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

Xynthia Kavelaars\textsuperscript{a,*}, Joris Mulder\textsuperscript{a}, and Maurits Kaptein \textsuperscript{b}

\textsuperscript{a} Department of Methodology and Statistics, Tilburg University, The Netherlands
\textsuperscript{b} Jheronimus Academy of Data Science, The Netherlands

Abstract

The effects of a treatment may differ between patients with different characteristics. Addressing such treatment heterogeneity is crucial to identify which patients benefit from a treatment, but can be complex in the context of multiple correlated binary outcomes. The current paper presents a novel Bayesian method for estimation and inference for heterogeneous treatment effects in a multivariate binary setting. The framework is suitable for prediction of heterogeneous treatment effects and superiority/inferiority decision-making within subpopulations, while taking advantage of the size of the entire study sample.

We introduce a decision-making framework based on Bayesian multivariate logistic regression analysis with a Pólya-Gamma expansion. The obtained regression coefficients are transformed into differences between success probabilities of treatments to allow for treatment comparison in terms of point estimation and superiority and/or inferiority decisions for different (sub)populations. Procedures for a priori sample size estimation under a non-informative prior distribution are included in the framework.

A numerical evaluation demonstrated that a) average and conditional treatment effect parameters could be estimated unbiasedly when the sample is large enough; b) decisions based on a priori sample size estimation resulted in anticipated error rates. Application to the International Stroke Trial dataset revealed a heterogeneous treatment effect: The model showed conditional treatment effects in opposite directions for patients with different levels of blood pressure, while the average treatment effect among the trial population was close to zero.

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

\textsuperscript{*}Contact: X.M. Kavelaars x.m.kavelaars@tilburguniversity.edu
1 Introduction

Clinical trials often focus on average treatment effects (ATEs) among the study population when comparing treatments. Average treatment effects can be sufficiently insightful when the effects of a treatment are relatively homogeneous over the trial population. However, average effects may give a limited, or even erroneous, impression when the effects of a treatment are heterogeneous and thus differ between patients with different characteristics. In that case, treatment effects conditional on a specific subpopulation are more informative and contribute to a better understanding of the potential of the treatment. Addressing potential treatment heterogeneity in the evaluation of treatments is crucial to a) identify which patients are likely to benefit from a treatment; and b) optimize treatment results of individual patients via personalized treatment assignment [8, 9, 29].

The current paper focuses on estimating heterogeneous treatment effects in the context of multiple (correlated) binary outcome variables. Often multiple types of clinical events (e.g. development of cancer metastases), functional measures (e.g. memory decline), and patient symptoms (e.g. fatigue) are assessed within a single clinical trial, making multiple binary outcome variables common in medical research [6, 24]. Combining multiple outcome variables in treatment evaluation can provide extensive multidimensional insights into the effects of a treatment. Furthermore, correspondence between statistical and clinical decision-making can be improved, since multiple treatment effects can be combined and weighted in various ways to provide a single statistical decision regarding treatment superiority or inferiority [e.g. 19, 17, 15]. Compared to univariate analysis, performing a multivariate analysis takes correlations between outcome variables into account and potentially reduces decision errors. More specifically, correlations influence the sample sizes required for decision-making with prespecified error rates and provoke under- or overpowerment when falsely omitted [6, 12]. In sum, combining multiple outcome variables and treatment heterogeneity in statistical analysis has the potential to reveal different outcome patterns for different patient profiles.

We illustrate the importance of addressing treatment heterogeneity in the context of multiple outcome variables with the International Stroke Trial [23, 10]. The International Stroke Trial (IST) investigated the effects of two anti-thrombotic agents (Aspirin and/or Heparin) on the occurrence of several adverse short-term events (e.g. recurrent stroke, embolism) and long-term patient status (e.g. dependency in daily life, recovery, or death) among subjects who had an initial stroke. The average treatment differences in the trial were very small, so one might conclude that treatment with one of these drugs was marginally effective. However, these findings show how specific characteristics of patients (e.g. age or blood pressure) and/or disease (e.g. type of stroke) potentially interact with the treatment and produce different risks and
perspectives for patients with different profiles. Average treatment effects do, for example, not reveal whether treated patients with low or high blood pressure have better prospects in terms of short-term damage risk and/or long-term recovery potential than other patients. Clearly, potential heterogeneous effects as these would have clinically relevant implications and advocate the development of more personalized treatment policies.

Although theoretically relevant in many contemporary trials, decision-making under treatment heterogeneity in the multivariate context is considerably more complex compared to the non-heterogeneous and/or univariate setting. First, such a setting demands a multivariate analysis method to obtain accurate decision error rates. For accurate estimation of conditional treatment effects, the analysis should not only include the correlation, but should also be flexible enough to deal with correlations that differ over subpopulations. The latter is not evident in multivariate analysis methods for binary outcome variables: Some methods impose the marginal correlation structure of the trial population on subpopulations (e.g. multivariate probit models by Chib [2] or Rossi et al. [22]). Second, the interpretation of treatment effects can be complex in multivariate non-linear models. Whereas univariate logit and probit regression analyses are often interpreted in terms of odds ratios and average marginal effects respectively, multivariate alternatives or generalizations are not straightforward to interpret [13, 18, 2, 22].

To capture the complexity of heterogeneous, multivariate treatment effects, we extend a Bayesian multivariate Bernoulli framework for superiority decision-making [12]. The existing model is interpreted in terms of differences between (multivariate) success probabilities and allows for estimation as well as inference regarding average and conditional treatment effects. We expand the framework with a Bayesian multivariate logistic regression analysis to incorporate potential treatment heterogeneity via the inclusion of covariates. The proposed method is able to model treatment effects and correlations on a subpopulation level and is suitable for estimation and inference among other populations than the trial population. Along with the regression model, we include a procedure to compute sample sizes for decision-making with prespecified frequentist error rates.

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 apply the methods to data
from the International Stroke Trial and conclude the paper with a discussion.

2 Decision-framework

2.1 Multivariate logistic regression

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

$$H = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ 1 & 1 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$ (1)

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

Joint response $y_i$ follows a multivariate Bernoulli distribution based on joint response probabilities $\phi_i$:

$$p(y_i) = \exp \left[ \log(\phi_i^1) \prod_{k=1}^{K} y_i^k + \log(\phi_i^2) \prod_{k=1}^{K-1} y_i^k (1 - y_i^K) + \cdots + \log(\phi_i^Q) \prod_{k=1}^{K-1} (1 - y_i^k) y_i^K + \log(\phi_i^Q) \prod_{k=1}^{K} (1 - y_i^k) \right].$$ (2)

Equation (2) can be recognized as a special case of the multinomial distribution, that treats dependent multivariate binary responses as independent multinomial joint response combinations.

The multinomial distribution in Equation (2) serves as a likelihood of unconditional joint response $y_i$ and is conjugate with a Dirichlet prior. Posterior joint response probabilities $\phi_i$ are known to follow a Dirichlet distribution and can be drawn directly. The multinomial distribution in Equation (2) can be used to represent the likelihood of conditional joint response $y_i | x_i$ as well. In this case, the probability is reflected by function, $\phi_i^q(x_i)$, that maps the dependency of joint response probabilities on covariates $x_i$ via
a multinomial logistic function:

\[
\phi^q_i(x_i) = \frac{\exp[\psi^q_i(x_i)]}{\sum_{r=1}^{Q-1} \exp[\psi^r_i(x_i)] + 1}
\]  

(3)

for response categories \(q = 1, \ldots, Q - 1\). In Equation (3) \(\psi^q_i(x_i)\) reflects the linear predictor of response category \(q\) and subject \(i\):

\[
\psi^q_i(x_i) = \beta^q_0 + \beta^q_1 x_{i1} + \cdots + \beta^q_p x_{ip}.
\]  

(4)

Here, \(x_{ip}\) can be a treatment indicator, a patient characteristic, or an interaction between these. Vector \(\beta^q = (\beta^q_0, \beta^q_1, \ldots, \beta^q_p)\) 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. \(\beta^Q = 0\).

The likelihood of response data follows from taking the product over \(n\) individual joint response probabilities from Equation (3) of \(Q\) response categories:

\[
l(y|\beta, x) = \prod_{i=1}^{n} \prod_{q=1}^{Q-1} \left( \frac{\sum_{r=1}^{Q-1} \exp[\psi^r_i(x_i)]}{\sum_{r=1}^{Q-1} \exp[\psi^r_i(x_i)] + 1} \right)^{I(y_i = H_q)} \left( \frac{1}{\sum_{r=1}^{Q-1} \exp[\psi^r_i(x_i)] + 1} \right)^{I(y_i = H_Q)}
\]  

(5)

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

\[
p(\beta|y) \propto p(y|\beta, x)p(\beta),
\]  

(6)

where \(p(\beta)\) reflects the prior distribution of the unknown parameters before observing the data. Posterior sampling can be done with a Gibbs sampling algorithm based on a Polya-Gamma expansion. Computational details of this procedure can be found in Appendix A.

2.2 Transformation of regression coefficients to treatment differences of (sub)populations

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

In the first step, the posterior sample of regression coefficients $\beta$ is transformed to a treatment effect in terms of marginal joint response probabilities $\phi_T(x)$ for each treatment $T \in \{0, 1\}$. Vector $\phi_T(x)$ is marginalized over individual joint response probabilities $\phi^q_T(x_i)$ of all subjects $i$ in treatment $T$. Thus, the vector of covariates $x_i$ should at least include a treatment indicator $T_i$, resulting in a more specific version of the general linear predictor from Equation 4):

$$
\psi^q_i(x_i) = \beta^q_0 + \beta^q_1 T_i + \cdots + \beta^q_P x_{iP}.
$$

(7)

Linear predictor $\psi^q_i(x_i)$ is then transformed to individual joint response probability $\phi^q_i(x_i)$ via the multinomial logistic function in Equation 3:

$$
\phi^q_i(x_i) = \frac{\exp[\psi^q_i(x_i)]}{\sum_{r=1}^{Q-1} \exp[\psi^r_i(x_i)] + 1}
$$

(3 revisited)

Marginalization over treatment $T$ can be done in several ways, depending on the specified (sub)population of interest. We discuss three different procedures to find $\phi_T(x)$ in Section 3.

2. **Joint response probabilities** $\phi_T(x)$ **to multivariate success probabilities** $\theta_T(x)$:

The next step in the transformation involves the conversion from joint response probabilities $\phi_T(x)$ to multivariate success probabilities of individual outcome variables $\theta_T(x)$. Especially when the number of outcome variables increases, success probabilities are more straightforward in their interpretation than joint response probabilities. The relation between both quantities is additive: Success probability $\theta^k_T$ on outcome $k$ and treatment $T$ equals the sum of a selection of elements of $\phi_T$, denoted by matrix $U_k$:

$$
\theta^k_T(x) = \sum_{q=1}^{Q} \phi^q_T(x) I(H_q \in U_k).
$$

(8)

Selection $U_k$ consists of the $2^{K-1}$ rows of $H$ that have their $k^{th}$ element equal to 1. If we drop the dependency on $x$ for notational simplicity of the following example, two outcome variables (i.e.
$k \in \{1, 2\}$ would imply that

$$
H = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad U_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \text{and} \quad U_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.
$$

Hence, success probabilities $\theta_T = (\theta^1_T, \theta^2_T) = (\phi^1_T + \phi^2_T, \phi^1_T + \phi^3_T)$.

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

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

$$
\delta^k(x) = \theta^1_k(x) - \theta^0_k(x).
$$

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

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

2.3 **Posterior decision-making**

Decisions rely on estimated treatment effects and their uncertainties. More formally, the multivariate case has complete parameter space $S \subset (-1, 1)^K$, which is divided into a rejection region $S_R$ and an non-rejection region $S_N$. Rejection region $S_R$ reflects the part of the parameter space that indicates the treatment difference of interest, while the non-rejection region $S_N$ 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 [28]. We consider the following three (commonly used) decision types:

1. superiority with region $S_R \in S_S$, where the treatment is better;
2. inferiority with region $S_R \in S_I$, where the treatment is worse;
3. two-sided with rejection region $S_R \in \{S_S, S_I\}$, where the treatment can be either better or worse.
We would conclude superiority and/or inferiority when the posterior probability that treatment difference \( \delta(x) \) lies in the rejection region exceeds a prespecified decision threshold, \( p_{\text{cut}} \):

\[
p(\delta(x) \in S_R | y) > p_{\text{cut}}.
\] (10)

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 \( \delta(x) \). Equation (10) is then applied in practice as:

\[
\frac{1}{L} \sum_{(i)=1}^{L} \delta^{(i)}(x) \in S_R | y > p_{\text{cut}}.
\] (11)

In a situation with multiple outcome variables, superiority and inferiority can be defined in multiple ways, resulting in different rejection regions. Although not intended as an exhaustive overview, we list a few possible rules and graphically present their rejection regions in Figure 1.
1. **Any rule**: The Any rule results in superiority or inferiority when the difference between success probabilities is larger or smaller than zero respectively on at least one of the outcome variables \(26\). The superiority and inferiority spaces are defined as:

\[
S_{S}^{\text{Any}} = \delta(x) \in \{ \delta(x) | \max_{1 < k < K} \delta^k(x) > 0 \} \mid y
\]

\[
S_{I}^{\text{Any}} = \delta(x) \in \{ \delta(x) | \min_{1 < k < K} \delta^k(x) < 0 \} \mid y.
\]

2. **All rule**: The All rule results in superiority or inferiority when the difference between success probabilities is larger or smaller than zero respectively on all of the outcome variables \(25\). The superiority and inferiority spaces are defined as:

\[
S_{S}^{\text{All}} = \delta(x) \in \{ \delta(x) | \min_{1 < k < K} \delta^k(x) > 0 \} \mid y
\]

\[
S_{I}^{\text{All}} = \delta(x) \in \{ \delta(x) | \max_{1 < k < K} \delta^k(x) < 0 \} \mid y.
\]

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:

\[
S_{S}^{\text{Comp}}(w) = \{ \delta(x) | \delta(w, x) > 0 \} \mid y
\]

\[
S_{I}^{\text{Comp}}(w) = \{ \delta(x) | \delta(w, x) < 0 \} \mid y
\]

where \(w = (w^1, \ldots, w^K)\) reflect weights of \(K\) treatment differences, \(\delta(w, x) = \sum_{k=1}^{K} w^k \delta^k(x), 0 \leq w^k \leq 1\), and \(\sum_{k=1}^{K} w^k = 1\) \(12\).

2.4 **Sample size computations**

To control decision error rates, methods for a priori sample size estimation are available for variables that follow a multivariate Bernoulli distribution and are eligible for large sample approximation by a (multivariate) normally distributed latent variable \(26, 25, 3\). When combined with a non-informative prior distribution, these procedures have shown to accurately control Type I rate \(\alpha\) and Type II error rate \(\beta\) in a Bayesian multivariate Bernoulli - Dirichlet-model on multivariate response data \(12\). Each of the presented decision
rules in Section 2.3 has an individual procedure to compute sample sizes, as discussed below. These equations provide insight in the required number of observations in absence of prior information and in the influence of the correlation on the sample size. They also allow for verification that correlated outcome variables might result in smaller sample sizes than uncorrelated outcome variables under some conditions detailed in [6] and [12]. 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 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 $\Psi_K$ with mean $\mathbf{0}$ and correlation matrix $\Sigma$:

$$1 - \beta = \Psi_K(c^1, \ldots, c^K),$$

(15)

In Equation 15, $c^k$ for outcome $k$ is defined by the decision rule of interest. Further, the off-diagonal elements of $\Sigma$ denote (estimated) pairwise correlations between outcome variables. For the Any rule,

$$c^k = z(1 - \alpha) - \frac{(\theta^k_1 - \theta^k_0)}{\sqrt{\theta^k_1(1 - \theta^k_1) + \theta^k_0(1 - \theta^k_0)}} n.$$  

(16)

For the All rule,

$$c^k = -z(1 - \alpha) + \frac{(\theta^k_1 - \theta^k_0)}{\sqrt{\theta^k_1(1 - \theta^k_1) + \theta^k_0(1 - \theta^k_0)}} n.$$  

(17)

In Equations 16 and 17, $n$ 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. Note that the Any rule (Equation 16) applies a Bonferroni correction for multiple testing.

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. 25:

1. Plug in estimates of $\theta^k_1$ in Equation 16 or 17

2. Plug in a starting value for $n$ in Equation 16 or 17 and calculate the power via Equation 15
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 [3, Chapter 4]. Plugging in estimates of weighted success probabilities per treatment $T$, $\theta^w_T$, results in:

$$n = \left[\frac{1}{\theta^w_1 (1 - \theta^w_1) + \theta^w_0 (1 - \theta^w_0)} \left[ \frac{z_{1-\alpha} + z_{1-\beta}}{\theta^w_1 - \theta^w_0} \right] \right]^2,$$

(18)

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

### 3 Capturing treatment heterogeneity

In the proposed framework, treatment heterogeneity can be captured by joint response probabilities that reflect conditional treatment effects and thus depend on the characteristics of a subpopulation of interest. We describe two ways to represent subpopulations: by fixed covariate values or by a prespecified interval of the covariate distribution(s). Both representations have their own applications. Specific values of covariates may be relevant when we wish to investigate treatment effects based on individual patients or on patient populations that can be accurately represented by a single number of the covariate (such as a mean or a level of a discrete variable). Intervals of covariate distributions may be more sensible when a number of consecutive covariate values are sufficiently exchangeable to estimate a marginal treatment effect among a population specified by this range. Although such intervals can be specified for discrete covariates as well, their use is particularly reasonable with continuous covariates, as intervals are inherently consistent with the idea of continuity.

We will discuss procedures for fixed values as well as intervals in more detail in the remainder of this subsection. In these discussions, we use a linear predictor $\psi^q_i(x)$ (cf. Equation 4) that distinguishes between treatments via a treatment indicator and allows for interaction between the treatment and a covariate. For such a model that includes a single population characteristic $x$, $x = (x, T, x \times T)$ and $\psi^q_i(x)$ is defined as:

$$\psi^q_i(x) = \beta^q_0 + \beta^q_1 T + \beta^q_2 x + \beta^q_3 x \times T.$$

(19)
3.1 Fixed values of covariate

For a patient population with fixed values of patient covariates, a posterior sample of joint response probabilities \( \phi_T(x) \) can be found by plugging in a vector of fixed covariate values \( x \) in linear predictor \( \psi_T(x) \). Subsequently applying the multinomial logistic link function in Equation 3 to each \( \psi_T(x) \) results in joint response probability \( \phi_T(x) \) for treatment \( T \). Applying these steps to the regression coefficients \( \beta(l) \) of each posterior draw \( (l) \) results in a sample of posterior joint response probabilities. The procedure is presented in Algorithm 1.

\[
\text{Algorithm 1 Transformation of posterior regression coefficients to posterior joint response probabilities based on fixed covariate values.}
\]

Define \( x = x_2, \ldots, x_P \) as a vector of covariate values of interest

Let \( \beta_Q = (0, \ldots, 0) \)

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 \( \psi_{(l)}^{(q)}(x) = \beta_0^{(l)} + \beta_1^{(l)}T + \beta_2^{(l)}x + \beta_3^{(l)}x \times T \)
5: Compute \( \phi_{(l)}^{(q)}(x) = \frac{\exp\left[\psi_{(l)}^{(q)}(x)\right]}{\sum_{r=1}^{Q-1} \exp\left[\psi_{(l)}^{(r)}(x)\right] + 1} \)
6: end for
7: end for
8: end for

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. We discuss two different computational approaches towards marginalization over the covariate distribution.

**Numerical marginalization** When the probability of a covariate interval is known or can be estimated with reasonable accuracy, treatment effects can be marginalized over an interval \( [x^L, x^H] \). Algorithm 2 presents the procedure for an interval of a single covariate.

The integral of the continuous covariate does not have a closed-form solution, but can be computed numerically for each draw from the posterior distribution of regression coefficients. Numerical marginalization can deal with populations that are fully defined by intervals as well as populations described by a mix of covariate values and intervals. Note however that defining multiple continuous covariates by their intervals requires multidimensional integration, which steeply increases the computational effort with every additional
integral. Computationally, the method does not depend on the size of a subsample in stratification of covariate data.

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

Define the interval of interest of covariate \( x \) as \([x_L, x_H]\)

Let \( \beta_Q = (0, \ldots, 0) \)

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:       if \( x \) is continuous then
5:         Define \( \psi_T^{(q)}(x) = \beta_{0}^{(q)} + \beta_{1}^{(q)} T + \beta_{2}^{(q)} x + \beta_{3}^{(q)} x \times T\)
6:         Compute \( \phi_T^{(q)}(x) = \int_{x_L}^{x_H} \left( \frac{\exp[\psi_T^{(q)}(x)]}{\sum_{r=1}^{Q-1} \exp[\psi_T^{(r)}(x)] + 1} \right) p(x) \, dx \)
7:       end if
8:       if \( x \) is discrete then
9:         for \( j \leftarrow x_L : x_H\) do
10:            Compute \( \psi_T^{(q)}(x_j) = \beta_{0}^{(q)} + \beta_{1}^{(q)} T + \beta_{2}^{(q)} x_j + \beta_{3}^{(q)} x_j \times T\)
11:           end for
12:       end if
13:       end for
14:   end for
15: end for

Empirical marginalization When probabilities of covariate intervals are unknown, a sample of covariate data can be used as input for marginalization. Empirical marginalization involves repeating procedure 3.1 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 \( \phi_T^{(l)}(x) \) for draw \((l)\). The procedure is presented in Algorithm 3. 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 \( x \) 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.
Algorithm 3 Transformation of posterior regression coefficients to posterior joint response probabilities based on empirical marginalization.

\[
\text{Let } \beta^Q = (0, \ldots, 0) \\
\text{1: } \text{for draw } (l) \leftarrow 1 : L \text{ do} \\
\text{2: } \text{for subject } i \leftarrow 1 : n \text{ do} \\
\text{3: } \text{for joint response } q \leftarrow 1 : Q \text{ do} \\
\text{4: } \text{Compute } \psi^q_i(l)(x_i) = \beta^q_1 x_i + \beta^q_2 x_i + \beta^q_3 x_i \times T_i \\
\text{5: } \text{Compute } \phi^q_i(l)(x_i) = \frac{\exp[\psi^q_i(l)(x_i)]}{\sum_{r=1}^{Q-1} \exp[\psi^r_i(l)(x_i)] + 1} \\
\text{6: } \text{for } T \leftarrow 0 : 1 \text{ do} \\
\text{7: } \text{Compute } \phi^q_T(x) = \frac{1}{\sum_{i=1}^n I(T_i = T)} \phi^q_i(l)(x_i) I(T_i = T) \\
\text{8: } \text{end for} \\
\text{9: } \text{end for} \\
\text{10: } \text{end for} \\
\text{11: } \text{end for}
\]

4 Numerical evaluation

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

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

4.1 Setup

4.1.1 Conditions

The performance of the framework was evaluated in a treatment comparison based on two outcome variables and one covariate. We varied the procedure to compute conditional treatment effects, the effect size, the (sub)population of interest, the procedure to compute the posterior distribution, and the decision rule. Each of these factors will be discussed in the following paragraphs.
**Procedure to estimate joint response probabilities** We used the three regression-based procedures from Section 3 to find the posterior samples of joint response probabilities for two populations of interest defined by:

1. **Fixed covariate values**

2. **An interval of the covariate distribution**, where we explore two computational methods:
   
   (a) **Numerical marginalization**
   
   (b) **Empirical marginalization**

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

3. **Unconditional multivariate Bernoulli - Dirichlet model**

   We used the unconditional multivariate Bernoulli model in [12]. This model relies on response data and can be used via stratification in the estimation of conditional treatment effects. Samples of treatment-specific joint response probabilities $\phi_T$ could be drawn directly from a posterior Dirichlet distribution with parameters $\alpha^0_T = \alpha^0 + \{\sum_{i=1}^{n} I(T_i = T)I(y_i = H_q)\}_{q=1}^{Q}$, where $\alpha^0$ is a vector of $Q$ prior hyperparameters.

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

1. **Conditions 1.1 & 1.2**: A homogeneous treatment effect, with average and conditional treatment differences of zero. This scenario aims to demonstrate the Type I error rate under a least favorable treatment difference for the Any and Compensatory rules in the trial as well as the subpopulation.

2. **Conditions 2.1 & 2.2**: A heterogeneous treatment effect, with an average treatment difference of zero and a conditional treatment effect larger than zero.

3. **Conditions 3.1 & 3.2**: A heterogeneous treatment treatment effect, with average and conditional treatment differences larger than zero. The conditional treatment difference is larger than the average treatment difference. The effect size is chosen to compare power of different methods, when the sample size should not lead to underpowerment for any of the approaches to the estimation of conditional treatment effects.
4. **Conditions 4.1 & 4.2**: A heterogeneous treatment effect on one of the outcomes with both average and conditional treatment differences larger than zero. The conditional treatment difference is smaller than the average treatment effect. The effect size is chosen such that the expected sample size after stratification of the study sample is smaller than the required sample for evaluation of the conditional treatment effect and aims to investigate the statistical power of regression-based methods when stratification leads to underpowered decisions. Further, this effect size reflects the least favorable treatment difference for a right-sided test of the All rule and should result in a Type I error rate equal to the chosen level of $\alpha$.

For each of these four effect sizes, we varied the measurement level of the covariate and created a model with a binary covariate and a model with a continuous covariate. This resulted in the eight data generating mechanisms presented in Table 1.

Table 1: Parameters of average treatment effects in the trial and conditional treatment effects in a subpopulation, by data-generating mechanism (DGM).

| DGM | Covariate | Average treatment effect | Conditional treatment effect |
|-----|-----------|--------------------------|-----------------------------|
|     |           | $(\delta_1, \delta_2)$  | $\delta(w)$ $p_{\nu_k, \nu_l}$ | $(\delta_1, \delta_2)$  | $\delta(w)$ $p_{\nu_k, \nu_l}$ |
| 1.1 | Discrete  | (0.000, 0.000) 0.000 -0.160 | (0.000, 0.000) 0.000 -0.200 |
| 1.2 | Continuous| (0.000, 0.000) 0.000 -0.163 | (0.000, 0.000) 0.000 -0.207 |
| 2.1 | Discrete  | (0.000, 0.000) 0.000 -0.154 | (0.250, 0.150) 0.200 -0.200 |
| 2.2 | Continuous| (0.000, 0.000) 0.000 -0.157 | (0.116, 0.069) 0.092 -0.206 |
| 3.1 | Discrete  | (0.150, 0.050) 0.100 -0.124 | (0.400, 0.300) 0.350 -0.200 |
| 3.2 | Continuous| (0.151, 0.050) 0.101 -0.131 | (0.276, 0.169) 0.223 -0.210 |
| 4.1 | Discrete  | (0.400, 0.000) 0.200 -0.194 | (0.200, 0.000) 0.100 -0.200 |
| 4.2 | Continuous| (0.401, -0.000) 0.200 -0.194 | (0.323, 0.000) 0.162 -0.205 |

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

1. **Trial population:**

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

2. **Subpopulation:**

   We assessed the conditional treatment effect among patients scoring low on the covariate. The low subpopulation of the binary covariate was described by a value of zero. Note that this subpopulation
could not be assigned a range, since subsetting a binary variable inherently results in a single value. As a consequence, analytical marginalization (procedure 3.2) reduces to the procedure for fixed covariate values (procedure 3.1). For the continuous covariate, we specified two different subpopulations. One subpopulation had a value of one standard deviation below the mean, while the other subpopulation was used in the marginalization approaches and defined by a range that entailed all values between the mean and one standard deviation below the mean.

**Decision rules and sample size** We applied the three decision rules in 4.1.2

1. Any rule
2. All rule
3. Compensatory rule with equal weights \((w = (0.50, 0.50))\)

We computed sample sizes per treatment group via the procedures in Section 2.4 for conditions with non-zero true average treatment effects. If the true average treatment difference was equal to zero, we used \(n = 2,000\). The required sample sizes are presented in Table 2 where we also included a) the required sample size for the conditional treatment effect in the subpopulation; and b) the sample size after stratification of the trial population. The sample size after stratification is the expected size in subpopulation analysis of a) response data in the reference approach; and b) covariate data in empirical marginalization.

| Table 2: Required sample sizes to evaluate the average treatment effect (ATE) and conditional treatment effect (CTE) and expected sample sizes of the subpopulation after stratification (Sub). |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| DGM   | Any ATE | Any CTE | Sub | All ATE | All CTE | Sub | Compensatory ATE | Compensatory CTE | Sub |
|-------|---------|---------|-----|---------|---------|-----|-----------------|-----------------|-----|
| 1.1   | -       | -       | 1000| -       | -       | 1000| -               | -               | 1000|
| 1.2   | -       | -       | 683 | -       | -       | 683 | -               | -               | 683 |
| 2.1   | -       | 45      | 1000| -       | 136     | 1000| -               | 30              | 1000|
| 2.2   | -       | 215     | 683 | -       | 658     | 683 | -               | 143             | 683 |
| 3.1   | 154     | 14      | 77  | 1234    | 34      | 617 | 134             | 9               | 67  |
| 3.2   | 152     | 36      | 52  | 1219    | 107     | 417 | 131             | 24              | 45  |
| 4.1   | 21      | 93      | 11  | -       | -       | 1000| 29              | 122             | 15  |
| 4.2   | 21      | 33      | 8   | -       | -       | 683 | 29              | 45              | 10  |

Note. Bold-faced subsamples are smaller than required for estimation of the CTE.
4.1.2 Procedure

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

$$\psi_i^q(x_i) = \beta_0^q + \beta_1^q T_i + \beta_1^q x_i + \beta_2^q x_i \times T_i.$$ (20)

To generate response data, we first applied the multinomial logistic link function (Equation 3) 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 $\phi_i(x_i)$.

Prior distribution For the multivariate logistic regression analysis, we used multivariate normally distributed prior with means $b^q = \mathbf{0}$ and precision matrix $B^0_q = \text{diag}(1e^{-2}, \ldots, 1e^{-2})$ for all regression coefficients. Prior covariances between regression coefficients were set at zero, implying that regression coefficients were independent a priori. For the reference approach, we used an improper prior with hyperparameters $\alpha^0 = 0$.

Gibbs sampling The regression coefficients in response categories $1, \ldots, (Q-1)$ were estimated via the Gibbs sampler in Appendix A. We ran three MCMC-chains via the Gibbs sampler introduced in Section 2 with $L = 10,000$ iterations plus 1,000 burnin iterations. Convergence diagnostics implied that there were no signals of non-convergence when the sample size was large enough. Multivariate Gelman-Rubin convergence diagnostics were below $< 1.10$ for most of the conditions. We noticed signs of non-convergence (Gelman-Rubin statistic 1.10 to 1.32) in a few datasets generated under mechanisms 4.1 and 4.2 with small sample sizes (i.e. belonging to the Any and Compensatory rules). We generated extra data to replace the datasets with questionable convergence.

Transformation and decision-making We applied the procedures in Sections 2.2 and 2.3 to arrive at a decision. For the analytical marginalization approach, we estimated the covariate distributions from each trial data as proportions (binary covariate) or sample means and standard deviations (continuous covariate). In empirical 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 $p_{\text{cut}} = 1 - \alpha = 0.95$ (Compensatory and All rules) and a for multiple tests corrected $p_{\text{cut}} =$
\[ 1 - \frac{a}{k} = 0.975 \text{ (Any rule)} \]

4.2 Results

**Bias**  Mean estimates of regression coefficients were asymptotically unbiased, implying that bias was negligible (< .01) in conditions with a sufficiently large sample. We observed some bias in conditions with smaller samples (DGM 3.1, 3.2, 4.1, and 4.2 under the Any and Compensatory decision rules). Of these conditions, bias was most prominent in data generating mechanisms 4.1 and 4.2 under the sample sizes used for the Any \( (n = 21) \) and Compensatory \( (n = 29) \) rules. The histograms of median regression coefficient for one of these conditions (DGM 4.2, Compensatory rule) are shown in Figure 2, revealing that some regression coefficients were skewed in the extreme direction.

The bias in regression coefficients is not necessarily problematic for our actual parameters of interest (success probabilities and differences between them), as transfer to these transformed quantities was not inherent. Even when regression coefficients were slightly biased (DGMs 3.1 and 3.2 under sample sizes of the Any and Compensatory rules), success probabilities and treatment differences could be estimated without bias (absolute bias < 0.025), similar to the conditions without biased regression coefficients. In these cases, the multinomial logistic transformation needed to obtain joint responses (Equation 3) appeared to normalize the skewed posterior samples of regression coefficients. More severe bias in conditions with smaller sample sizes was not fully corrected in the transformation steps: Two of the three regression-based methods resulted in biased average treatment differences in conditions with a small sample, as shown in Table 3. Treatment effect estimation based on fixed values and analytical marginalization under DGMs 4.1 and 4.2 resulted in treatment differences with absolute biases up to 0.077 for the Any and Compensatory rules. Bias appeared slightly more severe when the covariate was discrete, compared to a continuous covariate. The reference and empirical marginalization approaches could estimate parameters without bias, regardless of sample size.

**Decision error rates**  Probabilities to conclude superiority of average treatment effects are presented in Table 4. Decisions resulted in appropriate Type I error rates around .05 for each of the posterior distribution types under a least favorable scenario of no effect (i.e. DGM 1.1, 1.2, 2.1, 2.2 of Any and Compensatory rules, and 4.1 and 4.2 of the All rule) and the proportions of correct superiority conclusions (i.e. power) were close to the targeted .80. In general, regression-based methods performed comparable to the reference approach. In conditions with discrete covariates, the populations of the reference approach and fixed value regression were identical. When covariates were continuous, conditional treatment effects were estimated with the
Table 3: Comparison of bias in average and conditional treatment differences by sample size.

|                | Average Treatment Effect | Conditional Treatment Effect |
|----------------|--------------------------|-----------------------------|
|                | Method                   | Any            | All            | Comp           | Any            | All            | Comp           |
| Discrete covariate | δ(x)                      | δ(w, x)     | δ(x)                      | δ(w, x)     | δ(x)                      | δ(w, x)     |
| Reference      | (-0.004, -0.001)         | -0.002       | ( 0.000, 0.000)         | 0.000       | ( 0.001, 0.000)         | 0.000       |
| Empirical      | (-0.009, -0.004)         | -0.007       | ( 0.000, 0.000)         | 0.000       | (-0.002, -0.002)        | -0.002      |
| Analytical     | ( 0.072, -0.021)         | 0.025        | ( 0.016, 0.000)         | 0.008       | ( 0.066, -0.010)        | 0.028       |
| Value          | ( 0.077, -0.026)         | 0.025        | ( 0.001, -0.000)        | 0.000       | ( 0.068, -0.014)        | 0.027       |
| Conditional Treatment Effect | δ(x)                      | δ(w, x)     | δ(x)                      | δ(w, x)     | δ(x)                      | δ(w, x)     |
| Reference      | (-0.002, -0.008)         | -0.005       | (-0.001, -0.000)        | -0.000      | (-0.000, -0.003)        | -0.001      |
| Empirical      | -                        | -            | -                        | -           | -                        | -           |
| Analytical     | -                        | -            | -                        | -           | -                        | -           |
| Value          | ( 0.011, -0.002)         | 0.004        | (-0.001, -0.000)        | -0.000      | ( 0.010, 0.003)         | 0.007       |
| Continuous covariate | δ(x)                      | δ(w, x)     | δ(x)                      | δ(w, x)     | δ(x)                      | δ(w, x)     |
| Reference      | (-0.005, -0.004)         | -0.004       | (-0.000, -0.000)        | -0.000      | (-0.001, -0.003)        | -0.002      |
| Empirical      | (-0.014, -0.010)         | -0.012       | (-0.000, -0.000)        | -0.000      | (-0.006, -0.007)        | -0.007      |
| Analytical     | (-0.019, -0.011)         | -0.015       | (-0.000, -0.000)        | -0.000      | (-0.010, -0.008)        | -0.009      |
| Value          | ( 0.042, -0.026)         | 0.008        | ( 0.001, -0.000)        | 0.000       | ( 0.035, -0.018)        | 0.008       |

Comp = Compensatory rule
Figure 2: Histograms of median regression coefficients fitted for application of the Compensatory rule under DGM 4.2.
population specified by an interval, similar to empirical and analytical marginalization. Both comparisons showed similar decision error rates. Note that the power of the Compensatory rule in scenario’s 4.1 and 4.2 was slightly above .80 in regression-based methods, suggesting that the method was less robust to such small samples compared to the reference approach.

The results of conditional treatment effects in the subpopulations are presented in Table 5. Similar to the trial population, Type I error rates were around the targeted .05 under the least favorable scenarios of no effect (DGM 1.1, 1.2 for Any and Compensatory rules) for all estimation methods. With respect to power, empirical and analytical marginalization performed similar to each other, and higher than the reference approach. Discrete covariates allow for comparison of the fixed-values approach to the reference approach and show that the fixed values approach is less powerful than the reference approach.

5 Application

We applied the proposed method to a subset of data from the International Stroke Trial [10], which compared several short-term (14 days) and long-term effects (six months) of Aspirin, Heparin, Aspirin + Heparin and no treatment among \( n = 19,435 \) subjects who recently had a stroke. We selected participants who were alive after six months and were treated with either Aspirin + medium / high-dose Heparin or Aspirin only. We compared 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. The subset of data we fitted the model on contained \( n = 5,657 \) participants, of which \( n_{HA} = 1,859 \) were in the Heparin + Aspirin group (treatment = 1) and \( n_A = 3,798 \) subjects were in the Aspirin group (treatment = 0). The average blood pressure at randomisation was 160.04 mmHg (SD = 27.16) and similar in both treatment groups.

5.1 Method

We fitted a multivariate logistic model with the following linear predictor:

\[
\psi^q_i(Bp_i) = \beta_0^q + \beta_1^q T_i + \beta_2^q Bp_i + \beta_3^q Bp_i \times T_i.
\] (21)

We sought to examine the weighted and multivariate treatment differences between Aspirin + Heparin vs. Aspirin on dependency after six months (\( \delta_{dep} \)) and recurring stroke within 14 days (\( \delta_{strk} \)). We applied the
Table 4: Proportions of superiority decisions (p) and their standard errors (SE) for ATEs by data-generating mechanism (DGM), estimation method, and decision rule.

| DGM | Reference | Empirical | Analytical | Value |
|-----|-----------|-----------|------------|-------|
|     | p        | SE        | p          | SE    | p    | SE    | p   | SE    |
| 1.1 | 0.050    | (0.007)   | 0.058      | (0.007) | 0.054 | (0.007) | 0.054 | (0.007) |
| 1.2 | 0.044    | (0.006)   | 0.053      | (0.007) | 0.044 | (0.006) | 0.043 | (0.006) |
| 2.1 | 0.053    | (0.007)   | 0.055      | (0.007) | 0.052 | (0.007) | 0.052 | (0.007) |
| 2.2 | 0.044    | (0.006)   | 0.049      | (0.007) | 0.049 | (0.007) | 0.045 | (0.007) |
| 3.1 | 0.797    | (0.013)   | 0.817      | (0.012) | 0.824 | (0.012) | 0.808 | (0.012) |
| 3.2 | 0.786    | (0.013)   | 0.816      | (0.012) | 0.822 | (0.012) | 0.805 | (0.013) |
| 4.1 | 0.770    | (0.013)   | 0.815      | (0.012) | 0.842 | (0.012) | 0.842 | (0.012) |
| 4.2 | 0.787    | (0.013)   | 0.836      | (0.012) | 0.818 | (0.012) | 0.813 | (0.012) |

Rule = Any

| DGM | Reference | Empirical | Analytical | Value |
|-----|-----------|-----------|------------|-------|
|     | p        | SE        | p          | SE    | p    | SE    | p   | SE    |
| 1.1 | 0.001    | (0.001)   | 0.002      | (0.001) | 0.000 | (0.000) | 0.000 | (0.000) |
| 1.2 | 0.000    | (0.000)   | 0.000      | (0.000) | 0.001 | (0.001) | 0.000 | (0.000) |
| 2.1 | 0.002    | (0.001)   | 0.002      | (0.001) | 0.003 | (0.002) | 0.003 | (0.002) |
| 2.2 | 0.003    | (0.002)   | 0.004      | (0.002) | 0.004 | (0.002) | 0.002 | (0.001) |
| 3.1 | 0.823    | (0.012)   | 0.835      | (0.012) | 0.831 | (0.012) | 0.822 | (0.012) |
| 3.2 | 0.788    | (0.013)   | 0.799      | (0.013) | 0.802 | (0.013) | 0.813 | (0.012) |
| 4.1 | 0.048    | (0.007)   | 0.046      | (0.007) | 0.048 | (0.007) | 0.049 | (0.007) |
| 4.2 | 0.039    | (0.006)   | 0.040      | (0.006) | 0.040 | (0.006) | 0.041 | (0.006) |

Rule = All

| DGM | Reference | Empirical | Analytical | Value |
|-----|-----------|-----------|------------|-------|
|     | p        | SE        | p          | SE    | p    | SE    | p   | SE    |
| 1.1 | 0.052    | (0.007)   | 0.056      | (0.007) | 0.058 | (0.007) | 0.058 | (0.007) |
| 1.2 | 0.045    | (0.007)   | 0.052      | (0.007) | 0.044 | (0.006) | 0.045 | (0.007) |
| 2.1 | 0.063    | (0.008)   | 0.071      | (0.008) | 0.056 | (0.007) | 0.055 | (0.007) |
| 2.2 | 0.053    | (0.007)   | 0.065      | (0.008) | 0.067 | (0.008) | 0.052 | (0.007) |
| 3.1 | 0.814    | (0.012)   | 0.852      | (0.011) | 0.850 | (0.011) | 0.818 | (0.012) |
| 3.2 | 0.790    | (0.013)   | 0.831      | (0.012) | 0.831 | (0.012) | 0.835 | (0.012) |
| 4.1 | 0.819    | (0.012)   | 0.842      | (0.012) | 0.865 | (0.011) | 0.865 | (0.011) |
| 4.2 | 0.816    | (0.012)   | 0.837      | (0.012) | 0.823 | (0.012) | 0.824 | (0.012) |

Note. Bold-faced proportions represent correct rejections (i.e. power).
Table 5: Proportions of superiority decisions for CTEs by data-generating mechanism (DGM), estimation method, and decision rule.

| Rule = Any | Reference | Empirical | Analytical | Value |
|------------|-----------|-----------|------------|-------|
| DGM        | p  | SE    | p  | SE    | p  | SE    | p  | SE    |
| 1.1        | 0.059 | (0.007) | 0.064 | (0.008) | 0.055 | (0.007) | 0.055 | (0.007) |
| 1.2        | 0.048 | (0.007) | 0.060 | (0.008) | 0.055 | (0.007) | 0.055 | (0.007) |
| 2.1        | 1.000 | (0.000) | 1.000 | (0.000) | 1.000 | (0.000) | 1.000 | (0.000) |
| 2.2        | 0.919 | (0.009) | 0.998 | (0.001) | 0.999 | (0.001) | 1.000 | (0.000) |
| 3.1        | 1.000 | (0.000) | 1.000 | (0.000) | 1.000 | (0.000) | 1.000 | (0.000) |
| 3.2        | 0.233 | (0.013) | 0.542 | (0.016) | 0.559 | (0.016) | 0.175 | (0.012) |

| Rule = All | Reference | Empirical | Analytical | Value |
|------------|-----------|-----------|------------|-------|
| DGM        | p  | SE    | p  | SE    | p  | SE    | p  | SE    |
| 1.1        | 0.000 | (0.000) | 0.000 | (0.000) | 0.001 | (0.001) | 0.001 | (0.001) |
| 1.2        | 0.000 | (0.000) | 0.001 | (0.001) | 0.001 | (0.001) | 0.001 | (0.001) |
| 2.1        | 1.000 | (0.000) | 1.000 | (0.000) | 1.000 | (0.000) | 1.000 | (0.000) |
| 2.2        | 0.827 | (0.012) | 0.991 | (0.003) | 0.990 | (0.003) | 1.000 | (0.000) |
| 3.1        | 1.000 | (0.000) | 1.000 | (0.000) | 1.000 | (0.000) | 1.000 | (0.000) |
| 3.2        | 1.000 | (0.000) | 1.000 | (0.000) | 1.000 | (0.000) | 1.000 | (0.000) |
| 4.1        | 0.053 | (0.007) | 0.047 | (0.007) | 0.047 | (0.007) | 0.049 | (0.007) |
| 4.2        | 0.052 | (0.007) | 0.047 | (0.007) | 0.047 | (0.007) | 0.049 | (0.007) |

| Rule = Compensatory | Reference | Empirical | Analytical | Value |
|---------------------|-----------|-----------|------------|-------|
| DGM                 | p  | SE    | p  | SE    | p  | SE    | p  | SE    |
| 1.1                 | 0.060 | (0.008) | 0.057 | (0.007) | 0.051 | (0.007) | 0.051 | (0.007) |
| 1.2                 | 0.058 | (0.007) | 0.053 | (0.007) | 0.051 | (0.007) | 0.056 | (0.008) |
| 2.1                 | 1.000 | (0.000) | 1.000 | (0.000) | 1.000 | (0.000) | 1.000 | (0.000) |
| 2.2                 | 1.000 | (0.000) | 1.000 | (0.000) | 1.000 | (0.000) | 1.000 | (0.000) |
| 3.1                 | 1.000 | (0.000) | 1.000 | (0.000) | 1.000 | (0.000) | 1.000 | (0.000) |
| 3.2                 | 0.967 | (0.006) | 1.000 | (0.000) | 1.000 | (0.000) | 1.000 | (0.000) |
| 4.1                 | 0.253 | (0.014) | 0.273 | (0.014) | 0.273 | (0.014) | 0.273 | (0.014) |
| 4.2                 | 0.380 | (0.015) | 0.589 | (0.016) | 0.587 | (0.016) | 0.231 | (0.013) |

Note. Bold-faced proportions represent correct rejections (i.e. power).
three procedures from Subsection 3 (fixed values of covariates, empirical marginalization, and analytical marginalization) to assess the treatment difference 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 six different regions of blood pressure, where regions were defined as $0 \text{ SD} < Bp \leq 1 \text{ SD}$, $1 \text{ SD} < Bp \leq 2 \text{ SD}$, and $Bp > 2 \text{ SD}$ below and above the mean.

We specified a diffuse multivariate normally distributed prior with means $b^0 = 0$ and precision matrix $B^0 = \text{diag}(1e^{-2}, \ldots, 1e^{-2})$ for all regression coefficients, except the reference category ($\text{Strk}_{14} = 0, \text{Dep}_6 = 0$). Prior covariances between regression coefficients were set at zero, implying that regression coefficients were independent a priori. We ran three MCMC-chains via our proposed Gibbs sampler with 20,000 iterations plus 10,000 burnin iterations. Traceplots showed that chains mixed properly and the multivariate Gelman-Rubin convergence statistic had a value of 1.000, implying that there were no signals of non-convergence.

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

5.2 Results

Results are presented in Table 6 for different intervals and in Table 7 for fixed values of blood pressure. Among the trial population, the regression-based and reference approaches resulted in similar treatment difference estimates and posterior probabilities. Treatment differences were close to zero and each of the decision rules resulted in the conclusion that it does not matter whether Aspirin was administered alone or in combination with Heparin.
Table 6: Average and conditional treatment differences and their posterior probabilities of exceeding zero (pp) in the IST-data, by range of blood pressure.

| Method   | \( \delta(B_P) \)   | \( \delta(w, B_P) \)   | Compensatory |
|----------|----------------------|------------------------|--------------|
| Reference| (0.005, -0.014)      | (0.861, 0.152)         | 0.183        |
| Empirical| (0.004, -0.014)      | (0.825, 0.152)         | 0.178        |
| Numerical| (0.004, -0.015)      | (0.825, 0.144)         | 0.171        |

| Method   | \( \delta(B_P) \)   | \( \delta(w, B_P) \)   | Compensatory |
|----------|----------------------|------------------------|--------------|
| Reference| (-0.011, 0.101)      | (0.419, 0.776)         | 0.759        |
| Empirical| (0.021, 0.082)       | (0.929, 0.988)         | 0.993        |
| Numerical| (0.021, 0.084)       | (0.927, 0.989)         | 0.993        |

| Method   | \( \delta(B_P) \)   | \( \delta(w, B_P) \)   | Compensatory |
|----------|----------------------|------------------------|--------------|
| Reference| (0.003, 0.063)       | (0.571, 0.964)         | 0.964        |
| Empirical| (0.012, 0.040)       | (0.931, 0.959)         | 0.969        |
| Numerical| (0.012, 0.042)       | (0.932, 0.962)         | 0.972        |

| Method   | \( \delta(B_P) \)   | \( \delta(w, B_P) \)   | Compensatory |
|----------|----------------------|------------------------|--------------|
| Reference| (0.017, 0.002)       | (0.992, 0.526)         | 0.625        |
| Empirical| (0.005, 0.001)       | (0.880, 0.524)         | 0.572        |
| Numerical| (0.006, 0.004)       | (0.891, 0.604)         | 0.648        |

| Method   | \( \delta(B_P) \)   | \( \delta(w, B_P) \)   | Compensatory |
|----------|----------------------|------------------------|--------------|
| Reference| (0.003, -0.073)      | (0.610, 0.005)         | 0.006        |
| Empirical| (0.001, -0.037)      | (0.565, 0.008)         | 0.010        |
| Numerical| (0.001, -0.034)      | (0.597, 0.013)         | 0.015        |

| Method   | \( \delta(B_P) \)   | \( \delta(w, B_P) \)   | Compensatory |
|----------|----------------------|------------------------|--------------|
| Reference| (-0.014, -0.061)     | (0.122, 0.059)         | 0.046        |
| Empirical| (-0.002, -0.073)     | (0.351, 0.001)         | 0.001        |
| Numerical| (-0.002, -0.072)     | (0.352, 0.001)         | 0.001        |

| Method   | \( \delta(B_P) \)   | \( \delta(w, B_P) \)   | Compensatory |
|----------|----------------------|------------------------|--------------|
| Reference| (0.032, -0.013)      | (0.812, 0.434)         | 0.487        |
| Empirical| (-0.006, -0.118)     | (0.274, 0.001)         | 0.001        |
| Numerical| (-0.005, -0.112)     | (0.279, 0.001)         | 0.001        |

Note. \( > \) = superiority concluded, \( < \) = inferiority concluded
Table 7: Conditional treatment differences and their posterior probabilities of exceeding zero (pp) in the IST-data, by value of blood pressure.

| Value  | $\delta(Bp)$ | pp       | Any | All | $\delta(w, Bp)$ | pp       | Compensatory |
|--------|--------------|----------|-----|-----|----------------|----------|--------------|
| -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    | >            |

Note. $>$ = superiority concluded, $<$ = inferiority concluded.

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

A comparison between different estimation methods shows that regression-based methods performed highly similar. Their estimates of multivariate treatment difference $\delta$, weighted treatment difference $\delta(\mathbf{w})$, posterior probabilities, and superiority/inferiority decisions were comparable. Stratification of response data (i.e. the reference approach) resulted in relatively similar estimates and posterior probabilities in the center of the distribution of blood pressure (e.g. between $-1$ SD and $+1$ SD), but deviated from the regression-based approach in the tails. The amount of evidence was lower: Similar or more extreme treatment differences resulted in less extreme posterior probabilities, as shown in the Very Low, High, and Very High subpopulations (Table 6). Moreover, absolute treatment differences demonstrated a less stable relation between blood pressure and treatment differences, as shown in Figure 3. The different behavior in the tails of the covariate distribution might be explained by the combination of a smaller sample size after stratification. It should also be noted that the low number of events per joint response category left some cells without observations, having prior information only. If we would be certain that the model held properly over the entire range of blood pressure, regression-based methods appear more stable and more powerful in the tails of the covariate distribution.
6 Discussion

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

This application also highlighted the need to consider model assumptions in real data. A comparison of estimation methods showed diffuse results in the tails of the covariate distribution. Additional efforts may be undertaken to verify that the chosen generalized linear model fits the data well enough.

The model demonstrated its performance in settings with limited numbers of outcome variables and...
covariates. It is not clear how the model performs with more outcome and/or predictor variables. However, increasing numbers of variables lead to a steep increase in the number of parameters, which may pose computational challenges to the approach. The Gibbs sampling procedure may become unstable when the sample size is too small compared to the number of parameters. Moreover, numerical integration with multiple continuous variables requires repeatedly solving a multivariable integral, which can be a tedious process.

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. In line with our observations, small-sample bias in regression coefficients is a well-documented property of nonlinear regression methods in general [3, 16]. Although some bias in regression coefficients disappeared during transformation to joint response probabilities, success probabilities, and treatment differences, the mechanism is not yet fully understood. Hence, more light may be shed on circumstances for inheritance of distributional properties in the (non-linear) multinomial logistic transformation to obtain more elaborate insights in the minimum number of observations required for satisfactory model performance. Larger effect sizes (i.e. smaller sample sizes), complexity of the model (i.e. number of parameters), and events per variable are candidate factors to interact in their effects on model performance in small samples [11]. Therefore, optimum sample sizes in these regression-based decision approaches remain to be investigated more elaborately.

7 Conclusion

The presented Bayesian method aimed to capture treatment heterogeneity in multiple correlated binary outcome variables, via estimation of average and conditional treatment effects among the trial population and subpopulations respectively. The framework was built upon three major components: a multivariate logistic regression analysis, a subsequent transformation of regression coefficients to intuitively attractive multivariate success probabilities and differences between them, and a procedure to make decisions regarding treatment superiority or inferiority. When the sample is sufficiently large, treatment effects can be estimated unbiasedly and decisions regarding average and conditional treatment effects can be made with targeted error rates and a priori estimated sample sizes. The method is particularly useful in prediction of treatment
effects and decision-making within subpopulations, while taking advantage of the size of the entire study sample and while properly incorporating the uncertainty in a principled probabilistic manner using the full posterior distribution.

**Acknowledgments**

We thank The International Stroke Trial Collaborative Group for making the data from the International Stroke Trial publicly available.

**Author contributions**

XK performed the analyses and drafted the manuscript. JM and MK verified analytical methods, supported the drafting of the manuscript and supervised the project. All authors critically read and approved the manuscript.

**Financial disclosure**

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

**Conflict of interest**

The Authors declare that there is no conflict of interest.

**Availability of data and materials**

The International Stroke Trial data that support the findings of this study are available as electronic supplementary materials to the publication of Sandercock et al. [23] with the identifier(s) [http://doi.org/10.1186/1745-6215-12](http://doi.org/10.1186/1745-6215-12).
A copy of the retrieved dataset and R code used to generate results in Sections 4 and 5 can be found on GitHub [https://github.com/XynthiaKavelaars/Bayesian-multivariate-logistic-regression](https://github.com/XynthiaKavelaars/Bayesian-multivariate-logistic-regression).
References

[1] Ming-Hui Chen and Joseph G. Ibrahim. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, feb 2000. doi: 10.1214/ss/1009212673.

[2] Siddhartha Chib. Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, dec 1995. doi: 10.1080/01621459.1995.10476635.

[3] Shein-Chung Chow, Jun Shao, Hansheng Wang, and Yuliya Lokhnygina. *Sample Size Calculations in Clinical Research: Third edition*. Chapman and Hall/CRC, aug 2017. ISBN 9781315183084. doi: 10.1201/9781315183084.

[4] Bin Dai, Shilin Ding, Grace Wahba, et al. Multivariate bernoulli distribution. *Bernoulli*, 19(4):1465–1483, 2013.

[5] David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993. doi: 10.1093/biomet/80.1.27.

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

[7] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, dec 2008. doi: 10.1214/08-aaoas191.

[8] Jeffrey J. Goldberger and Alfred E. Buxton. Personalized medicine vs guideline-based medicine. *JAMA*, 309(24):2559, jun 2013. doi: 10.1001/jama.2013.6629.

[9] Margaret A. Hamburg and Francis S. Collins. The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304, jul 2010. doi: 10.1056/nejmp1006304.

[10] International Stroke Trial Collaborative Group. 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, may 1997. doi: 10.1016/s0140-6736(97)04011-7.

[11] Valentijn M. T. Jong, Marinus J. C. Eijkemans, Ben Calster, Dirk Timmerman, Karel G. M. Moons, Ewout W. Steyerberg, and Maarten Smeden. Sample size considerations and predictive performance of
multinomial logistic prediction models. Statistics in Medicine, 38(9):1601–1619, jan 2019. doi: 10.1002/sim.8063.

[12] Xynthia Kavelaars, Joris Mulder, and Maurits Kaptein. Decision-making with multiple correlated binary outcomes in clinical trials. Statistical Methods in Medical Research, 29(11):3265–3277, jul 2020. doi: 10.1177/0962280220922256.

[13] Henrick J. Malik and Bovas Abraham. Multivariate logistic distributions. The Annals of Statistics, 1(3):588–590, may 1973. doi: 10.1214/aos/1176342430.

[14] Maarten Marsman and Eric-Jan Wagenmakers. Three insights from a bayesian interpretation of the one-sided p value. Educational and Psychological Measurement, 77(3):529–539, oct 2016. doi: 10.1177/0013164416669201.

[15] Thomas A Murray, Peter F Thall, and Ying Yuan. Utility-based designs for randomized comparative trials with categorical outcomes. Statistics in medicine, 35(24):4285–4305, 2016.

[16] Szilard Nemes, Junmei Miao Jonasson, Anna Genell, and Gunnar Steineck. Bias in odds ratios by logistic regression modelling and sample size. BMC Medical Research Methodology, 9(1), jul 2009. doi: 10.1186/1471-2288-9-56.

[17] Peter C O’Brien. Procedures for comparing samples with multiple endpoints. Biometrics, pages 1079–1087, 1984.

[18] Sean M. O’Brien and David B. Dunson. Bayesian multivariate logistic regression. Biometrics, 60(3):739–746, aug 2004. doi: 10.1111/j.0006-341x.2004.00224.x.

[19] Stuart J. Pocock, Nancy L. Geller, and Anastasios A. Tsiatis. The analysis of multiple endpoints in clinical trials. Biometrics, 43(3):487, sep 1987. doi: 10.2307/2531989.

[20] Dale Poirier. Jeffreys, prior for logit models. Journal of Econometrics, 63(2):327–339, aug 1994. doi: 10.1016/0304-4076(93)01556-2.

[21] Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. Journal of the American statistical Association, 108(504):1339–1349, 2013.
[22] Peter E Rossi, Greg M Allenby, and Rob McCulloch. *Bayesian statistics and marketing*. John Wiley & Sons, 2005.

[23] Peter AG Sandercock, Maciej Niewada, and Anna Członkowska. The international stroke trial database. *Trials*, 12(1), apr 2011. doi: 10.1186/1745-6215-12-101.

[24] Wietske C. M. Schimmel, Eline Verhaak, Patrick E. J. Hanssens, Karin Gehring, and Margriet M. Sitskoorn. A randomised trial to compare cognitive outcome after gamma knife radiosurgery versus whole brain radiation therapy in patients with multiple brain metastases: research protocol CAR-study b. *BMC Cancer*, 18(1), feb 2018. doi: 10.1186/s12885-018-4106-2.

[25] Takashi Sozu, Tomoyuki Sugimoto, and Toshimitsu Hamasaki. Sample size determination in clinical trials with multiple co-primary binary endpoints. *Statistics in medicine*, 29:2169–2179, September 2010. ISSN 1097-0258. doi: 10.1002/sim.3972.

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

[27] Sheena G Sullivan and Sander Greenland. Bayesian regression in SAS software. *International Journal of Epidemiology*, 42(1):308–317, dec 2012. doi: 10.1093/ije/dys213.

[28] Don van Ravenzwaaij, Rei Monden, Jorge N. Tendeiro, and John P. A. Ioannidis. Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC Medical Research Methodology*, 19(1), mar 2019. doi: 10.1186/s12874-019-0699-7.

[29] Molin Wang, Donna Spiegelman, Aya Kuchiba, Paul Lochhead, Sehee Kim, Andrew T. Chan, Elizabeth M. Poole, Rulla Tamimi, Shelley S. Tworoger, Edward Giovannucci, Bernard Rosner, and Shuji Ogino. Statistical methods for studying disease subtype heterogeneity. *Statistics in Medicine*, 35(5):782–800, dec 2015. doi: 10.1002/sim.6793.
A Details of posterior computation

Starting from the likelihood of individual $K$-variate response $y_i$ (Equation 3), the likelihood of $n$ $K$-variate responses follows from taking the product over $n$ individual joint response probabilities in $Q$ response categories:

$$l(y_i | \beta, x) = \prod_{i=1}^{n} \frac{\exp [\psi_q^i(x)]}{\sum_{r=1}^{Q-1} \exp [\psi^r_i(x)] + 1} \cdot I(y_i = q)$$

Following Polson et al. [21], we introduce the Pólya-gamma variable by rewriting the multivariate likelihood in Equation 22 as a series of binomial likelihoods. The likelihood of $y$ conditional on the parameters of the $q^{th}$ response category, $\beta^q$, then equals:

$$l(y | \beta^q, \beta^{-q}) = \prod_{i=1}^{n} \left( \frac{\exp [\eta^q_i(x)]}{\exp [\eta^q_i(x)] + 1} \right) I(y_i = q) \frac{1}{\exp [\eta^q_i(x)] + 1}^{1 - I(y_i = q)}$$

where $-q$ refers to all rows in $H$ not having index $q$ and $\eta^q_i(x) = \psi^q_i(x) - \ln \left( \sum_{m \neq q} \exp [\psi^m_i(x)] \right)$.

The Polya-Gamma transformation to a Gaussian distribution relies on the following equality [21]:

$$\frac{\exp [\eta^q_i(x)]}{\exp [\eta^q_i(x)] + 1} = 2 \exp \left[ (y_i - 1) \eta^q_i(x) \right] \int_{0}^{\infty} \exp \left[ -\omega^q_i(x)^2 / 2 \right] p(\omega^q_i) d\omega^q_i$$

where $\omega^q_i$ has a Polya-Gamma distribution, i.e. $p(\omega^q_i) \sim PG(1, \psi^q_i(x))$.

If we use the equality in Equation 24, the binomial likelihood in Equation 23 can be transformed to a multivariate Gaussian likelihood by including an auxiliary Pólya-Gamma variable $\omega^q_i$ [21]:
\begin{equation}
\begin{aligned}
l(y|q, \beta_q) &= \prod_{i=1}^{n} \frac{\exp[\eta_q^i(x)]}{\exp[\eta_q^i(x)] + 1} \\
&= \prod_{i=1}^{n} 2\exp \left[ (y_i - \frac{1}{2})\eta_q^i(x) \right] \int_{0}^{\infty} \exp \left[ -\frac{(\eta_q^i(x))^2}{2} \right] p(\omega_q^i) d\omega_q^i \\
&= \prod_{i=1}^{n} \exp \left[ \kappa_q^i \omega_q^i (\eta_q^i(x) - \frac{1}{2}(\eta_q^i(x))^2)\omega_q^i \right] P(\omega_q^i | 1, 0) \\
&\propto \exp \left[ \frac{1}{2}(2\kappa_q^i \omega_q^i \eta_q^i(x) - \omega_q^i (\eta_q^i(x))^2) \right] \\
&\propto \exp \left[ -\frac{1}{2}(\kappa_q^i - \eta_q^i)^T (x) \Omega_q^i (\kappa_q^i - \eta_q^i(x)) \right] \\
&= \exp \left[ -\frac{1}{2}(\kappa_q^i - X\beta_q + \ln(\sum_{m \neq q} \exp(X\beta^m)))^T \Omega_q^i (\kappa_q^i - X\beta_q + \ln(\sum_{m \neq q} \exp(X\beta^m))) \right],
\end{aligned}
\end{equation}

where \( \kappa_q^i = \frac{l(y_i = H_q^i)}{\omega_q^i} - \frac{1}{2}, \kappa_q = (\kappa_1^q, \ldots, \kappa_n^q), \omega_q = (\omega_1^q, \ldots, \omega_n^q), \) and \( \Omega_q = \text{diag}(\omega_q). \)

### A.0.1 Prior distribution

The Gaussian likelihood in Equation 25 is conditionally conjugate with a normal prior distribution on regression coefficients \( \beta_q^i \):

\begin{equation}
\beta_q^i \sim N(b_q^i, B_q^0)
\end{equation}

where \( b_q^i \) is the vector of prior means of regression coefficient vector \( \beta_q^i \) and \( B_q^0 \) is a \( P \times P \) symmetric square matrix reflecting the prior precision of regression coefficients \( \beta_q^i \). 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. 27, 1]. We discuss the specification of informative prior means \( b_q^i \) in terms of joint response probabilities \( \phi \) in Appendix B. In absence of prior information, a non- or weakly informative prior distribution can be used, such as specified in 27, 20.
A.0.2 Posterior distribution

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

\[
p(\beta | y) \propto p(y | \beta, x)p(\beta),
\]

(27)

The combination of a Polya-Gamma transformed Gaussian like likelihood (Equation 25) and a normal prior distribution (Equation 26) respectively is proportional to a normally distributed posterior distribution, conditionally on Polya-Gamma variables in \( \omega^q \) [21]:

\[
p(\beta^q | Y, \Omega^q) \propto p(y | \beta^q, \omega^q)p(\beta^q)
\]

\[
\propto \exp \left[ -\frac{1}{2} \left( \kappa^q - X\beta^q + \ln \left( \sum_{m\neq q} \exp \left[ X\beta^m \right] \right) \right)^T \Omega^q (\kappa^q - X\beta^q + \ln \left( \sum_{m\neq q} \exp \left[ X\beta^m \right] \right)) \right] \times \\
\exp \left[ -\frac{1}{2} (\beta^q - b^q)^T (B^q)^{-1} (\beta^q - b^q) \right] \\
\propto N \left( V^q (X^T \Omega^q (\kappa^q + \ln \left( \sum_{m \neq q} \exp \left[ X\beta^m \right] \right)) + (B^q)^{-1} b^q), V^q \right)
\]

where \( V^q = (X^T \Omega^q X + (B^q)^{-1})^{-1} \). Similarly, subject-specific variable \( \omega^q_i \) follows a Polya-Gamma distribution that depends on regression coefficients \( \beta^q \) via linear predictor \( \psi^q_i \).

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

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

\[
\beta^q | \omega^q \sim N(m^q, V^q)
\]

(29)

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

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

\[
c = \left\{ \ln \left( \sum_{m \neq q} \exp \left[ \psi^m_i (x_i) \right] \right) \right\}_{i=1}^n.
\]
2. Sample $\omega^n|\beta^n$ as a vector of $n$ draws $\omega^q_i|\beta^n$ from a Pólya-Gamma distribution:

$$
\omega^q_i|\beta^n \sim PG(1, \psi^q_i - \ln \sum_{m \neq q} \exp[\psi^m_i(x)]).
$$

(30)

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

38
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 \( \psi \) with one covariate and an interaction between the treatment and the covariate:

\[
\psi^q_T = \beta_0^q + \beta_1^q T + \beta_2^q x + \beta_3^q x \times T
\]  

(31)

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

\[
\phi_{T}^{11}(x) = \rho_T(x) \sqrt{\theta_T^1(x) [1 - \theta_T^1(x)]} \theta_T^2(x) [1 - \theta_T^2(x)] + \theta_T^1(x) \theta_T^2(x)
\]

(32)

\[
\phi_{T}^{01}(x) = \theta_T^1(x) - \phi_{T}^{11}(x)
\]

\[
\phi_{T}^{10}(x) = \theta_T^2(x) - \phi_{T}^{11}(x)
\]

\[
\phi_{T}^{00}(x) = 1 - \theta_T^1(x) - \theta_T^2(x) + \phi_{T}^{11}(x)
\]

For each response category \( q \), joint responses \( \phi_T^q \) can be transformed to linear predictor \( \psi_T^q \) using the multinomial logistic link function in Equation 3.

Solving these linear predictors for \( \beta^q \) results in the following definitions of the elements in \( \beta^q \):

\[
\beta_0^q = \frac{x_H \psi_0^q(x_L) - x_L \psi_0^q(x_H)}{x_H - x_L}
\]

(33)

\[
\beta_1^q = \frac{x_H [\psi_1^q(x_L) - \psi_1^q(x_L)] + x_L [\psi_1^q(x_H) - \psi_1^q(x_H)]}{x_H - x_L}
\]

\[
\beta_2^q = \frac{\psi_2^q(x_H) - \psi_2^q(x_L)}{x_H - x_L}
\]

\[
\beta_3^q = \frac{\psi_3^q(x_H) - \psi_3^q(x_L) - \psi_3^q(x_L) + \psi_3^q(x_L)}{x_H - x_L}
\]

39
Table 8: Example of means of the prior distribution of regression coefficients

|       | $q = 1$ | $q = 2$ | $q = 3$ | $q = 4$ |
|-------|---------|---------|---------|---------|
| $\beta_0^q$ | -0.000 | 0.766   | 0.766   | 0.000   |
| $\beta_1^q$ | 0.000   | 0.000   | 0.000   | 0.000   |
| $\beta_2^q$ | 1.902   | 0.781   | 1.121   | 0.000   |
| $\beta_3^q$ | -3.804  | -1.562  | -2.241  | 0.000   |

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

$\theta_L^1 = (0.60, 0.70), \rho_L^1 = -0.30$

$\theta_H^1 = (0.40, 0.30), \rho_H^1 = -0.30$

$\theta_L^0 = (0.40, 0.30), \rho_L^0 = -0.30$

$\theta_H^0 = (0.60, 0.70), \rho_H^0 = -0.30$,

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