

Bayesian Multivariate Logistic Regression for Superiority and Inferiority Decision-Making under Observable Treatment Heterogeneity

Xynthia Kavelaars^{a,b} , Joris Mulder^a, and Maurits Kaptein^c

^aDepartment of Methodology and Statistics, Tilburg University; ^bDepartment of Theory, Methodology and Statistics, Open University of the Netherlands; ^cEindhoven University of Technology, Mathematics and Computer Science

ABSTRACT

The effects of treatments may differ between persons with different characteristics. Addressing such treatment heterogeneity is crucial to investigate whether patients with specific characteristics are likely to benefit from a new treatment. The current paper presents a novel Bayesian method for superiority 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 a Pólya-Gamma expansion; b) a transformation procedure to transfer obtained regression coefficients to a more intuitive multivariate probability scale (i.e., success probabilities and the differences between them); and c) a compatible decision procedure for treatment comparison with prespecified decision error rates. 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 toward 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 dependent variables; Bayesian analysis; Pólya-Gamma; subgroup analysis

1. Introduction

The current paper focuses on estimating treatment effects among populations, subpopulations, and individual patients in the context of two-arm randomized controlled trials (RCTs) with multiple (correlated) binary dependent 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; and b) to inform the prescription of treatments to patients in (clinical) practice (Food & Drug Administration, 2016). Although RCTs are broadly applicable to experimental research in general, we focus on the health domain and use the word “treatment” to refer to psychological and medical interventions in the broad sense. These interventions include - but are not limited to - behavioral therapies,

pharmacological support, and other experimental types of care.

Such RCTs often assess multiple types of (clinical) events (e.g., quitting substance abuse, death), functional measures (e.g., memory decline, ability to walk), or disease symptoms (e.g., fatigue, anxiety) (Food & Drug Administration, 2017). Studying multiple dependent variables in RCTs is useful, since multiple dependent variables provide multidimensional insights into the effects of a treatment and since analyzing multiple dependent variables together has the potential to improve the connection between clinical and statistical decision-making. More specifically, multiple effects of the intervention can be combined and weighted in various ways to provide a single statistical decision regarding superiority or inferiority, similar to decisions regarding treatment prescription made by therapists or clinicians (e.g., Pocock et al., 1987; O’Brien, 1984; Murray et al., 2016). Whereas performing multiple

univariate analyses on individual dependent variables is a common strategy to deal with data from multiple dependent variables, a single multivariate analysis is often preferable from a statistical point of view (Senn & Bretz, 2007; Ristl et al., 2019; Food & Drug Administration, 2017; Murray et al., 2016). Multivariate analysis takes the correlation between dependent variables into account and therefore 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; 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. In this case, patients react relatively similarly to the treatment. However, average effects may give a limited, or even erroneous, impression when the actual effects of a treatment are heterogeneous and thus interact with characteristics of patients. In that case, patients differ in their reactions to the treatment. Taking characteristics of patients into account in the estimation of treatment effects (i.e., estimating conditional average treatment effects; CATEs) can then contribute to a better understanding of the treatment's potential for an individual patient. Despite efforts to provide statistical methodology to model CATEs (e.g., Wang et al., 2015; Yang et al., 2021; Jones et al., 2011), investigating these effects is not the standard yet: Thall noted 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 conditional effects in the evaluation of treatments is crucial to a) identify how likely a specific patient will benefit from a treatment; and b) optimize treatment results of individual patients via personalized treatment assignment (Goldberger & Buxton, 2013; Hamburg & Collins, 2010; Wang et al., 2015; Simon, 2010).

An example of a trial with multiple dependent variables 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 were based on the assumption that specific characteristics of patients (e.g., sex or age) and/or disease (e.g., type of stroke or functional status after stroke) did not interact with the treatment to produce different effects for different patients. Average treatment effects could, 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, hypothetical heterogeneous effects as these would have clinically and psychologically relevant implications and advocate the development of more personalized treatment policies.

While multivariate treatment effects for patients with specific characteristics are theoretically relevant for many contemporary RCTs contributing to the personalization of treatments, decision-making under treatment heterogeneity in the multivariate context is considerably more complex compared to the non-heterogeneous and/or univariate setting. Generalizations to the heterogeneous and multivariate context 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 dependent variables (i.e., a multivariate analysis method) to obtain accurate decision error rates (e.g., Sozu et al., 2010, 2016; Kavelaars et al., 2020). Ignoring or misspecifying a non-zero correlation can result in over- or underestimation of the required sample size and thus affects the statistical power of the analysis. For accurate inference regarding conditional average 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 dependent 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 (i.e., treatment effects on the individual dependent variables) is recommended in treatment comparison, demanding any multivariate method to return interpretable univariate effects (Food & Drug

Administration, 2017; O'Brien & Dunson, 2004). Some existing multivariate models lack insight into marginal distributions (e.g. Malik & Abraham, 1973). Third, some multivariate methods estimate a single regression parameter to capture the relation between a covariate and all dependent variables (e.g., O'Brien & Dunson, 2004; Rossi et al., 2005). The latter assumes that all dependent variables vary identically over the full support of the covariate. In other words, all relations between the covariate and the outcome variable have the same size and direction. Clearly, such an assumption may be too strict to hold in practice.

In order to deal with 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 multivariate analysis model to estimate unknown 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 with prespecified error rates. The analysis procedure has advantages over several other approaches, as it relies on a multinomial distribution and therefore has the flexibility to model univariate effects and correlations between dependent variables. 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 decision procedure is suitable for Bayesian inference and can be flexibly applied with several decision rules for multiple dependent variables. Noteworthy is a decision rule with a compensatory mechanism, that can weigh dependent variables by their importance and has a natural compensatory mechanism that can balance positive and negative treatment effects. With this decision procedure, decisions regarding treatment superiority can be made with targeted decision error rates (i.e., Type I and Type II errors) and a priori computed sample sizes.

Kavelaars et al. (2020) proposed a multivariate Bernoulli model for multivariate Bernoulli outcomes to estimate average treatment effects and to make decisions based on multivariate treatment effects. In the current paper we propose a more flexible modeling framework for multivariate Bernoulli outcomes using Bayesian multivariate logistic regression models. This extension allows us to model and estimate multivariate treatment effects for different (sub)populations based on available covariate information. Moreover, to

make decisions about multivariate treatment effects for these different subpopulations, we extend the decision procedure of Kavelaars et al. (2020) to the new multivariate logistic regression model. Additionally, sample size recommendations are provided for estimating and decision-making under this framework.

Note that the proposed multivariate modeling framework aims to estimate heterogeneous multivariate treatment effects that are caused by observed covariate information and to make decisions about treatment superiority. Thereby, the aim is different from mixture modeling which aims to capture unobserved (treatment) heterogeneity using latent variables (either discrete or continuous). Mixture models use response data to cluster respondents based on their patterns of outcome data (e.g., patterns of symptoms), where each cluster has an individual distribution that forms a constituent of the mixture (McLachlan et al., 2019). The proposed regression model does not include latent variables (either discrete or continuous) to capture unobserved heterogeneity. Instead, multivariate (logistic) regression uses observed covariate information to define patient groups of interest, often based on theoretical (such as accepted cutoff values for high and low blood pressure) or statistical (such as those respondents with more extreme scores than one standard deviation below or above the mean) grounds. Subgroups are thus bounded by criteria specified by the researcher, rather than by response patterns in the data.

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 $k \in \{1, \dots, K\}$, where $y_i^k \in \{0, 1\}$, $0 = \text{failure}$ and $1 = \text{success}$. Vector $\mathbf{y}_i = (y_i^1, \dots, y_i^K)$ is the multivariate (or joint) binary response vector of

subject i on K dependent variables and has configuration $\mathbf{H}_q$, which is one of the $Q = 2^K$ possible response combinations of length K given in the q^{th} row of matrix $\mathbf{H}$:

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & \dots & 1 & 1 \\ 1 & 1 & \dots & 1 & 0 \\ & & \dots & & \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix} \quad (1)$$

The probability of 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 $\phi_i^q = p(y_i = \mathbf{H}_q)$ 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 $\phi_i^q(\mathbf{x}_i)$ 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:

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

for response categories $q = 1, \dots, Q - 1$. In Equation 2, $\psi_i^q(\mathbf{x}_i)$ reflects the linear predictor of response category q and subject i :

$$\psi_i^q(\mathbf{x}_i) = \beta_0^q + \beta_1^q x_{i1} + \dots + \beta_p^q x_{ip}. \quad (3)$$

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 2 of Q response categories:

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

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

$$p(\boldsymbol{\beta}^q | \mathbf{y}) \propto p(\mathbf{y} | \boldsymbol{\beta}^q) p(\boldsymbol{\beta}^q), \quad (5)$$

where $p(\boldsymbol{\beta}^q)$ 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 A.

2.2. Transformation to treatment differences

In contrast to several other regression analyses, the obtained multinomial regression coefficients have no straightforward interpretation. 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 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 as a covariate, so that we can estimate conditional effects for patients with different values of blood pressure, resulting in the following model:

$$\psi_i^q(\mathbf{x}_i) = \beta_0^q + \beta_1^q T_i + \beta_2^q b p_i + \beta_2^q b p_i T_i, \quad (6)$$

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

1. **Regression coefficients $\boldsymbol{\beta}$ to joint response probabilities $\boldsymbol{\phi}_T(\mathbf{x})$:** In the first step, the posterior sample of regression coefficients $\boldsymbol{\beta}$ is transformed to a treatment effect in terms of joint response probabilities $\boldsymbol{\phi}_{T_i}(\mathbf{x}_i)$ for each treatment $T \in \{0, 1\}$. Linear predictor $\psi_i^q(\mathbf{x}_i)$ is then transformed to individual joint response probability $\phi_i^q(\mathbf{x}_i)$ via the multinomial logistic function in Equation 2:

$$\phi_i^q(\mathbf{x}_i) = \frac{\exp[\psi_i^q(\mathbf{x}_i)]}{\sum_{r=1}^{Q-1} \exp[\psi_i^r(\mathbf{x}_i)] + 1} \quad (2 \text{ revisited})$$

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:

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

This probability can be computed for the other joint response combinations as well. Note that we are in fact interested in joint response probability $\phi_T(\mathbf{x})$, which reflects a treatment effect among a (sub)population defined by $\mathbf{x}$. This notation is more general than the joint response probability of an individual patient with covariates $\mathbf{x}_i$. The population can be reflected by an individual patient (e.g., with a systolic blood pressure of 100) in some situations, while other cases target the entire study population (e.g., no restriction on systolic blood pressure) or a subpopulation of interest (e.g., with a systolic blood pressure above 150). These variations have slightly different computational procedures, which we discuss in more detail in Section 3.

2. **Joint response probabilities $\phi_T(\mathbf{x})$ to multivariate success probabilities $\theta_T(\mathbf{x})$:** The next step in the transformation involves the conversion from joint response probabilities $\phi_T(\mathbf{x})$ to multivariate success probabilities of individual dependent variables $\theta_T(\mathbf{x})$. Especially when the number of dependent 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 ϕ_T , denoted by matrix $\mathbf{U}_k$:

$$\theta_T^k(\mathbf{x}) = \sum_{q=1}^Q \phi_T^q(\mathbf{x}) I(\mathbf{H}_q \in \mathbf{U}_k). \quad (8)$$

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 dependent variables from the IST are the following combinations, where we drop the dependency on $\mathbf{x}$ for notational simplicity.

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

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

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

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

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

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

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

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

$$\begin{aligned}\delta^{strk} &= \theta_{H+A}^{strk} - \theta_A^{strk} \\ \delta^{dep} &= \theta_{H+A}^{dep} - \theta_A^{dep}.\end{aligned}\quad (11)$$

Applying the three above-mentioned steps to each draw of the posterior sample of β , results in a posterior sample of multivariate treatment difference $\delta(\mathbf{x})$. 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 to generalize the conclusion to the specified (sub)population, as outlined in the next subsection.

2.3. Posterior decision-making

Decisions rely on estimated treatment effects, such as differences between success probabilities, and their uncertainties. More formally, multivariate treatment difference δ has complete parameter spaces $\mathcal{S} \subset [-1, 1]^K$, 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, where we would conclude that the treatments differ. 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 consider evidence sufficiently strong and would conclude superiority and/or inferiority when the posterior probability that treatment difference $\delta(\mathbf{x})$ lies in the rejection region exceeds a prespecified decision threshold, p_{cut} :

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

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 $\delta(\mathbf{x})$. Equation 12 is then applied in practice as:

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

In a situation with multiple dependent 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 D1. 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, as presented by the Food and Drug Administration 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). A more elaborate comparison of these rules can be found in Kavelaars et al. (2020).

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 dependent variables (Sozu et al., 2016). The superiority and inferiority spaces are defined as:

$$\begin{aligned} \mathcal{S}_S^{Any} &= \delta(\mathbf{x}) | \max_{1 \leq k \leq K} \delta^k(\mathbf{x}) > 0 \\ \mathcal{S}_I^{Any} &= \delta(\mathbf{x}) | \min_{1 \leq k \leq K} \delta^k(\mathbf{x}) < 0. \end{aligned} \quad (14)$$

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

$$\begin{aligned} \mathcal{S}_S^{All} &= \delta(\mathbf{x}) | \min_{1 \leq k \leq K} \delta^k(\mathbf{x}) > 0 \\ \mathcal{S}_I^{All} &= \delta(\mathbf{x}) | \max_{1 \leq k \leq K} \delta^k(\mathbf{x}) < 0. \end{aligned} \quad (15)$$

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:

$$\begin{aligned} \mathcal{S}_S^{Comp}(\mathbf{w}) &= \delta(\mathbf{x}) | \delta(\mathbf{w}, \mathbf{x}) > 0 \\ \mathcal{S}_I^{Comp}(\mathbf{w}) &= \delta(\mathbf{x}) | \delta(\mathbf{w}, \mathbf{x}) < 0 \end{aligned} \quad (16)$$

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).

2.4. Sample size computations

Decisions resulting from analysis with the Bayesian multivariate logistic regression framework are based on a posterior probability. In absence of prior information, the Bayesian posterior probability has a direct relation with the frequentist p-value: The Bayesian posterior probability equals $1 - p$ and behaves according to the well-known relationship between effect size, sample size, and decision error rates (Marsman & Wagenmakers, 2016). This allows for control of decision error rates (Type I and Type II-error) via a priori computed sample sizes. Methods to compute required sample sizes 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). 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 with a non-informative prior distribution (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 into the required number of observations in absence of prior information and in the influence of the correlation on the sample sizes needed to

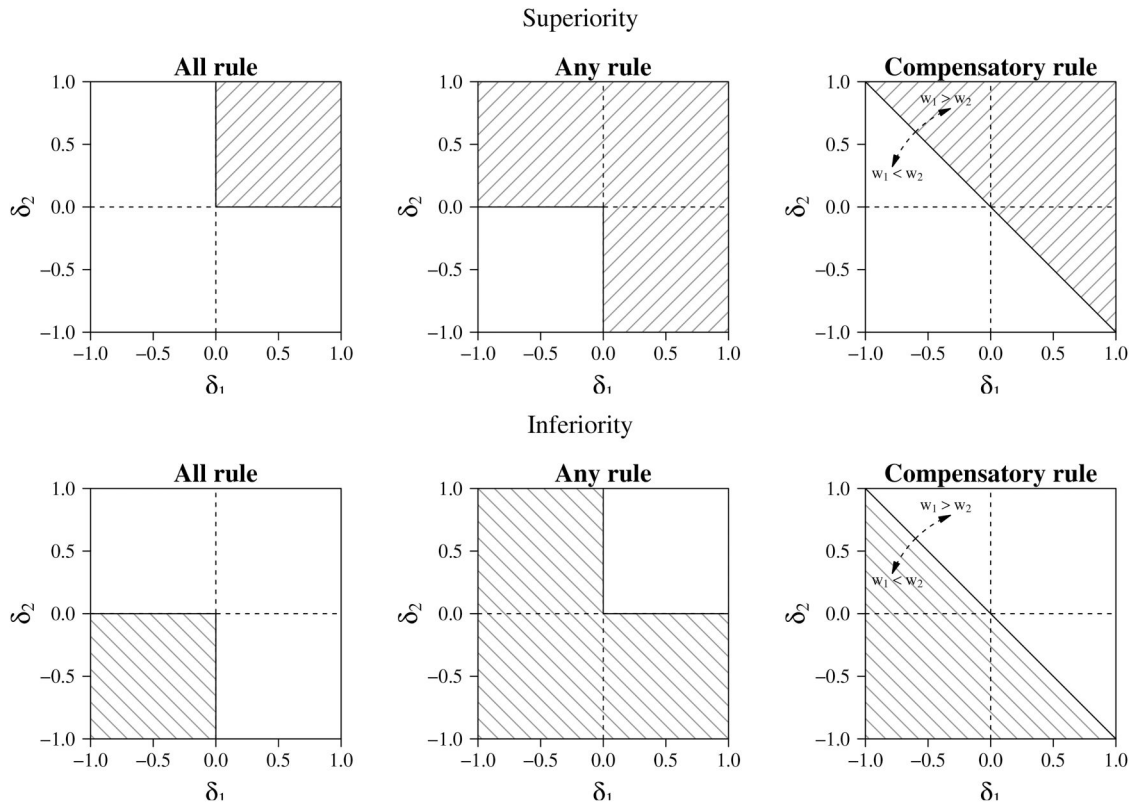


Figure D1. Bivariate superiority and inferiority spaces for the All, Any, and Compensatory ($w = 0.50, 0.50$) rules.

obtain targeted decision error rates. For notational simplicity, we discard the dependence on $\mathbf{x}$ 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 Ψ_K with mean $\mathbf{0}$ and correlation matrix Σ (Sozu et al., 2016):

$$1 - \beta = \Psi_K(c^1, \dots, c^K). \quad (17)$$

In Equation 17, c^k for outcome k is defined by the decision rule of interest. Further, the off-diagonal elements of Σ denote (estimated) pairwise correlations between dependent variables.

For the Any rule,

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

For the All rule,

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

In Equations 18 and 19, n is the sample size per treatment and $z_{(\cdot)}$ refers to the selected $(1 - \frac{\alpha}{K})$ or $(1 - \alpha)$ quantile from the univariate normal distribution. Further, θ_1^k and θ_0^k refer to success probabilities on outcome variable k of treatments 1 and 0 respectively.

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_1^k and θ_0^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 , θ_T^w , results in:

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

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

2.4.3. Correlation, sample size, and statistical power

We illustrate the relation between the sample size, the statistical power and the correlation between dependent variables with an example. We computed required sample sizes to obtain 80% statistical power for the following bivariate ($K = 2$) and trivariate ($K = 3$) outcomes, where we used different correlations ($\rho_{\theta^k, \theta^l}$) and multivariate treatment differences (δ).

1. **S2:** $K = 2$, $\rho_{\theta^k, \theta^l} \in -0.20, 0.00, 0.20$,
 $\delta = (0.20, 0.10)$
2. **S3:** $K = 3$, $\rho_{\theta^k, \theta^l} \in -0.20, 0.00, 0.20$,
 $\delta = (0.20, 0.10, 0.20)$
3. **L2:** $K = 2$, $\rho_{\theta^k, \theta^l} \in -0.40, 0.00, 0.40$,
 $\delta = (0.30, 0.20)$
4. **L3:** $K = 3$, $\rho_{\theta^k, \theta^l} \in -0.40, 0.00, 0.40$,
 $\delta = (0.30, 0.20, 0.30)$

Scenarios S2 and S3 had a smaller multivariate treatment difference δ and weaker non-zero correlations than scenarios L2 and L3.

Table D1 shows the the required sample sizes for these scenarios as well as the anticipated statistical power for right-sided superiority decision-making under two scenarios:

1. When sample sizes computations are based on the true multivariate treatment difference δ and the true correlation between dependent variables $\rho_{\theta^k, \theta^l}$. This scenario aims to highlight that statistical power can be targeted when sample size

computations follow the true data generating mechanism.

2. When sample sizes computations are based on the true multivariate treatment difference δ and uncorrelated dependent variables (i.e., $\rho_{\theta^k, \theta^l} = 0$). This scenario provides insight in anticipated error rates where the correlation is not taken into account in sample size computations. This situation is equivalent to performing multiple univariate analyses on correlated dependent variables.

These probabilities are computed by plugging in true treatment differences and correlations, while using either the required sample size (scenario 1) or the sample size for uncorrelated data (scenario 2) in Equations (17)–(20)).

This illustration provides five takeaways. First, larger effect sizes (L2 and L3) result in smaller required samples than smaller effect sizes (S2 and S3) respectively. Second, adding an additional dependent variable has the potential to reduce sample sizes. Required sample sizes are larger for a three-dimensional outcome (S3 and L3) than for a two-dimensional outcome (S2 and L2). Third, the required sample size depends on the correlation between dependent variables. Compared to uncorrelated dependent variables, the Any and Compensatory rules require fewer observations when dependent variables are negatively correlated, whereas positively correlated dependent variables require more observations. Consequently, when sample size computations do not take non-zero correlations into account, statistical power will be larger or smaller than targeted respectively. Fourth, the relation between correlation and required sample size is different for different decision rules. Compared to the Any and Compensatory rules, the All rule shows the opposite relation between the direction of the correlation and the required sample size. Here, positively correlated dependent variables require a smaller number of observations than uncorrelated or negatively correlated dependent variables. Moreover, the All rule appears less sensitive to the correlation than the other rules. Sample sizes are not very different and statistical power under independence is still close to the targeted .80. Fifth, the effect of the correlation on required sample size and statistical power is larger in the scenarios where non-zero correlations are stronger (L2 and L3). In these scenarios, the discrepancy between the targeted power of 0.80 and the actual power is larger for non-zero correlations. Further, the differences between presented sample sizes for negatively correlated, uncorrelated,

Table D1. Example of required sample sizes (n) for analysis with correlated data and anticipated probabilities to conclude superiority when sample size computations use the true correlation (p_T) vs. assume uncorrelated dependent variables (p_U) under four different data-generating mechanisms (DGMs).

DGM	$\rho < 0$			$\rho = 0$			$\rho > 0$		
	n	p_T	p_U	n	p_T	p_U	n	p_T	p_U
All rule									
S2	307	0.801	0.801	307	0.801	0.801	307	0.801	0.801
L2	77	0.801	0.801	77	0.803	0.803	76	0.803	0.808
S3	307	0.800	0.800	307	0.801	0.801	307	0.801	0.801
L3	79	0.801	0.800	79	0.804	0.804	77	0.803	0.812
Any rule									
S2	76	0.801	0.825	81	0.803	0.803	85	0.802	0.783
L2	27	0.811	0.862	31	0.807	0.807	35	0.801	0.756
S3	51	0.807	0.850	57	0.804	0.804	64	0.805	0.760
L3	18	0.821	0.918	23	0.809	0.809	29	0.804	0.714
Compensatory rule									
S2	53	0.798	0.845	61	0.801	0.801	68	0.799	0.760
L2	18	0.807	0.884	23	0.794	0.794	29	0.799	0.714
S3	24	0.796	0.923	37	0.804	0.804	49	0.802	0.699
L3	5	0.826	0.996	14	0.798	0.798	24	0.807	0.608

and positively correlated dependent variables is larger compared to the scenarios with less strong correlation (S2 and S3).

These takeaways are in line with detailed discussions in Sozu et al. (2010, 2016); Food and Drug Administration (2017); Kavelaars et al. (2020).

3. Estimating conditional average treatment effects

In the proposed framework, treatment heterogeneity can be captured by joint response probabilities that reflect conditional average treatment effects and thus depend on prespecified characteristics of a subpopulation of interest. We describe two ways to represent subpopulations: by fixed covariate values or by a pre-specified interval of the covariate distribution(s). Both representations have their own applications. Fixed 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 to estimate conditional average treatment effects based on fixed values and based on intervals in more detail in the remainder of this subsection. In these discussions, we use a linear predictor $\psi_i^q(\mathbf{x})$ (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 z , $\mathbf{x} = (z, T, zT)$ and $\psi_T^q(\mathbf{x})$ is defined as:

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

3.1. Fixed values of covariate

For a patient population with fixed values of patient covariates, a posterior sample of joint response probabilities $\phi_T(\mathbf{x})$ can be found by plugging in a vector of fixed covariate values $\mathbf{x}$ in linear predictor $\psi_T^q(\mathbf{x})$. Subsequently applying the multinomial logistic link function in Equation 2 to each $\psi_T^q(\mathbf{x})$ results in joint

response probability $\phi_T^{(l)}(\mathbf{x})$ for treatment T . Applying these steps each posterior draw (l) of regression coefficients $\beta^{(l)}$ 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 $\phi_T^{(l)}(\mathbf{x})$ for draw (l). The procedure is presented in Algorithm 2 in to Appendix C. 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. Numerical evaluation

The current section presents an evaluation of the performance of the proposed multivariate logistic regression procedure. 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 decision procedure results in an (in)correct superiority conclusion to learn about decision error rates when sample sizes are estimated a priori.

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

4.1. Setup

4.1.1. Conditions

The performance of the framework was evaluated in a treatment comparison based on one covariate and two dependent variables. In [Appendix D](#), we present an evaluation of the performance with three dependent variables. Six aspects were varied: the analysis procedure, the effect size, measurement level of the covariate, the correlation between dependent variables, the (sub)population, and the decision rule. Each of these factors will be discussed in the following paragraphs.

4.1.1.1. Analysis procedure. We present three Bayesian analysis procedures:

(1) **Multivariate logistic regression analysis (mLR):** We analyzed the generated data via the proposed Bayesian multivariate logistic regression model presented in [Section 2](#).

The performance of the mLR-model was compared to two reference approaches:

(2) **Multivariate Bernoulli analysis (mB):** To demonstrate the gain of a multivariate regression approach over multivariate subgroup analysis (i.e., multivariate stratified analysis), we fitted the unconditional Bayesian multivariate Bernoulli model in Kavelaars et al. (2020) to the data as well. Whereas the multivariate Bernoulli model takes the correlation between dependent variables into account, the multivariate Bernoulli model computes conditional average treatment effects via stratified multivariate analysis: the multivariate Bernoulli model only uses the response data from observations that belong to the (sub)population of interest. Hence, the estimation of ATEs uses the full sample of response data, whereas CATEs are estimated based on a subsample of response data. Samples of treatment-specific joint response probabilities ϕ_T could be drawn directly from a posterior Dirichlet distribution with parameters $\alpha_T^n = \alpha^0 + \{\sum_{i=1}^n I(T_i = T)I(y_i = H_q)\}_{q=1}^Q$, where α^0 is a vector of Q prior hyperparameters.

(3) **Univariate logistic regression (uLR):** To demonstrate the added value of a multivariate model over multiple univariate models, we fitted Bayesian univariate logistic regression models from Polson et al. (2013) to the individual dependent variables for the scenario with two dependent variables. This univariate model is a special case of the multivariate model presented in [Section 2](#) and [Appendix A](#). While these regression-based models use the full sample of data to estimate conditional average treatment effects among subpopulations, they cannot capture correlations between dependent variables.

4.1.1.2. Datagenerating mechanisms: effect size, measurement level of covariate, and correlation.

We included treatment differences of four different sizes that varied in heterogeneity:

1. **Effect size 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. **Effect size 2.1 & 2.2:** A heterogeneous treatment effect, with an average treatment difference of zero and a conditional treatment effect larger than zero.
3. **Effect size 3.1 & 3.2:** A heterogeneous treatment effect, with one average and both 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 average treatment effects. The effect size of the conditional average treatment effect reflects the least favorable average treatment effect 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 if the sample size is sufficiently large.
4. **Effect size 4.1 & 4.2:** A heterogeneous treatment effect, with one average and both 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. Similar to effect size 3.1/3.2, the effect size of the conditional average treatment effect reflects the least favorable effect 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 if the sample size is sufficiently large.

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

Table D2. Parameters of average treatment effects (ATEs) in the trial and conditional average treatment effects (CATEs) in a subpopulation for two outcome variables.

ES		ATE (δ_1, δ_2)			CATE (δ_1, δ_2)		
		$\delta(\mathbf{w})$	$\rho_{\theta^k, \theta^l}$		$\delta(\mathbf{w})$	$\rho_{\theta^k, \theta^l}$	
1.1	D	(0.000, 0.000)	0.000	-0.160 0.030 0.220	(0.000, 0.000)	0.000	-0.200 0.000 0.200
1.2	C	(0.000, 0.000)	0.000	-0.163 0.028 0.219	(0.000, 0.000)	0.000	-0.207 0.002 0.208
2.1	D	(0.000, 0.000)	0.000	-0.154 0.037 0.229	(0.250, 0.150)	0.225	-0.200 0.000 0.200
2.2	C	(0.000, 0.000)	0.000	-0.157 0.036 0.228	(0.116, 0.069)	0.104	-0.206 0.003 0.207
3.1	D	(0.100, 0.000)	0.075	-0.152 0.040 0.232	(0.300, 0.200)	0.275	-0.200 0.000 0.200
3.2	C	(0.100, 0.000)	0.075	-0.155 0.038 0.231	(0.196, 0.093)	0.170	-0.205 0.003 0.206
4.1	D	(0.350, 0.000)	0.263	-0.197 0.000 0.197	(0.200, 0.000)	0.150	-0.200 0.000 0.200
4.2	C	(0.350, 0.000)	0.263	-0.197 0.000 0.197	(0.288, 0.000)	0.216	-0.202 0.000 0.202

Es = Effect size, D = Discrete covariate, C = Continuous covariate.

continuous covariate. Further, we specified three pairwise correlations for the dependent variables: a negative correlation ($\rho_{\theta^k, \theta^l} = -.20$), no correlation ($\rho_{\theta^k, \theta^l} = .00$), and a positive correlation ($\rho_{\theta^k, \theta^l} = .20$). These pairwise correlations were identical for all dependent variable pairs and were specified for the conditional average treatment effects ($x = 0$ and $x = 1$ for the dichotomous covariate; at $x = -1$ and $x = 1$ for the continuous covariate). The correlation structures and effect sizes of the conditional average treatment effects determine, together with the probability distribution of the covariates, the correlation and effect size of average treatment effects.

These four effect sizes, two measurement levels of the covariate, and three correlation structures resulted in the $4 \times 2 \times 3 = 24$ data generating mechanisms (DGMs) presented in Table D2.

4.1.1.3. Treatment effects and (sub)populations.

We estimated three different treatment effects:

1. An average treatment effect (ATE) among the trial population. The trial population with a discrete covariate was defined by a binomially distributed covariate with a probability of 0.50. The trial population with a continuous covariate was defined by a covariate that followed a standard normal distribution.

2. A conditional average treatment effect (CATE) among a subpopulation defined by a sample of covariate an interval of a continuous covariate. This treatment effect was also estimated among patients scoring low on the covariate, but this time the subpopulation was defined as all values between the mean and one standard deviation below the mean. Note that the discrete covariate could not be assigned an interval, since subsetting a binary variable inherently results in a single value.
3. A conditional average treatment effect (CATE) among a subpopulation defined by a fixed value of a covariate. The treatment effect was estimated among patients scoring low on the covariate and was described by a value of 0 (discrete covariate) or -1 (continuous covariate).

4.1.1.4. 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 unequal weights ($\mathbf{w} = (0.75, 0.25)$)

We computed sample sizes per treatment group via the procedures from Subsection 2.4 for conditions with non-zero true average treatment effects targeting at a power of 0.80 and a right-sided α of 0.05. 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 average treatments effects. As a result, the power of conditional treatment effects was not targeted at 0.80, but should exceed this target when the required sample size for a CATE was larger than the sample size for an ATE. Similarly, the power of CATEs with a sample size smaller than the ATE sample size should be lower than .80. The required sample sizes are presented in Table D3. In these tables, we also included a) the required sample size for the conditional average 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 a stratified analysis approach; and b) covariate data in empirical marginalization.

4.1.2. Procedure

4.1.2.1. Data generation. For each data generating mechanism and each unique (decision-rule specific)

Table D3. Required sample sizes to evaluate the average treatment effect (ATE) and conditional treatment effect (CATE) for two outcome variables.

ES	ρ_{k^0, k^1}	All			Any			Compensatory		
		ATE	CATE	Sub	ATE	CATE	Sub	ATE	CATE	Sub
1.1	< 0	–	–	500	–	–	500	–	–	500
	≈ 0	–	–	500	–	–	500	–	–	500
	> 0	–	–	500	–	–	500	–	–	500
1.2	< 0	–	–	342	–	–	342	–	–	342
	≈ 0	–	–	342	–	–	342	–	–	342
	> 0	–	–	342	–	–	342	–	–	342
2.1	< 0	–	136	500	–	45	500	–	32	500
	≈ 0	–	136	500	–	48	500	–	36	500
	> 0	–	136	500	–	51	500	–	40	500
2.2	< 0	–	658	342	–	215	342	–	154	342
	≈ 0	–	649	342	–	229	342	–	175	342
	> 0	–	644	342	–	245	342	–	196	342
3.1	< 0	–	77	500	381	29	191	309	21	155
	≈ 0	–	77	500	385	31	193	349	23	175
	> 0	–	76	500	387	33	194	388	26	194
3.2	< 0	–	358	342	379	81	130	307	56	105
	≈ 0	–	358	342	383	86	131	347	65	119
	> 0	–	356	342	386	91	132	386	73	132
4.1	< 0	–	–	500	28	93	14	22	73	11
	≈ 0	–	–	500	28	93	14	25	83	13
	> 0	–	–	500	28	94	14	28	93	14
4.2	< 0	–	–	342	28	43	10	22	34	8
	≈ 0	–	–	342	28	44	10	25	39	9
	> 0	–	–	342	28	44	10	28	43	10

Sub = expected size of subsample.

Bold-faced subsamples are smaller than required for estimation of the CATE.

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 :

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

To generate response data, we first applied the multinomial logistic link function (Equation 2) to each true linear predictor $\psi_i(x_i)$ to obtain joint response probabilities $\phi_i(x_i)$ for each subject i . Next, we sampled response vector $y_i|x_i$ from a multinomial distribution with probabilities $\phi_i(x_i)$.

4.1.2.2. Prior distribution. We specified diffuse prior distributions. This is motivated by the idea that obtained the posterior distributions are then completely based on the information in the data, which is a common choice in default Bayesian analyses. For the multivariate logistic regression analysis, we set multivariate normally distributed prior with means $\mathbf{b}^q = \mathbf{0}$ and variance matrix $\mathbf{B}^{0q} = \text{diag}(10, \dots, 10)$ for all regression coefficients. Prior covariances between regression coefficients were set at zero, implying that regression coefficients were independent a priori. For the univariate logistic regression analysis we used univariate normally distributed priors with a mean of 0 and a variance of 10 for all parameters. The specified variance parameters of regression coefficients were

motivated by work of Gelman et al. (2008). These authors recommend to choose a variance parameter that results in realistic support for the probability parameter after non-linear transformations in logistic regression and has sufficient information to stabilize posterior computations. For the mB reference approach, we used a Dirichlet prior distribution with hyperparameters $\alpha^0 = \mathbf{0.01}$. This prior is close to the improper Haldane prior ($\alpha = \mathbf{0}$), which is considered the least informative prior distribution for bi- or multinomially distributed data, results in a posterior mean equal to the maximum likelihood estimator, and corresponds to a uniform prior on the log-odds scale (Tuyl et al., 2008; Kerman, 2011). The small deviation from the Haldane prior makes the prior distribution proper and ensures that cell probabilities can be sampled from the Dirichlet distribution when cells have no observations (Kavelaars et al., 2020).

4.1.2.3. Gibbs sampling. The regression coefficients in response categories $1, \dots, (Q - 1)$ were estimated via the Gibbs sampler detailed in Appendix A. We ran two MCMC-chains with $L = 10,000$ iterations plus 1,000 burnin iterations. We visually inspected traceplots of MCMC-chains and used multivariate Gelman-Rubin convergence diagnostics to assess convergence (Gelman & Rubin, 1992; Brooks & Gelman, 1998). As these traceplots showed satisfactory overlap

between chains and the convergence diagnostics were all between 1.00 and 1.10, we concluded that there were no issues with convergence.

4.1.2.4. 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 & Wagenmakers, 2016; Kavelaars et al., 2020; Sozu et al., 2016).

4.1.3. Software

We conducted our analyses in R (R Core Team, 2020). We drew variables from the multivariate normal, Pólya-Gamma, and Dirichlet distributions with the MASS, pgdraw and MCMCpack packages respectively (Venables & Ripley, 2002; Makalic & Schmidt, 2016; Martin et al., 2011). We used the coda package to explore MCMC chains (Plummer et al., 2006). The simulation procedure was parallelized using the foreach and doParallel packages (Microsoft & Weston, 2020a, 2020b). LaTeXTables were created with the xtable package (Dahl et al., 2019).

4.2. Results

4.2.1. Bias

Bias of multivariate and weighted treatment differences was negligible ($< |.01|$) in most conditions, implying that Bayesian multivariate logistic regression analysis was generally able to reproduce true treatment effects. However, the estimation of average treatment effects under effect sizes 4.1 and 4.2 resulted in slightly biased treatment differences for the Any and Compensatory rules. As shown in Table D4, these absolute biases ranged up to $|0.04|$. These biases were produced in both univariate and multivariate logistic regression analysis, but not in multivariate Bernoulli analysis. Conditional average treatment effects were estimated with comparable patterns of bias and a maximum of $|0.025|$. This bias showed up in conditions with a small sample, which is a well-documented property of logistic regression in general (Nemes et al., 2009).

This bias in treatment differences could be traced back to bias in regression coefficients. Mean estimates of regression coefficients were asymptotically unbiased, implying that bias was negligible ($< |0.01|$) in conditions with a sufficiently large sample. We

Table D4. Bias in average treatment differences of effect size (ES) 4.1 and 4.2 by decision rule.

All rule				
ES	$\rho_{\theta^k, \theta^l}$	uLR (δ^1, δ^2)	mB (δ^1, δ^2)	mLR (δ^1, δ^2)
4.1	< 0	(0.000, 0.000)	(−0.002, 0.001)	(0.000, −0.002)
	≈ 0	(0.000, 0.000)	(0.000, 0.000)	(−0.001, −0.001)
	> 0	(−0.001, 0.000)	(−0.001, −0.001)	(−0.001, 0.000)
4.2	< 0	(−0.002, 0.000)	(0.002, −0.001)	(0.000, −0.002)
	≈ 0	(−0.001, −0.001)	(−0.001, 0.001)	(−0.002, −0.001)
	> 0	(−0.001, −0.001)	(0.001, −0.001)	(−0.001, 0.001)
Any rule				
ES	$\rho_{\theta^k, \theta^l}$	uLR (δ^1, δ^2)	mB (δ^1, δ^2)	mLR (δ^1, δ^2)
4.1	< 0	(−0.013, 0.001)	(−0.001, 0.005)	(−0.024, −0.011)
	≈ 0	(−0.009, −0.002)	(0.001, 0.001)	(−0.023, −0.016)
	> 0	(−0.006, 0.001)	(0.002, 0.004)	(−0.028, −0.007)
4.2	< 0	(−0.018, −0.009)	(−0.005, 0.003)	(−0.031, −0.019)
	≈ 0	(−0.014, 0.003)	(−0.002, −0.001)	(−0.032, −0.011)
	> 0	(−0.018, −0.002)	(−0.001, 0.005)	(−0.030, −0.008)
Compensatory rule				
ES	$\rho_{\theta^k, \theta^l}$	uLR $\delta(w)$	mB $\delta(w)$	mLR $\delta(w)$
4.1	< 0	−0.006	0.000	−0.030
	≈ 0	−0.012	0.000	−0.019
	> 0	−0.004	−0.006	−0.015
4.2	< 0	−0.018	−0.003	−0.040
	≈ 0	−0.017	−0.003	−0.029
	> 0	−0.013	0.003	−0.024

uLR = univariate logistic regression.

mB = multivariate Bernoulli.

mLR = multivariate logistic regression.

observed some bias in conditions with smaller samples (ES 3.1, 3.2, 4.1, and 4.2 under the Any and Compensatory decision rules). We can conclude that bias in regression coefficients was not necessarily problematic for our actual parameters of interest, namely success probabilities and differences between them. Even when regression coefficients had a small bias ($< |0.20|$ on the log-odds scale), success probabilities and treatment differences could be estimated without bias ($< |0.01|$), similar to the conditions without biased regression coefficients. This was the case for ES 3.1 and 3.2 under sample sizes of the Any and Compensatory rules. Only more severe bias of regression coefficients ($< |0.57|$ on the log-odds scale) in conditions with smaller sample sizes was not fully corrected in the transformation steps. This was seen in ES 4.1 and 4.2 under sample sizes of the Any and Compensatory rules.

4.2.2. Decision error rates

4.2.2.1. Average treatment effects. Probabilities to conclude superiority of average treatment effects are presented in Table D5. Decisions resulted in appropriate Type I error rates around 0.05 for each of the posterior distribution types under a least favorable scenario of no effect (i.e., ES 1.1, 1.2, 2.1, 2.2 of Any and Compensatory rules) and the proportions of

Table D5. Proportions of superiority decisions for ATEs with two outcome variables by data-generating mechanism, correlation, and decision rule.

ES	$\rho < 0$			$\rho = 0$			$\rho > 0$		
	uLR	mB	mLR	uLR	mB	mLR	uLR	mB	mLR
Rule = All									
1.1	0.000	0.004	0.000	0.000	0.005	0.001	0.004	0.005	0.007
1.2	0.003	0.002	0.001	0.000	0.005	0.002	0.006	0.006	0.005
2.1	0.000	0.001	0.003	0.002	0.004	0.004	0.006	0.005	0.008
2.2	0.002	0.003	0.000	0.007	0.002	0.005	0.003	0.005	0.010
3.1	0.064	0.051	0.046	0.066	0.046	0.056	0.054	0.043	0.046
3.2	0.051	0.048	0.055	0.050	0.057	0.050	0.049	0.052	0.061
4.1	0.059	0.051	0.042	0.059	0.044	0.044	0.052	0.046	0.053
4.2	0.051	0.058	0.045	0.045	0.041	0.051	0.044	0.049	0.053
Rule = Any									
1.1	0.052	0.046	0.054	0.060	0.064	0.060	0.059	0.047	0.050
1.2	0.054	0.055	0.043	0.035	0.042	0.050	0.038	0.053	0.049
2.1	0.063	0.053	0.059	0.055	0.044	0.045	0.052	0.049	0.049
2.2	0.059	0.055	0.062	0.059	0.045	0.062	0.046	0.048	0.060
3.1	0.807	0.802	0.789	0.810	0.812	0.806	0.796	0.787	0.791
3.2	0.814	0.790	0.807	0.819	0.811	0.791	0.811	0.803	0.815
4.1	0.804	0.756	0.781	0.816	0.775	0.787	0.808	0.780	0.777
4.2	0.790	0.749	0.793	0.806	0.774	0.770	0.781	0.754	0.785
Rule = Compensatory									
1.1	0.049	0.056	0.054	0.059	0.069	0.050	0.076	0.047	0.048
1.2	0.045	0.041	0.056	0.045	0.040	0.051	0.063	0.047	0.055
2.1	0.053	0.040	0.053	0.069	0.054	0.048	0.076	0.051	0.053
2.2	0.051	0.048	0.054	0.059	0.040	0.061	0.057	0.048	0.058
3.1	0.757	0.821	0.813	0.824	0.802	0.815	0.815	0.801	0.794
3.2	0.779	0.804	0.838	0.802	0.811	0.804	0.836	0.801	0.815
4.1	0.794	0.795	0.774	0.805	0.799	0.810	0.858	0.781	0.790
4.2	0.759	0.786	0.771	0.820	0.792	0.806	0.815	0.798	0.792

uLR = Univariate logistic regression.

mB = Multivariate Bernoulli.

mLR = Multivariate logistic regression.

Bold-faced entries have effect sizes that should lead to a superiority conclusion.

correct superiority conclusions (i.e., power) were close to the targeted 0.80 under a priori estimated sample sizes when the true effect was larger than zero (i.e., ES 3.1, 3.2, 4.1, 4.2 of Any and Compensatory rules). These results showcase that a priori computed sample sizes result in adequate statistical decisions.

In general, multivariate logistic regression (mLR) performed comparable to stratified multivariate analysis (mB) in the estimation of average treatment effects: Type I-error rates of mB were around 0.05 and statistical power was close to the targeted 0.80 as well. Compared to univariate logistic regression analysis (uLR), statistical power of multivariate logistic regression (mLR) appeared less sensitive to the correlation of the data. Effect sizes 3.1/3.2 and 4.1/4.2 under the Compensatory rule demonstrate most clearly how power of uLR increased when the correlation moved from negative to positive, with uncorrelated data reaching the targeted .80. For these conditions, the sample size which the uLR model was fitted on was smaller and larger respectively than needed for an analysis that assumes uncorrelated data. The difference between uLR and mLR was relatively subtle however, which is in line with the pattern of required sample sizes in Table D3. This table shows that differences in required sample sizes for different

correlations were relatively small under most data-generating mechanisms. This implies that the effect of using an incorrect sample size on statistical power is relatively limited under the data-generating mechanisms in the simulation study, in contrast with the scenarios presented in Table D1.

4.2.2.2. Conditional average treatment effects. The results of conditional treatment effects in the subpopulations are presented in Table D6. Similar to average treatment effects, Type I error rates were around the targeted 0.05 under the least favorable scenarios of no effect (ES 1.1, 1.2 for Any and Compensatory rules) for all estimation methods. The proportion to conclude superiority correctly was above 0.80 in all scenarios with a sample size exceeding the required sample size for CATEs. In the scenarios where the sample size for CATEs was lower than required (4.1 and 4.2 for the Any and Compensatory rules and 2.2 and 3.2 for the All rule), the power was below 0.80.

A comparison of estimations methods for the continuous covariate revealed that multivariate logistic regression (mLR) was generally more powerful than the stratified multivariate analysis (mB) approach when the covariate was continuous. These effects are prominent in ES 2.2 and 3.2 (All rule) as well as ES

Table D6. Proportions of superiority decisions for CATEs with two outcome variables by data-generating mechanism, correlation, and decision rule.

ES	$\rho < 0$			$\rho = 0$			$\rho > 0$		
	mB	mLR-S	mLR-V	mB	mLR-S	mLR-V	mB	mLR-S	mLR-V
Rule = All									
1.1	0.002	–	0.000	0.006	–	0.001	0.009	–	0.004
1.2	0.000	0.000	0.001	0.004	0.002	0.004	0.007	0.003	0.004
2.1	0.999	–	0.997	0.999	–	0.998	1.000	–	0.999
2.2	0.484	0.873	0.998	0.537	0.854	1.000	0.529	0.880	1.000
3.1	1.000	–	1.000	1.000	–	1.000	1.000	–	1.000
3.2	0.790	0.972	1.000	0.801	0.979	1.000	0.804	0.982	1.000
4.1	0.050	–	0.040	0.042	–	0.036	0.045	–	0.048
4.2	0.051	0.045	0.054	0.052	0.053	0.059	0.046	0.056	0.060
Rule = Any									
1.1	0.054	–	0.050	0.064	–	0.039	0.051	–	0.052
1.2	0.053	0.038	0.054	0.057	0.055	0.056	0.063	0.048	0.048
2.1	1.000	–	1.000	1.000	–	1.000	1.000	–	1.000
2.2	0.933	1.000	1.000	0.913	0.999	1.000	0.904	0.999	1.000
3.1	1.000	–	1.000	1.000	–	1.000	1.000	–	1.000
3.2	0.932	0.999	1.000	0.939	0.998	1.000	0.899	0.999	1.000
4.1	0.251	–	0.266	0.251	–	0.242	0.233	–	0.230
4.2	0.336	0.508	0.181	0.305	0.522	0.183	0.308	0.512	0.174
Rule = Compensatory									
1.1	0.061	–	0.047	0.076	–	0.033	0.048	–	0.039
1.2	0.040	0.040	0.043	0.062	0.057	0.056	0.057	0.046	0.048
2.1	1.000	–	1.000	1.000	–	1.000	1.000	–	1.000
2.2	0.980	1.000	1.000	0.969	1.000	1.000	0.945	0.999	1.000
3.1	1.000	–	1.000	1.000	–	1.000	1.000	–	1.000
3.2	0.951	1.000	1.000	0.953	1.000	1.000	0.945	1.000	1.000
4.1	0.283	–	0.326	0.292	–	0.319	0.287	–	0.316
4.2	0.390	0.504	0.190	0.354	0.534	0.183	0.359	0.537	0.232

mB = Multivariate Bernoulli.

mLR-S = Multivariate logistic regression –sample.

mLR-V = Multivariate logistic regression –value.

Bold-faced entries have effect sizes that should lead to a superiority conclusion.

4.2 (Any and Compensatory rules). The statistical power of stratified multivariate analysis (mB) and multivariate logistic regression analysis (mLR) did not differ for the discrete covariate, as demonstrated under ES 2.1 and 3.1 (All rule) as well as ES 4.1 (Any and Compensatory rules).

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 and interval of the covariate) 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: $Bp < -1SD$ (Low), and $Bp > 1SD$ (High).

Similar to the Numerical evaluation, we specified a diffuse multivariate normally distributed prior with means $\mathbf{b}^q = \mathbf{0}$ and variance matrix $\mathbf{B}^0 = \text{diag}(10, \dots, 10)$ 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

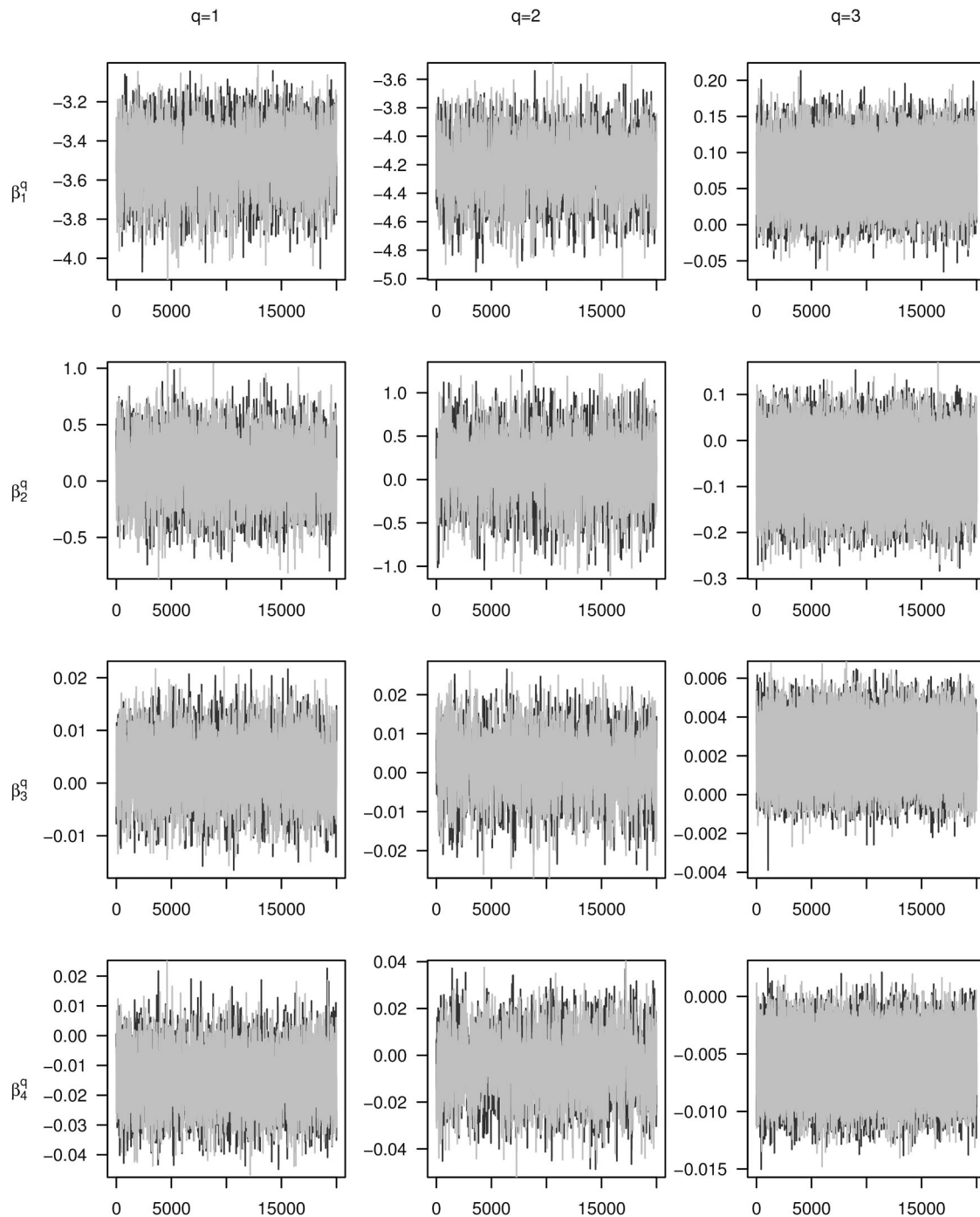


Figure D2. Traceplot of MCMC chains for the application of Bayesian multivariate logistic regression to the IST data.

proposed Gibbs sampler with 20,000 iterations plus 10,000 burnin iterations. Similar to the simulation study, we used traceplots and multivariate Gelman-Rubin convergence diagnostics to assess convergence (Gelman & Rubin, 1992; Brooks & Gelman, 1998). Traceplots (Figure D2) 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 longterm outcome was three times more relevant for the decision than the shortterm outcome. Since θ_T reflects failure

Table D7. Average and conditional average treatment effects (ATE and CATE respectively) and their posterior probabilities (pp) in the IST data, by interval of blood pressure (Bp). Superiority or inferiority was concluded when $>$ or $<$ respectively.

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

mB = Multivariate Bernoulli analysis.

mLR = Multivariate logistic regression.

Table D8. Conditional average treatment effects in the IST data, by value of blood pressure (Bp). Superiority or inferiority was concluded when $>$ or $<$ respectively.

Value	$\delta(\text{Bp})$	pp	Any	All	$\delta(w, \text{Bp})$	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	$>$

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 $\alpha = 0.05$ was performed with a decision threshold $p_{\text{cut}} = 1 - \frac{\alpha}{2} = 0.975$ (Compensatory and All rules) and a for multiple tests corrected $p_{\text{cut}} = 1 - \frac{\alpha}{2K} = 0.9875$ (Any rule).

5.2. Results

Results are presented in Table D7 for different intervals and in Table D8 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 did 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 toward higher failure probabilities on both dependent 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 multivariate logistic regression (mLR) and stratified multivariate analysis (mB) of response data 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 D3. If the regression-based 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 Bayesian multivariate logistic regression framework for analysis and decision-making with multiple correlated dependent variables. The framework is suitable to capture treatment heterogeneity among (groups of) patients that are distinguishable by observed covariate information (i.e., conditional average treatment effects) and to estimate overall treatment effects among the full population (i.e., average treatment effects) under a wide range of scenarios. In general, the proposed regression models were able to reproduce point estimates of average and conditional treatment differences correctly and resulted in decisions with anticipated error rates among the trial population and among subpopulations – as long as the sample was sufficiently large. Further, anticipated decision error rates were found under a priori sample size estimation for different correlation structures (namely negatively correlated, uncorrelated, and positively correlated dependent variables) and for two- and three-dimensional dependent variables. The illustration with the International

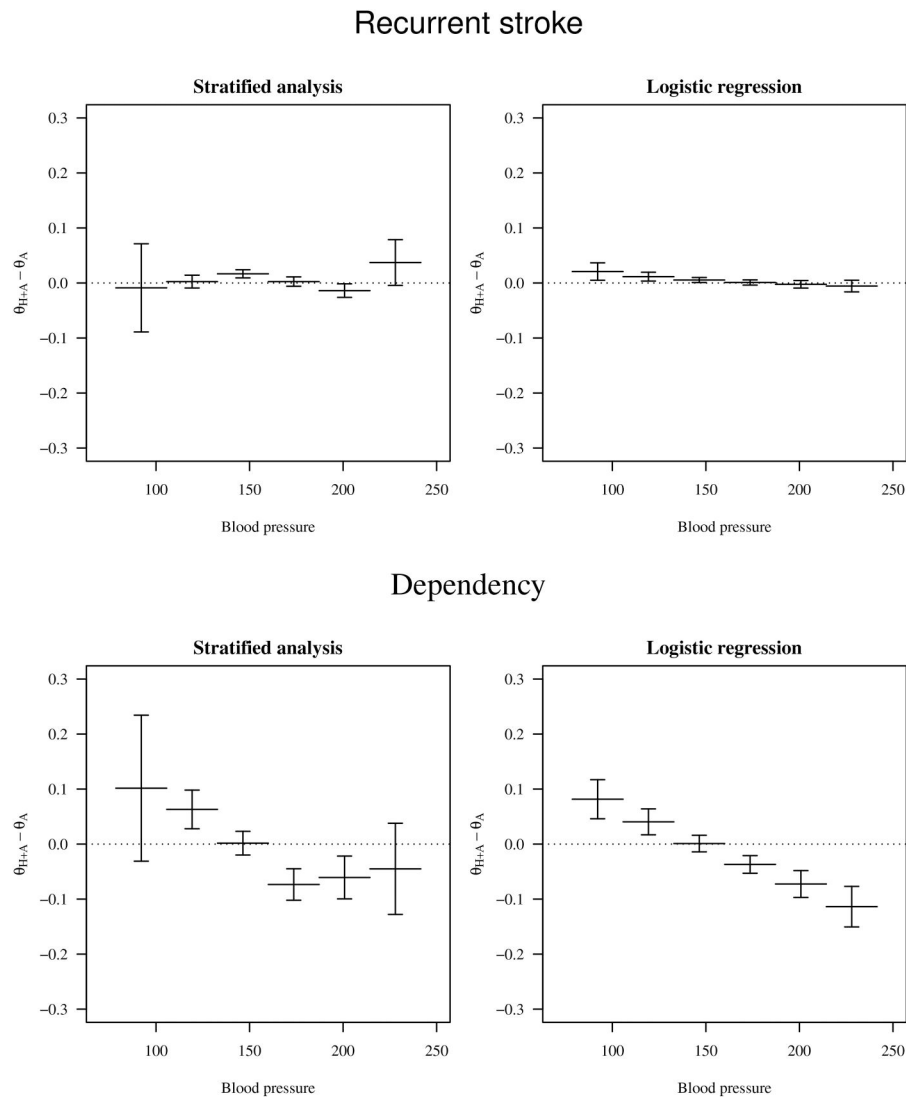


Figure D3. Comparison of CATEs and their standard deviations per interval of blood pressure after stratified multivariate analysis (mB) and multivariate logistic regression (mLR). Each interval reflects one standard deviation.

Stroke Dataset demonstrated how conditional average treatment effects could provide a more in-depth understanding of results beyond average treatment effects.

Compared to other approaches, the Bayesian multivariate logistic regression framework showed favorable properties. Decisions were more powerful than those obtained by multivariate stratified analysis when covariates were continuous, since they were based on information from the full sample rather than a subsample. Moreover, the Bayesian multivariate logistic regression model was more effective in targeting statistical power compared to multiple univariate logistic regression analyses when the correlation between dependent variables was non-zero. Whereas these effects were relatively subtle in the simulation study, the illustrative example in [Section 2.4.3](#) showcased that

they are more prominent when correlations are further from zero.

An advantage of the proposed multivariate logistic regression approach is its flexibility 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 dependent 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 dependent variables and (continuous) covariates.

In practice, researchers are encouraged to consider model assumptions in real 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. Further, researchers who aim to target decision error rates have to decide which treatment effect should be leading in the actual choice of sample size. Under treatment heterogeneity, average and (multiple) conditional average treatment effects have different effect sizes by definition, resulting in different sample sizes and raising the question which considerations meaningfully guide this choice.

Theoretically, the framework lends itself for use under a much wider range of scenarios than showcased in this paper. Each of the elements – modeling, transformation, decision-making – can be replaced by an alternative, resulting in a large number of variations. Some variations, such as a less computationally intensive analysis model, a wider range of prior distributions, and interim monitoring as an alternative to decision-making with a priori estimated sample sizes, were presented already (Kavelaars et al., 2020). Here, we mention two additional suggestions to elaborate the framework. First, in addition to the presented transformations to success probabilities and treatment, transformations to other associations between treatment and outcome, such as relative risks and risk ratios, may be of interest and are worth investigating. Second, other hypotheses than superiority and inferiority, such as non-inferiority or equality decision-making, can be relevant to be included in the framework as well (see for a discussion Van Ravenzwaaij et al., 2019). More flexible formulations of hypotheses and another perspective on the assessment of evidence can be achieved via the computation of Bayes factors (see for an introduction e.g., Mulder & Wagenmakers, 2016).

Other than the abovementioned variations, 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 current paper provided tools to compute required sample sizes and to control decision error rates, if researchers are able to estimate effect sizes with reasonable accuracy prior to the study and when sample sizes are sufficiently large. When sample sizes were relatively small, bias was introduced. 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.

Further, another interesting direction for future research would be to extend the proposed multivariate logistic regression model for estimating average and conditional average treatment effects and for decision-making with (discrete or continuous) latent variables to capture unexplained heterogeneity. This extension falls outside of the scope of the current paper which focuses on modeling treatment heterogeneity caused by observed covariate information.

Lastly, causal inference is less straightforward in (stratified) subgroup analysis as conditioning upon covariates might interfere with randomization (European Medicine Agency, 2019; Food & 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.

Article information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grant 406.18.505 from the Dutch Research Council.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

References

- Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4), 434–455.
- Chen, M.-H., & Ibrahim, J. G. (2000). Power prior distributions for regression models. *Statistical Science*, 15(1), 46–60. <https://doi.org/10.1214/ss/1009212673>
- 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>
- Chow, S.-C., Shao, J., Wang, H., & Lokhnygina, Y. (2017). *Sample size calculations in clinical research* (3rd ed.). Chapman and Hall/CRC.
- Chuang-Stein, C., Stryzak, 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.
- 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., & 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>
- 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. <https://doi.org/10.1093/biomet/80.1.27>
- 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).
- 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).
- 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>
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472. <https://doi.org/10.1214/ss/1177011136>
- Goldberger, J. J., & Buxton, A. E. (2013). Personalized medicine vs guideline-based medicine. *JAMA*, 309(24), 2559–2560. <https://doi.org/10.1001/jama.2013.6629>
- Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine. *The New England Journal of Medicine*, 363(4), 301–304. <https://doi.org/10.1056/NEJMp1006304>
- Hoogland, J., Int'Hout, J., Belias, M., Rovers, M. M., Riley, R. D., E Harrell, F., 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>
- 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., & Branson, M. (2011). Bayesian models for subgroup analysis in clinical trials. *Clinical Trials (London, England)*, 8(2), 129–143. <https://doi.org/10.1177/1740774510396933>

- 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>
- Kerman, J. (2011). Neutral noninformative and informative conjugate beta and gamma prior distributions. *Electronic Journal of Statistics*, 5, 1450–1470. <https://doi.org/10.1214/11-EJS648>
- Lipkovich, I., Dmitrienko, A., & B, R. (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: 1611.06649v3.
- 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>
- McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and Its Application*, 6(1), 355–378. <https://doi.org/10.1146/annurev-statistics-031017-100325>
- Microsoft, & Weston, S. (2020). *doParallel: Foreach Parallel adaptor for the 'parallel' package* R package version 1.0.16.
- Microsoft, & Weston, S. (2020). *foreach: Provides foreach looping construct* R package version 1.5.1.
- Mulder, J., & Wagenmakers, E.-J. (2016). Editors' introduction to the special issue "bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology*, 72, 1–5. <https://doi.org/10.1016/j.jmp.2016.01.002>
- 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), 56. <https://doi.org/10.1186/1471-2288-9-56>
- O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, 40(4), 1079–1087.
- 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>
- 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>
- 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](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*, 43(3), 487. <https://doi.org/10.2307/2531989>
- 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](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.
- Ristl, R., Urach, S., Rosenkranz, G., & Posch, M. (2019). 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*. John Wiley & Sons.
- Sandercock, P. A., Niewada, M., & Czlonkowska, A. (2011). The international stroke trial database. *Trials*, 12, 101. 1). <https://doi.org/10.1186/1745-6215-12-101>
- 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>
- Simon, R. (2010). Clinical trials for predictive medicine: New challenges and paradigms. *Clinical Trials (London, England)*, 7(5), 516–524. <https://doi.org/10.1177/1740774510366454>
- Sozu, T., Sugimoto, T., & Hamasaki, T. (2010). Sample size determination in clinical trials with multiple co-primary binary endpoints. *Statistics in Medicine*, 29(21), 2169–2179. <https://doi.org/10.1002/sim.3972>
- 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>
- 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), 23–30. <https://doi.org/10.2307/2532599>
- 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>
- Tuyl, F., Gerlach, R., & Mengersen, K. (2008). A comparison of Bayes-laplace, Jeffreys, and other priors: The case of zero events. *The American Statistician*, 62(1), 40–44. <https://doi.org/10.1198/000313008X267839>
- 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), 71. <https://doi.org/10.1186/s12874-019-0699-7>

- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. 4th ed. Springer. ISBN 0-387-95457-0.
- 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>
- 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 (London, England)*, 2(5), 387–393. <https://doi.org/10.1191/1740774505cn112oa>
- 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 (London, England)*, 18(5), 570–581. <https://doi.org/10.1177/17407745211028588>
- 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 (London, England)*, 4(2), 140–153. <https://doi.org/10.1177/1740774507076936>

Appendix 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:

$$l(\mathbf{y}|\beta, \mathbf{x}) = \prod_{i=1}^n \prod_{q=1}^{Q-1} \left(\frac{\exp[\psi_i^q]}{\sum_{r=1}^{Q-1} \exp[\psi_i^r] + 1} \right)^{I(y_i=q)} \left(\frac{1}{\sum_{r=1}^{Q-1} \exp[\psi_i^r] + 1} \right)^{I(y_i=Q)}. \quad (\text{A1})$$

Following Polson et al. (Polson et al., 2013), we introduce the Pólya-gamma variable by rewriting the multivariate likelihood in Equation A1 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:

$$l(\mathbf{y}|\beta^q, \beta^{-q}) = \prod_{i=1}^n \left(\frac{\exp[\eta_i^q]}{\exp[\eta_i^q] + 1} \right)^{I(y_i=q)} \left(\frac{1}{\exp[\eta_i^q] + 1} \right)^{1-I(y_i=q)} \quad (\text{A2})$$

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

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

$$\frac{\exp[\eta_i^q]}{\exp[\eta_i^q] + 1} = 2 \exp \left[\left(y_i - \frac{1}{2} \right) \eta_i^q \right] \int_0^\infty \exp \left[\frac{-\omega_i^q \eta_i^{q2}}{2} \right] p(\omega_i^q) d\omega_i^q \quad (\text{A3})$$

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

If we use the equality in Equation A3, the binomial likelihood in Equation A2 can be transformed to a multivariate Gaussian likelihood by including an auxiliary Pólya-Gamma variable ω_i^q (Polson et al., 2013):

$$\begin{aligned} l(\mathbf{y}|\beta^q, \beta^{-q}) &= \prod_{i=1}^n \frac{\exp[\eta_i^q]}{\exp[\eta_i^q] + 1} \\ &= \prod_{i=1}^n 2 \exp \left[\left(y_i - \frac{1}{2} \right) \eta_i^q \right] \int_0^\infty \exp \left[\frac{-\omega_i^q \eta_i^{q2}}{2} \right] p(\omega_i^q) d\omega_i^q \\ &= \prod_{i=1}^n \exp \left[\kappa_i^q \omega_i^q \eta_i^q - \frac{1}{2} (\eta_i^q)^2 \omega_i^q \right] \text{PG}(\omega_i^q | 1, 0) \\ &\propto \exp \left[\frac{1}{2} (2\kappa^q \omega^q \eta^q - \omega^q (\eta^q)^2) \right] \\ &\propto \exp \left[-\frac{1}{2} (\kappa^q - \eta^q)^T \mathbf{\Omega}^q (\kappa^q - \eta^q) \right] \\ &= \exp \left[-\frac{1}{2} \left(\kappa^q - \mathbf{X}\beta^q + \ln \left[\sum_{m \neq q} \exp(\mathbf{X}\beta^m) \right] \right)^T \right. \\ &\quad \left. \mathbf{\Omega}^q (\kappa^q - \mathbf{X}\beta^q + \ln \left[\sum_{m \neq q} \exp(\mathbf{X}\beta^m) \right]) \right], \end{aligned} \quad (\text{A4})$$

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

A.0.1. Prior distribution

The Gaussian likelihood in Equation A4 is conditionally conjugate with a normal prior distribution on regression coefficients β^q :

$$\beta^q \sim N(\mathbf{b}^q, \mathbf{B}^{0q}) \quad (\text{A5})$$

where $\mathbf{b}^q$ is the vector of prior means of regression coefficient vector β^q and $\mathbf{B}^{0q}$ 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 & Greenland, 2012; Chen & 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:

$$p(\beta|\mathbf{y}) \propto p(\mathbf{y}|\beta, \mathbf{x}) p(\beta), \quad (\text{A6})$$

The combination of a Poly-Gamma transformed Gaussian likelihood (Equation A4) and a normal prior distribution (Equation A5) respectively is proportional to a normally distributed posterior distribution, conditionally on Poly-Gamma variables in ω^q (Polson et al., 2013):

$$\begin{aligned}
p(\beta^q | \mathbf{Y}, \Omega^q) &\propto p(\mathbf{y} | \beta^q, \omega^q) p(\beta^q) \\
&\propto \exp \left[-\frac{1}{2} \left(\kappa^q - \mathbf{X} \beta^q + \ln \left[\sum_{m \neq q} \exp [\mathbf{X} \beta^m] \right] \right)^T \Omega^q (\kappa^q - \mathbf{X} \beta^q \right. \\
&\quad \left. + \ln \left[\sum_{m \neq q} \exp [\mathbf{X} \beta^m] \right] \right) \times \\
&\quad \exp \left[-\frac{1}{2} (\beta^q - \mathbf{b}^q)^T (\mathbf{B}^q)^{-1} (\beta^q - \mathbf{b}^q) \right] \\
&\propto N(\mathbf{V}^q (\mathbf{X}^T \Omega^q \mathbf{X} + \ln [\sum_{m \neq q} \exp [\mathbf{X} \beta^m]]) + (\mathbf{B}^q)^{-1} \mathbf{b}^q, \mathbf{V}^q)
\end{aligned} \tag{A7}$$

where $\mathbf{V}^q = (\mathbf{X}^T \Omega^q \mathbf{X} + (\mathbf{B}^q)^{-1})^{-1}$. 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 $q = 1, \dots, Q - 1$, 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 $\mathbf{V}^q$.

$$\beta^q | \omega^q \sim N(\mathbf{m}^q, \mathbf{V}^q) \tag{A8}$$

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

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

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

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

$$\omega_i^q | \beta^q \sim PG(1, \psi_i^q - \ln \sum_{m \neq q} \exp [\psi_i^m]). \tag{A9}$$

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

Appendix 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:

$$\psi_T^q = \beta_0^q + \beta_1^q T + \beta_2^q x + \beta_3^q x \times T \tag{B1}$$

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 $\theta_T(x)$ and correlations $\rho_T(x)$ can be transformed to joint response probabilities $\phi_T(x)$ via the following set of equations:

$$\begin{aligned}
\phi_T^{11}(x) &= \rho_T(x) \sqrt{\theta_T^1(x) [1 - \theta_T^1(x)] \theta_T^2(x) [1 - \theta_T^2(x)]} \\
&\quad + \theta_T^1(x) \theta_T^2(x) \\
\phi_T^{10}(x) &= \theta_T^1(x) - \phi_T^{11}(x) \\
\phi_T^{01}(x) &= \theta_T^2(x) - \phi_T^{11}(x) \\
\phi_T^{00}(x) &= 1 - \theta_T^1(x) - \theta_T^2(x) + \phi_T^{11}(x)
\end{aligned} \tag{B2}$$

For each response category q , joint responses ϕ_T^q can be transformed to linear predictor ψ_T^q 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 :

$$\begin{aligned}
\beta_0^q &= \frac{x^H \psi_0^q(x^L) - x^L \psi_0^q(x^H)}{x^H - x^L} \\
\beta_1^q &= \frac{x^H [\psi_1^q(x^L) - \psi_0^q(x^L)] + x^L [\psi_0^q(x^H) - \psi_1^q(x^H)]}{x^H - x^L} \\
\beta_2^q &= \frac{\psi_0^q(x^H) - \psi_0^q(x^L)}{x^H - x^L} \\
\beta_3^q &= \frac{\psi_1^q(x^H) - \psi_0^q(x^H) - \psi_1^q(x^L) + \psi_0^q(x^L)}{x^H - x^L}
\end{aligned} \tag{B3}$$

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

$$\begin{aligned}
\theta_1^L &= (0.60, 0.70), \rho_1^L = -0.30 \\
\theta_1^H &= (0.40, 0.30), \rho_1^H = -0.30 \\
\theta_0^L &= (0.40, 0.30), \rho_0^L = -0.30 \\
\theta_0^H &= (0.60, 0.70), \rho_0^H = -0.30,
\end{aligned}$$

then the regression coefficients would be as presented in Table D9.

Table D9. 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 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.

Let $\beta^Q = (0, \dots, 0)$

1. **for** draw $(l) \leftarrow 1 : L$ **do**
2. **for** treatment $T \leftarrow 0 : 1$ **do**
3. **for** joint response $q \leftarrow 1 : Q$ **do**
4. Compute $\psi_T^{q(l)} = \beta_0^{q(l)} + \beta_1^{q(l)} T + \beta_2^{q(l)} x + \beta_3^{q(l)} x \times T$
5. Compute $\phi_T^{q(l)} = \frac{\exp [\psi_T^{q(l)}]}{\sum_{r=1}^{Q-1} \exp [\psi_T^{r(l)}] + 1}$
6. **end for**
7. **end for**
8. **end for**

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

Let $\beta^Q = (0, \dots, 0)$

1. **for** draw $(l) \leftarrow 1 : L$ **do**
2. **for** subject $i \leftarrow 1 : n$ **do**
3. **for** joint response $q \leftarrow 1 : Q$ **do**
4. Compute $\psi_i^{q(l)} = \beta_1^{q(l)} T_i + \beta_2^{q(l)} x_i + \beta_3^{q(l)} x_i \times T_i$
5. Compute $\phi_i^{q(l)} = \frac{\exp[\psi_i^{q(l)}]}{\sum_{r=1}^{Q-1} \exp[\psi_i^{r(l)}] + 1}$
6. **for** $T \leftarrow 0 : 1$ **do**
7. Compute $\phi_T^{q(l)}(\mathbf{x}) = \frac{1}{\sum_{i=1}^n I(T_i=T)} \phi_i^{q(l)} I(T_i=T)$
8. **end for**
9. **end for**
10. **end for**
11. **end for**

Appendix D. Numerical evaluation with three outcome variables

In the current section, we present an evaluation of the BMLR framework with three dependent variables.

D.1. Setup

The evaluation largely follows the setup of the simulation with two dependent variables (Section 4). Aspects that differ from this simulation will be discussed here.

D.1.1. Analysis

We presented the results of Bayesian trivariate logistic regression analysis and compared it to a multivariate Bernoulli procedure.

D.1.2. Effect size

We presented the results of the Bayesian trivariate logistic regression analysis for a selection of effect sizes, namely 1.1 and 3.1. Using the three correlation structures ($\rho < 0$, $\rho \approx 0$, and $\rho > 0$) for each of the effect sizes resulted in the six data generating mechanisms presented in Table D10.

Table D10. Parameters of average treatment effects (ATEs) in the trial and conditional average treatment effects (CATEs) in a subpopulation for tree outcome variables.

ES	ATE		CATE			
	$(\delta_1, \delta_2, \delta_3)$	$\delta(\mathbf{w})$	$\rho_{\theta^k, \theta^l}$	$(\delta_1, \delta_2, \delta_3)$	$\delta(\mathbf{w})$	$\rho_{\theta^k, \theta^l}$
1.1 D	(0.000, 0.000, 0.000)	0.000	−0.160	(0.000, 0.000, 0.000)	0.000	−0.200
			0.030			0.000
			0.220			0.200
3.1 D	(0.100, 0.000, 0.100)	0.075	−0.152	(0.300, 0.200, 0.300)	0.275	−0.200
			0.040			0.000
			0.232			0.200

ES = Effect size, D = Discrete covariate.

Table D11. Required sample sizes to evaluate the average treatment effect (ATE) and conditional treatment effect (CATE) for three outcome variables.

ES	$\rho_{\theta^k, \theta^l}$	All			Any			Compensatory		
		ATE	CATE	Sub	ATE	CATE	Sub	ATE	CATE	Sub
1.1	< 0	–	–	500	–	–	500	–	–	500
	≈ 0	–	–	500	–	–	500	–	–	500
	> 0	–	–	500	–	–	500	–	–	500
3.1	< 0	–	79	500	234	20	117	153	9	77
	≈ 0	–	79	500	255	23	128	218	14	109
	> 0	–	78	500	276	26	138	284	19	142

Sub = expected size of subsample.

D.1.3. Sample size

Similar to the Numerical evaluation, we applied the All, Any, and Compensatory rules. We assigned the Compensatory rule unequal weights $\mathbf{w} = (0.50, 0.25, 0.25)$.

The required sample sizes for three outcome variables are computed via the procedure described in Section 2.4, targeting at a Type I error rate of 0.05 and a power of .80. The sample sizes are presented in Table D11.

D.1.4. Decision rule

We performed a right-sided (superiority) test aiming at a Type I-error rate of $\alpha = 0.05$. We used a decision threshold $p_{cut} = 1 - \alpha = 0.95$ (Compensatory and All rules) and a for multiple tests corrected $p_{cut} = 1 - \frac{\alpha}{K} = 0.981$ (Any rule) (Marsman & Wagenmakers, 2016; Kavelaars et al., 2020; Sozu et al., 2016).

D.1.5. Procedure

To stabilize computations, we used 20,000 iterations for the multivariate Bernoulli model.

D.2. Results

Table D12. Proportions of superiority decisions for three outcome variables by data-generating mechanism, correlation, and decision rule.

ES	Type	$\rho < 0$		$\rho = 0$		$\rho > 0$	
		mB	mLR	mB	mLR	mB	mLR
Rule = All							
1.1	ATE	0.000	0.000	0.000	0.001	0.004	0.002
3.1	ATE	0.045	0.045	0.058	0.048	0.044	0.058
1.1	CATE	0.000	0.000	0.000	0.000	0.001	0.000
3.1	CATE	1.000	1.000	1.000	1.000	1.000	1.000
Rule = Any							
1.1	ATE	0.049	0.056	0.045	0.050	0.046	0.056
3.1	ATE	0.814	0.822	0.796	0.775	0.815	0.775
1.1	CATE	0.047	0.050	0.042	0.037	0.063	0.032
3.1	CATE	1.000	1.000	1.000	1.000	1.000	1.000
Rule = Compensatory							
1.1	ATE	0.048	0.068	0.050	0.052	0.052	0.063
3.1	ATE	0.781	0.826	0.788	0.757	0.787	0.776
1.1	CATE	0.051	0.043	0.056	0.029	0.053	0.035
3.1	CATE	1.000	1.000	1.000	1.000	1.000	1.000

mB = Multivariate Bernoulli.

mLR = Multivariate logistic regression.

Bold-faced entries should lead to a superiority conclusion.