

$$
\begin{aligned}
l(\mathbf{y}|\beta^q, \beta^{-q}) &= \prod_{i=1}^{n} \frac{\exp\left[\eta_i^q\right]}{\exp\left[\eta_i^q\right]+1} \\
&= \prod_{i=1}^{n} 2\exp\left[(y_i - \frac{1}{2})\eta_i^q\right] \int_0^\infty \exp\left[\frac{-\omega_i^q \eta_i^{q2}}{2}\right] p(\omega_i^q) d\omega_i^q \\
&= \prod_{i=1}^{n} \exp\left[\kappa_i^q \omega_i^q \eta_i^q - \frac{1}{2}(\eta_i^q)^2 \omega_i^q\right] PG(\omega_i^q|1,0) \\
&\propto \exp\left[\frac{1}{2}(2\kappa^q \omega^q \eta^q - \omega^q (\eta^q)^2\right] \\
&\propto \exp\left[-\frac{1}{2}(\kappa^q - \eta^q)^T \mathbf{\Omega}^q (\kappa^q - \eta^q)\right] \\
&= \exp\left[-\frac{1}{2}(\kappa^q - \mathbf{X}\beta^q + \ln[\sum_{m\neq q} \exp(\mathbf{X}\beta^m)])^T \mathbf{\Omega}^q (\kappa^q - \mathbf{X}\beta^q + \ln[\sum_{m\neq q} \exp\left[\mathbf{X}\beta^m\right]]\right],
\end{aligned}
\tag{26}
$$

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

A.0.1 Prior distribution

The Gaussian likelihood in Equation 26 is conditionally conjugate with a normal prior distribution on regression coefficients β^q:

$$
\beta^q \sim N(\mathbf{b}^q, \mathrm{B}^{0q}) \tag{27}
$$

where $\mathbf{b}^q$ is the vector of prior means of regression coefficient vector β^q and B^{0q} is a $P \times P$ symmetric square matrix reflecting the prior precision of regression coefficients β^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. Sullivan and Greenland, 2012; Chen and Ibrahim, 2000). We discuss the specification of informative prior means $\mathbf{b}^q$ in terms of joint response probabilities ϕ in the next Appendix.

A.0.2 Posterior distribution

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

$$
p(\beta|\mathbf{y}) \propto p(\mathbf{y}|\beta, \mathbf{x}) p(\beta), \tag{28}
$$

The combination of a Polya-Gamma transformed Gaussian likelihood (Equation 26) and a normal prior distribution (Equation 27) respectively is proportional to a normally distributed posterior distribution, conditionally on Polya-Gamma variables in ω^q (Polson et al., 2013):

$$
\begin{aligned}
p(\beta^q|\mathbf{Y}, \mathbf{\Omega}^q) \propto & p(\mathbf{y}|\beta^q, \omega^q)p(\beta^q) \\
& \propto \exp\left[-\frac{1}{2}(\kappa^q - \mathbf{X}\beta^q + \ln[\sum_{m\neq q}\exp\left[\mathbf{X}\beta^m\right]])^T\mathbf{\Omega}^q(\kappa^q - \mathbf{X}\beta^q + \ln[\sum_{m\neq q}\exp\left[\mathbf{X}\beta^m\right]])\right] \times \\
& \exp\left[-\frac{1}{2}(\beta^q - \mathbf{b}^q)^T(\mathrm{B}^q)^{-1}(\beta^q - \mathbf{b}^q)\right] \\
& \propto N\left(\mathbf{V}^q(\mathbf{X}^T\mathbf{\Omega}^q(\kappa^q + \ln[\sum_{m\neq q}\exp\left[\mathbf{X}\beta^m\right]]) + (\mathrm{B}^q)^{-1}\mathbf{b}^q), \mathbf{V}^q\right)
\end{aligned}
\tag{29}
$$

where $\mathbf{V}^q = (\mathbf{X}^T\mathbf{\Omega}^q\mathbf{X} + (\mathrm{B}^q)^{-1})^{-1}$. Similarly, subject-specific variable ω_i^q follows a Polya-Gamma distribution that depends on regression coefficients β^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 β. Specifically, the sampling procedure involves iterating L times over the following two steps for $q = 1, \ldots, Q-1$, while keeping β^Q fixed at zero:

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

$$
\beta^q|\omega^q \sim N(\mathbf{m}^q, \mathbf{V}^q) \tag{30}
$$

$$
\begin{aligned}
\text{where} \quad [\mathbf{V}^q]^{-1} = & \quad \mathbf{X}\mathbf{\Omega}^q\mathbf{X} + [\mathbf{V}^{0q}]^{-1} \\
\mathbf{m}^q = & \quad \mathbf{V}^q(\mathbf{X}(\kappa^q + \mathbf{\Omega}^q\mathbf{c}) + [\mathbf{V}^{0q}]^{-1}\mathbf{m}^{0q}) \\
\mathbf{c} = & \quad \left\{\ln\left(\sum_{m\neq q}\exp\left[\psi_i^m\right]\right)_{i=1}^{n}\right\}.
\end{aligned}
$$

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

$$
\omega_i^q|\beta^q \sim PG(1, \psi_i^q - \ln\sum_{m\neq q}\exp\left[\psi_i^m\right]). \tag{31}
$$

The Gibbs sampling procedure results in a sample of L sets of regression coefficients from the posterior

distribution of β.

B 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:

$$\psi_T^q = \beta_0^q + \beta_1^q T + \beta_2^q x + \beta_3^q x \times T \tag{32}$$

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

$$\begin{aligned}
\phi_T^{11}(x^{\cdot}) &= \rho_T(x^{\cdot})\sqrt{\theta_T^1(x^{\cdot})\left[1-\theta_T^1(x^{\cdot})\right]\theta_T^2(x^{\cdot})\left[1-\theta_T^2(x^{\cdot})\right]} + \theta_T^1(x^{\cdot})\theta_T^2(x^{\cdot}) \\
\phi_T^{10}(x^{\cdot}) &= \theta_T^1(x^{\cdot}) - \phi_T^{11}(x^{\cdot}) \\
\phi_T^{01}(x^{\cdot}) &= \theta_T^2(x^{\cdot}) - \phi_T^{11}(x^{\cdot}) \\
\phi_T^{00}(x^{\cdot}) &= 1 - \theta_T^1(x^{\cdot}) - \theta_T^2(x^{\cdot}) + \phi_T^{11}(x^{\cdot})
\end{aligned} \tag{33}$$

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 2.

Solving these linear predictors for β^q results in the following definitions of the elements in β^q:

$$\begin{aligned}
\beta_0^q &= \frac{x^{\mathrm{H}}\psi_0^q(x^L) - x^{\mathrm{L}}\psi_0^q(x^H)}{x^{\mathrm{H}} - x^{\mathrm{L}}} \\
\beta_1^q &= \frac{x^{\mathrm{H}}\left[\psi_1^q(x^L) - \psi_0^q(x^L)\right] + x^{\mathrm{L}}\left[\psi_0^q(x^H) - \psi_1^q(x^H)\right]}{x^{\mathrm{H}} - x^{\mathrm{L}}} \\
\beta_2^q &= \frac{\psi_0^q(x^H) - \psi_0^q(x^L)}{x^{\mathrm{H}} - x^{\mathrm{L}}} \\
\beta_3^q &= \frac{\psi_1^q(x^H) - \psi_0^q(x^H) - \psi_1^q(x^L) + \psi_0^q(x^L)}{x^{\mathrm{H}} - x^{\mathrm{L}}}
\end{aligned} \tag{34}$$

Table 8 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

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

$$\theta_1^L = (0.60, 0.70), \rho_1^L = -0.30$$

$$\theta_1^H = (0.40, 0.30), \rho_1^H = -0.30$$

$$\theta_0^L = (0.40, 0.30), \rho_0^L = -0.30$$

$$\theta_0^H = (0.60, 0.70), \rho_0^H = -0.30,$$

then the regression coefficients would be as presented in Table 8.

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

Algorithm 1 Transformation of posterior regression coefficients to posterior joint response probabilities based on fixed covariate values.

Define $\mathbf{x} = x_2, \ldots, x_P$ as a vector of covariate values of interest
Let $\beta^Q = (0, \ldots, 0)$
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 $\phi_T^{q(l)} = \dfrac{\exp\left[\psi_T^{q(l)}\right]}{\sum_{r=1}^{Q-1} \exp\left[\psi_T^{r(l)}\right] + 1}$
    end for
  end for
end for

Algorithm 2 Transformation of posterior regression coefficients to posterior joint response probabilities based on empirical marginalization.

Let $\beta^Q = (0, \ldots, 0)$
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 $\phi_i^{q(l)} = \dfrac{\exp\left[\psi_i^{q(l)}\right]}{\sum_{r=1}^{Q-1} \exp\left[\psi_i^{r(l)}\right] + 1}$
      for $T \leftarrow 0 : 1$ **do**
        Compute $\phi_T^{q(l)}(\mathbf{x}) = \dfrac{1}{\sum_{i=1}^{n} I(T_i = T)} \phi_i^{q(l)} I(T_i = T)$
      end for
    end for
  end for
end for

D Observed bias in regression coefficients

The simulation study 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 3, 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 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 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.

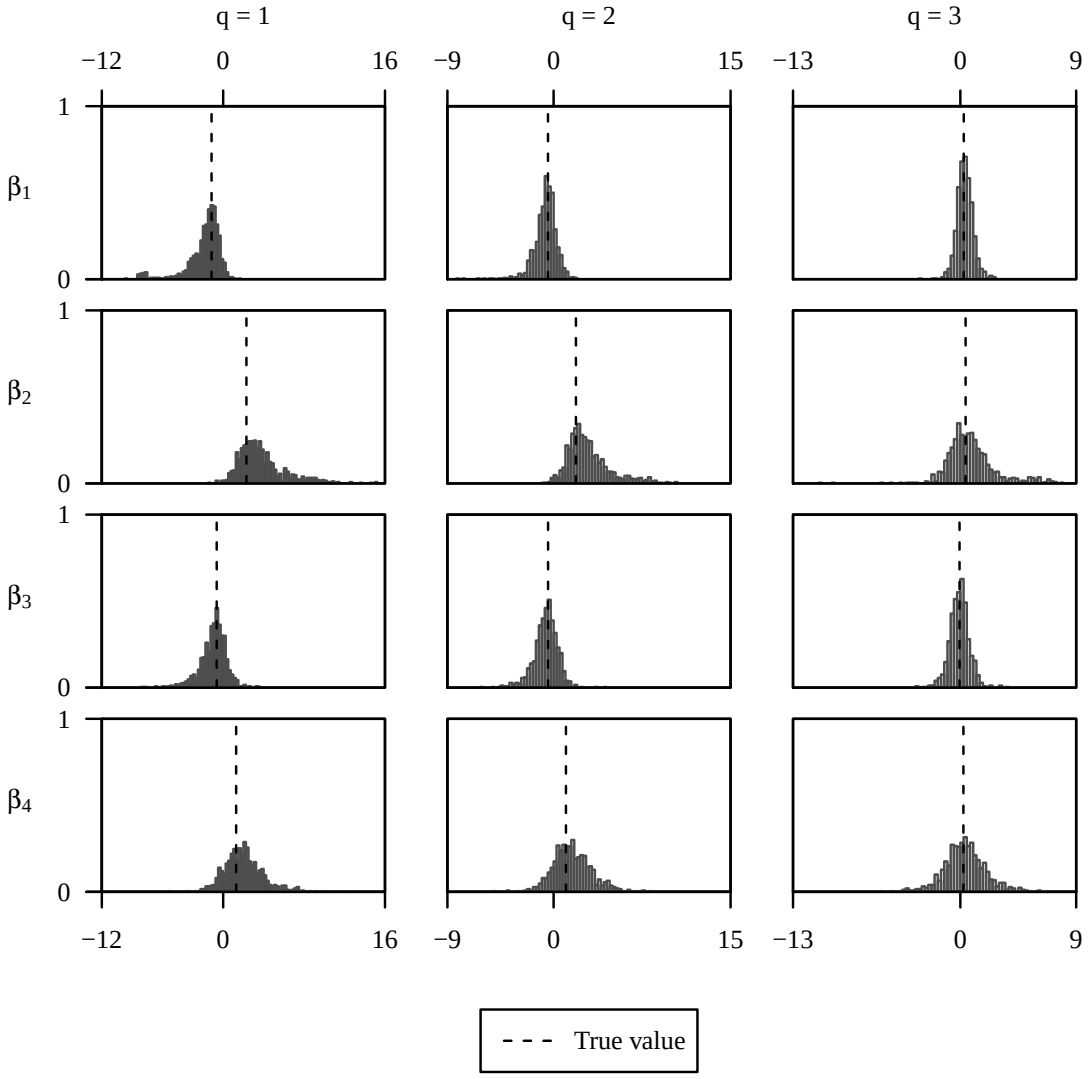

Figure 3