Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

A Bayesian approach to multiple testing, analysis, and study design for response variables of mixed types

Ethan M. Alt*[1] | Matthew A. Psioda[2] | Joseph G. Ibrahim[2]

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, Massachusetts, USA

[2]Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina, USA

Correspondence
*Ethan M. Alt, 75 Francis Street, Boston, MA, USA. Email: ealt1@bwh.harvard.edu

Summary

There has been increased interest in the design and analysis of studies consisting of multiple response variables of mixed types. For example, in clinical trials, it is desirable to establish efficacy for a treatment effect in primary and secondary outcomes. In this paper, we develop Bayesian approaches for hypothesis testing and study planning for data consisting of multiple response variables of mixed types with covariates. We assume that the responses are correlated via a Gaussian copula, and that the model for each response is, marginally, a generalized linear model (GLM). Taking a fully Bayesian approach, the proposed method enables inference based on the joint posterior distribution of the parameters. Under some mild conditions, we show that the joint distribution of the posterior probabilities under any Bayesian analysis converges to a Gaussian copula distribution as the sample size tends to infinity. Using this result, we develop an approach to control the type I error rate under multiple testing. Simulation results indicate that the method is more powerful than conducting marginal regression models and correcting for multiplicity using Bonferroni-Holm Method. We also develop a Bayesian approach to sample size determination in the presence of response variables of mixed types, extending the concept of probability of success (POS) to multiple response variables of mixed types.

KEYWORDS:
bayesian inference, multiplicity, probability of success, copula

1 | INTRODUCTION

There has been increased interest in jointly analyzing data, particularly of mixed types. For example, in clinical trials, it is common to measure multiple response variables (i.e., clinical outcomes) per subject, which may be continuous, binary, or a count, among others. Although these outcomes are known to be dependent, mixed data types are typically analyzed using marginal models, i.e., ignoring dependence between the response variables. Hypothesis testing in the presence of multiplicity involves computing adjusted p-values from these marginal models. Existing approaches to account for multiple comparisons, such as the Bonferroni-Holm method[1], may have conservative type I error rates under positive dependence and, hence, suboptimal power. Consequently, studies utilizing marginal models require larger sample sizes to achieve a desired level of power than would be required if the correlations were taken into account.

In this paper, we propose to model responses of mixed types using a Gaussian copula to account for dependence between response variables. We focus specifically on the vector generalized linear model[2]. Because any copula may be utilized to model

[0]**Abbreviations:** APP, asymptotic power prior FWER, family-wise error rate; GC, Gaussian copula; GLM, generalized linear model HMC, Hamiltonian Monte Carlo .

the dependence between response variables, we refer to the model as the Gaussian copula generalized linear model (GCGLM) to avoid ambiguity. In order to incorporate uncertainty surrounding parameter values and because frequentist inference of the GCGLM is unstable in the presence of multiple discrete response variables[3,4], we focus on Bayesian inference. Importantly, the Bayesian joint modeling approach enables inference on the joint posterior distribution of the parameters of interest, and we show that the asymptotic joint distribution of posterior probabilities under the null hypothesis, and under some mild conditions, is itself a Gaussian copula when the model is correctly specified. Utilizing the joint asymptotic distribution, we develop a Bayesian approach to hypothesis testing that asymptotically guarantees a family-wise error rate (FWER) of precisely level α (i.e., the FWER is not conservative). We further develop a step-down procedure similar to the Bonferroni-Holm procedure to ascertain which hypotheses may be rejected.

We further show how the GCGLM may be utilized to determine sample size using a multivariate version of probability of success (POS)[5], which is also called assurance[6]. The proposed method enables practitioners to plan a study that yields, for example, a high probability to reject a null hypothesis corresponding to a parameter of primary interest *and* at least one null hypothesis corresponding to multiple parameters of secondary interest. In clinical trials, for example, regulators suggest that studies be planned to provide adequate power for tests of both the primary and secondary endpoints, across which the type I error rates must be controlled. However, this is seldom done in practice, in part due to the lack of familiarity of statistical tools that can jointly model response variables, which allow for the generation of future trial data of mixed types that are correlated.

The rest of the paper is organized as follows. In Section 2, we review some Frequentist and Bayesian approaches for the multiple comparisons problem. In Section 3, we introduce the GCGLM and discuss prior elicitation, where we develop priors for the GCGLM in the contexts of no prior information (i.e., a noninformative prior) and when a similar, previous study is available, using an approximation to the power prior[7], known as the "asymptotic power prior"[8]. In Section 4, we show that the asymptotic distribution of the posterior probabilities under the null hypothesis follows a Gaussian copula. Exploiting this fact, we develop a Bayesian approach to hypothesis testing that provides type I error control at exactly level α. In Section 5, we develop a method to compute probability of success (POS) to robustly determine the sample size in the presence of multiple response variables of mixed types. In Section 6.2, we present results from an extensive simulation study showing the type I error rates and power for our proposed analysis method and compare it with the Bonferroni-Holm procedure with analyses performed by marginal GLMs. In Section 6.1, we apply our POS method to determine sample size for a future trial using a real historical data set. In Section 8, we close with some discussion.

2 | REVIEW OF METHODS HANDLING MULTIPLICITY

In this section, we review the multiple testing literature. Because the aim of this paper is to account for correlations between outcomes (and, hence, dependence between tests), we focus mostly on methods that account for dependence between tests. Throughout this section, we assume a statistical model is parameterized by a J-dimensional vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1 = (\theta_{11}, \dots, \theta_{1K})'$ is a K-dimensional vector consisting of the parameters of interest and $\boldsymbol{\theta}_2$ is a vector of dimension $J-K$ containing nuisance parameters. We assume a hypothesis test of the form $H_0 : \cap_{k=1}^K H_{0k}$ versus $H_1 : \cup_{k=1}^K H_{1k}$, where $H_{0k} = \{\boldsymbol{\theta} : \theta_{1k} = \tau_k\}$ and $H_{1j} = \{\boldsymbol{\theta} : \theta_{1k} > \tau_k\}$ and we suppose without loss of generality that $\tau_k = 0$ for $k = 1, \dots, K$. In words, H_0 is a global null hypothesis stating that all parameters of interest are zero, and H_1 is a global alternative hypothesis stating that *at least one* of the parameters of interest is positive. Finally, we assume it is of interest not only to reject H_0, but also to ascertain which sub-hypotheses comprising H_0 may be rejected while controlling the FWER, which is the probability of incorrectly rejecting *at least one* null sub-hypothesis given that the global null hypothesis is true. We specify a type I error rate of level α. For frequentist approaches, we let T_k and p_k denote, respectively, the test statistic and the p-value for testing the null hypothesis H_{0k}, $k = 1, \dots, K$.

One of the most utilized methods for multiplicity adjustment is the Bonferroni-Holm procedure[1], which is advocated for use by regulators in clinical trials[9]. The Holm method is a step-down approach that first orders the p-values in ascending order, say, $p_{(1)}, \dots, p_{(K)}$. If $p_{(1)} \geq \alpha/K$, the procedure terminates and H_0 cannot be rejected. If $p_{(1)} < \alpha/K$, then $H_{0(k)}$ and hence H_0 are rejected. The procedure continues, testing the null hypothesis $H_{0(k)}$ only if $H_{0(1)}, \dots H_{0(k-1)}$ are rejected and rejects $H_{0(k)}$ if $p_{(k)} < \alpha/(K-k+1)$. The Holm procedure is uniformly more powerful (UMP) compared to the ubiquitous Bonferroni procedure, which tests each null hypothesis at level α/K. However, the Holm procedure ignores any correlations in the p-values, and can be quite conservative when the parameters in the global null hypothesis exhibit positive dependence (see Figure 1 in the Supporting Information for an illustration).

In order to reduce the conservatism of the Holm procedure, step-down approaches utilizing resampling methods (e.g., a bootstrap approach) have been proposed. One of the more popular methods, which we refer to as the Westfall-Young method,[10] resampling the test statistics under H_0 to estimate correlation under the null. However, the method relies on an assumption called "subset pivotality," i.e., that the distributions $\max_{k\in I} T_k|H_0^{(I)}$ and $\max_{k\in I} T_k|H_0$ are identical for every subset $I \subset \{1,\dots,K\}$, where $H_0^{(I)} = \cap_{k\in I} H_{0k}$. The Westfall-Young method has been extended to allow for weaker assumptions[11]. Both methods lead to a single decision rule (i.e., reject or do not reject), but the strength of evidence cannot be quantified. To improve on this, an algorithm was developed to compute adjusted p-values based on the Westfall-Young method[12]. In addition, an empirical Bayes (EB) resampling approach for false discovery rate (FDR) control was proposed[13]. In simple terms, resampling methods attempt to approximate the joint distribution of the test statistics under the global null hypothesis H_0, finding a critical value that depends on the correlation of the test statistics. However, this process is somewhat counterintuitive since dependence is ignored in the modeling stage, ultimately to try to recapture information on dependence in a post hoc fashion.

Recently, some approaches have focused on utilizing joint models to account for correlations in outcomes directly. Factor analysis methods have been used to model correlations for the general linear model[14]. However, this method is not utilizable when one is in possession of outcomes of mixed types. A Bayesian model averaging (BMA) approach has also been proposed[15]. The BMA approach takes a weighted average of p-values and is shown to be UMP over the Bonferroni-Holm procedure. However, such an approach is a hybrid approach, taking frequentist p-values and casting them in a Bayesian light. Moreover, the approach requires the elicitation of prior model probabilities, which may be subjective.

In addition to frequentist approaches, there have been advances in Bayesian approaches to multiplicity, which typically first involve estimating a joint posterior distribution for the parameters of interest. A hierarchical Dirichlet process prior has been proposed to account for correlations between parameters[16]. However, this method was only developed for normal response variables. On the other hand, some suggest a shared random effect between outcomes of mixed types with the random effect entering on the linear predictor scale[17], but noted doing so can make it difficult or impossible to compare a resulting joint model with the marginal models. Finally, a Bayesian approach for the multivariate normal model has been proposed using the Westfall-Young method[18]. However, a general method for both jointly analyzing data of mixed types that provides an approach to control FWER has, to our knowledge, not yet been developed. This paper fills this gap by (1) proposing a Bayesian joint modeling approach that enables inference on outcomes of mixed types; and (2) deriving the asymptotic joint distribution of the posterior probabilities under the global null hypothesis H_0, exploiting the asymptotic distribution to develop a step-down procedure that asymptotically guarantees FWER control at precisely level α.

3 | BAYESIAN INFERENCE FOR GAUSSIAN COPULA GENERALIZED LINEAR MODELS

In this section, we review the Gaussian copula generalized linear model (GCGLM), which is a joint model for response variables of mixed types. We discuss prior elicitation in the context of a single or two data set settings.

3.1 | The GCGLM

Suppose we possess a data set where $\boldsymbol{J}$ response variables are recorded per individual. That is, the data are $D = \{(\boldsymbol{y}_i, \boldsymbol{x}_i), i = 1,\dots,n\}$, where $\boldsymbol{y}_i = (y_{i1},\dots y_{ij})'$ is a $\boldsymbol{J}$-dimensional vector of response variables and $\boldsymbol{x}_i = (x_{i1},\dots,x_{ip})'$ is a p-dimensional vector consisting of all covariates measured for subject i, . We assume that, marginally, y_{ij} is an observation from a generalized linear model (GLM), i.e.,

$$f_j(y_{ij}|\boldsymbol{\beta}_j, \phi_{ij}) \propto \exp\left\{\frac{1}{a(\phi_{ij})}\left[y_{ij}\theta_{ij} - b_j(\theta_{ij})\right] + c_j(y_{ij}, \phi_{ij})\right\}, \tag{1}$$

where $\theta_{ij} = \theta_j(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_j)$ and θ_j is the θ-link function for model j, $\boldsymbol{x}_{ij}$ is a p_j-dimensional covariate vector for subject i and outcome j, which may include an intercept term and some or all components of $\boldsymbol{x}_i$, $\boldsymbol{\beta}_j$ is a p_j-dimensional vector of regression coefficients for outcome j, ϕ_{ij} is a dispersion parameter (which may be fixed for some models), and the functions b_j and c_j index the particular density or mass function for the GLM. For notational convenience, we assume $a(\phi_{ij}) = \phi_{ij} = \phi_j$ for $i = 1,\dots,n$.

Let $\boldsymbol{y}_i = (y_{i1},\dots,y_{iJ})'$ denote the $\boldsymbol{J}$-dimensional response vector from individual i, $i = 1,\dots,n$ and let $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j, \phi_j)', j = 1,\dots,\boldsymbol{J}$. Although GLMs include a wide range of data types, e.g., continuous (normal and gamma), binary (Bernoulli), and count (Poisson), we cannot write the joint likelihood as a product of marginal likelihoods because the components of $\boldsymbol{y}_i$ are correlated. In order to account for within-subject correlation, we utilize a Gaussian copula regression approach[2]. Specifically,

let $F_{ij}(\cdot) = F_j(\cdot|\boldsymbol{x}_{ij}, \boldsymbol{\theta}_j)$ denote the cumulative distribution function (CDF) for outcome j of subject i, where F_j denotes the CDF based on the j^{th} outcome (e.g., the Bernoulli CDF if the j^{th} outcome is binary). Note that the CDF F_{ij} depends on the covariate vector $\boldsymbol{x}_{ij}$ and the parameters of the j^{th} regression model $\boldsymbol{\theta}_j$. Using a Gaussian copula, the joint CDF for the i^{th} subject is given by

$$F(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}; \boldsymbol{\Gamma}) = C\{F_{i1}(y_{i1}), \dots, F_{iJ}(y_{iJ})|\boldsymbol{\Gamma}\},$$

where

$$C(u_1, \dots, u_J|\boldsymbol{\Gamma}) = \Phi_J\{\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_J)|\boldsymbol{\Gamma}\}, \tag{2}$$

$\Phi_J(\cdot|\boldsymbol{\Gamma})$ is the J-dimensional multivariate Gaussian CDF with zero mean and correlation matrix $\boldsymbol{\Gamma}$ and $\Phi^{-1}(\cdot)$ is the standard normal quantile function. When all J outcomes are continuous, the density function of the Gaussian copula is[2] $c(\boldsymbol{u}|\boldsymbol{\Gamma}) = |\boldsymbol{\Gamma}|^{-J/2} \exp\left\{-\frac{1}{2}\boldsymbol{v}'(\boldsymbol{\Gamma}^{-1} - \boldsymbol{I})\boldsymbol{v}\right\}$, where $\boldsymbol{v} = (\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_J))'$. The joint likelihood when all response variables are continuous, obtained by partial differentiation, is given by

$$L(\boldsymbol{\theta}, \boldsymbol{\Gamma}|D) = \prod_{i=1}^{n} c(F_{i1}(y_{i1}), \dots, F_{iJ}(y_{iJ})|\boldsymbol{\Gamma}) \prod_{j=1}^{J} f_{ij}(y_{ij}),$$

where $f_{ij}(\cdot) = f_j(\cdot|\boldsymbol{x}_{ij}, \boldsymbol{\theta}_j)$ is the density function corresponding to F_{ij}. Frequentist inference is thus straightforward using optimization methods.

When at least one of the J outcomes is discrete, optimization is much more difficult. Suppose that the first J_1 of the J outcomes are continuous and the remaining $J_2 = J - J_1$ are discrete. Obtaining the mixed density function requires $n \times 2^{J_1}$ differences of the joint CDF (2), which is computationally prohibitive even for moderate values of J_1. Furthermore, evaluation of a multivariate normal CDF at a single data point $\boldsymbol{y}_i$ typically relies on Monte Carlo estimation when $J > 2$, and Monte Carlo estimation may be highly inaccurate.

3.2 | Bayesian inference for the GCGLM

Because of the difficulties in optimizing the likelihood of the GCGLM, a Bayesian latent variable approach has been developed to copula estimation[4,19], which we now review. Suppose initially that all J margins are discrete. The augmented likelihood of the GCGLM is given by

$$f(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\Gamma}) = \prod_{i=1}^{n} \left[c(\boldsymbol{u}_i|\boldsymbol{\Gamma}) \prod_{j=1}^{J} I(a_{ij} \leq u_{ij} < b_{ij}) \right], \tag{3}$$

where $a_{ij} = F_{ij}(y_{ij} - 1)$, $b_{ij} = F_{ij}(y_{ij})$, and $\boldsymbol{u} = (u_{i1}, \dots, u_{iJ}) \in (0, 1)^J$ are latent variables. Consider the transformation $v_j = \Phi^{-1}(u_j)$. It can be shown that the joint density under the transformation is

$$f(\boldsymbol{y}, \boldsymbol{v}|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\Gamma}) = \phi_J(\boldsymbol{v}|\boldsymbol{\Gamma}) \prod_{j=1}^{J} I(\Phi^{-1}(a_j) \leq v_j < \Phi^{-1}(b_j)),$$

where $\phi_J(\cdot|\boldsymbol{\Gamma})$ is the J-dimensional multivariate Gaussian density function with zero mean and correlation matrix $\boldsymbol{\Gamma}$. Thus, the joint density of $f(\boldsymbol{y}, \boldsymbol{v})$ is proportional to a truncated multivariate normal density, where the points of truncation depend on the observed data[19].

Suppose now that the some of the margins are continuous and others are discrete. We suppose without loss of generality that the first J_1 margins are continuous and the remaining $J_2 = J - J_1$ margins are discrete. Let $\boldsymbol{u}_{Ci} = (u_{i1}, \dots, u_{i,J_1})'$ and let $\boldsymbol{u}_{Di} = (u_{i,J_1+1}, \dots, u_{iJ})'$ so that $\boldsymbol{u}_i = (\boldsymbol{u}'_{Ci}, \boldsymbol{u}'_{Di})'$. Note that the components of $\boldsymbol{u}_{Ci}$ are not actually latent because the CDF F_{ij} is a one-to-one function, and $u_{ij} = F_{ij}(y_{ij})$ may be computed deterministically for $j = 1, \dots, J_1$. It has been shown[3] that the augmented likelihood in this case may be written as

$$f(\boldsymbol{y}, \boldsymbol{u}_D|\boldsymbol{\theta}, \boldsymbol{\Gamma}) \propto \prod_{i=1}^{n} \left\{ c(\boldsymbol{u}_i|\boldsymbol{\Gamma}) \left[\prod_{j=1}^{J_1} f_{ij}(y_{ij}) \right] \left[\prod_{j=J_1+1}^{J} I(a_{ij} \leq u_{ij} \leq b_{ij}) \right] \right\}, \tag{4}$$

where a_{ij} and b_{ij} are defined analogously to the fully discrete augmented likelihood (3).

3.3 | A default, noninformative prior for the GCGLM

We now discuss a default prior for the GCGLM that is noninformative. Specifically, we will utilize the conjugate prior for GLMs[20] for the regression coefficients, half-Cauchy priors for the dispersion parameters, and, for the correlation matrix, the "LKJ density"[21].

Unlike the normal linear model, it can be difficult to specify a noninformative prior for other GLMs. For example, a normal prior on the intercept of a logistic regression model induces a prior over the probability of response for the case where all covariates are zero. To illustrate, suppose $\beta_0 \sim N(0, 10^2)$. On the probability scale, the prior indicates that the response probability when all the covariates are zero is approximately 0 or 1.

As a solution to this issue, we assume that the regression coefficients between outcomes are independent a priori and utilize the conjugate prior for GLMs[20], which we refer to as the CI prior. The CI prior for the regression model of outcome j is given by

$$\pi_{\text{CI}}(\boldsymbol{\beta}_j|\tau_j, \lambda_j, \boldsymbol{\mu}_{0j}) \propto \prod_{i=1}^{n} \exp\left\{\lambda_j \tau_j \left[\mu_{0ij}\theta_{ij} - b_j(\theta_{ij})\right]\right\}, \tag{5}$$

where $\tau_j = \phi_j^{-1}$, $\boldsymbol{\mu}_{0j} = (\mu_{01j}, \ldots, \mu_{0nj})'$ is a prior prediction (or guess) for the mean response $E(\boldsymbol{y}_j)$, and $\lambda_j \in [0, 1]$ is a precision parameter that controls the level of influence the prior has on the posterior. If we wish for the prior to be noninformative, we may specify small values for λ_j, e.g., $\lambda_j = 0.10$ for $j = 1, \ldots, J$. For normal and Bernoulli models, we may elicit $\boldsymbol{\mu}_0 = 0\boldsymbol{J}_n$ and $\boldsymbol{\mu}_0 = 0.5\boldsymbol{J}_n$, respectively, where $\boldsymbol{J}_n$ is a n-dimensional vector of ones. For gamma and Poisson distributions, it is somewhat more difficult to choose a value for $\boldsymbol{\mu}_0$. However, $\boldsymbol{\mu}_0$ may be elicited on the basis of similar studies or expert opinion. For example, the average number of tumors per patient across cancers in a retrospective study was 3.3[22], so that we may elicit $\boldsymbol{\mu}_0 = 3.3\boldsymbol{J}_n$.

In addition, we require a prior for the dispersion parameters. We may assume that the dispersion parameters are independent a priori and specify a half-Cauchy prior on each dispersion parameter, which is given by

$$\pi_{\text{HC}}(\phi_j|\phi_{0j}, \xi_{0j}) \propto \left(1 + \left[\frac{\phi_j - \phi_{0j}}{\xi_{0j}}\right]^2\right)^{-1}, \phi_j > 0, j \in \mathcal{Q}, \tag{6}$$

where $\phi_{0j} \in \mathbb{R}$ is a location parameter, $\xi_{0j} > 0$ is a scale parameter, and $\mathcal{Q}$ contains the indices of the J outcomes that have random dispersion parameters. The prior in (6) is a proper prior whose expectation and variance do not exist. Many advantages of using the half-Cauchy prior for scale parameters in hyperpriors (i.e., priors for the hyperparameters) have been shown[23]. Although our prior is non-hierarchical, the prior in (6) guarantees posterior propriety and will not have meaningful influence in the posterior distribution of the dispersion parameters when ξ_{0j} is large and the sample size is not too small, sharing some of the desirable traits of using the half-Cauchy priors for hyperpriors. In the absence of prior information, we may take $\phi_{0j} = 0$ and $\xi_{0j} = 20$, which is approximately uniform over $(0, 10)$.

Finally, we must specify a prior for the correlation matrix of the Gaussian copula. We utilize the LKJ density, which is given by $\pi_{\text{LKJ}}(\boldsymbol{\Gamma}|\eta) \propto |\boldsymbol{\Gamma}|^{\eta-1}, \boldsymbol{\Gamma} \in \mathcal{P}_J^+$, where $\mathcal{P}_J^+$ is the space of J-dimensional positive definite correlation matrices and $\eta > 0$ is a hyperparameter. When $\eta = 1$, the density corresponds to a uniform prior over the space of positive definite correlation matrices. If $\eta \in (0, 1)$, the density has a trough at the identity matrix, favoring nonzero correlations. If $\eta \in (1, \infty)$, the density has a peak at the identity matrix, favoring independence. In the absence of prior knowledge about the correlations between the outcomes, taking $\eta = 1$ is reasonable.

For the default prior, we assume that the regression coefficients, dispersion parameters, and copula parameters are independent a priori. This gives the joint prior

$$\pi(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\Gamma}|\lambda, \boldsymbol{\mu}_0, \boldsymbol{\phi}) = \left[\prod_{j=1}^{J} \pi_{\text{CI}}(\boldsymbol{\beta}_j|\lambda_j, \boldsymbol{\mu}_{0j}, \phi_j)\right] \left[\prod_{k\in\mathcal{Q}} \pi_{\text{HC}}(\phi_k|\phi_{0k}, \xi_{0k})\right] \pi_{\text{LKJ}}(\boldsymbol{\Gamma}|\eta). \tag{7}$$

As described above, to allow the posterior to depend more on the data than the prior, it is advisable to take λ_j to be small (e.g., $\lambda_j = 0.10$), ξ_{0k} large (e.g., $\xi_{0k} = 20$), and $\eta = 1$. The prior in (7) is a noninformative proper prior, guaranteeing that the posterior density is proper.

3.4 | Prior elicitation with historical data

We now discuss prior elicitation for the GCGLM when we are in possession of historical data. Specifically, we will use an approximation to the power prior (PP) known as the asymptotic power prior (APP)[7,8].

Suppose that we possess historical data $D_0 = \{(y_{0ij}, \boldsymbol{x}_{0ij}), i = 1, \ldots, n_0, j = 1, \ldots, J\}$. The PP may be interpreted as a discounted likelihood of the historical data, i.e.,
$\pi_{\text{PP}}(\boldsymbol{\theta}, \boldsymbol{\Gamma}|a_0, D_0) \propto \left[L(\boldsymbol{\theta}, \boldsymbol{\Gamma}|D_0)\right]^{a_0} \pi_0(\boldsymbol{\theta}, \boldsymbol{\Gamma})$, where $a_0 \in [0, 1]$ controls for the level of influence that the historical data has on the posterior and where $\pi_0(\boldsymbol{\theta}, \boldsymbol{\Gamma})$ is referred to as an initial prior. While the PP is straightforward to implement in marginal models, it is unclear how it can be implemented in the augmented likelihood (4) due to the Gaussian copula density function, which is parameterized by a correlation matrix. Although $\boldsymbol{J}$ independent power priors could be specified (one for each marginal regression model), such a prior would ignore correlations between outcomes a priori, which is counterintuitive since the end goal is to jointly model the outcomes.

The PP converges to a multivariate normal density function with mean equal to the mode and covariance matrix equal to the inverse of the negative Hessian matrix based on the historical data D_0 [8]. Thus, we may utilize the APP, which is given by

$$\pi_{\text{APP}}(\boldsymbol{\theta}, \boldsymbol{\Gamma}|a_0, \boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0, \eta, \pi_0) \propto \left[\phi_q(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0)\right]^{a_0} \pi_{\text{LKJ}}(\boldsymbol{\Gamma}|\eta) \tag{8}$$

where $\boldsymbol{\theta}_0$ and $\boldsymbol{\Sigma}_0$ are, respectively, approximations to the mean and covariance of the regression parameters of the GCGLM based on the historical data and where $\phi_q(\cdot|\boldsymbol{m}, \boldsymbol{C})$ is the q-dimensional multivariate normal density function with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{C}$. The hyperparameters $(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0)$ may be obtained by drawing samples from the GCGLM of the historical data and computing the mean (or mode) and covariance matrix of the samples pertaining to the regression parameters.

Note that the APP (8) only depends on the correlation matrix through the initial prior π_0. Eliciting the power prior for only some parameters in a model is closely related to the "partial borrowing power prior" [8] since the prior is only informative for the regression parameters. The prior in (8) is thus an approximation to an asymptotic power prior with partial borrowing.

3.5 | Posterior inference in the GCGLM

We now discuss how posterior samples may be obtained for the GCGLM. We propose to use the efficient No U-Turn Sampler (NUTS) [24], which requires only the joint distribution of the parameters and latent variables up to a normalizing constant.

We may write the joint posterior distribution as

$$p(\boldsymbol{\theta}, \boldsymbol{\Gamma}, \boldsymbol{u}|\boldsymbol{y}) = f(\boldsymbol{y}, \boldsymbol{u}_D)\pi(\boldsymbol{\theta}, \boldsymbol{\Gamma}), \tag{9}$$

where $f(\boldsymbol{y}, \boldsymbol{u}_D)$ is the augmented likelihood (4) and where $\pi(\boldsymbol{\theta}, \boldsymbol{\Gamma})$ is a prior over the regression parameters and correlation matrix, which may be specified as the noninformative prior in (7) or, if historical data is available, the APP (8).

Posterior samples of (9) may be obtained via Markov chain Monte Carlo (MCMC). For example, [19] provide an efficient Gibbs sampling algorithm. However, since the latent variables are obtained utilizing the full conditional distributions, posterior samples under that algorithm exhibit high autocorrelation. Instead, one may utilize the NUTS algorithm [24], which is a Hamiltonian Monte Carlo (HMC) algorithm. Because HMC only depends on the joint posterior distribution (as opposed to the full conditionals), posterior samples are typically less autocorrelated than those of the Gibbs sampling algorithm.

4 | JOINT DISTRIBUTION OF POSTERIOR PROBABILITIES AND TYPE 1 ERROR CONTROL

In this section, we derive the asymptotic joint distribution of the posterior probabilities. In particular, we show that the asymptotic distribution is a Gaussian copula. We exploit this fact to develop a method to account for multiple comparisons that provides strict FWER control. Specifically, we develop a step-down procedure that controls FWER while allowing practitioners to determine which hypotheses in the global null hypothesis may be rejected.

4.1 | Asymptotic joint distribution of posterior probabilities

We now show that, under the global null hypothesis, the asymptotic joint distribution of the posterior probabilities is a Gaussian copula distribution. We state this formally in the following theorem:

Theorem 1. Suppose that, for a given model, we have parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$, where $\boldsymbol{\theta}_1$ is a K-dimensional vector giving the parameters of interest and $\boldsymbol{\theta}_2$ is a vector of nuisance parameters. Let the global null hypothesis be expressed as $H_0 : \cap_{k=1}^{K} H_{0k}$, where $H_{0k} = \{\boldsymbol{\theta} : \theta_{1k} = \delta_k\}$ and δ_k is a target value. We may assume without loss of generality that $\delta_k = 0$. Let $\pi_{H_0}^{(U)}(\boldsymbol{\theta}) \propto$

$\tilde{\pi}^{(v)}(\boldsymbol{\theta})1\{\boldsymbol{\theta}_1 \in \Theta_1\}$ be a validation prior[25] (also referred to as a sampling prior[26]) on $\boldsymbol{\theta}$, where $\tilde{\pi}^{(v)}(\boldsymbol{\theta})$ is a proper prior and where $\Theta_1 = \{\boldsymbol{\theta} : \theta_k = 0, k = 1, \ldots, K\}$. Assume that the data, $\boldsymbol{y}$, are generated from the prior predictive distribution of the validation prior, i.e., $f(\boldsymbol{y}) = \int f(\boldsymbol{y}|\boldsymbol{\theta})\pi^{(v)}_{H_0}(\boldsymbol{\theta})d\boldsymbol{\theta}$. Let $\pi^{(f)}(\boldsymbol{\theta})$ be a non-degenerate "fitting prior" (also referred to as an "analysis prior") that results in a proper posterior density given by $p\big(\boldsymbol{\theta}|\boldsymbol{y}, \pi^{(f)}\big) \propto L(\boldsymbol{\theta}|\boldsymbol{y})\pi^{(f)}(\boldsymbol{\theta})$. Let $\hat{\boldsymbol{\theta}}_1$ denote the MLE and let $\hat{\boldsymbol{\Sigma}}_1$ denote the corresponding inverse of the average Fisher information matrix evaluated at the MLE, i.e., $\hat{\boldsymbol{\Sigma}}_1 = [n^{-1}I(\hat{\boldsymbol{\theta}}_1)]^{-1}$. If under $f(\boldsymbol{y})$, $\hat{\boldsymbol{\Sigma}}_1 \xrightarrow{p} \boldsymbol{\Sigma}^*_1$ and if $\sqrt{n}\hat{\boldsymbol{\theta}}_1 \xrightarrow{\text{d}} N_K(\boldsymbol{0}, \boldsymbol{\Sigma}^*_1)$, then

$$(P_1, \ldots, P_K) \xrightarrow{d} \text{GC}(\boldsymbol{\Gamma}^*_1) \text{ as } n \to \infty, \tag{10}$$

where $P_k = P(\theta_k > 0|\boldsymbol{y})$, $\boldsymbol{\Gamma}^*_1$ is the correlation matrix corresponding to $\boldsymbol{\Sigma}^*_1$, and the notation $\xrightarrow{\text{d}}$ denotes convergence in distribution.

In words, Theorem 1 says, under some relatively mild conditions, the asymptotic joint distribution of posterior probabilities converges in distribution to a Gaussian copula distribution. A formal proof of Theorem 1 is presented in Section A.1 of the appendix. In the sequel, we provide a heuristic argument for the proof.

By the Bernstein-von Mises theorem[27], we have that the posterior distribution of $\boldsymbol{\theta}_1$ is, approximately, a normal distribution with mean $\hat{\boldsymbol{\theta}}_1$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_1/n$. It follows that, for $k = 1, \ldots, K$, $P_k \approx \Phi(\sqrt{n}\hat{\theta}_k/\hat{\sigma}_{1k})$ where $\hat{\sigma}^2_{1k}$ is the k^{th} diagonal element of $\hat{\boldsymbol{\Sigma}}_1$ and where $\Phi(\cdot)$ is the univariate standard normal CDF. Hence, by the assumptions regarding the consistency of $\hat{\boldsymbol{\Sigma}}_1$ and the asymptotic normality of $\hat{\boldsymbol{\theta}}_1$, we have

$$(P_1, \ldots, P_K) \approx (\Phi(\sqrt{n}\hat{\theta}_1/\hat{\sigma}_{11}), \ldots, \Phi(\sqrt{n}\hat{\theta}_K/\hat{\sigma}_{1K})) \xrightarrow{\text{d}} \text{GC}(\boldsymbol{\Gamma}^*_1).$$

The conditions for Theorem 1 are relatively weak. For GLMs, it is well-known that, under H_0, $\sqrt{n}\hat{\boldsymbol{\theta}}$ converges to a normal distribution with mean $\boldsymbol{0}$ and asymptotic covariance matrix given by $\boldsymbol{\Omega}^*_1 = \lim_{n\to\infty}[n^{-1}I(\boldsymbol{0})]^{-1}$. Thus, the asymptotic normality of $\hat{\boldsymbol{\theta}}_1$ is ensured for GLMs. Second, for GLMs, $\hat{\boldsymbol{\theta}}_1$ converges in probability to $\boldsymbol{0}$ as $n \to \infty$ under H_0. Thus, $\hat{\boldsymbol{\Sigma}}_1$, which is the inverse of the average Fisher information matrix evaluated at $\hat{\boldsymbol{\theta}}_1$, converges in probability to $\boldsymbol{\Omega}^*_1$ by the continuous mapping theorem.

We note that under the special case that $K = 1$, we have $\hat{\boldsymbol{\theta}} = \hat{\theta}_1$ and $\Phi(\sqrt{n}\hat{\theta}_1/\hat{\sigma}_{11})$ is asymptotically uniformly distributed under the null hypothesis. Hence, Theorem 1 may be viewed as a generalization of the well-known result regarding the asymptotic distribution of a single posterior probability under the null hypothesis.

4.2 | FWER control

In this section, we describe how we may utilize the asymptotic distribution of the posterior probabilities (10) to develop a procedure that addresses multiple comparisons and provides strict FWER control. Unlike the previously discussed Frequentist approaches, the proposed method takes into account dependence between parameters, asymptotically resulting in a FWER of precisely α.

Suppose we wish to test $H_0 : \cap_{k=1}^K H_{0k}$ versus $H_1 : \cup_{k=1}^K H_{1k}$. The Holm method, reviewed in Section 2, may be conservative if the parameters exhibit positive dependence. Conversely, exploiting the joint distribution of the posterior probabilities, we may find, via simulation, the value γ^* such that $P(P_{(K)} > \gamma^*) \approx 1 - \gamma$, where $P_{(1)}, \ldots, P_{(K)}$ are the order statistics of $\boldsymbol{P} \sim \text{GC}(\boldsymbol{\Gamma})$. In general, the value γ^* will depend on both the number of parameters, K, and the estimated posterior correlation matrix, $\hat{\boldsymbol{\Gamma}}$. By contrast, the Holm procedure ignores dependence and only depends only on K. The global null hypothesis H_0 may then be rejected if the maximal posterior probability is larger than γ^*, i.e., if $\max_{1\leq j\leq K}\{P(\theta_{1j} \geq 0)\} \geq \gamma^*$.

This approach may be generalized to find out which hypotheses may be rejected in the alternative hypotheses $H_1 = \cup_{k=1}^K H_{1j}$. Let $\tilde{\boldsymbol{p}} = (\tilde{p}_1, \ldots, \tilde{p}_K)'$ denote the estimated posterior probabilities. Let $\boldsymbol{P} = (P_1, \ldots, P_K) \sim \text{GC}(\boldsymbol{\Gamma})$. For notational convenience, we assume without loss of generality that the components of $\boldsymbol{P}$ and $\tilde{\boldsymbol{p}}$ are in descending order. We reject H_{01} if $\tilde{p}_1 > \gamma^*_1$, where γ^*_1 solves $P(P_1 > \gamma^*_1) = 1-\gamma$. If H_{01} is not rejected, the procedure terminates. Otherwise, we construct $\boldsymbol{\Gamma}_2$, the $(K-1)\times(K-1)$ correlation matrix of $(P_2, \ldots, P_K)' \sim \text{GC}(\boldsymbol{\Gamma}_2)$ by removing the first row and first column of $\boldsymbol{\Gamma}$. We then reject H_{02} if $\tilde{p}_2 > \gamma^*_2$, where γ^*_2 solves $P(P_2 > \gamma^*) = 1-\gamma$. Proceeding inductively, if $H_{01} \ldots, H_{0,k-1}$ are rejected, we construct the $(K-k+1)\times(K-k+1)$ correlation matrix $\boldsymbol{\Gamma}_k$ by removing the first $k-1$ rows and columns from $\boldsymbol{\Gamma}$, rejecting the null hypothesis H_{0j} if $\tilde{p}_j > \gamma^*_k$, where γ^*_k solves $P(P_k > \gamma^*_k) = 1 - \gamma$. This step-down procedure is closely related to that of **(author?)** [1] where the evidence threshold changes at each iteration. However, unlike the Holm-Bonferroni procedure, the proposed method is based on the joint distribution of the posterior probabilities (which are correlated), rather than p-values from marginal models (which do not take

into account correlation between response variables). When the submatrices of $\boldsymbol{\Gamma}$ have positive elements, the threshold for the posterior probabilities will be smaller than those assuming independence. Hence, the proposed method is more powerful than using a Holm-like procedure applied directly to the posterior probabilities.

5 | MULTIVARIATE PROBABILITY OF SUCCESS WITH OUTCOMES OF MIXED TYPES

In this section, we discuss how to utilize the GCGLM to robustly determine sample size for a planned study, i.e., to compute a multivariate probability of success (POS), which is also referred to in the literature as assurance[6] and Bayesian expected power.

5.1 | The general methodology

We now review the concept of multivariate POS and show how the GCGLM may be utilized to determine a Bayesian sample size. In particular, we discuss how POS may be computed when it is desirable for a planned study to meet multiple success criteria.

To fix ideas, suppose we wish to plan a study having J endpoints based on possibly mixed types (e.g., continuous, binary, and count). Let the future data be represented by $D = \{(\boldsymbol{y}_i, \boldsymbol{x}_i), i = 1, \dots, n\}$, where $\boldsymbol{y}_i = (y_{i1}, \dots, y_{iJ})'$ is a J-dimensional vector of responses and $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})'$ is a p-dimensional vector consisting of all measured covariates for subject i . We assume that the joint distribution of $\boldsymbol{y}_i$ may be modeled by a GCGLM.

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_J)$, where $\boldsymbol{\theta}_j$ is a p_j-dimensional vector of regression parameters for outcome j, including a vector of regression coefficients $\boldsymbol{\beta}_j$ and possibly a dispersion parameter ϕ_j. Let $\Omega = \cap_{k=1}^{K} \Omega_k$ denote the study success criteria, where $\Omega_k = \cup_{l=1}^{L_k} \Omega_{kl}$. For example, suppose it is desirable to determine the sample size for a study that rejects the null hypothesis corresponding to the primary endpoint, denoted by β_{11}, and global null hypothesis corresponding to the two secondary endpoints, denoted respectively by β_{12} and β_{13}. Then $K = 2$, $L_1 = 1$, $L_2 = 2$, $\Omega_1 = \Omega_{11} = \{\boldsymbol{\beta} : \beta_{11} > 0\}$, and $\Omega_2 = \Omega_{21} \cup \Omega_{22}$, where $\Omega_{2j} = \{\boldsymbol{\beta} : \beta_{1j} > 0\}$, $j = 2, 3$.

Let $\pi^{(f)}(\boldsymbol{\theta}, \boldsymbol{\Gamma})$ denote a fitting prior, which is the prior that will be used to analyze the future trial data. Let $p(\boldsymbol{\theta}, \boldsymbol{\Gamma}|D, \pi^{(f)})$ denote the posterior density based on the likelihood in (4) with prior $\pi^{(f)}$. We compute the indicator function for whether there is sufficient evidence to claim that the treatment is efficacious based on the trial success criteria Ω, which is given by

$$1\{\text{success}|D\} = \prod_{k=1}^{K} 1\left\{ \max_{1 \le l \le L_k} P(\Omega_{kl}|D) \ge \gamma_k \right\}, \tag{11}$$

where, for $k = 1, \dots, K$, γ_k is an evidence threshold that controls the type I error rate based on Ω_k (e.g., in the example above, $\gamma_1 = 0.95$ and γ_2 is computed as described in Section 4).

Let $\pi^{(v)}(\boldsymbol{\theta}, \boldsymbol{\Gamma}, \boldsymbol{\alpha})$ denote a validation prior over the parameters governing the response and covariate distributions. We generate future data sets D via the predictive distribution with respect to $\pi^{(v)}$, i.e.,

$$f\left(\boldsymbol{y}, \boldsymbol{x}|\pi^{(v)}\right) = \int\int\int f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\Gamma}) f(\boldsymbol{x}|\boldsymbol{\alpha}) \pi^{(v)}(\boldsymbol{\theta}, \boldsymbol{\Gamma}, \boldsymbol{\alpha}) d\boldsymbol{\theta} d\boldsymbol{\Gamma} d\boldsymbol{\alpha}.$$

POS is defined as the marginal probability that the success criteria is satisfied with respect to the prior predictive distribution of the validation prior, i.e.,

$$\text{POS} = \int\int 1\{\text{success}|D\} f\left(\boldsymbol{y}, \boldsymbol{x}|\pi^{(v)}\right) d\boldsymbol{y} d\boldsymbol{x}. \tag{12}$$

5.2 | Multivariate probability of success with historical data

In this section, we assume that we possess possibly two historical data sets $D_{0k} = \{(\boldsymbol{y}_{0ki}, \boldsymbol{x}_{0ki}), i = 1, \dots, n_{0k}\}$, $k = 1, 2$. The historical data sets will be utilized to elicit the validation prior and, in the presence of two historical data sets, the fitting prior. We assume that the parameters for the response variables and the covariates are a priori independent, i.e., $\pi^{(v)}(\boldsymbol{\theta}, \boldsymbol{\Gamma}, \boldsymbol{\alpha}) = \pi_1^{(v)}(\boldsymbol{\theta}, \boldsymbol{\Gamma}) \pi_2^{(v)}(\boldsymbol{\alpha})$.

We utilize the more recent historical data set, denoted by D_{01}, to elicit the validation prior for the response variables. We assume that the responses of the historical data set are generated from a GCGLM. The validation prior for the response model

parameters is defined to be the posterior distribution with respect to the more recent historical data set D_{01}, i.e.,

$$\pi_1^{(v)}(\boldsymbol{\theta}, \boldsymbol{\Gamma}) \propto \left[\prod_{i=1}^{n_{01}} L(\boldsymbol{\theta}, \boldsymbol{\Gamma} | \boldsymbol{y}_{01i}, \boldsymbol{u}_{01i}) \right] \pi_{01}(\boldsymbol{\theta}, \boldsymbol{\Gamma}),$$

where $L(\boldsymbol{\theta}, \boldsymbol{\Gamma} | \boldsymbol{y}_{01i}, \boldsymbol{u}_{01i})$ is the augmented likelihood in (4) for subject i based on the data D_{01} and $\pi_{01}^{(v)}(\boldsymbol{\theta}, \boldsymbol{\Gamma})$ is an initial validation prior, which may be specified as the noninformative default joint prior in (7).

Because the covariates are dependent, it may be difficult to specify a validation prior for the covariate distribution. It has been proposed to utilize a factorization of the covariates[25]. For example, suppose we have p covariates $x_1, \dots, x_p$, where we suppose without loss of generality that we wish to generate x_1 first, then x_2, and so forth. Previous works have utilized a power prior for each conditional distribution[25], i.e., $\pi_2^{(v)}(\boldsymbol{\alpha}) = \left\{ \prod_{j=1}^{p} \prod_{k=1}^{2} \left[\prod_{i=1}^{n_{0k}} f(x_{0kij} | \tilde{\boldsymbol{x}}_{0kij}, \boldsymbol{\alpha}_j) \right]^{b_{0k}} \right\} \pi_{20}^{(v)}(\boldsymbol{\alpha})$, where $\tilde{x}_{0kij} = (x_{0ki1}, \dots, x_{0ki(j-1)})$ is a $j-1$-dimensional vector consisting of covariates preceding the j^{th} covariate in the order of generation, $\boldsymbol{\alpha}_j$ is the covariate vector for the j^{th} conditional distribution, and $b_{0k} \in [0, 1]$ is a power prior parameter, where values closer to 1 are specified when characteristics of participants in the planned trial are expected to be more similar to those in the historical data, and where $\pi_{20}^{(v)}(\boldsymbol{\alpha})$ is an initial prior for $\boldsymbol{\alpha}$. Note that if we are only in possession of one historical data set, we may set $b_{02} = 0$. While the validation prior $\pi_2^{(v)}$ is a valid distribution over the covariate parameters, it is not order invariant except in special cases. An alternative approach is to model the covariate distribution via a Gaussian copula. The primary advantages of this approach are that it is order invariant. Moreover, the correlations between covariates are explicitly modeled.

If we possess historical data, we may utilize the APP (8) to elicit the fitting prior $\pi^{(f)}(\boldsymbol{\theta}, \boldsymbol{\Gamma})$. Conversely, if no historical data are available, we may elicit the joint prior in (7) for the APP. Note that prior information (such as expert opinion) may be incorporated in the joint prior in (7) through the prior predictions $\boldsymbol{\mu}_0$, and informativeness may be adjusted through $\lambda_j, j = 1, \dots, J$.

5.3 | Computational development

We now provide an algorithm to compute POS. We also describe how to compute POS if, ultimately, a frequentist analysis of the trial data will be conducted. Finally, we provide a practical guide for how to compute POS in practice.

The algorithm to compute POS is summarized as follows.

1. Define n (the future trial sample size), α (the desired type I error rate), B (the number of future data sets to generate), M (the number of posterior samples), and $\Omega = \cap_{k=1}^{K} \Omega_k$ (the study success criteria).

2. Obtain a sample of size n from the prior predictive distribution of the validation prior (12), constructing future data $D = \{(\boldsymbol{y}_i, \boldsymbol{x}_i), i = 1, \dots, n\}$.

3. Using the fitting prior $\pi^{(f)}(\boldsymbol{\theta})$ and the data D from step (2), obtain a sample of size M from the posterior distribution $p(\boldsymbol{\theta} | D, \pi^{(f)})$. Record whether the future data D satisfy the trial success criteria, i.e., compute

$$1\{\text{success} | D\} = \prod_{k=1}^{K} 1 \left\{ \max_{1 \le l \le L_k} P(\Omega_{kl} | D) \ge \gamma_k \right\},$$

   where γ_k is determined by the method in Section 4.

4. Repeat steps (2)-(3) B times. POS is the the average of the B indicators from step (3).

We note that if a frequentist analysis of the data will ultimately be performed in the future trial, we may replace the indicator function in step (3) with $1\{\text{success} | D\} = \prod_{k=1}^{K} 1\{\hat{p}_k < \alpha_k\}$, where α_k is the desired type I error rate for test k and $\hat{p}_k$ is the (potentially adjusted) p-value associated with testing Ω_k. In the example described above, we have $\Omega_1 = \{\boldsymbol{\beta} : \beta_{11} > 0\}$ and $\Omega_2 = \cup_{j=2}^{3} \{\boldsymbol{\beta} : \beta_{1j} > 0\}$. Let $\hat{p}_{1j}$ denote the p-value for testing $H_0 : \beta_{1j} = 0$. Using the Holm procedure, we have $1\{\text{success} | D\} = 1\{\hat{p}_{11} < 0.05\} 1\{2 \min\{\hat{p}_{12}, \hat{p}_{13}\} \le 0.05\}$. Even if a frequentist analysis will ultimately be performed, the validation prior associated with the GCGLM is still useful, as it generates response variables of mixed types that are correlated. Thus, the method to compute multivariate POS is quite general, and it may be utilized to robustly determine the sample size even if there are constraints set on the analysis method (such as by regulators). The portion of the POS calculation which must be fully Bayesian is the validation prior, which has no frequentist analogue.

6 | DATA ANALYSIS, TYPE I ERROR RATE, AND POWER

In this section, we conduct an analysis of a data set from a clinical trial in multiple sclerosis (MS) obtained through the Vivli platform. We obtain posterior samples and posterior probabilities utilizing the proposed Bayesian approach and compare our proposed method with frequentist analogues assuming marginal models. We also conduct simulation studies to compare power and type I error rates of the proposed method to those based on a standard approach where multiplicity is corrected utilizing the method of Holm applied to p-values obtained from marginal GLMs.

6.1 | Data set analysis

The MOBILE study[28] was a phase II, randomized, double-blind, placebo-controlled clinical trial comparing prolonged release fampiridine against placebo in patients with progressive/relapse-remitting MS whose Expanded Disability Status Scale (EDSS) score was between 4.0-7.0. A total of 132 patients were randomized, of which 64 received placebo and 68 received the experimental treatment. We consider the same outcomes and endpoints in the phase III study, the ENHANCE study[29], which are slightly different than those for the MOBILE study. Unlike the phase III study, which utilized multiple imputation for missing data, for sake of simplicity, we conducted a complete case analysis, which resulted in a sample size of $n_0 = 127$.

The primary outcome ($\boldsymbol{y}_{01}$) of the phase III study is whether a patient achieved a mean reduction of at least 8 points from baseline on the multiple sclerosis walking scale (MSWS-12) over 24 weeks. The MSWS-12 score is based on a questionnaire consisting of 12 items, where each question is associated with a response ranging from 1 to 5 and the score is given by the sum of each of the 12 individual scores, which is then rescaled to be between 0 and 100. Higher values of the MSWS-12 score are more indicative of severe disease. One of the key secondary outcomes, ($\boldsymbol{y}_{02}$), is an indicator of whether the patient achieved a mean improvement from baseline of at least 15% of the timed up and go (TUG) score, which measures the time it takes for a patient from sitting to stand, walk a certain distance, and walk back to the chair. Another key secondary outcome, ($\boldsymbol{y}_{03}$), is the change from baseline in the Berg Balance Scale (BBS), measured as a continuous variable. The BBS is computed by taking the sum of 14 sub-scores, each ranging from 0 (cannot perform) to 4 (normal performance). A positive change of the BBS score indicates improvement.

All regression models included an intercept term and a treatment assignment indicator (0 = placebo, 1 = control). For $\boldsymbol{y}_{01}$, we assume a logistic regression model adjusting for baseline MSWS-12 score, baseline TUG score, age, and baseline EDSS score. We also assume a logistic regression model for $\boldsymbol{y}_{02}$, adjusting for baseline TUG score and screening EDSS score. For $\boldsymbol{y}_{03}$, we assume a normal linear model adjusting for the baseline values of EDSS and BBS. The phase III trial additionally included prior aminopyridine use in all models, but this covariate was not available in the phase II data. We write the vector of linear predictors for the j^{th} regression model (with response variable $\boldsymbol{y}_{0j}$) as $\boldsymbol{\eta}_j = \boldsymbol{J}\beta_{0j} + \boldsymbol{z}_0\beta_{1j} + \boldsymbol{X}'_{0j}\beta_{2j}$, where $\boldsymbol{J}$ is a n_0-dimensional vector of ones, β_{0j} is an intercept term, $\boldsymbol{z}_0$ is a n_0-dimensional vector of treatment indicators, β_{1j} is the treatment effect for outcome j, $\boldsymbol{X}_{0j}$ is the design matrix of the covariates for the j^{th} outcome, and $\boldsymbol{\beta}_{2j}$ is vector of the regression coefficients associated with the covariates. The primary effect of interest for the study is β_{11}, and β_{12} and β_{13} are the secondary effects of interest.

We elicited the proposed default prior (7) with $\lambda_j = 0.10$ for $j = 1, 2, 3$, $\boldsymbol{\mu}_{01} = \boldsymbol{\mu}_{02} = 0.50\boldsymbol{J}$, $\boldsymbol{\mu}_{03} = \boldsymbol{0}$, and $\eta = 1.0$. For the variance parameter of $\boldsymbol{y}_3$, denoted as σ_3^2, we elicited $\sigma_3^2 \sim \text{HC}(0.0, 20.0)$. Results of the data analysis are presented in Table 1. The joint Bayesian analysis and the marginal frequentist analyses yielded similar results. The GCGLM suggests that treatment could be more beneficial for the primary outcome (MSWS) and first secondary outcome (TUG) than the frequentist method, but it suggested less efficacy for the second secondary outcome (BBS).

	Bayesian			Frequentist		
Response	Post. Mean	Post. SD	Post. Prob.	MLE	SE	p-value
MSWS	0.86	0.40	0.9855	0.82	0.39	0.0182
TUG	0.71	0.40	0.9608	0.70	0.40	0.0405
BBS	1.53	0.82	0.9663	1.57	0.82	0.0285

TABLE 1 Comparison of the posterior distribution of the treatment effects for the GCGLM against frequentist analyses of marginal models. MSWS = Multiple Sclerosis Walking Scale, TUG = Timed Up and Go, BBS = Berg Balance Scale, Post. = posterior, SD = standard deviation, Prob. = probability, MLE = maximum likelihood estimate, SE = standard error.

The posterior mean of the correlations between the outcomes was $(\hat{\gamma}_{12}, \hat{\gamma}_{13}, \hat{\gamma}_{23}) = (0.29, 0.39, 0.33)$, indicating that the response variables are moderately positively correlated. The mean posterior correlation of the treatment effects corresponding to the secondary endpoints is 0.28, yielding an estimated threshold of $\gamma_{23} = 0.9736$. Note that the threshold γ_{23} is close to what a Holm-type method would utilize (e.g., 0.975 for two-way hypotheses and 0.983 for three-way hypotheses). In order to understand the benefit of the proposed joint Bayesian approach, we examine the power and type I error in simulation studies.

6.2 | Power and type I error rate

We now compare the power and FWER of the proposed Bayesian approach with a frequentist approach applying the Holm method to p-values of the marginal GLMs. We find that the FWER may be inflated in small samples, but otherwise the FWER attains its theoretical level of 0.05. Moreover, the results show that the proposed Bayesian approach is at least as powerful as the described Holm approach.

Let the phase II data be denoted by $D_0 = \{(\boldsymbol{y}_{0i}, \boldsymbol{x}_{0i}), i = 1, \dots, n_0\}$, where $\boldsymbol{y}_{0i} = (y_{0i1}, y_{0i2}, y_{0i3})'$ and $\boldsymbol{x}_{0i} = (x_{0i1}, \dots, x_{0i6})'$. The historical data will be utilized to elicit the validation prior for the covariate parameters $\boldsymbol{\alpha}$ and the regression parameters for the response variables $\boldsymbol{\theta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \boldsymbol{\beta}_3', \sigma_3^2)'$.

The covariates that we assume to be normally distributed include age, baseline multiple sclerosis impact scale (MSIS) score, and baseline EDSS score, represented respectively by x_{01}, x_{02}, and x_{03}. The three remaining covariates, the baseline values of MSWS, TUG, and BBS (denoted respectively by x_{04}, x_{05}, and x_{06}, are assumed to be log-normally distributed due to their respective skewness. The covariate x_{02} is not utilized in the regression models, but it is included due to its correlation with the other baseline measurements. The validation prior for the covariates is elicited as

$$\pi_2^{(v)}(\boldsymbol{\alpha}, \boldsymbol{\Omega}) \propto \pi_{\text{LKJ}}(\boldsymbol{\Omega}|1.0) \prod_{l=1}^{L} \left\{ \prod_{i=1}^{n_0} \left[g_l(x_{0il}|\mu_l, \sigma_l) c(\boldsymbol{u}_{0i}|\boldsymbol{\Omega}) \right] \pi_{0l}(\mu_l, \sigma_l) \right\},$$

where $g_l(\cdot)$ is the density for covariate l, μ_l and σ_l are, respectively, the location and scale parameters for density π_l, which is a normal distribution for $l = 1, 2, 3$ and the log-normal distribution for $l = 4, 5, 6$, $c(\cdot|\boldsymbol{\Omega})$ is the Gaussian copula density function with $L \times L$ correlation matrix $\boldsymbol{\Omega}$, $\boldsymbol{u}_i = (u_{i1}, \dots, u_{i6})$ is the vector of pseudo-latent variables, $u_{il} = G_l(x_{0il}|\boldsymbol{\alpha}_j)$, $i = 1, \dots, n_0$, $l = 1, \dots, L$, where G_l is the CDF associated with density g_l. Note that all 6 covariates here are continuous, so no latent variables must be generated in the sampling scheme. The initial prior was elicited independently as $\pi_{0l}(\mu_l, \sigma_l) \propto \pi_{\text{HC}}(\sigma_l|0.0, 20.0)$.

The validation prior for most of the regression parameters was taken to be the posterior density of the historical data utilizing the proposed Bayesian approach. A Bayesian version of power, known as Bayesian conditional expected power (BCEP)[30], was computed utilizing a validation prior that restricted the treatment effects to be positive. The validation prior for the type I error was the posterior distribution of the phase II data excluding the treatment assignment indicator as covariates (i.e., the marginal validation prior for the treatment effects is a point mass at $\boldsymbol{\beta}_1 = \boldsymbol{0}$). We generated future data sets with fixed levels of correlation, setting $\rho \in \{-0.40, -0.20, 0.0, 0.20, 0.40, 0.80\}$. The purpose of fixing the correlations is to accurately identify the the effect of correlations between outcomes (e.g., if uncertainty in correlations is permitted, the generated data sets might range from relatively low correlation to relatively high correlation).

In large samples with discrete response variables, obtaining a large amount of posterior samples can take a long time because there are nJ_d latent variables, where J_d is the number of discrete response variables. We thus utilize an asymptotic approximation to compute the posterior probabilities. Specifically, we obtain 2,000 samples after a 1,000 burn-in period, utilizing a normal CDF approximation via $P(\beta_{1j} > 0|D) \approx 1 - \Phi(0|\hat{\beta}_{1j}, \hat{\tau}_j)$ for $j = 1, 2, 3$, where $\hat{\beta}_{1j}$ and $\hat{\tau}_j$ are, respectively, the posterior mean and variance based on the 2,000 samples and where $\Phi(\cdot|a, b)$ is the normal CDF function with mean a and variance b. When the posterior samples are approximately symmetric and unimodal (as is often the case for GLM regression coefficients except in small samples), a normal CDF approximation should work well for the reasons described in Section 4. In Figure 4 of the Supporting Information, as justification for this approximation, we simulate data sets, plotting the approximated posterior probabilities using the normal CDF and the posterior probabilities obtained using full MCMC. The R^2 is approximately equal to 1, and the points lie on a 45-degree line, indicating that they are approximately equal.

We conduct hypothesis tests for global null hypotheses $H_0^{(I)} = \cap_{i \in I} H_{0i}$ against global alternative hypotheses $H_1^{(I)} = \cup_{i \in I} H_{1i}$ for $I \in \{\{1, 2, 3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}\}$, where $H_{0i} = \{\boldsymbol{\beta} : \beta_{1i} = 0\}$ and $H_{1i} = \{\boldsymbol{\beta} : \beta_{1i} > 0\}$. For example, $H_0^{\{1,2,3\}}$ is the null hypothesis that β_{11}, β_{12} and β_{13} are all zero, while H_{1I} is the alternative hypothesis that at least one of them is positive. For brevity, we focus on the null hypothesis $H_0^{\{1,2,3\}}$. Simulation results for the other cases are presented in Section 2 of the Supporting Information.

Figure 1 depicts the type I error rate and power corresponding to the null hypothesis $H_0^{\{1,2,3\}}$. The type I error rate is inflated for lower sample sizes. When outcomes are highly positive correlated, power is noticeably higher under the proposed Bayesian approach compared to using the method of Holm, even when the type I error rate is controlled.

When the sample sizes are large enough so that asymptotic approximations are accurate, the proposed Bayesian approach should perform no worse than utilizing a frequentist marginal model approach, and power gains should be more noticeable for more highly correlated parameters. Conversely, when the parameters are negatively correlated or independent and type I error is controlled, there should be no difference between the Bayesian and frequentist approaches because the threshold for negatively correlated parameters based on the Gaussian copula is the same as under independence.

In order to more convincingly demonstrate that power gains using the proposed joint Bayesian approach are not due to inflated FWER as a result of asymptotic approximations, we conduct a large-scale simulation with a multivariate normal model in Section 1 of the Supporting Information. Figure 1 of the Supporting Information indicates that FWER is 0.05 for the proposed Bayesian approach across all correlation structures. When the outcomes are positively correlated, the Holm approach is noticeably more conservative and less powerful than the proposed Bayesian approach. However, noticeable power gains are only achieved under the highest correlation scenario (i.e., $\rho = 0.8$). Further simulations indicate that the expected number of rejected null hypotheses is higher under the proposed Bayesian method when the response variables are positively correlated (see Figure 3 of the Supplementary Materials).

7 | PROBABILITY OF SUCCESS APPLICATION

We now illustrate how the phase II MOBILE study may be utilized to estimate the sample size for the phase III study. We note that the phase III study was successful and utilized a sample size of $n = 646$. The planned sample size of the study was $n^* = 590$[29], corresponding to roughly 90% power for a two-sided test at $\alpha = 0.05$ to detect a minimum of 14.5% absolute improvement in the on-treatment response rate. The purpose of this section is to show how the proposed approach could have been utilized to determine the sample size for the phase III study.

The validation prior for the covariates is identical to that utilized in Section 6.2. The validation prior for the outcome model is the posterior distribution of the phase II data using the GCGLM, described in Section 6.1. Using the methods described in Section 5, we generated $B = 10,000$ data sets from the prior predictive distribution of the validation prior with sample sizes $n \in \{450, 475, 500, \ldots, 650\}$. For this simulation study, it is of interest to reject both the null hypothesis corresponding to the primary endpoint $\left(\text{i.e., } H_0^{\{1\}}\right)$ and the global null hypothesis corresponding to the secondary endpoints $\left(\text{i.e., } H_0^{\{2,3\}}\right)$. A particular data set, D, satisfies the study success criteria if $1\{\text{success}|D\} = 1\{P(\beta_{11} > 0|D) \geq 0.95\} \times 1\{\max\{P(\beta_{12} > 0|D), P(\beta_{13} > 0|D)\} \geq \gamma^*\}$ is equal to 1, where γ^* is determined to control the type I error based on the posterior correlation of the regression coefficients as described in Section 4.

Results for the simulation are depicted in Figure 2. The first panel of Figure 2 indicates that a sample size of approximately $n = 570$ would be required in order to attain a 90% probability of having a successful trial based on the primary endpoint. We note the POS we have developed assumes no loss to follow-up, so it may be an underestimate of the trial's POS if dropout occurs. Although we do not consider it here, it is straightforward to incorporate dropout in the predictive data generation process to account for this possibility in POS calculations. The POS calculation for the primary endpoint for the joint model is essentially the same as that from using a marginal GLM.

However, the second column of Figure 2 indicates that the proposed Bayesian approach provides a higher probability that at least one of the null hypotheses corresponding to the secondary endpoints is rejected. This should be expected since, as mentioned in 6.1, the outcomes are positively correlated, and, hence, the regression coefficients corresponding to the treatment are also positively correlated.

The overall POS for rejecting the null hypotheses corresponding to the primary endpoint *and at least one* of the secondary endpoints is slightly higher for the proposed Bayesian approach in smaller samples and not discernible in larger samples. It is worth mentioning that, in this example, we are only considering a POS calculation concerning the global null hypothesis that two secondary endpoints are zero. In other applications (such as in statistical genetics), there may be many more parameters of interest. The POS utilizing the proposed Bayesian step-down procedure could be substantially higher than using a Holm-like procedure that ignores dependence since the threshold for the Bayesian approach is no higher (and oftentimes lower) than assuming independence.

8 | DISCUSSION

While the examples provided in this paper focus on the joint analysis of GLMs, we note that the method can be extended to joint analyses of other types of outcomes (e.g., time-to-event data and regression models outside the exponential family), with the only restriction being that the CDF of the model is computable. Although the proposed approach requires an assumption that the data generating process is a Gaussian copula, it does not require post hoc adjustments to test statistics in order to control the FWER. Rather, the proposed Bayesian approach exploits the posterior correlations of the parameters of interest to determine rejection criteria. In other words, the correlations between parameters are modeled directly. In small samples, type I error rates may be inflated for hypothesis tests concerning the individual parameters, so that the proposed method for multiple testing yields an even more inflated type I error rate. A simple and elegant approach to handling this issue is to elicit $\eta > 1$ in the LKJ prior for the correlation matrix, which will favor independence.

This paper presents several avenues for further research. First, it is worth exploring whether the proposed step-down approach is the most powerful mechanism that controls FWER based on the joint Bayesian model. While it was demonstrated that a Holm procedure utilizing frequentist *p*-values was more conservative, power gains were only noticeable for highly positively correlated parameters.

Second, the asymptotic joint distribution of the posterior probabilities is useful in other contexts. For example, in genome-wide association studies (GWAS), response variables consist of hundreds or thousands of single nucleotide polymorphisms (SNPs), presenting a particularly challenging multiplicity problem. Because FWER is quite stringent, most methods focus on controlling for the false discovery rate (FDR), which is defined to be the expected proportion of rejected null hypotheses that are false. The asymptotic joint distribution of the posterior probabilities may be exploited to construct a step-up procedure[31] to control for the FDR.

Finally, in settings where historical data is available, it is often useful to construct a prior based on the historical data that discounts the impact the historical data has on the posterior, such as done in the power prior[7]. However, it is not clear how to discount the Gaussian copula density in an intuitive way since correlation parameters are not free parameters in general.

ACKNOWLEDGMENTS

This publication is based on research using data from Biogen that has been made available through Vivli, Inc. Vivli has not contributed to or approved, and is not in any way responsible for, the contents of this publication.

Financial disclosure

None reported.

Conflict of interest

The authors declare no potential conflict of interests.

SUPPORTING INFORMATION

The following supporting information is available as part of the online article:

How to cite this article: Alt, E. M., M. A. Psioda, and J. G. Ibrahim (2022) A Bayesian approach to multiple testing, analysis, and study design for response variables of mixed types, *Stat. Med.*, *2012;00:1–6*.

APPENDIX

A MULTIPLICITY CONTROL

In this section, we provide a formal proof of Theorem 1 in Section 4.1 of the main text. We also provide a detailed algorithm for the step-down testing approach described in Section (4.2).

A.1 A proof of Theorem 1 in Section 4.1

Proof. By the Bernstein von-Mises theorem[27],

$$\sup_B |P^*(\boldsymbol{\theta} \in B|\boldsymbol{y}) - Q_n(B)| \to 0$$

as $n \to \infty$, where Q_n is the measure with respect to the multivariate normal density $\phi(\cdot|\hat{\boldsymbol{\theta}}_1, n^{-1}\hat{\boldsymbol{\Sigma}}_1)$. It thus follows that for $k = 1, \dots, K$,

$$\left| P(\theta_{1k} > 0|D) - \Phi\left(\sqrt{n}\hat{\theta}_{1k}/\hat{\sigma}_{1k}\right)\right| \xrightarrow{\text{p}} 0.$$

Hence, we must have

$$\left\| \begin{pmatrix} P(\theta_1 > 0|\boldsymbol{y}) \\ \vdots \\ P(\theta_K > 0|\boldsymbol{y}) \end{pmatrix} - \begin{pmatrix} \Phi(\sqrt{n}\hat{\theta}_{11}/\hat{\sigma}_{11}) \\ \vdots \\ \Phi(\sqrt{n}\hat{\theta}_{1K}/\hat{\sigma}_{1K}) \end{pmatrix} \right\| \xrightarrow{\text{p}} 0 \tag{A1}$$

□

In words, the relationship (A1) indicates that, for large samples, the posterior probabilities may be approximated accurately with a normal distribution.

Under the usual regularity conditions for MLEs, we have

$$\hat{\boldsymbol{\Sigma}}_1^{-1/2}\sqrt{n}\hat{\boldsymbol{\theta}}_1 \xrightarrow{d} N_K(\mathbf{0}, \boldsymbol{I}).$$

Let $\hat{\boldsymbol{D}}_1 = \text{diag}\{\sigma_{11}^2, \dots, \sigma_{1K}^2\}$ denote the diagonal elements of $\hat{\boldsymbol{\Sigma}}_1$, so that we may decompose $\hat{\boldsymbol{\Sigma}}_1 = \hat{\boldsymbol{D}}_1^{1/2}\hat{\boldsymbol{\Gamma}}_1\hat{\boldsymbol{D}}^{1/2}$. Then by the consistency of $\hat{\boldsymbol{\Sigma}}_1$, we have

$$\hat{\boldsymbol{D}}_1^{-1/2}\sqrt{n}\hat{\boldsymbol{\theta}}_1 \xrightarrow{\text{d}} N(\mathbf{0}, \boldsymbol{\Gamma}_1^*). \tag{A2}$$

The result follows by combining (A1) and (A2) with the continuous mapping theorem.

A.2 FWER control

In this section, we describe how we may utilize the asymptotic distribution of the posterior probabilities (10) to develop a procedure that addresses multiple comparisons and provides strict FWER control. Unlike the previously discussed Frequentist approaches, the proposed method takes into account dependence between parameters, asymptotically resulting in a FWER of precisely α.

Suppose we wish to test $H_0 : \cap_{k=1}^K H_{0k}$ versus $H_1 : \cup_{k=1}^K H_{1k}$. The Holm method, reviewed in Section 2, may be conservative if the parameters exhibit positive dependence. Conversely, exploiting the joint distribution of the posterior probabilities, we may find, via simulation, the value γ^* such that $P(P_{(K)} > \gamma^*) \approx 1 - \gamma$, where $P_{(1)}, \dots, P_{(K)}$ are the order statistics of $\boldsymbol{P} \sim \text{GC}(\boldsymbol{\Gamma})$. In general, the value γ^* will depend on both the number of parameters, K, and the estimated posterior correlation matrix, $\hat{\boldsymbol{\Gamma}}$. By contrast, the Holm procedure ignores dependence and only depends only on K. The global null hypothesis H_0 may then be rejected if the maximal posterior probability is larger than γ^*, i.e., if $\max_{1\leq j\leq K}\{P(\theta_{1j} \geq 0)\} \geq \gamma^*$.

This approach may be generalized to find out which hypotheses may be rejected in the alternative hypotheses $H_1 = \cup_{k=1}^K H_{1j}$. Let $\tilde{\boldsymbol{p}} = (\tilde{p}_1, \dots, \tilde{p}_K)'$ denote the estimated posterior probabilities. Let $\boldsymbol{P} = (P_1, \dots, P_K) \sim \text{GC}(\boldsymbol{\Gamma})$. For notational convenience, we assume without loss of generality that the components of $\boldsymbol{P}$ and $\tilde{\boldsymbol{p}}$ are in descending order. We reject H_{01} if $\tilde{p}_1 > \gamma_1^*$, where γ_1^* solves $P(P_1 > \gamma_1^*) = 1-\gamma$. If H_{01} is not rejected, the procedure terminates. Otherwise, we construct $\boldsymbol{\Gamma}_2$, the $(K-1)\times(K-1)$ correlation matrix of $(P_2, \dots, P_K)' \sim \text{GC}(\boldsymbol{\Gamma}_2)$ by removing the first row and first column of $\boldsymbol{\Gamma}$. We then reject H_{02} if $\tilde{p}_2 > \gamma_2^*$, where γ_2^* solves $P(P_2 > \gamma^*) = 1 - \gamma$. Proceeding inductively, if $H_{01} \dots, H_{0,k-1}$ are rejected, we construct the $(K - k + 1) \times$

$(K - k + 1)$ correlation matrix $\mathbf{\Gamma}_k$ by removing the first $k - 1$ rows and columns from $\mathbf{\Gamma}$, rejecting the null hypothesis H_{0j} if $\tilde{p}_j > \gamma_k^*$, where γ_k^* solves $P(P_k > \gamma_k^*) = 1 - \gamma$. This step-down procedure is closely related to the Bonferroni-Holm procedure [1], where the evidence threshold changes at each iteration. However, unlike the Holm-Bonferroni procedure, the proposed method is based on the joint distribution of the posterior probabilities (which are correlated), rather than p-values from marginal models (which do not take into account correlation between response variables). When the submatrices of $\mathbf{\Gamma}$ have positive elements, the threshold for the posterior probabilities will be smaller than those assuming independence. Hence, the proposed method is more powerful than using a Holm-like procedure applied directly to the posterior probabilities.

References

1. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 1979; 6(2): 65–70.

2. Song PXK, Li M, Yuan Y. Joint Regression Analysis of Correlated Data Using Gaussian Copulas. *Biometrics* 2009; 65(1): 60–68. doi: 10.1111/j.1541-0420.2008.01058.x

3. Smith MS, Khaled MA. Estimation of Copula Models With Discrete Margins via Bayesian Data Augmentation. *Journal of the American Statistical Association* 2012; 107(497): 290–303. doi: 10.1080/01621459.2011.644501

4. Smith MS. Bayesian Approaches to Copula Modelling. *arXiv:1112.4204 [stat]* 2013. doi: 10.1093/acprof:oso/9780199695607.001.0001

5. Alt EM, Psioda MA, Ibrahim JG. Bayesian multivariate probability of success using historical data with type I error rate control. *Biostatistics* 2021: kxab050. doi: 10.1093/biostatistics/kxab050

6. O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. *Pharmaceutical Statistics* 2005; 4(3): 187–201. doi: 10.1002/pst.175

7. Ibrahim JG, Chen MH. Power prior distributions for regression models. *Statistical Science* 2000; 15(1): 46–60. doi: 10.1214/ss/1009212673

8. Ibrahim JG, Chen MH, Gwon Y, Chen F. The power prior: theory and applications. *Statistics in Medicine* 2015; 34(28): 3724–3749. doi: 10.1002/sim.6728

9. FDA . Multiple Endpoints in Clinical Trials Guidance for Industry. 2018.

10. Westfall PH, Young SS. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. John Wiley & Sons . 1993.

11. Romano JP, Wolf M. Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing. *Journal of the American Statistical Association* 2005; 100(469): 94–108. doi: 10.1198/016214504000000539

12. Romano JP, Wolf M. Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters* 2016; 113: 38–40. doi: 10.1016/j.spl.2016.02.012

13. Dudoit S, Gilbert HN, Laan MJvd. Resampling-Based Empirical Bayes Multiple Testing Procedures for Controlling Generalized Tail Probability and Expected Value Error Rates: Focus on the False Discovery Rate and Simulation Study. *Biometrical Journal* 2008; 50(5): 716–744. doi: 10.1002/bimj.200710473

14. Causeur D, Kloareg M, Friguet C. Control of the FWER in Multiple Testing Under Dependence. *Communications in Statistics - Theory and Methods* 2009; 38(16-17): 2733–2747. doi: 10.1080/03610910902936281

15. Wilson DJ. The harmonic mean p -value for combining dependent tests. *Proceedings of the National Academy of Sciences* 2019; 116(4): 1195–1200. doi: 10.1073/pnas.1814092116

16. Berry DA, Hochberg Y. Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference* 1999; 82(1): 215–227. doi: 10.1016/S0378-3758(99)00044-0

17. McCulloch C. Joint modelling of mixed outcome types using latent variables.. *Statistical methods in medical research* 2008; 17(1): 53–73. doi: 10.1177/0962280207081240

18. Westfall PH, Krishen A, Young SS. Using prior information to allocate significance levels for multiple endpoints. *Statistics in Medicine* 1998; 17(18): 2107–2119. doi: 10.1002/(SICI)1097-0258(19980930)17:18<2107::AID-SIM910>3.0.CO;2-W

19. Pitt M, Chan D, Kohn R. Efficient Bayesian Inference for Gaussian Copula Regression Models. *Biometrika* 2006; 93(3): 537–554.

20. Chen MH, Ibrahim JG. Conjugate priors for generalized linear models. *Statistica Sinica* 2003; 13(2): 461–476.

21. Lewandowski D, Kurowicka D, Joe H. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* 2009; 100(9): 1989–2001. doi: 10.1016/j.jmva.2009.04.008

22. Amersi FF, McElrath-Garza A, Ahmad A, et al. Long-term survival after radiofrequency ablation of complex unresectable liver tumors. *Archives of Surgery (Chicago, Ill.: 1960)* 2006; 141(6): 581–587; discussion 587–588. doi: 10.1001/archsurg.141.6.581

23. Polson NG, Scott JG. On the Half-Cauchy Prior for a Global Scale Parameter. *Bayesian Analysis* 2012; 7(4): 887–902. doi: 10.1214/12-BA730

24. Hoffman MD, Gelman A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* 2014; 15: 1593–1623.

25. Ibrahim JG, Chen MH, Lakshminarayanan M, Liu GF, Heyse JF. Bayesian probability of success for clinical trials using historical data. *Statistics in Medicine* 2015; 34(2): 249–264. doi: 10.1002/sim.6339

26. Wang F, Gelfand AE. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science* 2002; 17(2): 193–208. doi: 10.1214/ss/1030550861

27. Vaart v. dAW. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic MathematicsCambridge: Cambridge University Press . 1998.

28. Hupperts R, Lycke J, Short C, et al. Prolonged-release fampridine and walking and balance in MS: randomised controlled MOBILE trial. *Multiple Sclerosis (Houndmills, Basingstoke, England)* 2016; 22(2): 212–221. doi: 10.1177/1352458515581436

29. Hobart J, Ziemssen T, Feys P, et al. Assessment of Clinically Meaningful Improvements in Self-Reported Walking Ability in Participants with Multiple Sclerosis: Results from the Randomized, Double-Blind, Phase III ENHANCE Trial of Prolonged-Release Fampridine. *CNS drugs* 2019; 33(1): 61–79. doi: 10.1007/s40263-018-0586-5

30. Psioda MA, Ibrahim JG. Bayesian clinical trial design using historical data that inform the treatment effect. *Biostatistics* 2019; 20(3): 400–415. doi: 10.1093/biostatistics/kxy009

31. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 1995; 57(1): 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

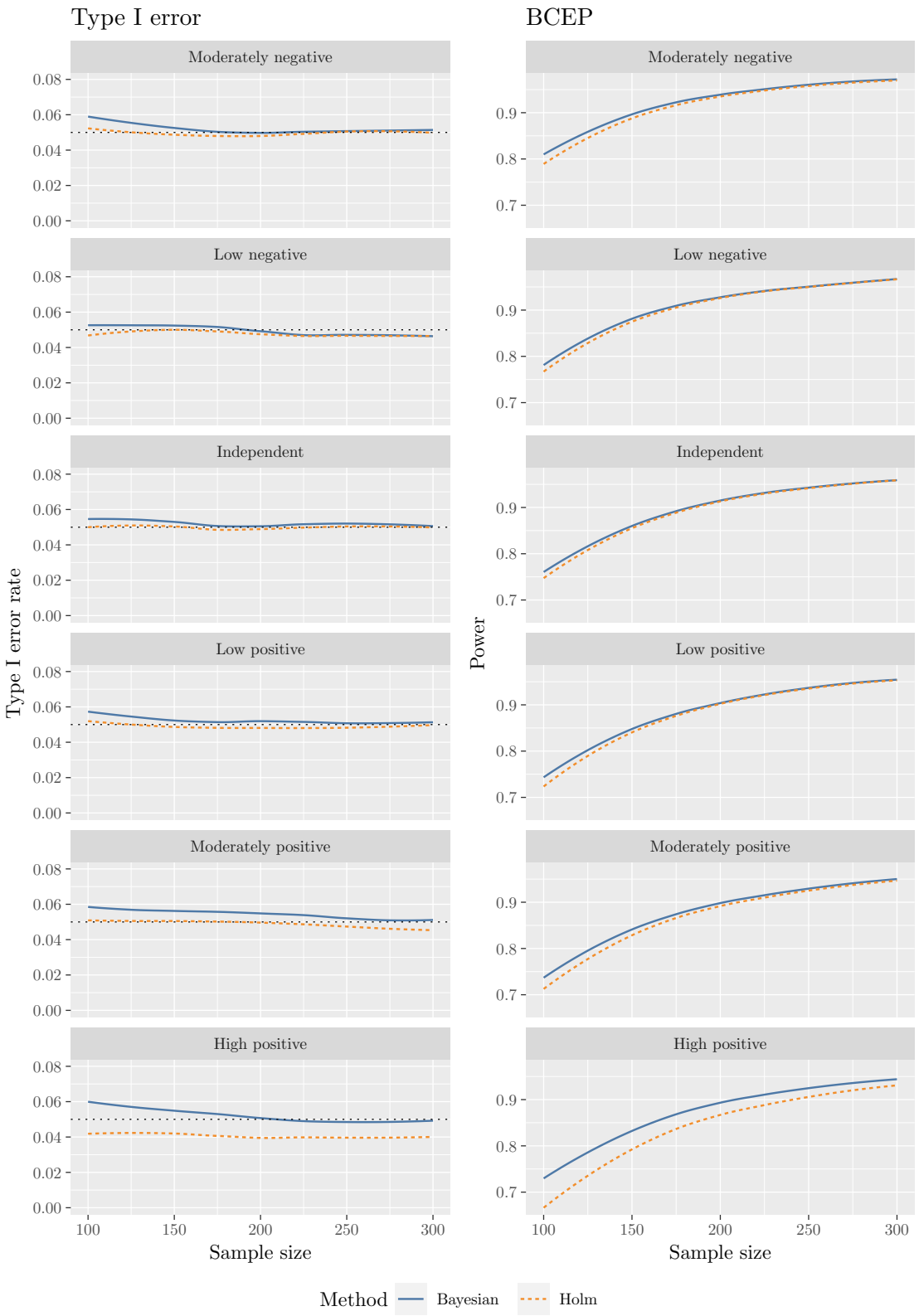

FIGURE 1 Estimated type I error rate and BCEP for $H_0^{\{1,2,3\}}$ versus $H_1^{\{1,2,3\}}$. The dotted black line is a horizontal line at 0.05

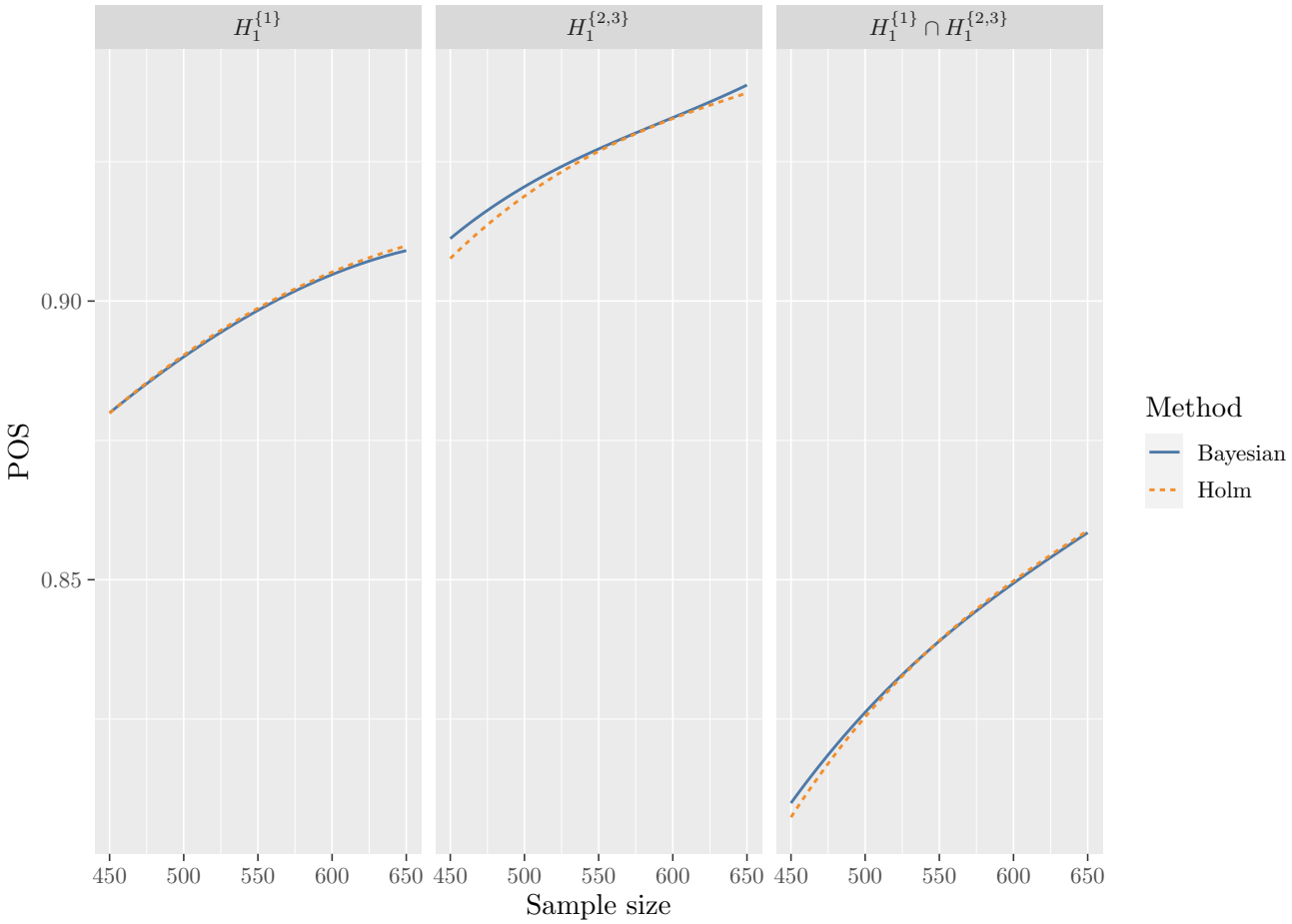

FIGURE 2 Probability of success based on the MOBILE trial