Simultaneous inference for multiple marginal generalized estimating equation models

Ristl, Robin; Hothorn, Ludwig; Ritz, Christian; Posch, Martin

Published in:
Statistical Methods in Medical Research

DOI:
10.1177/0962280219873005

Publication date:
2020

Document version
Publisher's PDF, also known as Version of record

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2020, Vol. 29(6) 1746–1762

DOI: 10.1177/0962280219873005
journals.sagepub.com/home/smm
SAGE

Simultaneous inference for multiple marginal generalized estimating equation models

Robin Ristl,[1] Ludwig Hothorn,[2] Christian Ritz[3] and Martin Posch[1]

Abstract

Motivated by small-sample studies in ophthalmology and dermatology, we study the problem of simultaneous inference for multiple endpoints in the presence of repeated observations. We propose a framework in which a generalized estimating equation model is fit for each endpoint marginally, taking into account dependencies within the same subject. The asymptotic joint normality of the stacked vector of marginal estimating equations is used to derive Wald-type simultaneous confidence intervals and hypothesis tests for multiple linear contrasts of regression coefficients of the multiple marginal models. The small sample performance of this approach is improved by a bias adjustment to the estimate of the joint covariance matrix of the regression coefficients from multiple models. As a further small sample improvement a multivariate *t*-distribution with appropriate degrees of freedom is specified as reference distribution. In addition, a generalized score test based on the stacked estimating equations is derived. Simulation results show strong control of the family-wise type I error rate for these methods even with small sample sizes and increased power compared to a Bonferroni-Holm multiplicity adjustment. Thus, the proposed methods are suitable to efficiently use the information from repeated observations of multiple endpoints in small-sample studies.

Keywords

Generalized estimating equations, multiple testing, multiple endpoints, dependent observations, small samples

1 Introduction

In empirical studies where for each subject multiple endpoints are observed, it is often of interest to identify predictive factors for several of these endpoints. To this end, regression models for the different endpoints can be defined to test respective null hypotheses on the model parameters. However, if for each endpoint, one (or more) hypotheses are tested, a multiple testing problem arises and adjustments for multiplicity are required to control, for example, the family wise type I error rate (FWER).

In this manuscript, we focus on settings where all or some of the multiple endpoints are measured repeatedly and derive multiple testing procedures and simultaneous confidence intervals that account for the correlation between the endpoints as well as the correlation between the repeated measurements of each endpoint. The endpoints may be on different scales (we particularly consider continuous, binary and count data), and the regression models may differ across endpoints. The proposed tests improve the Bonferroni test which is typically strictly conservative.

The testing procedures are based on generalized estimating equation (GEE) models[1] that are separately fitted for each endpoint. Thereby each model accounts for the dependencies between repeated observations of the according endpoint.

We use a representation of stacked estimating equations to show joint multivariate asymptotic normality of the regression coefficient estimators from different models and to estimate their covariance matrix, such that

[1]Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria
[2]Institute of Biostatistics, Leibniz University Hannover, Hannover, Germany
[3]Department of Nutrition, Exercise and Sports, Faculty of Science, University of Copenhagen, Copenhagen, Denmark

Corresponding author:
Robin Ristl, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria.
Email: robin.ristl@meduniwien.ac.at

parametric inference methods based on a multivariate normal approximation can be applied. The method generalizes the approach by Pipper et al.[2] who used stacked estimating equations for multiple generalized linear models and Rochon[3] who studied the case of repeated bivariate measurements comprising a continuous and a binary endpoint. Jensen et al.[4] applied a similar representation to multiple linear mixed models for repeatedly observed continuous endpoints. Also see Verbeke et al.[5] for a recent review on the analysis of multivariate longitudinal data.

While inference based on the multivariate normal approximation can be justified for large sample sizes by asymptotic arguments,[6] it may be inaccurate for smaller samples. In particular, the bias and variability of nuisance parameter estimates are neglected in purely asymptotic methods, resulting in too liberal inference procedures. To improve the small sample properties of Wald tests, we generalize bias-adjustment procedures proposed for covariance matrix estimators in single GEE models,[7] to the case of multiple marginal GEE models. Furthermore, similar to the studies by Hasler and Hothorn[8] and Pan and Wall,[9] we use multivariate t- and F-distributions to better control the type I error rate. We further propose a maximum-type generalized score test and show via simulation that it is a viable small sample alternative to the Wald test.

The paper is structured as follows: In Section 2, multiple marginal GEE models and a bias-adjusted covariance matrix estimator are introduced. In Section 3, we define Wald and score test statistics to test multiple linear contrasts and derive corresponding simultaneous confidence intervals. In Section 4, the proposed methods are applied to a retina disease study. Furthermore, in Section 5, we investigate the small sample properties of the proposed methods in a simulation study. Finally, in Section 6, we conclude with a discussion.

2 Multiple marginal GEE models

2.1 The statistical model

Assume that multiple endpoints (outcome variables) indexed by $m = 1,...,M$ are observed in subjects with index $i = 1,...,K$. Observations between different subjects are assumed to be independent. However, we allow for repeated observations, indexed by $j = 1, \ldots, n_{i^{(m)}}$, of the m-th endpoint within subjects such that $Y_{ij}^{(m)}$ denotes the j-th observation of endpoint m in subject i. Let $\boldsymbol{x}_{ij}^{(m)}$ be a row-vector of covariates that is of length $p^{(m)}$ and $\boldsymbol{\beta}^{(m)}$ a vector of regression coefficients.

Each endpoint $m = 1, \ldots, M$ is modeled with a separate generalized linear regression model for the mean $\mu_{ij}^{(m)} = E(Y_{ij}^{(m)}) = g^{(m)^{-1}}(\boldsymbol{x}_{ij}^{(m)}\boldsymbol{\beta}^{(m)})$ with link function $g^{(m)}$, where the variance of $Y_{ij}^{(m)}$ is modeled as $var(Y_{ij}^{(m)}) = v(\mu_{ij}^{(m)})\phi^{(m)}$, where v is a variance function depending only on $\mu_{ij}^{(m)}$ and $\phi^{(m)}$ is a scale parameter. In typical applications, the link function and the variance function are derived from the canonical representation of an exponential family model.[10] Throughout the manuscript, we assume that regression coefficients and nuisance parameters are unique to one model and not shared between any two models. The models for different endpoints are estimated independently, see Section 7 for a discussion on alternative approaches of joint estimation.

2.2 Generalized estimating equations

To account for dependencies between repeated observations within the same subject, the regression coefficients $\boldsymbol{\beta}^{(m)}$ and their covariance matrix are estimated based on the generalized estimating equation approach.[1] Thus, the estimate $\hat{\boldsymbol{\beta}}^{(m)}$ is given by the solution of the generalized estimating equation

$$\boldsymbol{U}^{(m)}(\boldsymbol{\beta}^{(m)}) = \sum_{i=1}^{K} \boldsymbol{U}_i^{(m)}(\boldsymbol{\beta}^{(m)}) = \boldsymbol{0} \tag{1}$$

with subject-wise contributions $\boldsymbol{U}_i^{(m)} = \boldsymbol{D}_i^{(m)^T}\boldsymbol{V}_i^{(m)^{-1}}\boldsymbol{S}_i^{(m)}$. Here, $\boldsymbol{S}_i^{(m)} = \boldsymbol{Y}_i^{(m)} - \boldsymbol{\mu}_i^{(m)}$ is a vector of residuals. $\boldsymbol{V}_i^{(m)} = \boldsymbol{A}_i^{(m)^{1/2}}\boldsymbol{R}_i^{(m)}(\boldsymbol{\alpha})\boldsymbol{A}_i^{(m)^{1/2}}\phi^{(m)}$ is a working covariance matrix with $\boldsymbol{A}_i^{(m)} = diag\left(v(\mu_{ij}^{(m)})\right)$ the diagonal matrix of variance functions and $\boldsymbol{R}_i^{(m)}$ the working correlation of $\boldsymbol{Y}_i^{(m)}$. $\boldsymbol{R}_i^{(m)}$ is parametrized via a parameter vector $\boldsymbol{\alpha}^{(m)}$ (which typically is of small dimension compared to the number of entries in $\boldsymbol{R}^{(m)}$). $\boldsymbol{D}_i^{(m)} = \left(\frac{\partial \boldsymbol{\mu}_i^{(m)}}{\partial \boldsymbol{\beta}^{(m)}}\right)^T$, and in an exponential family model with canonical link $\boldsymbol{D}_i^{(m)^T} = \frac{\partial \boldsymbol{\mu}_i^{(m)}}{\partial \boldsymbol{\beta}^{(m)}} = \boldsymbol{X}_i^{(m)^T}\boldsymbol{A}_i^{(m)}$, with $\boldsymbol{X}_i^{(m)^T} = \left(\boldsymbol{x}_{i1}^{(m)^T}, \ldots, \boldsymbol{x}_{in_i}^{(m)^T}\right)$.

Given $\boldsymbol{\beta}^{(m)}$, the parameters $\boldsymbol{\alpha}^{(m)}$ and $\phi^{(m)}$ may be consistently estimated from the residuals $\boldsymbol{S}_i^{(m)}, i = 1, \ldots, K$ by moment estimators.[1] Given $\boldsymbol{\alpha}^{(m)}$ and $\phi^{(m)}$, an estimate for $\boldsymbol{\beta}^{(m)}$ is found as solution to equation (1). Iteration of these two estimation steps results in a consistent estimate $\hat{\boldsymbol{\beta}}^{(m)}$ such that asymptotically $K^{1/2}(\hat{\boldsymbol{\beta}}^{(m)} - \boldsymbol{\beta}^{(m)})$ is multivariate normal (Theorem 2 in Liang and Zeger[1]). A consistent estimator for the covariance matrix of the limiting normal distribution as proposed in Liang and Zeger[1] is $\left(\frac{1}{K}\hat{\boldsymbol{H}}^{(m)}\right)^{-1} \frac{1}{K}\hat{\boldsymbol{B}}^{(m)}\left(\frac{1}{K}\hat{\boldsymbol{H}}^{(m)}\right)^{-1}$. Here, $\hat{\boldsymbol{B}}^{(m)} = \sum_{i=1}^{K} \boldsymbol{U}_i^{(m)}\boldsymbol{U}_i^{(m)}T$ and $\hat{\boldsymbol{H}}^{(m)} = -\sum_{i=1}^{K} \boldsymbol{D}_i^{(m)T}\boldsymbol{V}_i^{(m)^{-1}}\boldsymbol{D}_i^{(m)}$ both evaluated at $\hat{\boldsymbol{\beta}}^{(m)}$, with $\frac{1}{K}\hat{\boldsymbol{H}}^{(m)}$ converging to $\frac{1}{K}\boldsymbol{H}^{(m)} = \frac{1}{K}\frac{\partial \boldsymbol{U}^{(m)}(\boldsymbol{\beta}^{(m)})}{\partial \boldsymbol{\beta}^{(m)}}$. The asymptotic results for $\hat{\boldsymbol{\beta}}^{(m)}$ do not require that the working correlation $\boldsymbol{R}_i$ matches the true correlation of $\boldsymbol{Y}_i$; however, the efficiency of $\hat{\boldsymbol{\beta}}$ increases if $\boldsymbol{R}_i$ is close to the true correlation.

Note that in case the mean model is misspecified, $\hat{\boldsymbol{\beta}}^{(m)}$ will typically converge to a vector $\boldsymbol{\beta}^{(m)}$ that defines the model within the chosen mean structure that best approximates the true model in the sense of minimized Kullback-Leibler distance.[11,12] In that case, the proposed methods provide inference on the parameters of the approximating model.

2.3 Multiple marginal models

We are interested in simultaneous inference on the regression coefficient vectors $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(M)}$ and approximate the joint distribution of the stacked vector $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}^{(1)T}, \ldots, \hat{\boldsymbol{\beta}}^{(M)T})^T$ by a multivariate normal distribution based on the framework of Pipper et al.[2]

By equation (1), $\hat{\boldsymbol{\beta}}$ (together with the marginal model estimates for the nuisance parameters $\boldsymbol{\alpha}^{(m)}$ and $\phi^{(m)}, m = 1, \ldots, M$) is the solution to the stacked estimating equation

$$\boldsymbol{U} = \begin{pmatrix} \boldsymbol{U}^{(1)} \\ \vdots \\ \boldsymbol{U}^{(M)} \end{pmatrix} = \sum_{i=1}^{K} \begin{pmatrix} \boldsymbol{U}_i^{(1)} \\ \vdots \\ \boldsymbol{U}_i^{(M)} \end{pmatrix} = \sum_{i=1}^{K} \boldsymbol{U}_i = \boldsymbol{0} \tag{2}$$

Similar to the case of a single GEE model, for increasing number of subjects K, $\sqrt{K}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges to a multivariate normal distribution with mean zero and covariance matrix $\lim_p \left(\frac{1}{K}\frac{\partial \boldsymbol{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^{-1} \left(\frac{1}{K}\sum_{i=1}^{K} \boldsymbol{U}_i(\boldsymbol{\beta})\boldsymbol{U}_i(\boldsymbol{\beta})^T\right) \left(\frac{1}{K}\frac{\partial \boldsymbol{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^{-1}$ provided $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ and certain regularity conditions are met. Here lim_p denotes the limit in probability when K goes to infinity. Consistency of $\hat{\boldsymbol{\beta}}$ follows if $\hat{\boldsymbol{\beta}}^{(m)}$ is consistent for $\boldsymbol{\beta}^{(m)}$ for all $m = 1, \ldots, M$. The essential regularity conditions concern the derivatives of $\boldsymbol{U}$ with respect to $\boldsymbol{\beta}$ (see Chapter 5.3 in the book by Van der Vaart[6] and Chapter 9.1 in the book by Cox and Hinkley[13]). However, the matrix of first derivatives $\boldsymbol{H} = \frac{\partial \boldsymbol{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ is a block diagonal matrix of the matrices $\boldsymbol{H}^{(m)} = \frac{\partial \boldsymbol{U}^{(m)}(\boldsymbol{\beta}^{(m)})}{\partial \boldsymbol{\beta}^{(m)}}$. Hence, conditions such as existence of derivatives, a dominating function, expectation and a matrix-inverse are inherited if they are met by all marginal models.

An estimate of the covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{H}}^{-1}\hat{\boldsymbol{B}}\hat{\boldsymbol{H}}^{-1} = \begin{pmatrix} \hat{\boldsymbol{H}}^{(1)-1}\hat{\boldsymbol{B}}^{(1,1)}\hat{\boldsymbol{H}}^{(1)-1} & \ldots & \hat{\boldsymbol{H}}^{(1)-1}\hat{\boldsymbol{B}}^{(1,M)}\hat{\boldsymbol{H}}^{(M)-1} \\ \vdots & \ddots & \vdots \\ \hat{\boldsymbol{H}}^{(1)-1}\hat{\boldsymbol{B}}^{(M,1)}\hat{\boldsymbol{H}}^{(M)-1} & \ldots & \hat{\boldsymbol{H}}^{(M)-1}\hat{\boldsymbol{B}}^{(M,M)}\hat{\boldsymbol{H}}^{(M)-1} \end{pmatrix} \tag{3}$$

where $\hat{\boldsymbol{B}} = \sum_{i=1}^{K} \boldsymbol{U}_i(\hat{\boldsymbol{\beta}})\boldsymbol{U}_i(\hat{\boldsymbol{\beta}})^T$ is calculated from the stacked vectors $\boldsymbol{U}_i$. The components $\hat{\boldsymbol{B}}^{(m,m')} = \sum_{i=1}^{K} \boldsymbol{U}_i^{(m)}(\hat{\boldsymbol{\beta}}^{(m)})\boldsymbol{U}_i^{m'}(\hat{\boldsymbol{\beta}}^{(m')})^T$ correspond to the empirical correlation between the contributions of a subject to the estimating equations of models m, m'. $\hat{\boldsymbol{H}}$ is a block diagonal matrix with block elements $\hat{\boldsymbol{H}}^{(m)} = \frac{\partial \boldsymbol{U}^{(m)}(\hat{\boldsymbol{\beta}}^{(m)})}{\partial \boldsymbol{\beta}^{(m)}}$ that are the corresponding estimates from the marginal models. The resulting multiple model sandwich variance estimator maintains the marginal GEE sandwich variance estimators in the diagonal blocks, while the off diagonal blocks contain estimates for the covariances between estimated regression coefficients from different models.

2.3.1 Bias-adjusted covariance estimator

The covariance matrix estimator in GEE models is consistent but it is in general not unbiased. With small sample sizes, the variances may be underestimated which leads to an inflation of the type I error rate of hypothesis tests and to confidence intervals with coverage less than the nominal $1 - \alpha$ level. Based on the bias-adjusted estimator proposed by Mancl and DeRouen[7] (see also Wang et al.[14]) for a single GEE model, we derive a bias-adjusted covariance estimator for multiple models, given by

$$\hat{\boldsymbol{\Sigma}}_{adj} = \hat{\boldsymbol{H}}^{-1} \hat{\boldsymbol{B}}_{adj} \hat{\boldsymbol{H}}^{-1} \tag{4}$$

where $\hat{\boldsymbol{B}}_{adj} = \sum_{i=1}^{K} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} (\boldsymbol{I}_i - \hat{\boldsymbol{P}}_{ii})^{-1} \boldsymbol{S}_i \boldsymbol{S}_i^T (\boldsymbol{I}_i - \hat{\boldsymbol{P}}_{ii})^{-1} \boldsymbol{V}_i^{-1} \boldsymbol{D}_i$ and $\hat{\boldsymbol{P}}_{ii} = \boldsymbol{D}_i \hat{\boldsymbol{H}}^{-1} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1}$ and $\boldsymbol{I}_i$ is the identity matrix with matching dimension. See the supplemental material Section S.1 for details. The matrices $\boldsymbol{D}_i$, $\boldsymbol{V}_i$ and, consequently, $\hat{\boldsymbol{P}}_{ii}$ are block diagonal with block elements $\boldsymbol{D}_i^{(m)}$, $\boldsymbol{V}_i^{(m)}$, $\hat{\boldsymbol{P}}_{ii}^{(m)}$, $m = 1, \ldots, M$, respectively. $\hat{\boldsymbol{\Sigma}}_{adj}$ contains in the diagonal blocks the bias-adjusted variance matrix estimators that would result from separate marginal models and the off-diagonal blocks contain bias-adjusted estimates of the covariances between regression coefficient estimates from different models.

2.4 Missing data

Values of some $Y_{ij}^{(m)}$ or $x_{ij}^{(m)}$ may be missing in an actual data set. As for single GEE models,[1] $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ if some observations are missing completely at random (MCAR), i.e. missingness is completely independent from any missing or non-missing values of the included variables, and each model is fit with the respective available data. Also, inference based on the asymptotic normality of the stacked score vector $\boldsymbol{U}$ stays unaffected under MCAR. If a subject i has to be excluded entirely from the model for the m-th endpoint due to missing data, the realization of the respective contribution to the estimating equation is treated as $\boldsymbol{U}_i^{(m)} = \boldsymbol{0}$ in equations (1) and (2) and subsequent calculations. Note that this does not bias the estimated covariance matrix (3) or (4) as the effective sample size enters these equations in terms of the observed information $\hat{\boldsymbol{H}}^{-1}$ and not the number of clusters.

If data are missing at random (MAR), i.e. the missingness may depend on non-missing values of observed variables, residuals may be biased. Consequently, a model fit with available data may result in biased and inconsistent estimates. This bias may be counteracted by weighing observed residuals with the inverse probability of non-missingness of the given data point. Under MAR, these probabilities may in principle be estimated from the observed data. There are different ways to introduce weights in the generalized estimating equation.[15–18] Our software implementation, discussed in Section 6, allows for weights that resemble a scale factor for each observation, similar to the GENMOD procedure in SAS.[19] Here, subject-wise contributions to a correspondingly weighted generalized estimating equation are of the form $\boldsymbol{D}_i^{(m)^T} \boldsymbol{W}_i^{1/2} \boldsymbol{V}_i^{(m)^{-1}} \boldsymbol{W}_i^{1/2} \boldsymbol{S}_i^{(m)}$, where $\boldsymbol{W}_i$ is a diagonal matrix of weights for the observations of subject i. When all observations of a subject receive the same weight, this formulation is equivalent to the cluster-weighted GEE proposed by Fitzmaurice and Laird.[15] With an identity working correlation, it is equivalent to the weighted GEE proposed by Robins et al.[16]

3 Maximum-type tests and quadratic form tests for linear hypotheses

Consider a null hypothesis of the form $H_0 : \boldsymbol{L}\boldsymbol{\beta} - \boldsymbol{r} = \boldsymbol{0}$, where $\boldsymbol{L}$ is a matrix of linear constraints with c rows and number of columns equal to the length of $\boldsymbol{\beta}$ and $\boldsymbol{r}$ is a vector of matching dimension. Each row of this equation corresponds to an elementary null hypothesis $H_i : (\boldsymbol{L}\boldsymbol{\beta})_i = r_i, i = 1, \ldots, c$. Furthermore, assume that $\boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{r}$ has at least one solution in $\boldsymbol{\beta}$. In Section 3.1, we construct asymptotic hypothesis tests for the global null hypothesis H_0 that are based on the maximum of multivariate Wald statistics, and we propose adjustments of the tests for small samples. In Section 3.2, a maximum test based on score statistics is proposed. In Section 3.3, we discuss Wald and score tests for H_0 that are based on quadratic forms. In Section 3.4, the closed testing principle is applied to construct multiple testing procedures, allowing for decisions on intersection and elementary hypotheses with type I error rate control. Furthermore, simultaneous confidence intervals for $(\boldsymbol{L}\boldsymbol{\beta})_i, i = 1, \ldots, c$ corresponding to a single step multiple testing procedure are derived.

3.1 Maximum-type Wald test

The maximum test rejects H_0 if

$$\max_{i=1,\ldots,c}\left|\left(\left(\boldsymbol{L}\hat{\boldsymbol{\beta}}\right)_i - r_i\right)\Big/\hat{SE}_i\right| > q_{1-\alpha}$$

where we use the normal approximation $\sqrt{K}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{L}\boldsymbol{\beta}) \approx N(\boldsymbol{0}, K\boldsymbol{L}\hat{\boldsymbol{\Sigma}}\boldsymbol{L}^T)$ to define the critical value $q_{1-\alpha}$ as the solution of $P(\max_{i=1,\ldots,c}|Z_i| \le q_{1-\alpha}) = 1-\alpha$.[20] Here, $\boldsymbol{Z}$ denotes a c-dimensional multivariate normal variable with mean $\boldsymbol{0}$, unit variances and correlation structure given by $\boldsymbol{L}\hat{\boldsymbol{\Sigma}}\boldsymbol{L}^T$, such that $cov(\boldsymbol{Z}) = diag(\hat{\boldsymbol{SE}})^{-1}\boldsymbol{L}\hat{\boldsymbol{\Sigma}}\boldsymbol{L}^T diag(\hat{\boldsymbol{SE}})^{-1}$, where the vector of standard errors $\hat{\boldsymbol{SE}}$ is given by the square roots of the diagonal entries of $\boldsymbol{L}\hat{\boldsymbol{\Sigma}}\boldsymbol{L}^T$. Similarly, the p-value for the maximum test is defined as $P(\max_{i=1,\ldots,c}|Z_i| \ge \max_{i=1,\ldots,c}|((\boldsymbol{L}\hat{\boldsymbol{\beta}})_i - r_i)/\hat{SE}_i|)$.

$\boldsymbol{L}$ may be below full rank since the quantile $q_{1-\alpha}$ is also defined for a degenerate multivariate normal distribution with singular covariance matrix.

3.1.1 Small sample improvements

For small samples, the type I error rate of the above test can be considerably greater than the nominal level. As a first small sample improvement, the bias-adjusted covariance estimate $\hat{\boldsymbol{\Sigma}}_{adj}$ (equation (4)) may be used instead of $\hat{\boldsymbol{\Sigma}}$ (equation (3)). Furthermore, to also account for the variability of the covariance estimators, the critical value $q_{1-\alpha}$ of the multivariate normal distribution can be replaced by the critical value $t_{1-\alpha}$ of a multivariate t-distribution,[21] such that $P(\max_{i=1,\ldots,c}|T_i| \le t_{1-\alpha}) = 1-\alpha$, where $\boldsymbol{T}$ is distributed according to a c-dimensional multivariate t-distribution with correlation matrix $diag(\hat{\boldsymbol{SE}})^{-1}\boldsymbol{L}\hat{\boldsymbol{\Sigma}}\boldsymbol{L}^T diag(\hat{\boldsymbol{SE}})^{-1}$ and an appropriate number of degrees of freedom (df). See the earlier studies[7,8,22] for related approaches in the context of multiple contrast tests. As a simple method to choose the error degrees of freedom, we propose $df = \min_{m=1,\ldots,M}(K - p^{(m)})$ where $p^{(m)}$ is the number of regression coefficients in model m (compare with Munzel and Hothorn[23]). Alternative methods to choose degrees of freedom for multivariate comparisons are discussed in Section 7.

3.2 Maximum-type score test

We derive the maximum-type generalized score test as an approximation to the Wald test. By first order approximation, $\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{r} = \boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{L}\boldsymbol{\beta} \approx -\boldsymbol{L}\boldsymbol{H}^{-1}\boldsymbol{U}(\boldsymbol{\beta})$. Hence, tests for H_0 can be constructed based on the right hand side $-\boldsymbol{L}\boldsymbol{H}^{-1}\boldsymbol{U}(\boldsymbol{\beta})$ and its normal approximation under the null hypothesis, $N(\boldsymbol{0}, \boldsymbol{L}\boldsymbol{H}^{-1}\boldsymbol{B}\boldsymbol{H}^{-1}\boldsymbol{L}^T)$. Under a simple null hypothesis, the true $\boldsymbol{\beta}$ is known, under a composite null hypothesis, a restricted estimate $\tilde{\boldsymbol{\beta}}$, which satisfies $\boldsymbol{L}\tilde{\boldsymbol{\beta}} - \boldsymbol{r} = \boldsymbol{0}$, is plugged in. To estimate the limiting distribution covariance matrix, $\boldsymbol{H}$ and $\boldsymbol{B}$ are replaced by estimates $\tilde{\boldsymbol{H}} = -\boldsymbol{D}_i^{(m)^T}\boldsymbol{V}_i^{(m)^{-1}}\boldsymbol{D}_i^{(m)}$ evaluated at $\tilde{\boldsymbol{\beta}}^{(m)}$ and $\tilde{\boldsymbol{B}} = \sum_{i=1}^{K}\boldsymbol{U}_i(\tilde{\boldsymbol{\beta}})\boldsymbol{U}_i(\tilde{\boldsymbol{\beta}})^T$.

The maximum-type score test rejects H_0 if

$$\max_{i=1,\ldots,c}|(\boldsymbol{L}\tilde{\boldsymbol{H}}^{-1}\boldsymbol{U}(\tilde{\boldsymbol{\beta}}))_i/\tilde{SE}_i| > \tilde{q}_{1-\alpha}$$

Here $\tilde{SE}_i$ is the square root of the i-th diagonal element of $\boldsymbol{L}\tilde{\boldsymbol{H}}^{-1}\tilde{\boldsymbol{B}}\tilde{\boldsymbol{H}}^{-1}\boldsymbol{L}^T$. $\tilde{q}_{1-\alpha}$ satisfies $P(\max_{i=1,\ldots,c}|\tilde{\boldsymbol{Z}}_i| \le \tilde{q}_{1-\alpha}) = 1-\alpha$, where $\tilde{\boldsymbol{Z}}$ denotes a c-dimensional multivariate normal variable with mean $\boldsymbol{0}$, and covariance matrix given by $diag(\tilde{\boldsymbol{SE}})^{-1}\boldsymbol{L}\tilde{\boldsymbol{H}}^{-1}\tilde{\boldsymbol{B}}\tilde{\boldsymbol{H}}^{-1}\boldsymbol{L}^T diag(\tilde{\boldsymbol{SE}})^{-1}$.

For a single marginal GEE model, the restricted estimate $\tilde{\boldsymbol{\beta}}^{(m)}$ can be computed by the iterative restricted weighted least squares algorithm $\tilde{\boldsymbol{\beta}}^{(m,j+1)} = \tilde{\boldsymbol{\beta}}^{(m,j)} - \left(\frac{\partial \boldsymbol{U}}{\partial \boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}}^{(m,j)})\right)^{-1}(\boldsymbol{U}(\tilde{\boldsymbol{\beta}}^{(m,j)}) - \boldsymbol{L}^T\boldsymbol{\lambda}^{(j)})$, with the vector of Lagrange multipliers $\boldsymbol{\lambda}^{(j)} = -\left(\boldsymbol{L}\left(\frac{\partial \boldsymbol{U}}{\partial \boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}}^{(m,j)})\right)^{-1}\boldsymbol{L}^T\right)^{-1}(\boldsymbol{L}\tilde{\boldsymbol{\beta}}^{(m,j)} - \boldsymbol{r} - \boldsymbol{L}\left(\frac{\partial \boldsymbol{U}}{\partial \boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}}^{(m,j)})\right)^{-1}\boldsymbol{U}(\tilde{\boldsymbol{\beta}}^{(m,j)})\Big)$, compare with Rao and Toutenburg.[24] Here, the second superscript indicates the iteration number. Where $\hat{\boldsymbol{\beta}}$ can be understood to maximize a quasi-likelihood with first derivative $\boldsymbol{U}$, $\tilde{\boldsymbol{\beta}}$ maximizes the quasi-likelihood subject to the restriction of H_0.

If the null hypothesis $\boldsymbol{L\beta} - \boldsymbol{r} = \boldsymbol{0}$ has a block diagonal structure such that each constraint involves only parameters of one marginal model, we have

$$\boldsymbol{L\beta} = \begin{pmatrix} \boldsymbol{L}^{(1)} & & \\ & \ddots & \\ & & \boldsymbol{L}^{(M)} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^{(1)} \\ \vdots \\ \boldsymbol{\beta}^{(M)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{L}^{(1)}\boldsymbol{\beta}^{(1)} & & \\ & \ddots & \\ & & \boldsymbol{L}^{(M)}\boldsymbol{\beta}^{(M)} \end{pmatrix} \tag{5}$$

Then, the restricted estimate $\widetilde{\boldsymbol{\beta}}$ is a stacked vector of restricted estimates from the marginal models. We consider only null hypotheses covered by equation (5). Otherwise, the elements of $\widetilde{\boldsymbol{\beta}}$ needed to be estimated jointly for all models. However, contrasts between coefficients from different marginal models are rarely of interest, if they correspond to different units or scales of measurement. If outcomes are indeed measured at the same scale and units, they can be modelled together in one GEE model.

3.2.1 Small sample considerations

With the score test, nuisance parameters are estimated under the null hypothesis based on the restricted estimate $\widetilde{\boldsymbol{\beta}}$, which is less variable than $\hat{\boldsymbol{\beta}}$. (In the limit $cov(\sqrt{K}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) = cov(\sqrt{K}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{H}_p^{-1}\boldsymbol{L}^T(\boldsymbol{L}\boldsymbol{H}_p^{-1}\boldsymbol{L}^T)^{-1}\boldsymbol{L}\boldsymbol{H}_p$, where $\boldsymbol{H}_p = lim_p \frac{1}{K}\boldsymbol{H}$.) Consequently, we may expect that the nuisance parameters are estimated with less variability, too, and the type I error rate control with the score test is improved compared to the unadjusted Wald test. In principle, though, the Mancl and DeRouen bias adjustment can be extended to the estimate $\widetilde{\boldsymbol{B}}$. Instead of $\hat{\boldsymbol{P}}_{ii} = \boldsymbol{D}_i\hat{\boldsymbol{H}}^{-1}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}$, the adjustment utilizes $\widetilde{\boldsymbol{P}}_{ii} = \boldsymbol{D}_i(\boldsymbol{I} - \widetilde{\boldsymbol{H}}^{-1}\boldsymbol{L}^T(\boldsymbol{L}\widetilde{\boldsymbol{H}}^{-1}\boldsymbol{L}^T)^{-1}\boldsymbol{L})\widetilde{\boldsymbol{H}}^{-1}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}$. When calculating the score test, $\widetilde{\boldsymbol{B}}$ is replaced by $\widetilde{\boldsymbol{B}}_{adj} = \sum_{i=1}^{K}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{I}_i - \widetilde{\boldsymbol{P}}_{ii})^{-1}\boldsymbol{S}_i\boldsymbol{S}_i^T(\boldsymbol{I}_i - \widetilde{\boldsymbol{P}}_{ii})^{-1}\boldsymbol{V}_i^{-1}\boldsymbol{D}_i$. See supplemental material Section S.1 for the derivation. Also, a multivariate t reference distribution could be used. We will, however, focus on the unadjusted score test in the numeric simulations.

3.3 Quadratic form tests

Alternatives to the maximum-type tests may be derived based on the normal approximation of the multivariate Wald and score statistics. An example are quadratic form tests. The quadratic form Wald test rejects H_0 if

$$\left(\boldsymbol{L\hat{\beta}} - \boldsymbol{r}\right)^T\left(\boldsymbol{L\hat{\Sigma}L}^T\right)^{-1}\left(\boldsymbol{L\hat{\beta}} - \boldsymbol{r}\right) > Q_c^{(\chi^2)}(1-\alpha)$$

where $Q_c^{(\chi^2)}(1-\alpha)$ denotes the $1-\alpha$ quantile of the chi-squared distribution with c degrees of freedom.

As small sample improvement, $\hat{\boldsymbol{\Sigma}}$ may be replaced by $\hat{\boldsymbol{\Sigma}}_{adj}$. Further, $Q_c^{(\chi^2)}(1-\alpha)$ may be replaced by $cQ_{c,df}^{(F)}(1-\alpha)$ or $\frac{df}{df-c+1}cQ_{c,df-c+1}^{(F)}(1-\alpha)$, where $Q_{c,df}^{(F)}(1-\alpha)$ denotes the $1-\alpha$ quantile of an F-distribution with c numerator degrees of freedom and df denominator degrees of freedom, chosen in the same way as for the maximum test. The former option is analogous to an F-test, where the variability of an assumedly independent and chi-squared distributed single nuisance parameter is taken into account. The latter option is analogous to Hotelling's test which adjusts for the variability of an assumedly independent and Wishart distributed covariance matrix estimate. This approach was described by Kenward and Roger[25] for random effects models and by Pan and Wall[9] in the context of single GEE models.

The quadratic form generalized score test rejects H_0 if

$$\boldsymbol{U}(\widetilde{\boldsymbol{\beta}})^T\widetilde{\boldsymbol{H}}^{-1}\boldsymbol{L}^T\left(\boldsymbol{L}\widetilde{\boldsymbol{H}}^{-1}\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{H}}^{-1}\boldsymbol{L}^T\right)^{-1}\boldsymbol{L}\widetilde{\boldsymbol{H}}^{-1}\boldsymbol{U}(\widetilde{\boldsymbol{\beta}}) > Q_c^{(\chi^2)}(1-\alpha)$$

Similar to the maximum score test, this test is based on the normal approximation for the statistic $-\boldsymbol{L}\widetilde{\boldsymbol{H}}^{(-1)}\boldsymbol{U}(\widetilde{\boldsymbol{\beta}})$. For an alternative derivation see the study by Boos.[26]

If L is below full rank, a generalized matrix inverse may be applied in the calculation of the quadratic form statistics and c is replaced by the rank of L.

In terms of the multivariate space of $\boldsymbol{L\hat{\beta}} - \boldsymbol{r}$, the quadratic form test statistic and the maximum test statistic apply different metrics to measure deviations from the null vector. The quadratic form test is monotone in the non-centrality parameter $(\boldsymbol{L\beta} - \boldsymbol{r})^T(\boldsymbol{L\Sigma L}^T)^{-1}(\boldsymbol{L\beta} - \boldsymbol{r})$; however, it is in general not monotone in the observed effects

$|(\boldsymbol{L}\hat{\boldsymbol{\beta}})_i|$. In contrast, the maximum test is monotone in $|(\boldsymbol{L}\hat{\boldsymbol{\beta}})_i|$. In the setting of single regression models, a familiar application of quadratic form tests is testing a null hypothesis of equal effects between all, $c+1$ say, stages of some grouping variable. There, the elements of $\boldsymbol{L}\hat{\boldsymbol{\beta}}$ constitute a set of c arbitrarily selected between-group differences and a quadratic metric is appropriate to assess the overall deviation from the null hypothesis. If, however, the null hypothesis refers to a set of effects in multiple models, each elements of $\boldsymbol{L}\hat{\boldsymbol{\beta}}$ has an individual interpretation and asks for a test that is monotone in the individual observed effects. Therefore, a maximum-type test will typically be preferable when testing hypotheses across multiple models.

3.4 Multiple testing procedures

The tests considered so far test a global null hypothesis $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{r}$. Each row of this equation corresponds to an elementary null hypotheses $H_i : (\boldsymbol{L}\boldsymbol{\beta})_i = r_i, i = 1, \ldots, c$ while controlling the FWER in the strong sense requires an appropriate multiple testing procedure.

3.4.1 Single step procedure

Based on the maximum-type Wald test, a single step multiple testing procedure with strong FWER control at level α rejects H_i if $|((\boldsymbol{L}\hat{\boldsymbol{\beta}})_i - r_i)/\hat{SE}_i| > q_{1-\alpha}$. Corresponding simultaneous $1 - \alpha$ Wald confidence intervals for $\boldsymbol{L}\boldsymbol{\beta}$ are given by

$$(\boldsymbol{L}\hat{\boldsymbol{\beta}})_i \pm q_{1-\alpha}\hat{SE}_i \tag{6}$$

$q_{1-\alpha}$ may be replaced by $t_{1-\alpha}$ to use a multivariate t reference distribution. Note that the single step procedure based on the maximum-type score test may not control the FWER because its multivariate reference distribution is valid only under the global null hypothesis H_0.

3.4.2 Closed testing procedure

A general and more powerful multiple testing procedure can be constructed with the closed testing principle.[27] Let $I = \{1, \ldots, c\}$ denote the index set of the elementary hypotheses. According to the closed testing principle, an intersection hypothesis $\cap_{i\in S}H_i, S \subseteq I$ can be rejected with strong control of the FWER at level α if all intersection hypotheses $\cap_{i\in S'}H_i$ with $S \subseteq S'$ are rejected by a local level α test. Note that in the context of linear hypotheses, the intersection hypothesis $\cap_{i\in S}H_i$ corresponds to the set of linear contrasts $(\boldsymbol{L}\boldsymbol{\beta})_{i\in S} = (\boldsymbol{r})_{i\in S}$. Thus, to construct a multiple testing procedure, we define for each such intersection hypothesis a level α test given by one of the tests described above and decide on the intersection and elementary hypothesis according to the closed testing principle.

If the matrix $\boldsymbol{L}$ is not of full rank, some intersection hypotheses are equivalent. It is therefore not necessary, and in fact would reduce the power of the procedure, to test all intersection hypotheses in the closed testing procedure. Instead, the test of a hypothesis $\cap_{i\in S}H_i$ may be substituted by the test for an equivalent hypothesis $\cap_{i\in S'}H_i$ with $S \subset S'$. Shaffer[28] describes a general method to identify redundant intersection hypotheses. In the context of linear contrast tests, and under the assumption that $\boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{r}$ has at least one solution, two intersection hypotheses $\cap_{i\in S}H_i$ and $\cap_{i\in S'}H_i$ are equivalent if the corresponding contrast matrices $(\boldsymbol{L})_{i\in S}$ and $(\boldsymbol{L})_{i\in S'}$ define the same region in the parameter space. This is the case if $\text{rank}\big((\boldsymbol{L})_{i\in S}\big) = \text{rank}\Big(\big((\boldsymbol{L})^T_{i\in S}, (\boldsymbol{L})^T_{i\in S'}\big)^T\Big)$.

Multiplicity adjusted p-values for the test of $\cap_{i\in S}H_i$ are defined as the smallest family-wise significance level for which $\cap_{i\in S}H_i$ can be rejected using the closed testing procedure or, equivalently, as the maximum of local p-values for all local tests of $\cap_{i\in S'}H_i, S \subseteq S'$.

When using maximum-type tests, a weighted closed testing procedure as discussed by Xi et al.[29] may be applied to account for differences in importance of the tested hypotheses, depending on the study aims.

4 Example – A retina disease study

In a recent exploratory study, the association between two metric endpoints $Y^{(1)}$ and $Y^{(2)}$, both measuring retinal function, and three categorical variables $X^{(1)}, X^{(2)}, X^{(3)}$, each representing the condition of one of three retinal cell layers, was analyzed. $X^{(1)}$ and $X^{(2)}$ allow for three stages of deterioration in {0, 1, 2} and $X^{(3)}$ comprises two stages {0, 1}. Within each eye, the set of variables was measured at 29 to 51 distinct locations defined through a common grid. In total, the study data comprise observations from 1489 locations in 35 eyes of 18 patients. Six marginal

analysis models (7) to (12) were defined.

$$E(Y_{ij}^{(1)}) = \beta_0^{(1)} + \mathbb{1}_{\{X_{ij}^{(1)}=1\}}\beta_1^{(1)} + \mathbb{1}_{\{X_{ij}^{(1)}=2\}}\beta_2^{(1)} \tag{7}$$

$$E(Y_{ij}^{(1)}) = \beta_0^{(2)} + \mathbb{1}_{\{X_{ij}^{(2)}=1\}}\beta_1^{(2)} + \mathbb{1}_{\{X_{ij}^{(2)}=2\}}\beta_2^{(2)} \tag{8}$$

$$E(Y_{ij}^{(1)}) = \beta_0^{(3)} + \mathbb{1}_{\{X_{ij}^{(3)}=1\}}\beta_1^{(3)} \tag{9}$$

$$E(Y_{ij}^{(2)}) = \beta_0^{(4)} + \mathbb{1}_{\{X_{ij}^{(1)}=1\}}\beta_1^{(4)} + \mathbb{1}_{\{X_{ij}^{(1)}=2\}}\beta_2^{(4)} \tag{10}$$

$$E(Y_{ij}^{(2)}) = \beta_0^{(5)} + \mathbb{1}_{\{X_{ij}^{(2)}=1\}}\beta_1^{(5)} + \mathbb{1}_{\{X_{ij}^{(2)}=2\}}\beta_2^{(5)} \tag{11}$$

$$E(Y_{ij}^{(2)}) = \beta_0^{(6)} + \mathbb{1}_{\{X_{ij}^{(3)}=1\}}\beta_1^{(6)} \tag{12}$$

Here, $\mathbb{1}$ is the indicator function. Each model was fit using the GEE method with patient as clustering variable and specifying an exchangeable working correlation structure. Note that the robust variance estimation via the GEE approach was preferred over a mixed model since the true correlation structure is most likely too complicated to be explicitly modelled correctly.

Six null hypotheses, addressing the association between an outcome and one independent factor, are regarded in the study: $H_1 : \beta_1^{(1)} = \beta_2^{(1)} = 0$, $H_2 : \beta_1^{(2)} = \beta_2^{(2)} = 0$, $H_3 : \beta_1^{(3)} = 0$, $H_4 : \beta_1^{(4)} = \beta_2^{(4)} = 0$, $H_5 : \beta_1^{(5)} = \beta_2^{(5)} = 0$ and $H_6 : \beta_1^{(6)} = 0$. We illustrate the application of multivariate inference for the set of hypotheses $\{H_i, i = 1, \ldots, 6\}$ based on the joint distribution of the coefficients from all six models. Following the discussion at the end of Section 3.3, we use maximum tests. Define the contrast matrices $\boldsymbol{L}^{(1)} = \boldsymbol{L}^{(2)} = \boldsymbol{L}^{(4)} = \boldsymbol{L}^{(5)} = ((0, 1, 0)^T, (0, 0, 1)^T, (0, 1, -1)^T)^T$ and let $\boldsymbol{L}^{(3)} = \boldsymbol{L}^{(6)} = (0, 1)$. The right-hand side vector is $\boldsymbol{r} = \boldsymbol{0}$. Then, the intersection hypotheses $\cap_{i \in S} H_i, S \subseteq \{1, \ldots, 6\}$ correspond to $\boldsymbol{L}_S \boldsymbol{\beta}_S = \boldsymbol{0}$ where $\boldsymbol{L}_S$ is a block diagonal matrix composed of the matrices $\boldsymbol{L}^{(i)}, i \in S$ and $\boldsymbol{\beta}_S$ is the stacked vector of $\boldsymbol{\beta}^{(i)}, i \in S$. Each of these hypotheses is tested by a maximum-type Wald test, using the bias-adjusted covariance matrix estimate and a multivariate t-distribution with $df = K - 3 = 18 - 3 = 15$ degrees of freedom (or $df = K - 2 = 16$ if only H_3 and H_6 are involved) as reference distribution. Adjusted p-values resulting from the closed test for $\{H_i, i = 1, \ldots, 6\}$ are calculated as described in Section 3.4. For comparison, adjusted p-values according to the Bonferroni–Holm method[30] are also calculated.

Table 1 shows the unadjusted p-values of the separate maximum tests for $H_1, \ldots, H_6$, adjusted p-values resulting from the application of the Bonferroni-Holm method to the former unadjusted p-values and adjusted p-values calculated by applying the closed testing procedure outlined in Section 3.4 to the set of hypotheses $\{H_1, \ldots, H_6\}$. Hypotheses H_1, H_2 and H_3 are rejected at a family-wise 5% significance level with both multiplicity adjustments. Also, for H_4 and H_5, both methods give similar results and do not reject. The test for H_6 has a local p-value of 0.0237, and the Bonferroni–Holm adjustment results in an adjusted p-value of 0.0710, such that the hypothesis is not rejected with this procedure. In contrast, the closed test based on maximum tests across multiple marginal GEE models results in a multiplicity adjusted p-value of 0.0362, allowing for the rejection of H_6.

Table 1. Unadjusted and adjusted p-values for maximum-type Wald tests in the retina disease example.

Hypothesis	Unadjusted p	Holm	mmmGEE
H_1	<0.0001	<0.0001	<0.0001
H_2	<0.0001	<0.0001	<0.0001
H_3	0.0001	0.0002	0.0002
H_4	0.0773	0.0828	0.0819
H_5	0.0414	0.0828	0.0819
H_6	0.0237	0.0710	0.0363

Adjusted p-values are calculated using the Bonferroni-Holm method (Holm) and the closed testing procedure applied to contrasts across multiple marginal generalized estimating equation (GEE) models (mmmGEE).

5 Type I error rate and power comparisons in finite samples

A simulation study was performed to investigate the power and type I error rate of the proposed hypothesis tests as well as the coverage probability of simultaneous confidence intervals in settings with multiple differently scaled endpoints. The supplemental material Section S.2 contains a further simulation study assessing the coverage of simultaneous confidence intervals based on multiple marginal GEE models in a recently planned clinical trial in actinic keratosis with a continuous and a binary endpoint.

5.1 Data generating model

We considered scenarios with $M \in \{3, 6, 9, 12\}$ endpoints $Y_{ij}^{(m)}, m = 1, \ldots, M,$ with subjects indexed $i = 1, \ldots, K$ and repeated measurements indexed $j = 1, \ldots, n_i$. For each subject, n_i was randomly drawn from a discrete uniform distribution on {2, 3, 4}. In each scenario, one third of the endpoints were continuous $Y_{ij}^{(m)} \in \mathbb{R}, m = 1, \ldots, M/3,$ with a conditional normal distribution (conditional on the covariates $Group_i$ and $x_i^{(m)}$) with variance 1 and mean

$$\mu_{ij}^{(m)} = \beta_0^{(m)} + Group_i \beta_1^{(m)} + x_i^{(m)} \beta_2^{(m)} \tag{13}$$

One third were count data endpoints $Y_{ij}^{(m)} \in \{0, 1, 2, \ldots\}, m = M/3 + 1, \ldots, 2M/3,$ with a conditional negative binomial distribution with variance $\mu_{ij}^{(m)} + \mu_{ij}^{(m)^2}$ and mean structure

$$\log \mu_{ij}^{(m)} = \beta_0^{(m)} + Group_i \beta_1^{(m)} + x_i^{(m)} \beta_2^{(m)} \tag{14}$$

The final third were binary endpoints $Y_{ij}^{(m)} \in \{0, 1\}, m = 2M/3 + 1, \ldots, M,$ with a conditional Bernoulli distribution with mean structure.

$$\log \frac{\mu_{ij}^{(m)}}{1 - \mu_{ij}^{(m)}} = \beta_0^{(m} + Group_i \beta_1^{(m)} + x_i^{(m)} \beta_2^{(m)} \tag{15}$$

Here, *Group* is a binary variable in {0, 1}, e.g. indicating treatment versus control. In the simulation study, inference for the corresponding coefficients $\beta_1^{(m)}, m = 1, \ldots, M$ was studied. $x_i^{(m)}$ corresponds to a covariate that is specific for the m-th outcome and that is observed once for each patient. The vector $(x_i^{(1)}, \ldots, x_i^{(M)})$ was drawn from a multivariate normal distribution with zero mean vector, unit variances and all pair-wise correlations set to 0.4.

To simulate correlated observations $(Y_{i1}^{(1)}, \ldots, Y_{in_i}^{(1)}, \ldots, Y_{i1}^{(M)}, \ldots, Y_{in_i}^{(M)})$, for each subject, $i = 1, \ldots, K$ first a latent multivariate normal vector $\boldsymbol{\xi}_i = (\xi_{i1}^{(1)}, \ldots, \xi_{in_i}^{(1)}, \ldots, \xi_{i1}^{(M)}, \ldots, \xi_{in_i}^{(M)})$ with mean zero and unit variances was sampled. Elements of $\boldsymbol{\xi}_i$ corresponding to repeated observations of the same endpoint had pair-wise correlations of 0.75. The correlation between elements corresponding to different endpoints was 0, 0.25, 0.5 or 0.75 to model zero, weak, intermediate and strong correlations between endpoints. The intermediate correlation was our base setting used in most simulation scenarios. Observations on the continuous, count data and binary outcomes were then obtained by the quantile substitution $Y_{ij}^{(m)} = Q_{ij}^{(m)}(\Phi(\xi_{ij}^{(m)}))$. Here, Φ is the standard normal distribution function and $Q_{ij}^{(m)}$ is the quantile function of, depending on the type of endpoint, a normal distribution with mean $\mu_{ij}^{(m)}$ and variance 1, a negative binomial distribution with mean $\mu_{ij}^{(m)}$ and dispersion parameter 1 (resulting in a variance of $\mu_{ij}^{(m)} + \mu_{ij}^{(m)^2}$), or a Bernoulli distribution with mean $\mu_{ij}^{(m)}$.

The resulting pair-wise correlations between the marginal Wald or score statistics were close to the corresponding correlation between the latent variables or, for pairs involving non-continuous endpoints, slightly below the latent variable correlation.

5.2 Hypothesis tests and confidence intervals

We tested the global null hypothesis $H_0 : \beta_1^{(1)} = \cdots = \beta_1^{(M)} = 0$. For the scenario with $M = 3$ endpoints and intermediate correlations, we also tested the elementary hypotheses $H_1 : \beta_1^{(1)} = 0$, $H_2 : \beta_1^{(2)} = 0$ and

$H_3 : \beta_1^{(3)} = 0$, using the closed testing approach of Section 3. The estimates $\hat{\boldsymbol{\beta}}^{(1)}, \ldots, \hat{\boldsymbol{\beta}}^{(M)}$ were calculated from marginal GEE models with mean structures as defined in equations (13) to (15), canonical variance functions for linear, Poisson and logistic regression models, respectively, and subject as clustering variable. Note that inference in the Poisson type GEE models is valid in the presence of overdispersion due to the robust covariance estimation. The exchangeable working correlation structure was specified for all models.

We investigated the performance of the following hypothesis tests as described in Section 3: The quadratic form Wald test using the chi-squared, F or scaled F reference distribution, the maximum-type Wald test using the multivariate normal or multivariate t-distribution as reference distribution, the quadratic form score test with a chi-squared reference distribution and the maximum-type score test with a multivariate normal reference distribution. For the Wald statistics, all tests were calculated, both, with and without bias adjustment of the covariance matrix estimate. For comparison, we further included Bonferroni–Holm tests for the maximum-type Wald statistics using the $1\text{-}\alpha/2/M$ quantile of a univariate normal or univariate t-distribution as critical quantile.

Simultaneous confidence intervals according to equation (6) were calculated for scenarios with three endpoints and intermediate correlations for $\beta_1^{(1)}$, $\beta_1^{(2)}$ and $\beta_1^{(3)}$, with and without bias adjustment of the covariance matrix estimate and based on a critical quantile of either a multivariate normal or t-distribution.

For methods based on a multivariate t-distribution or an F-distribution, $df = K - 3$ error degrees of freedom were used. For all tests, the nominal type I error rate was $\alpha = 0.05$.

5.3 Simulation scenarios

We considered scenarios with $K = 40$ and $K = 100$ subjects, with $K/2$ subjects in each class of the *Group* variable. The true coefficients in the data generating models were $\boldsymbol{\beta}^{(m)} = (0, \beta_1^{(m)}, 0.25)$ for the continuous endpoints, $\boldsymbol{\beta}^{(m)} = (0, \beta_1^{(m)}, 0.25)$ for the count data endpoints and $\boldsymbol{\beta}^{(m)} = (-0.75, \beta_1^{(m)}, 0.25)$ for the binary endpoints. To investigate type I error rates, simulations were performed under the global null hypothesis where $\beta_1^{(1)} = \cdots = \beta_1^{(M)} = 0$. For the case $M = 3$, we studied the power under alternative hypotheses with an effect in (a) all three endpoints, (b) in endpoints 2 and 3 only and (c) in endpoint 3 only. For $M \in \{6, 9, 12\}$, we considered scenarios with an effect in all endpoints. To model an effect in the respective endpoint with $K = 40$ subjects, we set the coefficients $\beta_1^{(m)}$ to 0.75, 0.95 and 1.5 for continuous, count and binary endpoints, respectively. In scenarios with $K = 100$ subjects, the respective values for $\beta_1^{(m)}$ were 0.45, 0.6 and 0.9. This choice of parameters results in similar expectations of the marginal Wald and score statistics close to 2.5 for each endpoint under the respective marginal alternative in all scenarios. For an unadjusted single-endpoint test based on a standard normal distribution, this corresponds to a power of approximately 70%.

For each scenario, 10^5 simulation runs were performed, except for more computation intensive simulations addressing the effect of increasing numbers of endpoints, where 2×10^4 simulation runs were performed. The power for each test was calculated as the proportion of simulation runs in which H_0 was rejected.

5.4 Simulation results

Simulation results regarding the type I error rate of tests for H_0 with $M = 3$ and $M = 12$ and intermediate correlations are shown in Table 2. Using any of the Wald tests without small sample adjustments leads to severe inflation of the type I error rate, and the inflation is increasing with the number of endpoints and decreasing with the number of subjects. Among the studied scenarios, the type I error rate was up to 10% for the unadjusted maximum-type Wald test and up to 41% for the unadjusted quadratic form Wald test. Both, the bias adjustment of the covariance estimate and the distributional approximation via an F- or a multivariate t-distribution, are required to control the type I error rate at the nominal level. For the quadratic form Wald test, using a scaled F statistic in analogy to Hotelling's T^2 test is required to control the type I error rate across all scenarios.

The score statistics exhibit favorable properties in the simulation, with type I error rates very close to the nominal level. No small sample adjustment in terms of bias adjustment or refined distributional approximation is required.

Table 2. Type I error rate in the simulations with $M = 3$ endpoints and $M = 12$ endpoints with intermediate correlations when testing $H_0 : \beta_1^{(1)} = \cdots = \beta_1^{(M)} = 0$ using the methods described in Section 3 or a Bonferroni test.

				Type I error rate (%)			
				$M = 3$		$M = 12$	
Statistic	Type	Approximation	Bias adj.	$K = 40$	$K = 100$	$K = 40$	$K = 100$
Wald	Quadratic	Chi-squared	no	9.8	6.8	41.0	15.1
			yes	6.2	5.5	28.1	11.6
		F	no	7.6	6.1	30.0	12.1
			yes	4.7	4.9	19.1	9.0
		Scaled F	no	6.4	5.6	8.6	6.7
			yes	3.8	4.5	4.3	4.7
	Maximum	MVN	no	8.1	6.1	10.4	6.8
			yes	5.2	5.1	6.2	5.5
		MVT	no	6.5	5.6	7.6	6.0
			yes	4.0	4.6	4.3	4.7
		Normal-Bonferroni	no	7.3	5.5	8.4	5.3
			yes	4.7	4.5	4.9	4.2
		t-Bonferroni	no	5.8	5.0	5.6	4.5
			yes	3.5	4.0	3.0	3.5
Score	Quadratic	Chi-squared	no	4.6	4.9	2.2	4.0
	Maximum	MVN	no	5.0	5.0	4.6	4.9

The tests are based on quadratic form or maximum-type Wald or score statistics. The reference distributions are chi-squared, *F*, scaled *F*, MVN or MVT. All Wald tests were performed with and without bias adjustment (column 'Bias adj.'). The considered sample sizes were $K = 40$ and $K = 100$. The results are based on 10^5 simulation runs. MVN: multivariate normal; MVT: multivariate *t*.

We studied the power of those procedures for which type I error rate control was observed in the simulation. For $M = 3$ endpoints and intermediate correlations, the results for the test of $H_0 : \beta_1^{(1)} = \beta_1^{(2)} = \beta_1^{(3)} = 0$ under scenarios (a), (b) and (c) with effects in three, two and one endpoint are shown in Table 3. Throughout these scenarios, the multivariate maximum-type Wald test has a power advantage of some percentage points over the Bonferroni test. Furthermore, the score test is considerably more powerful than the Wald test and this holds for, both, the quadratic form statistics and the maximum statistics.

The quadratic form Wald test has more power to reject the global null hypothesis in scenario (b) and almost identical power in scenario (c) compared to scenario (a). This observation is in agreement with the discussion in Section 3.3. Under the simulation settings, the correlation matrix of $(\hat{\beta}_1^{(1)}, \hat{\beta}_1^{(2)}, \hat{\beta}_1^{(3)})$ is approximately $\boldsymbol{C} = ((1, 0.5, 0.5)^T, (0.5, 1, 0.5)^T, (0.5, 0.5, 1)^T)$. Thus, the non-centrality parameter of the quadratic form Wald test is approximately $(2.5, 2.5, 2.5)\boldsymbol{C}^{-1}(2.5, 2.5, 2.5)^T = 9.4$ under scenario (a), $(0, 2.5, 2.5)\boldsymbol{C}^{-1}(0, 2.5, 2.5)^T = 12.5$ under scenario (b) and $(0, 0, 2.5)\boldsymbol{C}^{-1}(0, 0, 2.5)^T = 9.4$ under scenario (c). This corresponds to a theoretical power of 73%, 86% and 73%, respectively, under a chi-squared approximation, which is in line with the simulation results in Table 3. For the quadratic form score test, similar results hold. Under the simulation settings the pair-wise correlations between the marginal score statistics were approximately 0.5, however the expectation of the score statistics were slightly different, with values approximately 2.4, 2.3 and 2.6 for the continuous, the count data and the binary endpoint, respectively. Hence, even scenario (c), in which the only effect is on the binary endpoint, results in more power than scenario (a) for the quadratic form score test.

We further studied for scenario (a), the power of closed testing procedures, which utilize the above tests for each intersection hypothesis, to reject particularly H_1, H_2 or H_3, as well as the power to reject at least one elementary hypothesis or all elementary hypotheses, see Table 4. For comparison, the closed testing procedure based on Bonferroni tests (which results in the Bonferroni-Holm procedure) is included. Similar to the results on the global test, the maximum-type Wald test is more powerful than the Bonferroni–Holm test and the score tests are for most decisions more powerful than the Wald tests.

The coverage probability of simultaneous confidence intervals for $(\beta_1^{(1)}, \beta_1^{(2)}, \beta_1^{(3)})$ did not depend on the actual values of the coefficients, up to simulation error. The observed values for scenario (a) are shown in Table 5.

Table 3. Power to reject $H_0 : \beta_1^{(1)} = \beta_1^{(2)} = \beta_1^{(3)} = 0$ with selected testing approaches that control the type I error rate (see Table 2 for details).

				Power (%)					
				K = 40			K = 100		
Statistic	Type	Approx.	Bias adj.	a	b	c	a	b	c
Wald	Quadratic	Scaled F	yes	67.5	83.7	63.1	70.5	86.8	70.5
Wald	Maximum	MVT	yes	77.7	71.2	53.7	78.8	72.9	57.8
Wald	Maximum	t-Bonferroni	yes	75.5	68.6	50.8	76.9	70.9	55.5
Score	Quadratic	Chi-squared	no	72.1	86.7	74.5	71.9	87.7	74.2
Score	Maximum	MVN	no	80.7	74.3	64.7	79.7	73.9	61.6

The power was calculated for sample sizes $K = 40$ and $K = 100$ in three different scenarios. In scenario (a), there was an effect in all three endpoints ($\beta_1^{(i)} \neq 0, i = 1, 2, 3$), in scenario (b) there was an effect in endpoints 2 and 3 ($\beta_1^{(1)} = 0$), and in scenario (c) there was an effect in endpoint 3 only ($\beta_1^{(1)} = \beta_1^{(2)} = 0$). See text for the exact values of non-zero coefficients. The results are based on 10^5 simulation runs. MVN: multivariate normal; MVT: multivariate *t*.

Table 4. Power of closed testing procedures to reject $H_1 : \beta_1^{(1)} = 0,\ H_2 : \beta_1^{(2)} = 0,\ H_3 : \beta_1^{(3)} = 0$, at least one elementary hypothesis (any H_i), or all elementary hypotheses.

K	Statistic	Type	Approx.	Bias adj.	H_1	H_2	H_3	any H_i	all H_i
40	Wald	Quadratic	scaled F	yes	53.1	54.0	54.3	67.1	41.4
40	Wald	Maximum	MVT	yes	59.0	59.4	60.6	77.7	42.2
40	Wald	Maximum	t-Bonferroni	yes	57.4	57.9	58.9	75.5	41.5
40	Score	Quadratic	Chi-squared	no	57.7	53.8	64.0	71.7	46.0
40	Score	Maximum	MVN	no	63.0	58.1	69.2	80.7	46.6
100	Wald	Quadratic	Scaled F	yes	54.3	56.3	57.8	70.1	43.1
100	Wald	Maximum	MVT	yes	58.8	60.8	62.8	78.8	43.2
100	Wald	Maximum	t-Bonferroni	yes	57.5	59.4	61.4	76.9	42.7
100	Score	Quadratic	Chi-squared	no	56.3	56.1	61.5	71.5	45.0
100	Score	Maximum	MVN	no	60.6	60.2	66.1	79.7	45.1

Results are shown for the simulation scenario (a) with an effect in all endpoints. The results are based on 10^5 simulation runs. MVN: multivariate normal; MVT: multivariate *t*.

Table 5. Simultaneous coverage probability of nominal 95% simultaneous confidence intervals for $(\beta_1^{(1)}, \beta_1^{(2)}, \beta_1^{(3)})$ for scenario (a) with sample sizes $K = 40$ and $K = 100$.

Approximation	Bias adj.	$K = 40$	$K = 100$
MVN	no	91.9	93.7
MVN	yes	94.7	94.8
MVT	no	93.4	94.3
MVT	yes	95.8	95.2

The intervals were calculated based on an approximating MVN or MVT distribution, with and without bias adjustment for the covariance matrix estimate. The results are based on 10^5 simulation runs. MVN: multivariate normal; MVT: multivariate *t*.

Both considered small sample adjustments are required to achieve a coverage probability of at least the nominal value of 95%. Similar results were observed in the simulation contained in the supplemental material Section S.2.

In a further simulation study, we investigated the impact of increasing the number of endpoints and increasing the correlation between endpoints for scenarios with $K = 40$ subjects and an effect in all endpoints under the alternative. We included only those tests that controlled the type I error rate.

The simulation results are shown in Figure 1. Here, we also included the case $M = 1$. To allow for an unambiguous comparison with the case of multiple endpoints of different types, we computed for $M = 1$ the results for models with a single continuous, count data and binary endpoint, respectively, and plotted the average power across these three models from a total of 2×10^4 simulations.

The benefit of the maximum-type Wald and score tests over the Bonferroni test becomes more pronounced for larger correlations and their power (under the considered alternative with an effect in all endpoints) is increasing with the number of endpoints. Under high correlations, the power of the maximum-type Wald and score test is approximately constant with an increasing number of endpoints, whereas the power of the Bonferroni test is decreasing. Note that in an extreme case of correlation 1, the maximum test would be identical to a test for a single endpoint, with no loss in power, whereas the Bonferroni test would correspond to a single-endpoint test at level α/M hence loosing power.

The power of the quadratic form Wald and score tests depends strongly on the correlation between endpoints. As seen in Table 3 and discussed in Section 3.3, if there is an effect in all endpoints and the endpoints are positively correlated, the direction of the effects is not the direction of deviations from the null hypothesis which are considered particularly large by the metric of these tests. This becomes more pronounced with an increasing number of endpoints and increasing correlation, and the power of the quadratic form tests decreases rapidly.

6 Software implementation

The proposed test procedures were implemented in the R-package 'mmmgee'[31] that is available from the CRAN repository.[32] The model fitting routines are based on those of the R-package 'geeM',[33] and multivariate normal or t distribution probabilities are calculated using the package 'mvtnorm'.[34]

The mmmgee package provides three main functions: geem2 fits marginal GEE models as described in Sections 2.1 and 2.2. mmmgee calculates the estimate (equation (3)) or the bias adjusted estimate (equation (4)) of the covariance matrix of a stacked vector of regression coefficients from multiple marginal GEE models fitted with geem2. mmmgee.test calculates the multiple hypothesis tests and simultaneous confidence intervals described in Section 3. The latter functions are applied to a list of models fitted with geem2. As a special case, the package may also be applied to test hypotheses within a single GEE model.

An instance of a simulated data set with $M = 3$, $K = 40$ and intermediate correlations as described in Section 5 is included in the package as exemplary data. The R code below invokes an analysis as used in the simulation studies in Section 5. Marginal GEE models are fit for the three endpoints using geem2. The function mmmgee.test is applied to test the global null hypothesis $H_0 : \beta_1^{(1)} = \beta_1^{(2)} = \beta_1^{(3)} = 0$ as well as the elementary hypotheses $H_1 : \beta_1^{(1)} = 0$, $H_2 : \beta_1^{(2)} = 0$ and $H_3 : \beta_1^{(3)} = 0$ in a closed testing procedure. In the example, a maximum-type Wald test using the bias adjusted covariance matrix estimate and a multivariate t reference distribution is requested.

```
library(mmmgee)
data(datasim)
mod1<-geem2(Y.lin~gr.lang+x1,id=id,data=datasim,
            family="gaussian",corstr="exchangeable")
mod2<-geem2(Y.poi~gr.lang+x2,id=id,data=datasim,
            family="poisson",corstr="exchangeable")
mod3<-geem2(Y.bin~gr.lang+x3,id=id,data=datasim,
            family="binomial",corstr="exchangeable")
Li<-matrix(c(0,1,0),nrow=1)
mmmgee.test(list(mod1,mod2,mod3),L=list(Li,Li,Li),
            statistic="Wald",type="maximum",biascorr=TRUE,
            asymptotic=FALSE,closed.test=TRUE)
```

The output includes the test statistic, degrees of freedom and p-value for the test of the global null hypothesis as described in Section 3.1. It further shows the estimated contrasts, which in this case correspond to $\hat{\beta}_1^{(1)}$, $\hat{\beta}_1^{(2)}$ and $\hat{\beta}_1^{(3)}$, the right hand side value of each contrast under the respective null hypothesis, the unadjusted p-values and multiplicity adjusted p-values according to the closed testing procedure of Section 3.4:

```
Hypothesis tests for linear contrasts in multiple marginal GEE models

Statistic: Maximum-type Wald statistic
Approximation: Multivariate t
Alternative: Undirected

Global test:
Max|T| = 3.368, df = 37, p-value = 0.005021

Closed testing procedure:
contrast estimate  rhs  p.unadj  p.adjusted
1      1    0.752    0  0.004586   0.009038
2      2    1.309    0  0.006406   0.009038
3      3    1.931    0  0.001778   0.005021
```

See the supplemental material S.3 for further examples.

7 Discussion

We proposed a general inference framework for multiple or multivariate outcomes. In particular, we considered the problem of testing multiple hypotheses, the need to account for dependencies of observations within the same subject, the lack of accuracy of asymptotic methods in small sample sizes, and we tried to give some advice on the choice between particular test statistics in the multivariate setting.

The approach is based on multiple marginal models[2] and requires a distributional model for the joint vector of parameter estimates from these models, but it is not required to assume a fully specified joint model of all outcomes. Marginal GEE models, accounting for dependent observations with respect to one endpoint, fit naturally in the multiple marginal model framework, since both concepts are based on estimating equations that are sums of independent contributions of different subjects and both utilize robust sandwich variance estimation. Note that usual generalized linear models and linear models may be seen as special cases of GEE models and can be included in the proposed framework. As alternative to using multiple marginal models, the correlation between endpoints could be utilized in a weighted estimation of regression coefficients, which may reduce their variance. This idea was studied by Fitzmaurice and Laird[15] and Rochon[3] for the special case of a continuous and a binary endpoint. However, as a simulation study by Teixeira-Pinto and Normand[35] suggests, the improvement is small, and the required additional nuisance parameters may introduce further variability in the case of small samples. We therefore focussed on estimation via marginal models.

The proposed method provides multiplicity adjustment when testing multiple hypotheses or when constructing simultaneous confidence intervals and it takes the correlation between the studied parameters into account. Hence, it is more efficient than commonly used methods that are based on the Bonferroni inequality. Furthermore, while the Bonferroni adjustment is applicable to maximum-type tests, the multivariate normal approximation of the parameter estimates allows for construction of more general test statistics.

Insufficient small sample accuracy of asymptotic methods is a frequent problem that particularly affects studies at early stages of research and studies in rare diseases.[36] For many asymptotic methods, a major improvement in terms of type I error rate control can be achieved by replacing normal with t-distributions and chi-squared with F-distributions. In general, the joint distribution of the $(\boldsymbol{L}\hat{\boldsymbol{\beta}})_i/\hat{SE}_i, i = 1, \ldots, c$ will not be exactly a multivariate t-distribution (see the book by Kotz and Nadarajah[21] for the definition of the multivariate t-distribution). Nonetheless, the multivariate t-distribution often provides a good approximation.[7,8,22] Also, the simulation results (see Section 5) suggest that the tests and confidence intervals based on $t_{1-\alpha}$ often have error rates close to the nominal level even for small sample sizes. In any case, the liberalism of the hypothesis tests and confidence intervals is reduced compared to the multivariate normal approximation, because $t_{1-\alpha} > z_{1-\alpha}$. Asymptotically, both approaches are identical.

However, with the exception of certain models under the assumption of normally distributed data, there is no direct way to determine the respective number of error degrees of freedom. We used a simple method essentially subtracting the number of parameters in a model from the number of subjects, with convincing results in the numeric simulations. Several other approaches have been proposed in different contexts and it may be worthwhile to include these in further research.

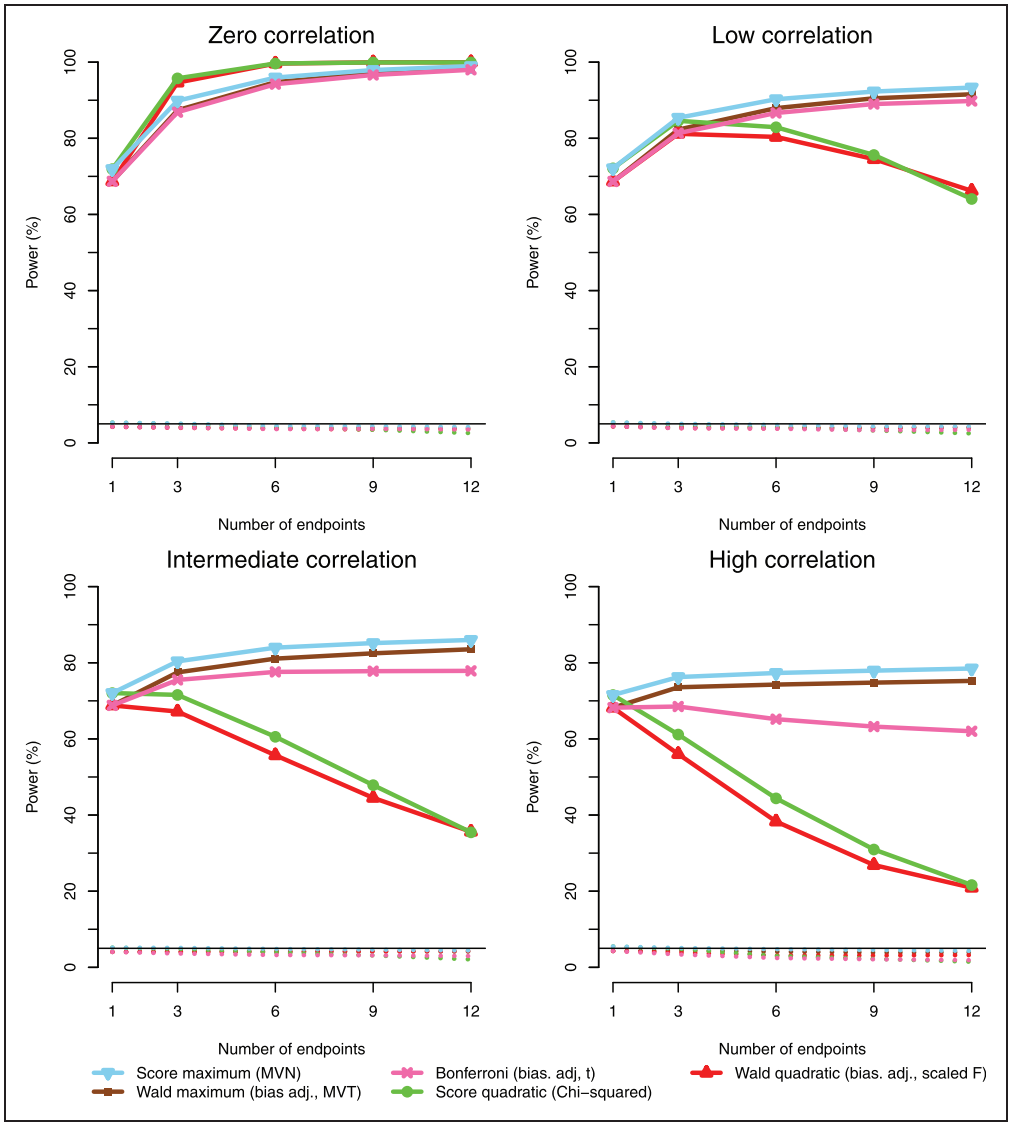

Figure 1. Power to reject the global null hypothesis $H_0 : \beta_1^{(1)} = \cdots = \beta_1^{(M)} = 0$ under an alternative with an effect in all M endpoints (solid lines) and type I error rate under H_0 (dotted lines) for scenarios with $K = 40$ subjects and increasing number of endpoints. The correlation between marginal Wald or score statistics is approximately 0.25, 0.5 and 0.75 in the scenarios with low correlation, intermediate correlation and high correlation. The nominal level of 0.05 is indicated by a horizontal line. The studied tests are listed in the legend. The information in parentheses shows that the bias adjustment for the covariance matrix was applied to all tests using Wald statistics; furthermore, the reference distributions with abbreviations as in Table 2 are indicated. The results are based on 2×10^4 simulation runs.

Pan and Wall[9] report good results for single GEE models using a Satterthwaite approximation. However, this method requires the estimation of the variance of the covariance matrix estimate. Alternatively, degrees of freedom may be calculated from the effective sample size, which is the number of independent observations that would result in the same efficiency as the observed sample of partially dependent observations.[22] This method requires that the covariance structure is correctly specified, which is otherwise not required for GEE models. In some cases, it may be reasonable to attribute different degrees of freedom to different tested contrast, e.g. if the number of parameters strongly differs between the marginal models. In that case, a method described by Hasler and Hothorn[8] may be utilized: For each tested contrast, a critical value is calculated from a multivariate t-distribution with common correlation matrix and contrast-specific degrees of freedom. The test decision is based on comparison of the individual test statistics with the respective critical values. Another way to improve the distributional normal approximation of a statistic is to directly calculate and adjust for the error of approximation. Kauermann and

Carroll[37] propose this solution for the case of univariate contrast tests in GEE models. The method could in principle be extended to multivariate tests.

The other small sample improvement we investigated is the bias adjustment of the covariance matrix estimate. We focused on the method of Mancl and DeRouen[7]; however, related methods proposed in the literature for single GEE models[14] may as well be extendable to the multiple marginal models approach.

We regarded the score test as an approximation to the Wald test. Note that generalized score tests with quadratic form test statistics may also be constructed from the score vector $\boldsymbol{U}$ and a generalized inverse of its asymptotic covariance matrix (which may be singular).[26] For the case of a linear hypothesis H_0, the resulting test is identical to the quadratic form score test motivated via approximation of the Wald test. The latter approach is, however, easily applicable to construct maximum type tests. In the numeric investigations, the score tests did not require adjustments to the asymptotic approximation to control the type I error rate and they were more powerful than the corresponding Wald tests. In contrast to Guo et al.,[38] who studied quadratic form score tests for single GEE models, we did not observe a conservative behaviour, but the type I error rate was controlled almost exactly at the nominal level. These results may not hold for all possible analysis scenarios but they suggest the score tests as viable small sample alternative to the Wald tests. Confidence sets for $\boldsymbol{L\beta}$ corresponding to a score test may in principle be found as set of all vectors $\boldsymbol{r}' \in \mathbb{R}^c$ such that $H_0 : \boldsymbol{L\beta} - \boldsymbol{r}' = \boldsymbol{0}$ is rejected.[38] To provide contiguous intervals or sets, the test statistic needs to be a convex function of $\boldsymbol{r}$. In contrast to the Wald statistic, the score statistic depends on $\boldsymbol{r}$ in a non-trivial way, as the statistic is based on a model fitted under the constraint $\boldsymbol{L\beta} - \boldsymbol{r} = \boldsymbol{0}$. Thus, the required convexity property needs to be checked for each given class of models, which may not be easily done except for some special cases.

We focused on two-sided inference. The extension to one-sided tests for null hypotheses of the form $H_0' = \cap_{i=1}^{c}\{(\boldsymbol{L\beta})_i \leq r_i\}$ and according one-sided confidence intervals is straight forward for maximum-type Wald tests (compare with Hothorn[20]). The least favorable configuration under H_0' is $\boldsymbol{L\beta} = \boldsymbol{r}$ since $\hat{\boldsymbol{\Sigma}}$ is estimated without restriction and the multivariate normal or t reference distribution is monotone in the assumed mean vector. Hence, evaluation of the one-sided Wald test under the configuration $\boldsymbol{L\beta} = \boldsymbol{r}$ is sufficient. To extend the maximum-type score test to one-sided hypotheses, the restricted estimate $\widetilde{\boldsymbol{\beta}}$ has to be calculated under the according inequality restriction which is subject of further research.

Acknowledgements

The authors gratefully acknowledge Dr Anna Ledolter and Dr Markus Ritter for their agreement to use the retina disease study as example and Dr Hannes Trattner and Dr Sonja Radakovic for their agreement to use the actinic keratosis study as example.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work has been funded by the FP7-HEALTH-2013-INNOVATION-1 project Advances in Small Trials Design for Regulatory Innovation and Excellence (ASTERIX) Grant Agreement No. 603160.

ORCID iD

Robin Ristl https://orcid.org/0000-0002-4163-9236

Supplemental Material

Supplemental material for this article is available online.

References

1. Liang KY and Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13–22.
2. Pipper CB, Ritz C and Bisgaard H. A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2012; **61**: 315–326.
3. Rochon J. Analyzing bivariate repeated measures for discrete and continuous outcome variables. *Biometrics* 1996; **52**: 740–750.

4. Jensen SM, Pipper CB and Ritz C. Evaluation of multi-outcome longitudinal studies. *Stat Med* 2015; **34**: 1993–2003.
5. Verbeke G, Fieuws S, Molenberghs G, et al. The analysis of multivariate longitudinal data: A review. *Stat Meth Med Res* 2014; **23**: 42–59.
6. Van der Vaart AW. *Asymptotic statistics*. Cambridge: Cambridge University Press, 2000.
7. Mancl LA and DeRouen TA. A covariance estimator for gee with improved small-sample properties. *Biometrics* 2001; **57**: 126–134.
8. Hasler M and Hothorn LA. Multiple contrast tests in the presence of heteroscedasticity. *Biom J* 2008; **50**: 793–800.
9. Pan W and Wall MM. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Stat Med* 2002; **21**: 1429–1441.
10. McCullagh P and Nelder JA. *Generalized linear models*. 2nd ed. London: Chapman & Hall, 1989.
11. Huber PJ et al. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 221–233. Berkeley: University of California Press.
12. White H. Maximum likelihood estimation of misspecified models. *Econometrica* 1982; **50**: 1–25.
13. Cox DR and Hinkley DV. *Theoretical statistics*. Boca Raton: CRC Press, 1979.
14. Wang M, Kong L, Li Z, et al. Covariance estimators for generalized estimating equations (gee) in longitudinal analysis with small samples. *Stat Med* 2016; **35**: 1706–1721.
15. Fitzmaurice GM and Laird NM. Regression models for a bivariate discrete and continuous outcome with clustering. *J Am Stat Assoc* 1995; **90**: 845–852.
16. Robins JM, Rotnitzky A and Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc* 1995; **90**: 106–121.
17. Preisser JS, Lohman KK and Rathouz PJ. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Stat Medicine* 2002; **21**: 3035–3054.
18. Chen B, Yi GY and Cook RJ. Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random. *J Am Stat Assoc* 2010; **105**: 336–353.
19. SAS Institute Inc. *SAS/STAT(R) 15.1 user's guide*. Cary, NC: SAS Institute Inc., 2018.
20. Hothorn T, Bretz F and Westfall P. Simultaneous inference in general parametric models. *Biom J* 2008; **50**: 346–363.
21. Kotz S and Nadarajah S. *Multivariate t-distributions and their applications*. Cambridge University Press, 2004.
22. Faes C, Molenberghs G, Aerts M, et al. The effective sample size and an alternative small-sample degrees-of-freedom method. *Am Stat* 2009; **63**: 389–399.
23. Munzel U and Hothorn LA. A unified approach to simultaneous rank test procedures in the unbalanced one-way layout. *Biom J* 2001; **43**: 553–569.
24. Rao CR and Toutenburg H. *Linear models*. Berlin: Springer, 1995.
25. Kenward MG and Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; **53**: 983–997.
26. Boos DD. On generalized score tests. *The American statistician* 1992; **46**: 327–333.
27. Marcus R, Eric P and Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**: 655–660.
28. Shaffer JP. Modified sequentially rejective multiple test procedures. *J Am Stat Assoc* 1986; **81**: 826–831.
29. Xi D, Glimm E, Maurer W, et al. A unified framework for weighted parametric multiple test procedures. *Biom J* 2017; **59**: 918–931.
30. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat* 1979; **6**: 65–70.
31. Ristl R. *mmmgee: simultaneous inference for multiple linear contrasts in GEE models* (R package version 1.20), https://CRAN.R-project.org/package=mmmgee (2019, accessed 22 August 2019).
32. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, www.R-project.org/ (2016, accessed 22 August 2019).
33. McDaniel LS, Henderson NC and Rathouz PJ. Fast pure R implementation of GEE: application of the matrix package. *R J* 2013; **5**: 181.
34. Genz A and Bretz F. *Computation of multivariate normal and t probabilities*. Lecture Notes in Statistics. Heidelberg: Springer-Verlag, 2009.
35. Teixeira-Pinto A and Normand SLT. Correlated bivariate continuous and binary outcomes: issues and applications. *Stat Med* 2009; **28**: 1753–1773.
36. Hee SW, Willis A, Smith CT, et al. Does the low prevalence affect the sample size of interventional clinical trials of rare diseases? an analysis of data from the aggregate analysis of clinicaltrials.gov. *Orphanet J Rare Dis* 2017; **12**: 44.
37. Kauermann G and Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. *J Am Stat Assoc* 2001; **96**: 1387–1396.
38. Guo X, Pan W, Connett JE, et al. Small-sample performance of the robust score test and its modifications in generalized estimating equations. *Stat Med* 2005; **24**: 3479–3495.