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

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

X.M. Kavelaars*[1,3], J. Mulder[1], and M.C. Kaptein[2]

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

[2]Jheronimus Academy of Data Science, 's Hertogenbosch, The Netherlands

[3]Department of Theory, Methodology, and Statistics, Open University of the Netherlands, Heerlen, The Netherlands

*E-mail: x.m.kavelaars@tilburguniversity.edu

Abstract

Background: In medical, social, and behavioral research we often encounter datasets with a multilevel structure and multiple correlated dependent variables. These data are frequently collected from a study population that distinguishes several subpopulations with different (i.e., heterogeneous) effects of an intervention. Despite the frequent occurrence of such data, methods to analyze them are less common and researchers often resort to either ignoring the multilevel and/or heterogeneous structure, analyzing only a single dependent variable, or a combination of these. These analysis strategies are suboptimal: Ignoring multilevel structures inflates Type I error rates, while neglecting the multivariate or heterogeneous structure masks detailed insights.

Methods: To analyze such data comprehensively, the current paper presents a novel Bayesian multilevel multivariate logistic regression model. The clustered structure of multilevel data is taken into account, such that posterior inferences can be made with accurate error rates. Further, the model shares information between different subpopulations in the estimation of average and conditional average multivariate treatment effects. To facilitate interpretation, multivariate logistic regression parameters are transformed to posterior success probabilities and differences between them.

Results: A numerical evaluation compared our framework to less comprehensive alternatives and highlighted the need to model the multilevel structure: Treatment comparisons based on the multilevel model had targeted Type I error rates, while single-level alternatives resulted in inflated Type I errors. Further, the multilevel model was more powerful than a single-level model when the number of clusters was higher. A re-analysis of the Third International Stroke Trial data illustrated how incorporating a multilevel structure, assessing treatment heterogeneity, and combining dependent variables contributed to an in-depth understanding of treatment effects. Further, we demonstrated how Bayes factors can aid in the selection of a suitable model.

Conclusion: The method is useful in prediction of treatment effects and decision-making within subpopulations from multiple clusters, 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.

1 Background

In medical, social, and behavioral research we often encounter datasets with a multilevel structure and multiple correlated dependent variables. An example of such a study is the Cognition and Radiation Study B [68, 67] that investigated whether local brain radiation (stereotactic radiosurgery) preserves cognitive functioning and quality of life better than whole brain radiation in cancer patients with multiple brain metastases. Patients were recruited from multiple hospitals and the treatment was executed in two treatment centers, giving the data a multilevel structure. Many other examples of such datasets can be found in a paper by Biswas and colleagues [3], who presented a nonexhaustive overview of hundreds of Bayesian trial protocols executed in a specialized center for cancer treatment. The authors noted that a) almost half of the reviewed studies were multicenter trials; and b) many studies were designed to assess effectiveness and side effects simultaneously, thus including at least two dependent variables.

Often, these multilevel, multivariate data are collected from a study population that consists of several subpopulations with potentially distinctive (i.e., heterogeneous) effects of an intervention. Examples of such studies are the two International Stroke Trials [International Stroke Trial (IST) and Third International Stroke Trial (IST-3); 24, 66, 77, 65], which investigated the effects of antiplatelet and antithrombotic treatments on various (neuro)psychological, functional and psychosocial dependent variables respectively. Both trials covered multiple treatment centers from multiple countries and included a variety of patient characteristics that could potentially predict treatment effects. We discuss the IST-3 in more depth as it serves as a running example throughout the paper. The IST-3 investigated the effects of an intravenous thrombolysis treatment on shortterm (e.g., recurrent stroke, functional deficits) and long-term (e.g., dependency, depression, pain) indicators of health status among patients who suffered from an acute ischaemic stroke. The IST-3 data revealed considerable variation in characteristics of patients and disease - such as subtype or severity of stroke, blood pressure, and age - that can be predictive of treatment effects and call for exploration of treatment heterogeneity to gain insight into subpopulation-specific effects [35].

All of the abovementioned trials made treatment comparisons in the context of Randomized Controlled Trials (RCTs): Randomized experiments in which an experimental or a control treatment is randomly assigned and administered to a random sample of patients. RCTs often aim to evaluate whether the experimental treatment is superior or (non-)inferior to the control condition and ultimately guide clinicians in evidence-based assignment of treatments and interventions [12].

Whereas RCTs are considered a golden standard for treatment comparison, their implementation is challenged by a growing demand for personalized treatment [9, 51, 20, 71]. That is, clinical practice relies more and more on the idea that different patients react differently to treatments. Treatment prescription is increas-

ingly guided by a trade-off between patient-specific risks and benefits, making the research context for these decisions multivariate and heterogeneous [49]. While demanding more complex methodology, personalization of treatments can impede the collection of sufficient data for rigorous treatment evaluation. Development of more targeted treatments limits eligibility for participation in trials, thereby making the recruitment of subjects more difficult. As a solution, trials more often span multiple treatment centers or countries. This adds another layer of complexity to the research context: clustered data that require multilevel analysis. To meet the methodological demands of these increasingly complex research problems, RCTs ideally provide a) a broad understanding of the treatment's effects on multiple dependent variables; and b) insights potential dependencies of treatment effects on characteristics of patients; and c) an accurate handling of clustered data structures. In practice, such comprehensive methods are less common, and often researchers resort to either ignoring the multilevel and/or heterogeneous structure, analyzing only a single dependent variable, or a combination of these. Below, we discuss how the abovementioned three aspects can be implemented in Randomized Controlled Trial methodology to support research in personalized treatment.

First, many RCTs evaluate more than one dependent variable, which are analysed separately in multiple univariate analyses [11]. As an example, the investigators of the IST-3 were primarily interested in living independently six months after stroke and secondarily in several other dependent variables, such as recurrent events, adverse reactions to the treatment, and mental health indicators. Analyzing dependent variables independently provides useful insights in treatment effects on each of these dependent variables individually, but discards available information about the relation between them. When the effects on individual dependent variables are complemented with information about their co-occurrences via multivariate analysis, a more detailed picture of treatment effects emerges. Multivariate analysis models relationships between dependent variables and can a) be helpful to detect outcome patterns that would be ignored when dependent variables are considered in isolation; and b) improve the accuracy of sample size computations and error rates in statistical decision-making [11, 75, 73, 33].

Second, incorporating patient and/or disease characteristics in treatment comparison can result in a considerable improvement of the practical value of RCTs. The IST-3 used a sample of diverse patients with different personal and disease characteristics. This variation contains valuable information regarding differences in treatment effects. For example, knowing whether patients with different weights or blood pressures have different chances of a recurrent stroke or independent living has the potential to inform treatment recommendations. When treatments have distinct effects on patients with different characteristics, treatment effects are considered heterogeneous among (sub)populations of patients. In this case, average treatment effects (ATEs) give a global idea of treatment results among the trial population, but have limited value in targeting treatments to specific patients with their individual (disease) characteristics [22, 43,

69]. Conditional average treatment effects (CATEs) among specific patient groups provide insight in the variation of treatment effects among the population and help to distinguish patients who ultimately benefit from the treatment from those who do not or may even experience adverse treatment effects. Unfortunately, subgroup-specific treatment comparisons are insufficiently implemented as part of standard trial methodology yet [76]. If subgroups are targeted at all, their effects are often analyzed independently via stratified (or subgroup) analysis. Such a subgroup analysis disregards information from related subgroups and suffers from suboptimal power due to subsetting. Modelling heterogeneity is a more powerful alternative that directly uses the relation between subgroups and allows subgroups to borrow strength from each other [29, 30, 32].

Third, multilevel data are characterized by observational units that are grouped in clusters. For example, the IST-3 spans multiple treatment centers and multiple countries. Reasons to use multilevel analysis can be both substantive and statistical. From a substantive perspective, multilevel analysis can be useful to explain differences between clusters, while using the information from the entire sample [82, 17]. Different trials may - for example - have overlapping but non-identical target populations that can be distinguished by covariate information and may contribute to the understanding of treatment effects. Statistically, differences between clusters should be taken into account for the sake of validity, even if these differences are not of direct interest [23, 61, 40]. Clustered data require specific analysis methods that are flexible enough to treat observations from different clusters as more similar to each other than to observations from other clusters. If observations within clusters are indeed more similar, the clustered structure is reflected in variance partitioning, where the within-cluster and the between-cluster variances are modelled separately. This induces a dependence between the observations within clusters when marginalizing over the cluster-specific effects. When clustered observations are treated as independent observations on the other hand, variance originating from differences between clusters is then erroneously attributed to differences between a manifold of observational units and the unique amount of information is overestimated. As a result, standard errors are overestimated, Type I error rates are inflated, and validity of statistical inference is compromised. The larger the variance between clusters relative to the variance between observational units within clusters, the larger the effect on standard errors. Properly modelling the multilevel structure of clustered data and allowing the parameters to vary over clusters is therefore crucial for accurate statistical decision-making [23, 61].

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

This framework consists of three coherent elements [32]:

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

The first element, the modelling procedure, assumes multivariate Bernoulli distributed dependent variables and assigns them a multinomial parametrization. A multinomial parametrization is helpful for two reasons, since it a) allows statisticians to draw and build upon existing, established multinomial techniques with tractable (conditional) posterior distributions; and b) has the flexibility to model correlations between dependent variables on the subpopulation level, which contributes to the accuracy of inference under treatment heterogeneity [8, 33, 32]. Several other multivariate modelling procedures, such as the multivariate probit model [5] or multivariate logistic regression models [37, 54], have a more restrictive correlation structure and are therefore theoretically less suitable to detect treatment heterogeneity with adequate error control. Moreover, the multivariate logistic regression model by Malik and Abraham [37] does not provide insight in the treatment effects on individual dependent variables. Copula structures have been proposed as promising multivariate alternatives as well, but these models can be difficult to apply to binary dependent variables [4, 52, 55]. The second element, the transformation procedure, builds upon the close relation between the multinomial and multivariate parametrizations to express results on the scale of (multivariate) success probabilities and differences between them, as a more intuitive alternative to multinomial (log-)odds. The transformed parameters provide understandable insights in the treatment's performance on the trial population (i.e., ATEs) as well as subpopulations of interest (i.e., CATEs). The third element, the decision procedure, conveniently uses the Bayesian nature of the modelling procedure, allowing for inference on the posterior samples of transformed parameters. Decisions can be made in several ways to flexibly combine and weigh multiple dependent variables into a single decision for a population of interest, while taking correlations between dependent variables into account.

The main contribution of the current paper is the extension of the single-level BMLR framework to the multilevel context. The novel Bayesian multilevel multivariate logistic regression (BMMLR) framework provides BMLR with a multilevel model component and adjusts the transformation and decision procedure accordingly, to make the framework suitable for the multilevel context, resulting in accurate type I errors. The remainder of the paper is structured as follows. Section 2 introduces the multilevel multivariate logistic regression model to obtain a sample from the posterior distribution of regression coefficients. Section

3 outlines how to transform the obtained regression coefficients to more interpretable treatment effect parameters. Section 4 discusses the decision procedure to use the treatment effect parameters for treatment comparison. Section 5 demonstrates the performance of the model numerically via simulation and in Section 6 the methodology is illustrated with data from the IST-3. The paper concludes with a discussion in Section 7.

2 BMMLR: Bayesian multilevel multivariate logistic regression

Consider the general case with $K \in \{1, \ldots, K\}$ binary dependent variables y_{ji}^k for subject $i \in \{1, \ldots, n_j\}$ in cluster $j \in \{1, \ldots, J\}$. Outcome y_{ji}^k is Bernoulli distributed with success probability θ_{ji}^k and multivariate vector of K dependent variables, $\boldsymbol{y}_{ji} = (y_{ji}^1, \ldots, y_{ji}^K)$ is multivariate Bernoulli distributed [8]. The multivariate Bernoulli distribution relies on a hybrid parameterization where a K-variate success probability in $\boldsymbol{\theta}_{ji} = (\theta_{ji}^1, \ldots, \theta_{ji}^K)$ is expressed in terms of $Q = 2^K$ multinomial joint response probabilities in $\boldsymbol{\phi}_{ji} = (\phi_{ji}^1, \ldots, \phi_{ji}^Q)$ [8]. The q^{th} joint response probability in $\boldsymbol{\phi}_{ji}$ corresponds to multinomial response combination $\boldsymbol{h}^q$, which has length K and is given in the q^{th} row of the matrix of joint response combinations denoted by $\boldsymbol{H}$:

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

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

2.1 Likelihood of the data

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

The relation between outcome vector $\boldsymbol{y}_{ji}$ and covariate vector $\boldsymbol{x}_{ji}$ is mapped with a multinomial logistic

function that expresses the probability of $\boldsymbol{y}_{ji}$ being in response category q, conditional on $\boldsymbol{x}_{ji}$:

$$\begin{aligned}\phi_{ji}^{q} &= p(\boldsymbol{y}_{ji} = \boldsymbol{h}^{q}|\boldsymbol{x}_{ji}) \\ &= \frac{\exp(\psi_{ji}^{q})}{\sum_{r=1}^{Q-1}\exp(\psi_{ji}^{r}) + 1},\end{aligned} \tag{2}$$

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

$$\psi_{ji}^{q} = \boldsymbol{x}_{ji}^{'}\boldsymbol{\gamma}_{j}^{q} \tag{3}$$

In Equation 3, regression coefficients for response category q, $\boldsymbol{\gamma}_{j}^{q} = (\gamma_{0j}^{q}, \ldots, \gamma_{(P-1)j}^{q})$ are unknown parameters of interest. Regression coefficients of response categories $1, \ldots, Q-1$ are estimated, while regression coefficients of response category Q are fixed at zero (i.e., $\boldsymbol{\gamma}_{j}^{Q} = \mathbf{0}$) to ensure identifiability of the model. The entire set of regression coefficients in cluster j is denoted with $\boldsymbol{\gamma}_{j}$.

A key aspect of multilevel models is that the regression coefficients $\boldsymbol{\gamma}_{j}^{q}$ are allowed to vary over clusters according to a common normal distribution on the second level. The common distribution for the random effects on the second level induces a dependency structure of the observations within clusters. The observations of diffferent individuals in the same clusters are assumed to be conditionally independent conditional on the cluster-specific random effects. The random effects distribution on the second level can be written as:

$$\begin{aligned}\gamma_{pj}^{q} &= \gamma_{p0}^{q} + u_{pj}^{q} \\ \boldsymbol{u}_{j}^{q} &= (u_{0j}^{q}, \ldots, u_{(P-1)j}^{q}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}^{q})\end{aligned} \tag{4}$$

Equation 4 consists of two elements that reflect the distributional parameters:

1. The parameter γ_{p0}^{q} is the common effect in the population and does not vary over clusters.

2. The random effect u_{pj}^{q} quantifies the cluster specific deviation from the common effect γ_{p0}^{q}.

Equation 4 can be adjusted to model cluster-specific predictors or cross-level interactions between cluster-level predictors and individual level-predictors. Further, Equation 4 can be extended to model mixed effects, which combine regression coefficients that vary over clusters, which are called random effects, and regression coefficients that are identical for all clusters, which are called fixed effects. More information on the specification of more complex linear predictors can be found in general resources on multilevel models, such as Hox et al. [23] or Gelman and Hill [17]. In general, it should be noted that each additional random effect

increases the number of parameters, affecting computational burden and estimation precision.

2.2 Posterior distribution of regression coefficients

The primary goal of BMMLR is estimating the joint posterior distribution of unknown regression coefficients $\boldsymbol{\gamma}_j^q$, their means $\boldsymbol{\gamma}^q$, and their covariance matrices $\boldsymbol{\Sigma}^q$ for category $q \in 1, \ldots, (Q-1)$. The posterior probability distribution of these parameters for category q is given by:

$$p(\boldsymbol{\gamma}_j^q, \boldsymbol{\gamma}^q, \boldsymbol{\Sigma}^q | \boldsymbol{y}) \propto p(\boldsymbol{y}_j | \boldsymbol{\gamma}_j^q) p(\boldsymbol{\gamma}_j^q | \boldsymbol{\gamma}^q, \boldsymbol{\Sigma}^q) p(\boldsymbol{\gamma}^q) p(\boldsymbol{\Sigma}^q), \tag{5}$$

where $\boldsymbol{\gamma}^q$ reflects the vector of average effects for category q, $\boldsymbol{\Sigma}^q$ is the covariance matrix of the effects across clusters for category q, and $\boldsymbol{\gamma}_j^q$ reflects the vector of cluster specific effects of cluster j for category q. The posterior probability distribution in Equation 5 is proportional to the product of three types of probability distributions:

1. The likelihood of the data quantifies the probability of the dependent variables conditional on cluster-specific regression coefficients, $p(\boldsymbol{y}_j | \boldsymbol{\gamma}_j^q)$, which is the multinomial logistic function given by Equation 2;

2. The probability distribution of the cluster-specific regression coefficients $\boldsymbol{\gamma}_j^q$ conditional on their means $\boldsymbol{\gamma}^q$ and covariance matrix $\boldsymbol{\Sigma}^q$ for category q, $p(\boldsymbol{\gamma}_j^q | \boldsymbol{\gamma}^q, \boldsymbol{\Sigma}^q)$;

3. The prior probability distributions of regression coefficient's means $\boldsymbol{\gamma}^q$, $p(\boldsymbol{\gamma}^q)$, and covariance matrix $\boldsymbol{\Sigma}^q$, $p(\boldsymbol{\Sigma}^q)$ for category q, before observing the data.

As the multinomial logistic function (Equation 2) does not have a (conditionally) conjugate prior distribution, the functional form of the posterior distribution is unknown and the regression coefficients cannot be sampled directly from the posterior distribution. In the Supplemental material, we present a Gibbs sampling algorithm based on a Pólya-Gamma auxiliary variable expansion of the likelihood proposed by Polson et al. [59]. The expanded likelihood has a Gaussian form and can be combined with normal prior distributions on regression coefficients $\boldsymbol{\gamma}^q$ and an inverse-Wishart distribution on covariance matrix $\boldsymbol{\Sigma}^q$. The parameters are known to have conditionally conjugate posterior distributions and allow for direct sampling from their multivariate normal and inverse-Wishart distributions respectively, resulting in MCMC chains of the joint posterior distribution in Equation 5. We also include a few comments on prior specification for the proposed Gibbs sampling procedure in the Supplemental material.

As an alternative to the proposed Gibbs sampling procedure, sampling from the posterior distribution(s) of multinomial logistic regression coefficients can theoretically be done with other standard MCMC-methods for non-conjugate prior-likelihood combinations, such as Metropolis-Hastings [e.g., 6, 13, 64, Ch. 3 and 5] or Hamiltonian Monte Carlo [e.g., 2, 78, 1] sampling algorithms.

3 Transformation of posterior regression coefficients to the probability scale

The output of the BMMLR model from Section 2 is an MCMC sample of posterior multinomial regression coefficients. These regression coefficients reflect the importance of a predictor on a specific joint response combination and represent - in exponentiated form - the odds compared to reference category Q. While these regression coefficients can be insightful in a truly multinomial research problem, they have no straightforward interpretation in multivariate treatment comparison where marginal effects on individual dependent variables play a central role [11].

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

We use the framework for transformation to the probability scale and decision-making with a posterior sample of multivariate treatment differences introduced in [33] and [32]. Technical details of these procedures are presented in Algorithm 1 in Appendix B. We use the remainder of this section to summarize and illustrate the procedure with a toy example from the IST-3-data, where we assume interest in the effect of Alteplase in the experimental condition (T_A) compared to no treatment in the control group (T_C).

Assume that we re-analyze a part of the IST-3 data using the BMMLR framework and take one of originally presented analyses as a starting point [77]. In the selected analysis, the researchers compared the effects of Alteplase vs. control on their primary outcome, long-term independent living after six months ($Indep6$), among subgroups of patients based on the severity of their initial stroke. In our example, we perform a multivariate analysis of the treatment effects on the primary outcome ($Indep6$) and one of the secondary (short-term) dependent variables: being stroke-free in the first seven days after the initial stroke ($Strk7$). We incorporate severity of the initial stroke as a predictor variable to study heterogeneity, using the grouping criteria from the original trial for the estimation of conditional average treatment effects. We aim

to investigate the average treatment effect among the trial population as specified by the original eligibility criteria for inclusion. We are also interested in a potential interaction between the treatment and stroke severity, and investigate the conditional average treatment effects among patients with various severities of stroke. To take the clustered structure of the data into account, we specified a BMMLR mixed-effects model with random slopes for the intercept and the main treatment effect, resulting in the following linear predictor:

$$\begin{aligned}\psi_{ji}^{q} &= \gamma_{0j}^{q} + \gamma_{1j}^{q} T_{ji} + \beta_{2}^{q} NIHSS_{ji} + \beta_{3}^{q} NIHSS_{ji} T_{ji} \\ \gamma_{0j}^{q} &= \gamma_{00}^{q} + u_{0j} \\ \gamma_{1j}^{q} &= \gamma_{10}^{q} + u_{1j}.\end{aligned} \tag{6}$$

In Equation 6, $\boldsymbol{x}_{ji} = (1, T_{ji}, NIHSS_{ji}, NIHSS_{ji}T_{ji})$ with treatment indicator T_{ji} and $NIHSS_{ji}$ being the stroke severity score of subject i in hospital j. The $Q = 4$ resulting joint response categories are $(\{Strk7 = 1, Indep6 = 1\}, \{Strk7 = 1, Indep6 = 0\}, \{Strk7 = 0, Indep6 = 1\}, \{Strk7 = 0, Indep6 = 0\})$, which we refer to as $(\{11\}, \{10\}, \{01\}, \{00\})$.

3.1 Transformation to cluster-specific (differences between) probabilities

The main quantity of interest, the (cluster-specific) marginal multivariate treatment difference, is defined as the difference between cluster-specific multivariate success probabilities of the two treatments:

$$\begin{aligned}\delta_{j}^{Strk7} &= \theta_{Aj}^{Strk7} - \theta_{Cj}^{Strk7} \\ \delta_{j}^{Indep6} &= \theta_{Aj}^{Indep6} - \theta_{Cj}^{Indep6}\end{aligned} \tag{7}$$

where subscripts Aj and Cj indicate cluster-specific parameters of the (experimental) Alteplase and control treatments respectively. The elements on the right-hand sides of Equation 7, success probabilities θ_{Tj}^{k}, are sums of the multinomial joint response probabilities of all response categories with a success on outcome k:

$$\begin{aligned}\theta_{Tj}^{Strk7} &= p(\boldsymbol{y}_j = \{11\}|T) + p(\boldsymbol{y}_j = \{10\}|T) = \phi_{Tj}^{1} + \phi_{Tj}^{2} \\ \theta_{Tj}^{Indep6} &= p(\boldsymbol{y}_j = \{11\}|T) + p(\boldsymbol{y}_j = \{01\}|T) = \phi_{Tj}^{1} + \phi_{Tj}^{3}\end{aligned} \tag{8}$$

The multinomial joint response probabilities $\boldsymbol{\phi}_{Tj}$ that form the elements of success probabilities $\boldsymbol{\theta}_{Tj}$ follow from plugging in posterior regression coefficients $\boldsymbol{\gamma}_{j}^{q}$ in the linear predictor (Equation 6) and the multinomial logistic link function (Equation 2) for prespecified covariates $\boldsymbol{x}_j$ and for the relevant response

category q.

$$\phi_{Tj}^{q} = = \frac{\exp\left(\psi_{Tj}^{q}\right)}{\sum_{r=1}^{Q-1} \exp\left(\psi_{Tj}^{r}\right) + 1}. \tag{9}$$

The information in covariate vector $\boldsymbol{x}_j$, which directly affects ψ_{Tj}^{q}, determines the treatment as well as the subpopulation of interest. Subpopulations can be defined as a value, such as a stroke severity score of one standard deviation below or above the mean, that can be plugged in directly into Equations 6 and 2. When interested in a subpopulation that is defined by an interval, such as the groups of stroke severity in the IST-3, the joint response probability is marginalized over the specified interval or averaged over a sample of observations in this interval. In the latter case, joint response probability ϕ_{Tj}^{q} is computed for each observed subject $i \in 1, \ldots, n_j$ via Equation 2. The joint response probability for each treatment T is then computed by averaging over all subjects i in treatment T and cluster j.

Since the model in Section 2 resulted in a sample of L posterior draws of each regression coefficient, multivariate treatment differences are computed for each draw (l) separately. The resulting posterior samples can be summarized with standard descriptive methods.

3.2 Pooling treatment effects over clusters

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

$$\boldsymbol{\delta} = \frac{\sum_{j=1}^{J} n_j \boldsymbol{\delta}_j}{\sum_{j=1}^{J} n_j} \tag{10}$$

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

4 Decision-making based on multivariate treatment effects

The obtained sample of posterior treatment differences can be used for statistical decision-making regarding treatment superiority and inferiority. The multivariate context has multiple options to define superiority and inferiority, leaving much flexibility to combine and prioritize dependent variables in a suitable way.

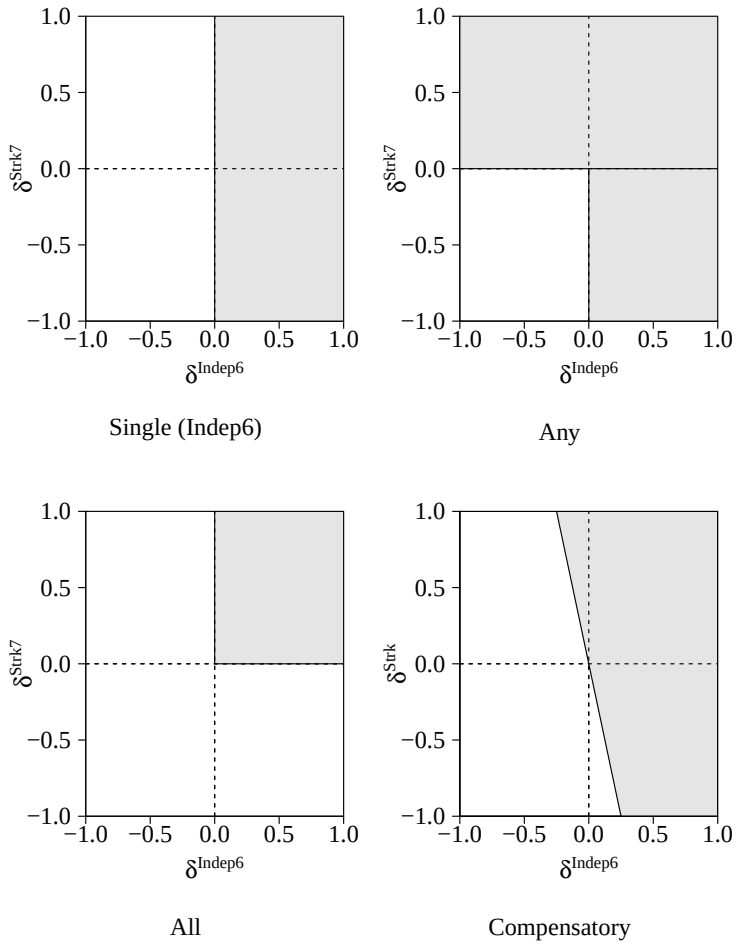

Figure 1: Superiority regions of four decision rules applied to the IST-3. The Compensatory rule has weights $\boldsymbol{w} = (0.20, 0.80)$.

We shortly discuss four different decision rules to give some idea of possibilities, without intending to be exhaustive or complete. The presented rules have different theoretical underpinnings and distinct statistical properties, such as acceptance regions, a priori estimated sample sizes, cutoff values, and error rates. The acceptance regions for superiority decisions of the four presented rules are graphically presented in Figure 1. More details to guide an informed choice for one of these decision rules in practice can be found in Kavelaars et al. [33].

Three of these rules originate from guidelines of the Food and Drug Administration (FDA) [11]. The FDA defines superiority as a treatment difference larger than zero on the primary outcome (which we refer to as "Single rule"), on all dependent variables ("All rule") or on any of the dependent variables ("Any rule"). The Single rule reduces the statistical analysis to a univariate problem, using only the treatment difference of independent living after 6 months as a primary outcome (Single rule). The All and Any rules make no

distinction in the importance of dependent variables and assume that the short-term and long-term outcome are either both required for superiority or inferiority (All rule), or are interchangeable (Any rule).

In practice, these rules can oversimplify decision-making. Secondary outcome variables often contribute to treatment evaluation as well, but are given a co-primary status in the All and Any rules or are not formally included in the statistical decision procedure when the Single rule is used [74, 73]. To handle outcomes that differ in relative importance, linear combinations of dependent variables with pre-assigned (importance) weights have been proposed as a flexible alternative [53, 49, 83, 75, 33]. We refer to a linear combination as a Compensatory rule, referring to its inherent mechanism that allows (weighted) positive and negative effects to compensate each other. The Compensatory rule allows the IST-3 data to consider the effects on the long-term much more important than the short-term effect without completely excluding the risk of a recurrent stroke from the final decision. In such a situation, we can assign the primary outcome ($Indep6$) - for example - four times more weight than the secondary outcome ($Strk7$) and consider Alteplase superior to no treatment if a lower chance of dependency is outweighed by a small increase in the risk of a recurrent stroke.

Evidence in favor of the decision rule can be quantified by the proportion posterior draws of the pooled treatment difference $\boldsymbol{\delta}$ that lie in the decision-rule specific acceptance region, denoted by $\mathcal{S}_R$. A conclusion is reached via comparison to p_{cut}, which is a cutoff value to balance the required amount of evidence with anticipated Type I error rates [38]:

$$p(\boldsymbol{\delta} \in \mathcal{S}_R) > p_{cut}. \tag{11}$$

In the multivariate logistic regression model, the probability in Equation 11 has no analytical solution. Therefore, decisions are made via the posterior MCMC-sample of L draws. Superiority is concluded when:

$$\frac{1}{L} \sum_{(l)=1}^{L} I(\boldsymbol{\delta}^{(l)} \in \mathcal{S}_R) > p_{cut}. \tag{12}$$

Similarly, inferiority is concluded when:

$$\frac{1}{L} \sum_{(l)=1}^{L} I(\boldsymbol{\delta}^{(l)} \in \mathcal{S}_R) < 1 - p_{cut}. \tag{13}$$

In Section 6, we demonstrate these decision with data from the IST-3 as part of an illustration of the BMMLR framework.

5 Numerical evaluation

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

1. how well the obtained regression coefficients and treatment effects correspond to their true values to examine bias;

2. how often the BMMLR framework results in an (in)correct superiority or inferiority conclusion to learn about decision error rates;

5.1 Setup

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

1. The BMMLR model presented in Section 2. We generated response data from a mixed effects model to include random effects while keeping the number of estimated parameters limited. We included an interaction between the treatment and the covariate as well, resulting in the following linear predictor:

$$\begin{aligned} \psi_{ji}^q &= \gamma_{0j}^q + \gamma_{1j}^q T_{ji} + \beta_2^q w_{ji} + \beta_3^q w_{ji} T_{ji} \\ \gamma_{0j}^q &= \gamma_{00}^q + u_{0j} \\ \gamma_{1j}^q &= \gamma_{10}^q + u_{1j}. \end{aligned} \tag{14}$$

   In line with previous notation, $\boldsymbol{x}_{ji} = (1, T_{ji}, w_{ji}, w_{ji}T_{ji})$ in Equation 14. Further, vector $\boldsymbol{\gamma}_j^q = (\gamma_{0j}^q, \gamma_{1j}^q)$ reflects random effects with multivariate normally distributed errors (i.e., $(u_{0j}^q, u_{1j}^q) \sim N(\mathbf{0}, \boldsymbol{\Sigma}^q)$) for the intercept and main effect of the treatment. Regression coefficients $\boldsymbol{\beta}^q = (\beta_2^q, \beta_3^q)$ reflect fixed effects for the covariate and covariate-by-treatment interaction.

2. Single-level Bayesian multivariate logistic regression model [BMLR; 32], as a first reference approach. For this model, we use a restricted version of Equation 14 with fixed regression coefficients only:

$$\psi_{ji}^q = \beta_0^q + \beta_1^q T_{ji} + \beta_2^q w_{ji} + \beta_3^q w_{ji} T_{ji}, \tag{15}$$

   MCMC chains were sampled with a simplified version of the Gibbs sampling procedure in Appendix A,

that iterates over $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$. The model shares information in the estimation of conditional treatment effects with sufficient power, but does not take the multilevel structure of the data into account.

3. Single-level unconditional Bayesian multivariate Bernoulli analysis [BMB; 33], as a second reference approach. Bayesian multivariate Bernoulli analysis relies on a conjugate multinomial likelihood and Dirichlet prior. MCMC draws are sampled directly from the posterior Dirichlet distribution with parameters $\sum_{j=1}^{J}\sum_{i=1}^{n_j} I(\boldsymbol{y}_{ji} = \boldsymbol{h}^q)+\alpha^{0q}$, where we assigned prior hyperparameters $\boldsymbol{\alpha}^0 = (0.01, 0.01, 0.01, 0.01)$. The approach can estimate homogeneous treatment effects accurately and fast, but cannot deal with multilevel data. Moreover, conditional treatment effects originate from subsampling, which is less powerful than regression due to the isolation from other information.

Effect size. We specified a heterogeneous treatment effect, with pooled average treatment differences of zero ($\boldsymbol{\delta} = (0, 0)$, $\delta(\boldsymbol{w}) = 0$) and pooled conditional treatment differences larger than zero ($\boldsymbol{\delta} = (0.25, 0.15)$, $\delta(\boldsymbol{w}) = 0.20$). This scenario aimed to demonstrate the Type I error rate among the trial population. It reflects a least favorable treatment difference for the Any and Compensatory rules and should therefore result in the targeted Type I error rate for these rules to be considered accurate. The conditional treatment effect provided insight in the power to conclude superiority among the subpopulation under consideration. Outcome variables were negatively correlated ($\rho_{ATE} = -.157$; $\rho_{CATE} = -.20$). The regression parameters used to generate these effects are presented in Table 1.

Table 1: True regression parameters used for data generation

	q_1	q_2	q_3	q_4
p_0 (Intercept)	0.000	0.433	0.433	0.000
p_1 (T_{ji})	0.000	0.000	0.000	0.000
p_2 (w_{ji})	1.027	0.601	0.427	0.000
p_3 ($w_{ji}T_{ji}$)	-2.055	-1.201	-0.854	0.000

For the BMMLR model, the covariance matrix of random effects, $\boldsymbol{\Sigma}^q$, was specified as:

$$\begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \tag{16}$$

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

Sample size. We varied the sample sizes at the cluster and subject level. Since there are no clear guidelines regarding sample size computations in multilevel multivariate logistic regression, we explored performance of the model for different numbers of clusters and different sample sizes within clusters. Specifically, we used

number of clusters $J \in \{10, 100\}$ and observations per cluster $n_j \in \{10, 100\}$ for each treatment, resulting in four different sample size combinations.

5.1.1 Procedure

Data generation. For each sample size, we sampled 1000 datasets under the mixed effects model in Equation 14 with the true regression parameters in Table 1. We assigned n_j participants to each treatment T and generated covariate x from a standard normal distribution. We sampled response vector $\boldsymbol{y}_{ji}$ from a multinomial distribution with probabilities $\boldsymbol{\phi}_{ji}$.

Gibbs sampling. Regression coefficients for the BMMLR and BMLR models were estimated via the Gibbs sampling procedure in Appendix A. We ran two MCMC-chains via the Gibbs sampler introduced in Section 2 with $L = 50,000$ iterations plus $10,000$ burn-in iterations. This large number of iterations aims to minimize the influence of the potentially high autocorrelations between parameters in multilevel models on the stationary distribution of the parameters. Autocorrelations were highest among random effect parameters $\boldsymbol{\gamma}_j$ and ranged between 0.107 and 0.781 at lag 1 and reduced to a range of $-0.012 - 0.276$ at lag 10. Further, following the guidelines in Gelman et al. [16], we ensured that the multivariate potential scale reduction factor was below 1.10.

Prior specification. For the multilevel model (BMMLR), we specified diffuse priors, which were multivariate normally distributed for regression coefficients:

$$\begin{aligned} (\beta_2^q, \beta_3^q) &\sim N([\begin{smallmatrix}0\\0\end{smallmatrix}], [\begin{smallmatrix}10 & 0\\0 & 10\end{smallmatrix}]) \\ (\gamma_{00}^q, \gamma_{10}^q) &\sim N([\begin{smallmatrix}0\\0\end{smallmatrix}], [\begin{smallmatrix}10 & 0\\0 & 10\end{smallmatrix}]) \end{aligned} \tag{17}$$

The specified variance matrices of regression coefficients were motivated by a paper of Gelman et al. [18], who recommend to choose a variance parameter that results in realistic support for the probability parameter after non-linear transformation in logistic regression. We specified an inverse-Wishart prior distribution for the covariance matrix:

$$\boldsymbol{\Sigma}^q \sim \mathcal{W}^{-1}(2, [\begin{smallmatrix}0.1 & 0\\0 & 0.1\end{smallmatrix}]).$$

The regression parameters $\boldsymbol{\beta}^q$ in the single-level regression model (BMLR) were the same as in the multivariate approach (i.e., independent normal priors with means of 0 and variances of 10).

Transformation and decision-making. We applied the procedures in Algorithm 1 to use the obtained MCMC-chains of posterior regression coefficients for superiority decision-making. We thinned the chains in the transformation procedure with a factor 10 to reduce the computational burden.

We considered two different effects:

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

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

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

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

We computed the probability to conclude superiority (p_{Sup}) as the proportion of posterior treatment differences in the superiority region via Equation 11. The targeted Type I-error rate of $\alpha = .05$ corresponded to decision threshold $p_{cut} = 1 - \alpha = 0.95$ (Compensatory and All rules) and a for multiple tests corrected threshold $p_{cut} = 1 - \frac{\alpha}{K} = 0.975$ (Any rule) [38, 33, 73].

5.1.2 Software

We conducted our analyses in `R` and made use of several existing packages [60]. Pólya-Gamma variables were drawn with the `pgdraw` package [36]. Further, we drew variables from the multivariate normal, truncated normal, and Dirichlet distributions with the `MASS`, `msm`, and `MCMCpack` packages respectively [80, 25, 39]. MCMC chains were diagnosed with the `coda` and `mcmcse` packages [57, 10]. We parallellized the simulation procedure with the `foreach` and `doParallel` packages [42, 41] and created LaTeX tables with the `xtable` package [7]. The `R` code used to generate results can be found on GitHub `https://github.com/XynthiaKavelaars/Bayesian-multilevel-multivariate-logistic-regression`.

5.2 Results

The current subsection presents the results of the simulation study. Presented decision error rates are in Table 2.

5.2.1 Bias

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

5.2.2 Decision error rates

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

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

Power The conditional treatment effect demonstrated the power to correctly conclude superiority for all three rules. Three results were highlighted. First, the multilevel model (BMMLR) is more powerful when the number of clusters is higher. The two conditions with an equal total sample size showed a .30 difference

in power under the All rule. The other rules showed the same patterns, but had too high proportions of superiority conclusions to clearly distinguish the sample size conditions: The power in the other conditions equaled or was close to the maximum of 1.000.

Second, the single-level regression model (BMLR) resulted in more superiority conclusions than the multilevel regression model, implying that the posterior distributions of treatment differences of the single-level regression model had smaller variances. Again, differences were best illustrated by the All rule and the condition with small sample sizes for the Any and Compensatory decision rules, as these proportions were well below the maximum. Similar to the Type I error rates, the differences between the proportions of superiority conclusions appeared to be subject to the number of clusters, as demonstrated by a comparison of the two conditions with an identical total sample size under the All rule. The multilevel model was less powerful than the single-level model when the number of clusters was low in particular, being in line with non-independence of clustered observations.

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

Table 2: Proportions of superiority decisions and standard errors by data-generating mechanism, estimation method, and decision rule.

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

p = proportion of superiority decisions
se = Standard errors

6 Illustration with IST-3 data

To illustrate the proposed framework with real data, we re-analyzed a subset of data from the Third International Stroke Trial using the BMMLR framework [77, 65]. The included 3,035 subjects in the IST-3 were recruited from 156 different hospitals in 12 different countries, resulting in multilevel data from patients clustered within hospitals and hospitals clustered within countries. We selected a two-level subset of 1,447 subjects from 75 hospitals in the United Kingdom with a known health and survival status at six months after the initial stroke and a known or predicted severity score of the initial stroke (NIH Stroke Score; NIHSS) at randomisation. The cluster sizes were skewed and ranged from 1 to 117, with a median cluster size of 7 (SD: 26.66). Of the selected subset of data, $n_A = 716$ subjects were in the Alteplase group (treatment = 1) and $n_C = 731$ subjects were in the control group (treatment = 0). We compared the effects of the two treatments on a) being stroke-free for seven days (0 = no; 1 = yes) and b) long-term independent living at six months (0 = no, 1 = yes), while taking the severity of the initial stroke into account. The NIHSS can range from 0 to 42 with a higher score indicating a more severe stroke. The average stroke severity score in the IST-3 was 13.12 (SD: 6.91) and comparable in both treatment groups.

6.1 Method

We fitted our model with random slopes for the intercept and the treatment effect. We sought to compare our multilevel model (BMMLR) to the two single-level models (BMLR and BMB) from the Numerical evaluation section in treatment comparison of Alteplase and control on dependency after six months (δ^{Indep6}) and recurrent stroke within seven days (δ^{Strk7}). The multilevel model (BMMLR) was fitted with the linear predictor in Equation 6 and the linear predictor of the single-level regression model (BMLR) was:

$$\psi_{ji}^{q} = \beta_0^q + \beta_1^q T_{ji} + \beta_2^q NIHSS_{ji} + \beta_3^q NIHSS_{ji} T_{ji} \tag{18}$$

Prior specification. For the regression coefficients in the multilevel model (BMMLR) and the single-level regression model (BMLR), we specified independent normal prior distributions with means of 0 and variances of 10. For covariance matrix Σ^q, we specified an improper uniform prior for the random effects covariance matrix for each category q, to enable testing for the presence of random effects in the model comparison step discussed later.

Gibbs sampling. We ran two MCMC-chains via the Gibbs samplers. Since the chains of regression coefficients were highly autocorrelated in the multilevel model (lag 10: $\boldsymbol{\beta}$: $0.47 - 0.59$; $\boldsymbol{\gamma}$: $0.62 - 0.80$, $\boldsymbol{\Sigma}$: $-0.01 - 0.38$), we sampled a large number of 500,000 iterations plus 10,000 burnin iterations. The

multivariate potential scale reduction factor was below 1.01 for all parameters, implying that there were no signals of non-convergence. We thinned MCMC-chains in follow-up posterior transformations with a factor 10 to reduce computational demands, resulting in inference based on $L = 50,000$ draws.

Transformation and decision-making. We applied the procedures in Algorithm 1 to the thinned MCMC-chains of posterior regression coefficients to make superiority decisions. We considered (conditional) average treatment effects among seven different (sub)populations:

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

The grouping criteria for CATEs of ranges were taken from the original IST-3 paper [77].

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

6.1.1 Model comparison

Since the true model of these real-world data is unknown, we followed up on the analysis with a comparison of model fit via Bayes factors. Bayes factors [26] quantify the relative evidence in the data between competing statistical models. Here we use default Bayes factors which avoid the need to manually specify prior distributions [48, 47, 81].

BMLR vs. BMB. To compare the two single-level models, we computed a Bayes factor on the probabilities that the regression coefficients of the covariate (β_2^q) and the interaction between the covariate and the interaction (β_3^q) was equal to zero for all $q \in q, \ldots, Q-1$ using the `BF()`-function from the `R`-package `BFpack` [48].

BMMLR vs. BMLR. To compare the proposed multilevel model (BMMLR) and the single-level model (BMLR), we computed empirical Bayes factors as proposed by Vieira-Generoso et al. [81], which tests whether the random effects are equal across clusters using uniform priors for the random effects covariance matrices. This test is executed separately for all six different random effects in the multilevel model.

Software. In addition to the software packages used in Section 5, we used `R` packages `haven` to import the dataset [84], `BFpack` [48] to compute Bayes factors for comparison of the two single-level models.

6.2 Results

6.2.1 Results of different (sub)populations

Table 3 show how different analysis models and different decision rules provide elaborate insights in the effects of Alteplase vs. control on a combination of dependent variables among different (sub)populations. Analysis of the selected data with the BMMLR, BMLR, and BMB models gave the following results.

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

Table 3: Average (ATE) and conditional average (CATE) treatment effects of the specified (sub)populations of the IST-3.

	$(\delta^{Strk7}, \delta^{Indep6})$	Pop	Any	All	$\delta(\boldsymbol{w})$	Pop	Comp
ATE		$n_A = 716$, $n_C = 731$					
BMMLR	$(-0.114, 0.029)$	$(0.000, 0.886)$	$<$	-	0.000	0.504	-
BMLR	$(-0.116, 0.033)$	$(0.000, 0.941)$	$<$	-	0.003	0.572	-
BMB	$(-0.117, 0.032)$	$(0.000, 0.911)$	$<$	-	0.003	0.549	-
CATE - Low range		$n_A = 99$, $n_C = 105$					
BMMLR	$(-0.078, -0.023)$	$(0.003, 0.317)$	$<$	-	-0.034	0.200	-
BMLR	$(-0.081, -0.016)$	$(0.004, 0.365)$	$<$	-	-0.029	0.225	-
BMB	$(-0.110, -0.036)$	$(0.019, 0.318)$	-	-	-0.051	0.207	-
CATE - Mid-Low range		$n_A = 327$, $n_C = 334$					
BMMLR	$(-0.090, 0.038)$	$(0.000, 0.884)$	$<$	-	0.013	0.679	-
BMLR	$(-0.092, 0.044)$	$(0.000, 0.937)$	$<$	-	0.017	0.752	-
BMB	$(-0.114, 0.045)$	$(0.001, 0.853)$	$<$	-	0.013	0.642	-
CATE - Mid-High range		$n_A = 237$, $n_C = 252$					
BMMLR	$(-0.139, 0.051)$	$(0.000, 0.992)$	$<$ & $>$	-	0.013	0.753	-
BMLR	$(-0.141, 0.054)$	$(0.000, 0.995)$	$<$ & $>$	-	0.015	0.783	-
BMB	$(-0.118, 0.047)$	$(0.006, 0.938)$	$<$	-	0.014	0.694	-
CATE - High range		$n_A = 53$, $n_C = 40$					
BMMLR	$(-0.183, 0.020)$	$(0.002, 0.980)$	$<$	-	-0.021	0.100	-
BMLR	$(-0.188, 0.021)$	$(0.001, 0.982)$	$<$	-	-0.021	0.100	-
BMB	$(-0.173, 0.019)$	$(0.069, 0.687)$	-	-	-0.019	0.327	-
CATE - Low value							
BMMLR	$(-0.078, -0.007)$	$(0.002, 0.440)$	$<$	-	-0.021	0.291	-
BMLR	$(-0.080, 0.000)$	$(0.002, 0.503)$	$<$	-	-0.016	0.328	-
CATE - High value							
BMMLR	$(-0.140, 0.052)$	$(0.000, 0.991)$	$<$ & $>$	-	0.014	0.751	-
BMLR	$(-0.142, 0.055)$	$(0.000, 0.994)$	$<$ & $>$	-	0.015	0.777	-

Pop = Posterior probability
$>$ = superiority concluded
$<$ = inferiority concluded

Conditional average treatment effects. The four conditional average treatment effects (CATEs) that reflected subpopulations as ranges sketched a more heterogeneous picture than the average treatment effects. Whereas all ranges showed a lower probability of being free of strokes after treatment with Alteplase, these probabilities increased with the severity of the stroke. Differences between success probabilities of the two treatments appeared to increase with severity of the stroke, such that Alteplase appeared to have the largest negative effect on being stroke-free when the severity of the initial stroke was highest. A more diffuse relation between stroke severity and treatment difference emerged on long-term independent living. Alteplase resulted in a slightly lower point estimate of the probability of independent living among patients with a Low stroke severity, but resulted in a higher estimated probability of independent living in all categories of more severe strokes. Patients in the Mid-Low and Mid-High ranges of stroke severity had the largest positive effect of Alteplase on independent living. The Low and High stroke severity patients had slightly higher weighted probabilities after Alteplase, while patients with a Mid-Low and Mid-High stroke severity had weighted probabilities close to zero. These non-zero point estimates were not unanimously supported by sufficient evidence to conclude superiority or inferiority. The All and Compensatory rules remained inconclusive for all models among all subpopulations. The BMMLR and BMLR were unanimous in their conclusions for the Any rule: Inferiority was concluded for patients with a Low, Mid-Low and High stroke severity, while both superiority and inferiority were concluded for patients with a Mid-High range stroke severity. The BMB model remained inconclusive in the Low and High ranges and concluded inferiority among patients with a Mid-Low or Mid-High stroke severity, according to the Any rule.

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

Model comparison Bayes factors [31] are computed to test whether there is evidence that a dependency structure is present in the data that is caused by the multilevel structure. The results are presented in Table 4. These results indicate that there is evidence that each of the six different random effects do not vary across clusters. This implies that the parsimonious single-level model (BMLR) is preferred over the multilevel model (BMMLR) for these specific data. This result is also in agreement with the obtained estimates which

are virtually identical under both models. Model comparison between the two single-level models (BMLR vs. BMB) resulted in a log-transformed Bayes factor of 16.348, reflecting strong evidence that the a regression model (BMLR) fitted the data better than the multivariate Bernoulli (BMB) model. We give a general recommendation on model selection in the Discussion section.

Table 4: Logarithmic transformations of Bayes factors of BMLR vs. BMMLR

	$q = 1$	$q = 2$	$q = 3$
NIHSS	5.769	5.642	11.238
NIHSS $\times$ Trt	5.653	6.181	8.555

6.2.2 Conclusions and discussion

Several conclusions regarding the BMMLR framework could be drawn from the presented results. First, multilevel analysis did not affect point estimates in the used subset of IST-3 data: BMMLR and BMLR models resulted in similar point estimates of $\boldsymbol{\delta}$ and $\delta(\boldsymbol{w})$, as expected from the negligible bias in the results of the simulation study. The posterior probabilities of the BMMLR and the BMLR model were similar and did not lead to different superiority or inferiority conclusions. A model comparison based on Bayes factors resulted in evidence in favor of a single-level model. It would be helpful to have information about clustering beforehand and we concluded that these results call for a proper method to quantify the degree of dependence among observations within clusters prior to the analysis. Such insights could help in clarifying the statistical urgency of a multilevel model and the appropriateness of a single-level model in advance.

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

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

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

7 Discussion

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

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

The presented framework accurately handled the multilevel structure of the data in the presence of heterogeneous treatment effects on multiple (correlated) binary dependent variables. A simulation study demonstrated that the proposed model indeed a) estimated average and conditional treatment effects in multilevel data without bias; and b) resulted in statistical decisions with targeted Type I error rates. A multilevel model was clearly superior for clustered data: Naive models that did not take the multilevel structure into account resulted in inflated Type I-error rates. Further, the logistic model promoted information-sharing between clusters and subpopulations, being a more powerful alternative than subgroup analysis to identify heterogeneous treatment effects. A re-analysis of the IST-3 provided another perspective on the data than the original paper [77]. Detailed insights as well as the varying treatment effects among subpopulations demonstrated the importance of a) a well-considered and specific decision rule; and b) the assessment of treatment heterogeneity. The statistical need for a multilevel model has not clearly become evident for this specific analysis. The results suggested that a substantive cluster structure in the data does not necessarily imply a relevant statistical dependency structure between observations. We demonstrated that an implied dependency structure can be tested using empirical Bayes factors [81]. If these Bayes factors provide evidence that none of the random effects varies, a single-level model gives a more parsimonious description

of the data. In case of evidence for the presence of random effects due to the multilevel structure in the data, the proposed multilevel multivariate model is preferred as it gives more accurate type I errors. If there is evidence that some of the random effects do not vary across clusters, it is recommended to fix these parameters to give a more parsimonious description of the data.

Application of the BMMLR framework is not limited to the presented analyses. Theoretically, the model can be adapted to the longitudinal setting, may be used to borrow strength from different trials, or may be extended to data with multiple levels of clustering for example. In practice, such extensions require additional exploration of the (computational) properties of the model, since MCMC sampling procedures appeared sensitive to the amount of autocorrelation and the number of parameters. In a related fashion, carefully choosing which random effects to include is helpful for smooth execution of multilevel analysis. The model has a large number of options regarding specification of the model, giving a lot of flexibility to model cluster effects precisely. This flexibility reduces parsimony however, as it easily increases the number of model parameters. While it is technically possible to expand the model, some care must be taken when adding many outcome variables and many covariates however. This would result in many more model parameters, which results in considerably less parsimonious description of the data and can intensify computations notably. Similarly, the multinomial setup is most suitable for a limited number of dependent variables. Increasing the number of dependent variables results in a large number of response categories, which may lead to sparsity issues.

Future research might advance the design of the BMMLR framework in multiple ways. First, a priori sample size computation and power analysis have priority in medical research. Sample sizes in logistic regression should not be too small and preferably take the success probability into account [28, 50]. In line with our findings, larger numbers of clusters generally appear to be more powerful than larger numbers of subjects within clusters [72], although a study into sample sizes for multilevel logistic regression analysis provided less clear results [46]. Expanding and refining knowledge regarding sample sizes in multilevel models aids in strategic experimental design [62, 45, 44] Additionally, ethical aspects, such as risks and burden of (potentially inferior) treatment, and practical considerations, such as limited access to (large numbers of) subjects, require more in-depth understanding of power and sample sizes. Especially in precision medicine – where treatments are targeted at specific patient populations - numbers of eligible subjects are limited and a priori power analysis helps to manage expectations in terms of duration.

Second, the methodology can be placed into a broader framework of Bayesian statistics. The framework can be extended with the computation of Bayes factors to aid in decision-making regarding superiority and inferiority as well, for example following the ideas presented in Van Ravenzwaaij et al. [79]. Further, the specification of prior distributions requires consideration. Specification of non-informative priors may not be

trivial. The general tendency to choose relatively large variance parameters for normally distributed prior distributions [18], does not necessarily work well with the proposed model. Covering a range far beyond realistic parameter values, can (negatively) affect the efficiency of the sampling procedure and even the resulting posterior distribution. Thus, concrete guidelines for the specification of non-informative priors would be helpful.

Third, pooling of treatment estimates can be done in several other ways than presented. In general, the pooled treatment effect over clusters is a weighted combination of cluster-specific estimates, where the weights aim to balance aspects that influence estimation and are imbalanced over clusters (e.g., cluster size or variance). Whereas we applied a cluster size-based approach, several advanced weighing procedures balance unequal variances within clusters via regularization methods [for overviews, see 34, 27, 14]. These weighing methods generally produce shrinkage to the mean a) when group level variance is smaller; and/or b) when sample sizes are smaller [17, p. 269]. Such weighing procedures have interesting balancing properties but are probably less suitable for trials with clusters of single subjects, such as IST-3. These clusters have no variance, should not be discarded or merged inconsiderately, and call for the exploration of suitable weighing procedures for such data.

Finally, the BMMLR framework and multilevel models for discrete data in general lack a standard way to quantify the degree of clustering and the corresponding need for a multilevel model. Often, the degree of clustering is quantified as the variance between clusters relative to the variance within clusters, expressed via an intraclass correlation coefficient (ICC). The computation of ICCs in binary data is not straightforward: The variance within clusters - and therefore the ICC - is a function of the predictors in the model and the ICC depends on the prevalence, requiring an alternative approximation to obtain an appropriate estimate of the ICC [63, 21, 56, 19]. We leave the extension of our framework in this direction for future research.

8 Conclusion

The presented Bayesian method aimed to capture a multilevel structure and treatment heterogeneity simultaneously in data with multiple correlated binary outcome variables and observed covariates. The framework was built upon three major components: a multivariate logistic regression analysis, a subsequent transformation of regression coefficients to the multivariate probability scale, 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 useful in prediction of treatment effects and decision-making within subpopulations from multiple clusters, 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.

Abbreviations

ATE	Average Treatment Effect
BMB	Bayesian multivariate Bernoulli
BMLR	Bayesian multivariate logistic regression
BMMLR	Bayesian multilevel multivariate logistic regression
CATE	Conditional Average Treatment Effect
FDA	Food and Drug Administration
IST	Internation Stroke Trial
IST-3	Third International Stroke Trial
NIHSS	National Institutes of Health Stroke Score
RCT	Randomized Controlled Trial

Declarations

Acknowledgements

We thank Peter Sandercock on behalf of The International Stroke Trial-3 Collaborative Group for making the data from the Third International Stroke Trial publicly available. We gratefully acknowledge The IST-3 Collaborative Group, the trial joint sponsors (The University of Edinburgh and the Lothian Health Board), and the chief funding agencies of the study: UK Medical Research Council, Health Foundation UK, Stroke Association UK, Research Council of Norway, Arbetsmarknadens Partners Forsakringsbolag (AFA) Insurances Sweden, Swedish Heart Lung Fund, The Foundation of Marianne and Marcus Wallenberg, Polish Ministry of Science and Education, the Australian Heart Foundation, Australian National Health and Medical Research Council (NHMRC), Swiss National ResearchFoundation, Swiss Heart Foundation, Assessorato alla Sanita, Regione dell'Umbria, Italy, and Danube University. We thank three reviewers for their helpful comments on an earlier draft of the manuscript.

Funding

The current work was supported by the Dutch Research Council (NWO) [no. 406.18.505].

Availability of data and materials

The Third International Stroke Trial data that support the findings of this study are available with the identifiers [https://doi.org/10.1016/S0140-6736(16)30414-7] and [http://doi.org/10.7488/ds/1350]. The R code used to generate results in Sections 5 and 6 can be found on GitHub https://github.com/XynthiaKavelaars/Bayesia

Ethics approval and consent to participate

Not applicable

Competing interests

The authors declare that they have no competing interests.

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.

References

[1] M. J. Betancourt and M. Girolami. *Hamiltonian Monte Carlo for Hierarchical Models*. 2013. DOI: 10.48550/ARXIV.1312.0906.

[2] M. Betancourt. *A Conceptual Introduction to Hamiltonian Monte Carlo*. 2017. DOI: 10.48550/ARXIV.1701.02434.

[3] S. Biswas, D. D. Liu, J. J. Lee, and D. A. Berry. "Bayesian clinical trials at the University of Texas MD Anderson cancer center". *Clinical Trials* 6.3 (2009), pp. 205–216.

[4] J. Braeken, F. Tuerlinckx, and P. De Boeck. "Copula Functions for Residual Dependency". en. *Psychometrika* 72.3 (2007), p. 393. ISSN: 1860-0980. DOI: 10.1007/s11336-007-9005-4. URL: https://doi.org/10.1007/s

[5] S. Chib. "Marginal Likelihood from the Gibbs Output". *Journal of the American Statistical Association* 90.432 (1995), pp. 1313–1321. DOI: 10.1080/01621459.1995.10476635.

[6] S. Chib and Y. Chen. "MCMC Methods for Fitting and Comparing Multinomial Response Models" (1998).

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

[8] B. Dai, S. Ding, G. Wahba, et al. "Multivariate bernoulli distribution". *Bernoulli* 19.4 (2013), pp. 1465–1483.

[9] D. Evans. "Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions". *Journal of Clinical Nursing* 12.1 (2003), pp. 77–84. DOI: 10.1046/j.1365-2702.2003.00662.x.

[10] J. M. Flegal, J. Hughes, D. Vats, N. Dai, K. Gupta, and U. Maji. *mcmcse: Monte Carlo Standard Errors for MCMC*. R package version 1.5-0. Riverside, CA, and Kanpur, India, 2021.

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

[12] Food and Drug Administration. *Non-Inferiority Clinical Trials to Establish Effectiveness: Guidance for Industry*. 2016.

[13] J. J. Forster. "Markov chain Monte Carlo exact inference for binomial and multinomial logistic regression models." *Statistics and Computing* 13.2 (2003), pp. 169–177. DOI: 10.1023/a:1023212726863.

[14] P. P. Gallo. "CENTER-WEIGHTING ISSUES IN MULTICENTER CLINICAL TRIALS". *Journal of Biopharmaceutical Statistics* 10.2 (2000), pp. 145–163. DOI: 10.1081/bip-100101019.

[15] A. Gelman. "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)". *Bayesian Analysis* 1.3 (2006), pp. 515–534. DOI: 10.1214/06-BA117A. URL: https://doi.org/10.1214/06-BA117A.

[16] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, and D. B. Dunson. *Bayesian Data Analysis*. Taylor & Francis Ltd., 2013. 675 pp. ISBN: 1439898200.

[17] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007. 648 pp. ISBN: 0521867061. URL: https://www.ebook.de/de/product/6522123/andrew_gelman_

[18] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. "A weakly informative default prior distribution for logistic and other regression models". *The Annals of Applied Statistics* 2.4 (2008), pp. 1360–1383. DOI: 10.1214/08-aoas191.

[19] H. Goldstein, W. Browne, and J. Rasbash. "Partitioning Variation in Multilevel Models". *Understanding Statistics* 1.4 (2002), pp. 223–231. DOI: 10.1207/s15328031us0104_02.

[20] R. Grol and J. Grimshaw. "From best evidence to best practice: effective implementation of change in patients' care". *The Lancet* 362.9391 (2003), pp. 1225–1230. DOI: 10.1016/s0140-6736(03)14546-1.

[21] M. Gulliford, G. Adams, O. Ukoumunne, R. Latinovic, S. Chinn, and M. Campbell. "Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data". *Journal of Clinical Epidemiology* 58.3 (2005), pp. 246–251. DOI: 10.1016/j.jclinepi.2004.08.012.

[22] M. A. Hamburg and F. S. Collins. "The Path to Personalized Medicine". *New England Journal of Medicine* 363.4 (2010), pp. 301–304. DOI: 10.1056/nejmp1006304.

[23] J. Hox, M. Moerbeek, and R. Van de Schoot. *Multilevel Analysis*. Taylor & Francis Ltd, 2017. 348 pp. ISBN: 1138121363.

[24] International Stroke Trial Collaborative Group. "The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke". *The Lancet* 349.9065 (1997), pp. 1569–1581. DOI: 10.1016/s0140-6736(97)04011-7.

[25] C. H. Jackson. "Multi-State Models for Panel Data: The msm Package for R". *Journal of Statistical Software* 38.8 (2011), pp. 1–29. URL: http://www.jstatsoft.org/v38/i08/.

[26] H. Jeffreys. *Theory of Probability*. Oxford, England: Clarendon Press, 1961.

[27] B. Jones, D. Teather, J. Wang, and J. A. Lewis. "A comparison of various estimators of a treatment difference for a multi-centre clinical trial". *Statistics in Medicine* 17.15-16 (1998), pp. 1767–1777. DOI: 10.1002/(sici)1097-0258(19980815/30)17:15/16<1767::aid-sim978>3.0.co;2-h.

[28] V. M. T. Jong, M. J. C. Eijkemans, B. Calster, D. Timmerman, K. G. M. Moons, E. W. Steyerberg, and M. Smeden. "Sample size considerations and predictive performance of multinomial logistic prediction models". *Statistics in Medicine* 38.9 (2019), pp. 1601–1619. DOI: 10.1002/sim.8063.

[29] M. Kaptein. "The use of Thompson sampling to increase estimation precision". *Behavior Research Methods* 47.2 (2014), pp. 409–423. DOI: 10.3758/s13428-014-0480-0.

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

[31] R. E. Kass and A. E. Raftery. "Bayes Factors". *Journal of the American Statistical Association* 90.430 (1995), pp. 773–795. DOI: 10.1080/01621459.1995.10476572.

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

[33] X. Kavelaars, J. Mulder, and M. Kaptein. "Decision-making with multiple correlated binary outcomes in clinical trials". *Statistical Methods in Medical Research* 29.11 (2020), pp. 3265–3277. DOI: 10.1177/0962280220922256.

[34] Z. Lin. "An issue of statistical analysis in controlled multi-centre studies: how shall we weight the centres?" *Statistics in Medicine* 18.4 (1999), pp. 365–373. DOI: 10.1002/(sici)1097-0258(19990228)18:4<365::aid-sim

[35] R. I. Lindley et al. "Alteplase for Acute Ischemic Stroke". *Stroke* 46.3 (2015), pp. 746–756. DOI: 10.1161/strokeaha.114.006573.

[36] E. Makalic and D. Schmidt. "High-Dimensional Bayesian Regularised Regression with the BayesReg Package". arXiv:1611.06649v3. 2016.

[37] H. J. Malik and B. Abraham. "Multivariate Logistic Distributions". *The Annals of Statistics* 1.3 (1973), pp. 588–590. DOI: 10.1214/aos/1176342430.

[38] M. Marsman and E.-J. Wagenmakers. "Three Insights from a Bayesian Interpretation of the One-Sided P Value". *Educational and Psychological Measurement* 77.3 (2016), pp. 529–539. DOI: 10.1177/0013164416669201.

[39] A. D. Martin, K. M. Quinn, and J. H. Park. "MCMCpack: Markov Chain Monte Carlo in R". *Journal of Statistical Software* 42.9 (2011), p. 22. URL: http://www.jstatsoft.org/v42/i09/.

[40] A. E. McGlothlin and K. Viele. "Bayesian Hierarchical Models". *JAMA* 320.22 (2018), p. 2365. DOI: 10.1001/jama.2018.17977.

[41] Microsoft and S. Weston. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.16. 2020. URL: https://CRAN.R-project.org/package=doParallel.

[42] Microsoft and S. Weston. *foreach: Provides Foreach Looping Construct*. R package version 1.5.1. 2020. URL: https://CRAN.R-project.org/package=foreach.

[43] R. Mirnezami, J. Nicholson, and A. Darzi. "Preparing for Precision Medicine". *New England Journal of Medicine* 366.6 (2012), pp. 489–491. DOI: 10.1056/nejmp1114866.

[44] M. Moerbeek, G. J. P. V. Breukelen, and M. P. F. Berger. "Optimal Experimental Designs for Multilevel Logistic Models". *Journal of the Royal Statistical Society: Series D (The Statistician)* 50.1 (2001), pp. 17–30. DOI: 10.1111/1467-9884.00257.

[45] M. Moerbeek, G. J. P. van Breukelen, and M. P. F. Berger. "Design Issues for Experiments in Multilevel Populations". *Journal of Educational and Behavioral Statistics* 25.3 (2000), p. 271. DOI: 10.2307/1165206.

[46] R. Moineddin, F. I. Matheson, and R. H. Glazier. "A simulation study of sample size for multilevel logistic regression models". *BMC Medical Research Methodology* 7.1 (2007). DOI: 10.1186/1471-2288-7-34.

[47] J. Mulder and J.-P. Fox. "Bayes Factor Testing of Multiple Intraclass Correlations". *Bayesian Analysis* 14.2 (2019). DOI: 10.1214/18-ba1115.

[48] J. Mulder, D. R. Williams, X. Gu, A. Tomarken, F. Böing-Messing, A. Olsson-Collentine, M. Meijerink, J. Menke, R. van Aert, J.-P. Fox, H. Hoijtink, Y. Rosseel, E.-J. Wagenmakers, and C. van Lissa. "BFpack: Flexible Bayes Factor Testing of Scientific Theories in R". *Journal of Statistical Software* 100.18 (2021), pp. 1–63. DOI: 10.18637/jss.v100.i18.

[49] T. A. Murray, P. F. Thall, and Y. Yuan. "Utility-based designs for randomized comparative trials with categorical outcomes". *Statistics in medicine* 35.24 (2016), pp. 4285–4305.

[50] S. Nemes, J. M. Jonasson, A. Genell, and G. Steineck. "Bias in odds ratios by logistic regression modelling and sample size". *BMC Medical Research Methodology* 9.1 (2009). DOI: 10.1186/1471-2288-9-56.

[51] P. C. Ng, S. S. Murray, S. Levy, and J. C. Venter. "An agenda for personalized medicine". *Nature* 461.7265 (2009), pp. 724–726. DOI: 10.1038/461724a.

[52] A. K. Nikoloulopoulos and D. Karlis. "Multivariate logit copula model with an application to dental data". *Statistics in Medicine* 27.30 (2008), pp. 6393–6406. DOI: 10.1002/sim.3449.

[53] P. C. O'Brien. "Procedures for comparing samples with multiple endpoints". *Biometrics* (1984), pp. 1079–1087.

[54] S. M. OBrien and D. B. Dunson. "Bayesian Multivariate Logistic Regression". *Biometrics* 60.3 (2004), pp. 739–746. DOI: 10.1111/j.0006-341x.2004.00224.x.

[55] A. Panagiotelis, C. Czado, and H. Joe. "Pair copula constructions for multivariate discrete data". *Journal of the American Statistical Association* 107.499 (2012), pp. 1063–1072.

[56] S. Paul, K. Saha, and U. Balasooriya. "An empirical investigation of different operating characteristics of several estimators of the intraclass correlation in the analysis of binary data". *Journal of Statistical Computation and Simulation* 73.7 (2003), pp. 507–523. DOI: 10.1080/0094965021000050883.

[57] M. Plummer, N. Best, K. Cowles, and K. Vines. "CODA: Convergence Diagnosis and Output Analysis for MCMC". *R News* 6.1 (2006), pp. 7–11. URL: https://journal.r-project.org/archive/.

[58] D. Poirier. "Jeffreys' prior for logit models". *Journal of Econometrics* 63.2 (1994), pp. 327–339. DOI: 10.1016/0304-4076(93)01556-2.

[59] N. G. Polson, J. G. Scott, and J. Windle. "Bayesian inference for logistic models using Pólya–Gamma latent variables". *Journal of the American statistical Association* 108.504 (2013), pp. 1339–1349.

[60] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2020. URL: https://www.R-project.org/.

[61] S. W. Raudenbush and A. S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*. SAGE PUBN, 2001. 512 pp. ISBN: 076191904X.

[62] S. W. Raudenbush and X. Liu. "Statistical power and optimal design for multisite randomized trials." *Psychological Methods* 5.2 (2000), pp. 199–213. DOI: 10.1037/1082-989x.5.2.199.

[63] M. S. Ridout, C. G. B. Demetrio, and D. Firth. "Estimating Intraclass Correlation for Binary Data". *Biometrics* 55.1 (1999), pp. 137–148. DOI: 10.1111/j.0006-341x.1999.00137.x.

[64] P. E. Rossi, G. M. Allenby, and R. McCulloch. *Bayesian Statistics and Marketing*. Wiley, 2005. DOI: 10.1002/0470863692.

[65] P. Sandercock, J. Wardlaw, R. Lindley, G. Cohen, and W. Whiteley. "The third International Stroke Trial (IST-3), 2000-2015. [dataset]". en (2016). DOI: 10.7488/DS/1350.

[66] P. A. Sandercock, M. Niewada, and A. Członkowska. "The International Stroke Trial database". *Trials* 12.1 (2011). DOI: 10.1186/1745-6215-12-101.

[67] W. C. M. Schimmel, E. Verhaak, P. E. J. Hanssens, X. M. Kavelaars, J. Mulder, M. C. Kaptein, M. M. Sitskoorn, and K. Gehring. "P01.06.B Interim results from CAR-Study B: An ongoing randomized trial on the effect of SRS or WBRT on cognitive performance in patients with 11-20 brain metastases". *Neuro-Oncology* 24.Supplement_2 (2022), pp. ii24–ii24. DOI: 10.1093/neuonc/noac174.078.

[68] W. C. Schimmel, E. Verhaak, P. E. Hanssens, K. Gehring, and M. 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 (2018), p. 218.

[69] N. J. Schork. "Personalized medicine: Time for one-person trials". *Nature* 520.7549 (2015), pp. 609–611. DOI: 10.1038/520609a.

[70] N. K. Schuurman, R. P. P. P. Grasman, and E. L. Hamaker. "A Comparison of Inverse-Wishart Prior Specifications for Covariance Matrices in Multilevel Autoregressive Models". *Multivariate Behavioral Research* 51.2-3 (2016), pp. 185–206. DOI: 10.1080/00273171.2015.1065398.

[71] R. Simon. "Clinical trials for predictive medicine: new challenges and paradigms". *Clinical Trials* 7.5 (2010), pp. 516–524. DOI: 10.1177/1740774510366454.

[72] T. A. B. Snijders. *Power and Sample Size in Multilevel Linear Models*. 2005. DOI: 10.1002/0470013192.bsa492.

[73] T. Sozu, T. Sugimoto, and T. Hamasaki. "Reducing unnecessary measurements in clinical trials with multiple primary endpoints". *Journal of biopharmaceutical statistics* 26.4 (2016), pp. 631–643.

[74] T. Sozu, T. Sugimoto, and T. Hamasaki. "Sample size determination in clinical trials with multiple co-primary binary endpoints." *Statistics in medicine* 29 (21 2010), pp. 2169–2179. ISSN: 1097-0258. DOI: 10.1002/sim.3972.

[75] T.-L. Su, E. Glimm, J. Whitehead, and M. Branson. "An evaluation of methods for testing hypotheses relating to two endpoints in a single clinical trial". *Pharmaceutical statistics* 11.2 (2012), pp. 107–117.

[76] P. F. Thall. "Bayesian cancer clinical trial designs with subgroup-specific decisions". *Contemporary Clinical Trials* 90 (2020), p. 105860. DOI: 10.1016/j.cct.2019.105860.

[77] The International Stroke Trial-3 Collaborative Group. "The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6 h of acute ischaemic stroke (the third international stroke trial [IST-3]): a randomised controlled trial". *The Lancet* 379.9834 (2012), pp. 2352–2363. DOI: 10.1016/s0140-6736(12)60768-5.

[78] S. Thomas and W. Tu. "Learning Hamiltonian Monte Carlo in R". *The American Statistician* 75.4 (2021), pp. 403–413. DOI: 10.1080/00031305.2020.1865198.

[79] D. Van Ravenzwaaij, R. Monden, J. N. Tendeiro, and J. P. A. Ioannidis. "Bayes factors for superiority, non-inferiority, and equivalence designs". *BMC Medical Research Methodology* 19.1 (2019). DOI: 10.1186/s12874-019-0699-7.

[80] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer, 2002. URL: http://www.stats.ox.ac.uk/pub/MASS4/.

[81] F. Vieira-Generoso, R. Leenders, D. McFarland, and J. Mulder. "A Bayesian actor-oriented multilevel relational event model with hypothesis testing procedures" (2023). DOI: 10.48550/ARXIV.2204.10676.

[82] K. Viele, S. Berry, B. Neuenschwander, B. Amzal, F. Chen, N. Enas, B. Hobbs, J. G. Ibrahim, N. Kinnersley, S. Lindborg, et al. "Use of historical control data for assessing treatment effects in clinical trials". *Pharmaceutical statistics* 13.1 (2014), pp. 41–54.

[83] J. Whitehead, M. Branson, and S. Todd. "A combined score test for binary and ordinal endpoints from clinical trials". *Statistics in medicine* 29.5 (2010), pp. 521–532.

[84] H. Wickham and E. Miller. *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. R package version 2.4.3. 2021. URL: https://CRAN.R-project.org/package=haven.

A Gibbs sampling procedure based on Pólya-Gamma expansion

A.1 Random effects model

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

$$p(\boldsymbol{\gamma}_j^q, \boldsymbol{\gamma}^q, \boldsymbol{\Sigma}^q | \boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\gamma}_j^q) p(\boldsymbol{\gamma}_j^q|\boldsymbol{\gamma}^q, \boldsymbol{\Sigma}^q) p(\boldsymbol{\gamma}^q) p(\boldsymbol{\Sigma}^q). \tag{19}$$

The multinomial logistic likelihood (Equation 2) can be expanded with a Pólya-Gamma auxiliary variable to suit a Gibbs sampling procedure. This expansion relies on the following equality [59]:

$$\begin{aligned} p((\boldsymbol{y}_j = \boldsymbol{h}^q)|\boldsymbol{\gamma}_j^q, \boldsymbol{\gamma}_j^{-q}, \boldsymbol{\omega}_j^q = & \frac{\exp\left(\boldsymbol{x}_{ji}\boldsymbol{\gamma}_j^q\right)}{\sum_{r=1}^{Q-1} \exp\left(\boldsymbol{x}_{ji}\boldsymbol{\gamma}_j^r\right) + 1}, \\ & \propto \exp\left[-\frac{1}{2}(\boldsymbol{\kappa}_j^q - \boldsymbol{\eta}_j^q)^T \boldsymbol{\Omega}_j^q (\boldsymbol{\kappa}_j^q - \boldsymbol{\eta}_j^q)\right], \end{aligned} \tag{20}$$

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

Equation 20 can be recognized as the kernel of a multivariate Gaussian likelihood of working variable $\boldsymbol{\kappa}_j^q$ [59]:

$$\boldsymbol{\kappa}_j^q \sim N\left(\boldsymbol{\eta}_j^q, \{\boldsymbol{\Omega}_j^q\}^{-1}\right) \tag{21}$$

Here, $\boldsymbol{\Omega}_j^q$ reflects the diagonal matrix of Pólya-Gamma distributed variables $\boldsymbol{\omega}_j^q = (\omega_{j1}^q, \ldots, \omega_{jn_j}^q)$. A Gibbs sampler can be constructed when the likelihood in Equation 21 is combined with multivariate normal prior distributions on random regression coefficients $\boldsymbol{\gamma}_j^q|\boldsymbol{\gamma}^q, \boldsymbol{\Sigma}^q$ and mean random regression coefficients $\boldsymbol{\gamma}^q$, and an inverse-Wishart prior distribution on covariance matrix $\boldsymbol{\Sigma}^q$:

$$\begin{aligned} \boldsymbol{\gamma}_j^q &\sim N(\boldsymbol{\gamma}^q, \boldsymbol{\Sigma}^q) \\ \boldsymbol{\gamma}^q &\sim N(\boldsymbol{g}^q, \boldsymbol{G}^q) \\ \boldsymbol{\Sigma}^q &\sim \mathcal{W}^{-1}(j^0, \boldsymbol{S}^q) \end{aligned} \tag{22}$$

The resulting Gibbs sampler consists of the following steps:

1. Sample mean regression coefficients:

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

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

2. Sample covariance matrices of regression coefficients:

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

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

3. For each j, sample random regression coefficients:

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

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

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

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

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

A.1.1 Deriving the likelihood function

The following equality forms the basis to rewrite the multinomial likelihood in Equation 2 as a Gaussian likelihood [59]:

$$
\begin{aligned}
p((\boldsymbol{y}_j = \boldsymbol{h}^q)|\boldsymbol{\gamma}_j, \boldsymbol{\omega}_j^q, \boldsymbol{x}_j) &= \frac{\exp\left(\boldsymbol{x}_{ji}\boldsymbol{\gamma}_j^q\right)}{\sum_{r=1}^{Q-1} \exp\left(\boldsymbol{x}_{ji}\boldsymbol{\gamma}_j^r\right) + 1}, \\
&= \prod_{i=1}^{n_j} 2 \, \exp\left[\kappa_{ji}^q \omega_{ji}^q \eta_{ji}^q\right] \int_0^\infty \exp\left[\frac{-\omega_{ji}^q (\eta_{ji}^q)^2}{2}\right] p(\omega_{ji}^q) d\omega_{ji}^q
\end{aligned}
\tag{23}
$$

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

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

$$
\begin{aligned}
p((\boldsymbol{y}_j = \boldsymbol{h}^q)|.) &= \prod_{i=1}^{n_j} 2 \, \exp\left[\kappa_{ji}^q \omega_{ji}^q \eta_{ji}^q\right] \int_0^\infty \exp\left[\frac{-\omega_{ji}^q (\eta_{ji}^q)^2}{2}\right] p(\omega_{ji}^q) d\omega_{ji}^q \\
&\propto \exp\left[\frac{1}{2}(\boldsymbol{\kappa}_j^q \boldsymbol{\omega}_j^q \boldsymbol{\eta}_j^q - \boldsymbol{\omega}_j^q (\boldsymbol{\eta}_j^q)^2\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\kappa}_j^q - \boldsymbol{\eta}_j^q\right)^T \boldsymbol{\Omega}_j^q \left(\boldsymbol{\kappa}_j^q - \boldsymbol{\eta}_j^q\right)\right],
\end{aligned}
\tag{24}
$$

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

$$
\boldsymbol{\kappa}_j^q \sim N\left(\boldsymbol{\eta}_j^q, \{\boldsymbol{\Omega}_j^q\}^{-1}\right). \tag{25}
$$

A.1.2 Deriving conditional posterior distributions

Random regression coefficients $\boldsymbol{\gamma}_j^q$ Using the likelihood in Equation 25 and prior distribution $\boldsymbol{\gamma}_j^q \sim N(\boldsymbol{\gamma}^q, \{\boldsymbol{\Sigma}^q\})$, the conditional posterior distribution of random regression coefficients $\boldsymbol{\gamma}_j^q$ is also a multivariate

normal distribution:

$$
\begin{aligned}
p(\boldsymbol{\gamma}_j^q|.) \propto & p(\boldsymbol{y}_j|\boldsymbol{\gamma}_j^q, \boldsymbol{\gamma}_j^{-q}, \boldsymbol{\omega}_j^q, \boldsymbol{x}) p(\boldsymbol{\gamma}_j^q) \\
& \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\kappa}_j^q - (\boldsymbol{\eta}_j^q)\right)^T \boldsymbol{\Omega}_j^q \left(\boldsymbol{\kappa}_j^q - (\boldsymbol{\eta}_j^q)\right)\right] \times \\
& \quad \exp\left[-\frac{1}{2}\left(\boldsymbol{\gamma}_j^q - \boldsymbol{\gamma}^q\right)^T \{\boldsymbol{\Sigma}^q\}^{-1} \left(\boldsymbol{\gamma}_j^q - \boldsymbol{\gamma}^q\right)\right] \\
& \propto \exp\Bigg[-\frac{1}{2}\Big(\{\boldsymbol{\gamma}_j^q\}^T(\{\boldsymbol{X}_j\}^T \boldsymbol{\Omega}_j^q \boldsymbol{X}_j + \{\boldsymbol{\Sigma}^q\}^{-1})\boldsymbol{\gamma}_j^q - 2\{\boldsymbol{\gamma}_j^q\}^T \\
& \qquad (\{\boldsymbol{X}_j\}^T \boldsymbol{\Omega}_j^q (\boldsymbol{\kappa}_j^q + \ln[\sum_{m \neq q} \exp(\boldsymbol{X}_j \boldsymbol{\gamma}_j^m)]) + \{\boldsymbol{\Sigma}^q\}^{-1}\boldsymbol{\gamma}^q)\Big)\Bigg] \\
& \propto \exp\Bigg[-\frac{1}{2} \\
& \qquad \left(\boldsymbol{\gamma}_j^q - \boldsymbol{V}_{\boldsymbol{\gamma}_j}^q (\{\boldsymbol{X}_j\}^T \boldsymbol{\Omega}_j^q (\boldsymbol{\kappa}_j^q + \ln[\sum_{m \neq q} \exp(\boldsymbol{X}_j \boldsymbol{\gamma}_j^m)]) + \{\boldsymbol{\Sigma}^q\}^{-1}\boldsymbol{\gamma}^q)\right)^T \\
& \qquad\quad \{\boldsymbol{V}_{\boldsymbol{\gamma}_j}^q\}^{-1} \\
& \qquad \left(\boldsymbol{\gamma}_j^q - \boldsymbol{V}_{\boldsymbol{\gamma}_j}^q (\{\boldsymbol{X}_j\}^T \boldsymbol{\Omega}_j^q (\boldsymbol{\kappa}_j^q + \ln[\sum_{m \neq q} \exp(\boldsymbol{X}_j \boldsymbol{\gamma}_j^m)]) + \{\boldsymbol{\Sigma}^q\}^{-1}\boldsymbol{\gamma}^q)\right)\Bigg] \\
& \sim N\left(\boldsymbol{V}_{\boldsymbol{\gamma}_j}^q (\boldsymbol{X}_j \boldsymbol{\Omega}_j^q (\boldsymbol{\kappa}_j^q + \ln[\sum_{m \neq q} \exp(\boldsymbol{X}_j \boldsymbol{\gamma}_j^m)]) + \{\boldsymbol{\Sigma}^q\}^{-1}\boldsymbol{\gamma}^q), \boldsymbol{V}_{\boldsymbol{\gamma}_j}^q\right)
\end{aligned}
\tag{26}
$$

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

Random mean $\boldsymbol{\gamma}^q$ When the posterior distribution of $\boldsymbol{\gamma}_j^q$ (Equation 26) is included as a likelihood and combined with a $N(\boldsymbol{g}^q, \{\boldsymbol{G}^q\}^{-1})$ prior distribution, the conditional posterior distribution of random mean

$\boldsymbol{\gamma}^q$ is another multivariate normal distribution:

$$
\begin{aligned}
p(\boldsymbol{\gamma}^q|.) \propto & \prod_{j=1}^{J} p(\boldsymbol{\gamma}_j^q|\boldsymbol{\gamma}^q, \boldsymbol{\Sigma}^q) p(\boldsymbol{\gamma}^q) \\
\propto & \prod_{j=1}^{J} \exp\left[-\frac{1}{2}(\boldsymbol{\gamma}_j^q - \boldsymbol{\gamma}^q)^T \{\boldsymbol{\Sigma}^q\}^{-1} (\boldsymbol{\gamma}_j^q - \boldsymbol{\gamma}^q)\right] \times \\
& \exp\left[-\frac{1}{2}(\boldsymbol{\gamma}^q - \boldsymbol{g}^q)^T \boldsymbol{G}^q (\boldsymbol{\gamma}^q - \boldsymbol{g}^q)\right] \\
\propto & \exp\left[-\frac{1}{2}(\{\boldsymbol{\gamma}^q\}^T \left(J\{\boldsymbol{\Sigma}^q\}^{-1}\right) \boldsymbol{\gamma}^q) - 2\{\boldsymbol{\gamma}^q\}^T \left(\{\boldsymbol{\Sigma}^q\}^{-1} \sum_{j=1}^{J} \boldsymbol{\gamma}_j^q\right)\right] \times \\
& \exp\left[-\frac{1}{2}\{\boldsymbol{\gamma}^q\}^T \boldsymbol{G}^q \boldsymbol{\gamma}^q - 2\{\boldsymbol{\gamma}^q\}^T \boldsymbol{G}^q \boldsymbol{g}^q\right] \\
\propto & \exp\left[-\frac{1}{2}\{\boldsymbol{\gamma}^q\}^T \left(J\{\boldsymbol{\Sigma}^q\}^{-1} + \boldsymbol{G}^q\right) \boldsymbol{\gamma}^q - \right. \\
& \left. 2\{\boldsymbol{\gamma}^q\}^T \left(\{\boldsymbol{\Sigma}^q\}^{-1} \sum_{j=1}^{J} \boldsymbol{\gamma}_j^q + \boldsymbol{G}^q \boldsymbol{g}^q\right)\right] \\
\propto & \exp\left[-\frac{1}{2}\left(\boldsymbol{\gamma}^q - \boldsymbol{V}_{\boldsymbol{\gamma}}^q \left(\{\boldsymbol{\Sigma}^q\}^{-1} \sum_{j=1}^{J} \boldsymbol{\gamma}_j^q + \boldsymbol{G}^q \boldsymbol{g}^q\right)\right)^T \{\boldsymbol{V}_{\boldsymbol{\gamma}}^q\}^{-1} \right. \\
& \left. \left(\boldsymbol{\gamma}^q - \boldsymbol{V}_{\boldsymbol{\gamma}}^q \left(\{\boldsymbol{\Sigma}^q\}^{-1} \sum_{j=1}^{J} \boldsymbol{\gamma}_j^q + \boldsymbol{G}^q \boldsymbol{g}^q\right)\right)\right] \\
\sim & N\left(\boldsymbol{V}_{\boldsymbol{\gamma}}^q \left(\{\boldsymbol{\Sigma}^q\}^{-1} \sum_{j=1}^{J} \boldsymbol{\gamma}_j^q + \boldsymbol{G}^q \boldsymbol{g}^q\right), \boldsymbol{V}_{\boldsymbol{\gamma}}^q\right),
\end{aligned}
\tag{27}
$$

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

Random variance $\boldsymbol{\Sigma}^q$ When the posterior distribution of $\boldsymbol{\gamma}_j^q$ (Equation 26) is included as a likelihood and combined with an inverse Wishart $\mathcal{W}^{-1}(j^0, \boldsymbol{S}^q)$ prior, the conditional posterior distribution of random

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

$$
\begin{aligned}
p(\boldsymbol{\Sigma}^q|.) &\propto p(\boldsymbol{\gamma}_j^q|\boldsymbol{\gamma}^q, \boldsymbol{\Sigma}^q) p\{\boldsymbol{\Sigma}^q\} \\
&\propto \prod_{j=1}^{J} |\boldsymbol{\Sigma}^q|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{\gamma}_j^q - \boldsymbol{\gamma}^q)^T \{\boldsymbol{\Sigma}^q\}^{-1} (\boldsymbol{\gamma}_j^q - \boldsymbol{\gamma}^q)\right] \times \\
&\quad |\boldsymbol{\Sigma}^q|^{\frac{1}{2}(j^0 + p^R + 1)} \exp\left[-\frac{1}{2} tr(\boldsymbol{S}^q \{\boldsymbol{\Sigma}^q\}^{-1})\right] \\
&\propto |\boldsymbol{\Sigma}^q|^{-\frac{1}{2}(j^0 + J + P^R + 1)} \times \\
&\quad \exp\left[-\frac{1}{2} tr\left(\left(\boldsymbol{S}^q + \sum_{j=1}^{J} (\boldsymbol{\gamma}_j^q - \boldsymbol{\gamma}^q)(\boldsymbol{\gamma}_j^q - \boldsymbol{\gamma}^q)^T\right) \{\boldsymbol{\Sigma}^q\}^{-1}\right)\right] \\
&\sim \mathcal{W}^{-1}\left(j^0 + J, \boldsymbol{S}^q + \sum_{j=1}^{J} (\boldsymbol{\gamma}_j^q - \boldsymbol{\gamma}^q)(\boldsymbol{\gamma}_j^q - \boldsymbol{\gamma}^q)^T\right).
\end{aligned}
\tag{28}
$$

A.2 Mixed effects model

A mixed effect model is defined as follows:

$$
\boldsymbol{\phi}_{ji}^q = f(\boldsymbol{x}_{ji}^F \boldsymbol{\beta}^q + \boldsymbol{x}_{ji}^R \boldsymbol{\gamma}_j^q) \tag{29}
$$

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

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

$$
\boldsymbol{\kappa}_j^q \sim N\left(\boldsymbol{\eta}_j^q, \{\boldsymbol{\Omega}_j^q\}^{-1}\right). \tag{30}
$$

Here, $\boldsymbol{\eta}_j^q = \boldsymbol{X}_j^F \boldsymbol{\beta}^q + \boldsymbol{X}_j^R \boldsymbol{\gamma}_j^q - \ln[\sum_{m \neq q} \exp(\boldsymbol{X}_j^F \boldsymbol{\beta}^m + \boldsymbol{X}_j^R \boldsymbol{\gamma}_j^m)]$. The likelihood in Equation 30 can be combined with the prior distributions in Equation 22, complemented with a multivariate normally distributed prior on $\boldsymbol{\beta}^q$:

$$
\boldsymbol{\beta}^q \sim N(\boldsymbol{b}^q, \boldsymbol{B}^q) \tag{31}
$$

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

1. Sample fixed regression coefficients:

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

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

2. Sample mean random regression coefficients:

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

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

3. Sample covariance matrices of random regression coefficients:

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

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

4. For each j, sample random regression coefficients:

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

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

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

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

A.3 A note on prior specification

A.3.1 Regression parameters

In the Gibbs sampling framework, regression coefficients are normally distributed with a mean and covariance matrix. We shortly discuss the role of these parameters below. The covariance matrix defines the spread of the distribution and therefore has a substantial influence on informativity: Small variance parameters increase prior information. When non-informativity is preferable, large variance parameters are not the simple answer, as they may destabilize computations in Bayesian logistic regression analysis [18]. Jeffreys's prior could be an option, but sufficiently stable computation is not guaranteed [58, 18]. The challenge is therefore to specify prior variance parameters that are both sufficiently small to support stable analysis and to give a realistic support of the parameter and at the same time sufficiently large to be considered vague.

The mean hyperparameters defines the center of the distribution and becomes increasingly influential on the posterior distribution when the variance of the distribution is small. The relevance of adequate mean hyperparameters therefore increases with the informativity of the analysis. It should be noted that prior information of mean regression coefficients is not always available in the required parametrization. Researchers may be more likely to have information available in terms of (success) probabilities rather than logistic regression parameters. Kavelaars et al. propose an approach to compute mean hyperparameters for the context of treatment comparison in the presence of a single patient characteristics, based on expected joint response probabilities [32].

A.3.2 Covariance matrices

The covariance matrix follows an inverse-Wishart distribution with parameters. Specifying a non-informative prior on covariance matrices and variance parameters in general is not straightforward [15, 70]. The informativity of the inverse-Wishart distribution is sensitive to the size of variance parameters: small variances make inverse-Wishart distributions more informative. Naively specifying standard prior hyperparameters without consideration of prior information or data at hand may result in an undesirably large prior influence. Weakly informative (data-based) prior specification may be superior, if not essential for computational stability [15].

B Procedure for transformation to the probability scale and decision-making

Algorithm 1 Procedure for statistical decision-making with posterior regression coefficients

1: **Step 1. Transform regression coefficients to treatment differences**

Let $\boldsymbol{\gamma}_j^Q = (0, \ldots, 0)$ and $\boldsymbol{x} = (1, T, w, \ldots)$

for draw $(l) \leftarrow 1 : L$ **do**

  for cluster $j \leftarrow 1 : J$ **do**

    Compute joint response probabilities

    for treatment $T \leftarrow 0 : 1$ **do**

      for joint response category $q \leftarrow 1 : Q$ **do**

        if Population of interest defined by a range of values of w **then**

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

        end if

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

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

        end if

      end for

      Compute multivariate success probabilities

      for outcome $k \leftarrow 1 : K$ **do**

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

        Compute multivariate treatment difference

$$\text{Compute } \delta_j^{k(l)} = \theta_{1j}^{k(l)} - \theta_{0j}^{k(l)}$$

      end for

    end for

  end for

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

end for

end for

Step 2. Make superiority decision

Define superiority region $\mathcal{S}_R$

Draw conclusion

32: **if** $\frac{1}{L}\sum_{(l)=1}^{L} I(\boldsymbol{\delta}^{(l)} \in \mathcal{S}_R) > p_{cut}$ **then** Conclude superiority

33: **else** Conclude non-superiority

34: **end if**