Computational Statistics and Data Analysis 64 (2013) 87–98

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Simultaneous confidence intervals for comparing margins of multivariate binary data

Bernhard Klingenberg [a,*,1], Ville Satopää [b]

[a] Department of Mathematics and Statistics, Williams College, Williamstown, MA 01267, United States

[b] Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104-6340, United States

ARTICLE INFO

Article history:
Received 17 August 2012
Received in revised form 12 February 2013
Accepted 13 February 2013
Available online 6 March 2013

Keywords:
Correlated binary responses
Familywise error rate
Marginal homogeneity
Multiple comparisons

ABSTRACT

In many applications two groups are compared simultaneously on several correlated binary variables for a more comprehensive assessment of group differences. Although the response is multivariate, the main interest is in comparing the marginal probabilities between the groups. Estimating the size of these differences under strong error control allows for a better evaluation of effects than can be provided by multiplicity adjusted P-values. Simultaneous confidence intervals for the differences in marginal probabilities are developed through inverting the maximum of correlated Wald, score or quasi-score statistics. Taking advantage of the available correlation information leads to improvements in the joint coverage probability and power compared to straightforward Bonferroni adjustments. Estimating the correlation under the null is also explored. While computationally complex even in small dimensions, it does not result in marked improvements. Based on extensive simulation results, a simple approach that uses univariate score statistics together with their estimated correlation is proposed and recommended. All methods are illustrated using data from a vaccine trial that investigated the incidence of four pre-specified adverse events between two groups and with data from the General Social Survey.

1. Introduction

In order to adequately capture the difference between two groups, often several potentially correlated variables are measured simultaneously with the hope to better and more completely describe and understand group differences or treatment effects. Here, we focus on the case where the response variables are all binary and observed from two independent groups. Such comparisons are frequently encountered in toxicity and safety evaluations of medicinal products (see, e.g., the recent articles by Huang et al., 2011 and Davidov and Peddada, 2011 and references therein), but occur in many other areas such as quality of life assessments, opinion surveys, psychiatric and behavioral research or in the social and political sciences.

Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iJ})^t$ denote the response vector of J correlated binary variables in group $i = 1, 2$, and let $\pi_i(\boldsymbol{a}) = P(Y_{i1} = a_1, \ldots, Y_{iJ} = a_J)$ be the corresponding joint distribution of the observed response vector $\boldsymbol{a} = (a_1, \ldots, a_J)^t$, with $a_j \in \{0, 1\}$ for $j = 1, \ldots, J$. In this paper, we consider J to be in the single or lower double digits. For each of the two groups, consider the contingency table of dimension 2^J (one cell for each possible response sequence) that shows the counts of how many subjects (out of n_i in group i) had a particular response sequence. The counts in this table follow a multinomial distribution with cell probabilities $\{\pi_i(\boldsymbol{a})\}$.

The distribution of the response vector will be identical in the two groups if $\pi_1(\boldsymbol{a}) = \pi_2(\boldsymbol{a})$ for all possible response sequences $\boldsymbol{a}$, but this is a very stringent condition. Often, descriptions of group differences or treatment effects on a mul-

* Corresponding author. Tel.: +1 413 5972467.
E-mail addresses: bklingen@williams.edu (B. Klingenberg), satopaa@wharton.upenn.edu (V. Satopää).

[1] Supplementary material available online demonstrates the use of R code for all proposed methods.

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.02.016

tivariate binary response focus on *marginal* success probabilities for variable j in group i, given by $\pi_i(j) = P(Y_{ij} = 1) = \sum_{\boldsymbol{a}:a_j=1} \pi_i(\boldsymbol{a})$. These capture one major aspect where the joint distributions could differ. The *homogeneity* of these marginal probabilities across the two groups (for each j) leads to a hypothesis that Agresti and Klingenberg (2005) called simultaneous marginal homogeneity, where the null hypothesis is $H_0 : \delta_j = \pi_1(j) - \pi_2(j) = 0$ for all $j = 1, \ldots, J$. When $J = 1$, this is the regular homogeneity hypothesis for proportions in a single 2×2 table (one variable) and here we discuss inference for the extension to the multivariate case (several correlated variables). In this paper, however, we are not interested in hypothesis testing and associated global P-values (for H_0) or individual multiplicity-adjusted P-values (for the j-th sub-hypothesis), but rather want to give interval estimates of the δ_j's. This leads to a more informative description of the size and magnitude as well as the practical importance of group differences, as measured by the marginal proportions. Hence, our goal is to construct simultaneous confidence intervals (SCIs) for all J differences $\delta_j,\ j = 1, \ldots, J$.

The term simultaneous refers to the collection (or family) of confidence intervals that we are going to construct and here we focus on controlling the familywise error rate (FWER), i.e. the probability that at least one of the J confidence intervals fails to cover the true difference of proportion. Other, less stringent criteria for error control (such as controlling the false coverage rate, Benjamini and Yekutieli, 2005, see also Mehrotra and Adewale, 2012 for an application to correlated binary data) can be chosen but will lead to a different methodology. We further try to improve on the commonly used Bonferroni correction for FWER control by taking advantage of the correlation information that exists in the test statistics that we invert to obtain the intervals. These correlations are induced by the natural association among the response variables $Y_{i1}, \ldots, Y_{ij}$. For instance, for the first example in Section 5, the estimated odds of success for one variable are 16 times higher when another variable also resulted in a success, and similar strong associations are observed for many other pairs of variables in both groups. We want to investigate how this correlation information can be used to our advantage when forming SCIs.

Applying no multiplicity adjustment naturally provides the most powerful procedure and the shortest possible confidence intervals. However, the consequences of not adjusting the intervals for the simultaneous comparisons leads to statistical statements with true error (or overall coverage) rates that are unknown. Therefore, it seems more prudent to control the FWER at some *known* level, e.g., 5% or, as recently argued by Mehrotra and Adewale (2012) a more reasonable 10% when simultaneous statements are sought. Presenting the results alongside the unadjusted ones allows for assessing the effects of adjusting for multiplicity.

Most research on (simultaneous) marginal homogeneity for comparing multivariate binary data in two groups focuses on hypothesis testing of H_0 or computation of multiplicity adjusted P-values or posterior odds under a variety of assumptions and settings (Pocock et al., 1987; Westfall and Young, 1989; Lehmacher et al., 1991; Lefkopoulou and Ryan, 1993; Berry and Berry, 2004; Mehrotra and Heyse, 2004; Pipper et al., 2012; Mehrotra and Adewale, 2012). Constructing SCIs for effect size estimation and interpretation of the magnitude of marginal differences is not discussed. There are, however, results for the one-sample case. Goodman (1965), Fitzpatrick and Scott (1987), Sison and Glaz (1995) and Chafaï and Concordet (2009) present methods to form SCIs for multinomial probabilities or their differences (Piegorsch and Richwine, 2001) based on a *single* multivariate sample. SCIs for the marginal probabilities in this one-sample case are presented in Westfall (1985) using an iterative bootstrap approach to estimate the FWER, while confidence intervals for contrasts of these marginal probabilities were discussed by Bhapkar and Somes (1976), see also Chapter 10 of Tamhane and Hochberg (1987) for related results. Here, we extend some of the results to the two-sample case.

In Section 2 we propose a general strategy for forming SCIs based on inverting the maximum of adjusted Wald or score statistics. Since a full score approach is computationally demanding unless J is small, we develop two approaches based on a local score statistic. Inversion of the Pearson and likelihood ratio statistics are also briefly mentioned but quickly discarded based on simulation results. In Section 3 we discuss model-based methods of forming SCIs, such as fitting a marginal model via restricted generalized estimating equations (GEE) and inverting the maximum of quasi-score statistics. A simulation study in Section 4 evaluates all proposed methods under various scenarios in terms of simultaneous coverage and power. Section 5 illustrates the methodology using an example with $J = 4$ pre-specified adverse events in a vaccine trial and with $J = 6$ questions on the performance of the US government on various issues, using data from the General Social Survey. Section 6 concludes with a discussion and further research questions.

2. Inverting tests for simultaneous marginal homogeneity

To construct SCIs we write the simultaneous marginal homogeneity hypothesis in multiparameter from $H_0 : \boldsymbol{\delta} = \boldsymbol{\delta}_0$ (or, equivalently $H_0 : \bigcap_{j=1}^{J} H_{0j}$ with $H_{0j} : \delta_j = \delta_{j0}$) for a given null vector $\boldsymbol{\delta}_0 = (\delta_{10}, \ldots, \delta_{J0})^t$ of the marginal differences. Inverting H_0 with respect to $\boldsymbol{\delta}_0$ will yield the SCIs. Note that the associations among the binary responses $\{Y_{ij}\}_{j=1}^{J}$ for a given subject induce dependences among the marginal sample proportions $\{\hat{\pi}_i(j)\}_{j=1}^{J}$ in group i. This correlation translates to the elements $\hat{\delta}_j = \hat{\pi}_1(j) - \hat{\pi}_2(j)$ of the vector $\hat{\boldsymbol{\delta}}$ of differences in the marginal sample proportions. All test statistics developed in this section are based on standardized versions of $\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0$, which is asymptotically multivariate normal (MVN) with covariance matrix $\boldsymbol{\Sigma}$ given by diagonal and off-diagonal elements

$$\Sigma_{jj} = \mathrm{Var}[\hat{\delta}_j] = \sum_{i=1}^{2} \pi_i(j)[1 - \pi_i(j)]/n_i$$

$$\Sigma_{jj'} = \mathrm{Cov}[\hat{\delta}_j, \hat{\delta}_{j'}] = \sum_{i=1}^{2} [\pi_i(j, j') - \pi_i(j)\pi_i(j')]/n_i, \quad j \neq j', \tag{1}$$

where $\pi_i(j, j') = P(Y_{ij} = 1, Y_{ij'} = 1)$ is the joint success probability for variables j and j' in group i.

2.1. Inverting maximum-type test statistics

Let $\hat{\boldsymbol{\Sigma}}$ be a consistent estimator for $\boldsymbol{\Sigma}$ and let $\boldsymbol{T} = (T_1, \ldots, T_J)^t = \mathrm{diag}[\hat{\boldsymbol{\Sigma}}]^{-1/2}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)$, where $\mathrm{diag}[\hat{\boldsymbol{\Sigma}}]^{-1/2}$ denotes a diagonal matrix with diagonal elements $\hat{\Sigma}_{jj}^{-1/2}$. Further, let $\boldsymbol{R} = \mathrm{diag}[\boldsymbol{\Sigma}]^{-1/2}\ \boldsymbol{\Sigma}\ \mathrm{diag}[\boldsymbol{\Sigma}]^{-1/2}$ be the correlation matrix for the vector $\boldsymbol{T}$ and consider $\max_j |T_j|$ as a test statistic for H_0. Note that $\boldsymbol{T}$ is asymptotic MVN, and we expect the normal approximation to be good when the normal approximation for each T_j is good.

To control the FWER at level α, a critical value c is computed from solving $P_{H_0}(\max_j |T_j| > c) = 1 - P_{H_0}(|T_1| \leq c, \ldots, |T_J| \leq c) = \alpha$. Note that the null distribution of $\max_j |T_j|$ and hence the critical value c may vary with $\boldsymbol{\delta}_0$, depending on how $\boldsymbol{\Sigma}$ (and $\boldsymbol{R}$) are estimated. In general, the set of vectors $\boldsymbol{\delta}_0$ for which H_0 is not rejected (i.e., for which $\max_j |T_j| \leq c$) forms a simultaneous $(1-\alpha)100\%$ confidence region. Projecting the convex hull of this region on its J axes results in simultaneous (equivariant) confidence intervals for the δ_j's.

Incorporating the correlation information yields smaller critical values than given by the Bonferroni procedure. For instance, the estimated pairwise correlations for the test statistics $T_1, \ldots, T_4$ for the example in Section 5 range from 0.20 to 0.44. This leads to a critical value of $c = 2.47$ (when $\alpha = 5\%$), compared to $c = 2.50$ with the Bonferroni procedure, resulting in coverage closer to the nominal level and potentially more power (see simulation results in Section 4). The next three sections present maximum statistics that differ in the way $\boldsymbol{\Sigma}$ is estimated.

2.1.1. Maximum of adjusted Wald statistics

The maximum likelihood estimates (MLEs) for $\pi_i(j)$ and $\pi_i(j, j')$ are the marginal and joint sample proportions $y_{ij+}/n_i = \sum_{k=1}^{n_i} y_{ijk}/n_i$ and $\sum_{k=1}^{n_i} y_{ijk}y_{ij'k}/n_i$, respectively, where $y_{ijk} = 1$ if subject k in group i recorded a success on variable j and $y_{ijk} = 0$ otherwise. Estimating $\boldsymbol{\Sigma}$ using these sample proportions leads to Wald inference, where each T_j is the usual Wald statistic for the difference of proportions. These, however, can perform inadequately in the univariate two group setting when inverted, a fact that extends to the multivariate setting (see Section 4). In the univariate case, much better performance (Agresti and Caffo, 2000) results from adding two pseudo-observations, one success and one failure, for each of the two groups. This leads to adjusted estimates $\hat{\pi}_i(j) = (y_{ij+} + 1)/(n_i + 2)$ for the marginal success probabilities. For the 2×2 table cross-classifying the outcomes of variables j and j' in group i, this corresponds to adding 0.5 to the count in each of the four possible cells (1, 1), (1, 0), (0, 1) and (0, 0) and motivates the adjusted estimator $\hat{\pi}_i(j, j') = (\sum_{k=1}^{n_i} y_{ijk}y_{ij'k} + 0.5)/(n_i + 2)$ for the joint success probability.

Let $\hat{\boldsymbol{\Sigma}}$ be the estimator for $\boldsymbol{\Sigma}$ using these adjusted marginal and joint sample proportions and set $\hat{\boldsymbol{T}} = \mathrm{diag}[\hat{\boldsymbol{\Sigma}}]^{-1/2}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)$. The critical value c is then equal to the equidistant two-sided $1 - \alpha$ quantile of the multivariate normal distribution with estimated correlation matrix $\hat{\boldsymbol{R}} = \mathrm{diag}[\hat{\boldsymbol{\Sigma}}]^{-1/2}\ \hat{\boldsymbol{\Sigma}}\ \mathrm{diag}[\hat{\boldsymbol{\Sigma}}]^{-1/2}$. The SCIs for the δ_j's from inverting the maximum adjusted Wald test can be explicitly given by

$$\hat{\delta}_j \pm c \sum_{i=1}^{2} \hat{\pi}_i(j)[1 - \hat{\pi}_i(j)]/n_i, \quad j = 1, \ldots, J.$$

2.1.2. Maximum of global score statistics

An alternative to the (adjusted) Wald approach is to invert the maximum of score statistics. Research in recent years (e.g., Newcombe and Nurminen, 2011) for the *univariate* two sample case has shown that the interval obtained by inverting the asymptotic score test has very reasonable performance (coverage close to nominal, small expected length) under a wide range of settings, including small sample sizes. For the multivariate case considered here, the score approach necessitates estimating $\pi_i(j)$ and $\pi_i(j, j')$ in (1) under the global null hypothesis H_0, using restricted ML methods. These restricted MLEs are obtained by maximizing the product of the two $(n_i, \{\pi_i(\boldsymbol{a})\})$, $i = 1, 2$ multinomial likelihoods with respect to all $2 \times 2^{J-1}$ parameters $\{\pi_i(\boldsymbol{a})\}$, under the restrictions that $\pi_1(j) - \pi_2(j) = \delta_j^0$, $j = 1, \ldots, J$. Lang (1996) developed refined methods for obtaining these restricted MLEs based on Lagrange multipliers and provides R (R Core Team, 2012) software to carry out the computations (Lang, 2004). For a link to this software, see the supplementary materials.

Let $\tilde{\pi}_i(j)$ and $\tilde{\pi}_i(j, j')$ (obtained from $\tilde{\pi}_i(\boldsymbol{a})$ by summation) denote these restricted marginal and joint MLEs, leading to $\tilde{\boldsymbol{\Sigma}}$ as the estimate of $\boldsymbol{\Sigma}$ under H_0. Note that the estimated distribution of $\tilde{\boldsymbol{T}} = \mathrm{diag}[\tilde{\boldsymbol{\Sigma}}]^{-1/2}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)$ and hence $\max_j |\tilde{T}_j|$ depends on $\boldsymbol{\delta}_0$ because $\tilde{\pi}_i(j)$ and $\tilde{\pi}_i(j, j')$ depend on it. This means that for each candidate null vector $\boldsymbol{\delta}_0$, the estimated null correlation matrix $\tilde{\boldsymbol{R}} = \mathrm{diag}[\tilde{\boldsymbol{\Sigma}}]^{-1/2}\ \tilde{\boldsymbol{\Sigma}}\ \mathrm{diag}[\tilde{\boldsymbol{\Sigma}}]^{-1/2}$ and corresponding critical value c change. To invert the test, we must conduct a search over a grid of reasonable $\boldsymbol{\delta}_0$ values, where for each grid point we need to find (i) the value of the observed test statistic $\tilde{t}_{\max}$ and (ii) whether the P-value $P_{H_0}(\max_j |\tilde{T}_j| > \tilde{t}_{\max}) \geq \alpha$. Step (i) involves restricted ML estimation and step (ii) the computation of one J-variate multivariate normal integral. The projection of the convex hull of those grid points $\boldsymbol{\delta}_0$ for which the P-value is $\geq \alpha$ yields the J SCIs under this full (or global) score approach.

2.1.3. Maximum of local score statistics

Because of the computational complexity (both in terms of restricted ML estimation and the grid search) of the global score approach for even moderate J, we consider an alternative where the T_j's are the regular univariate score statistics for testing each individual $H_{0j} : \pi_1(j) - \pi_2(j) = \delta_{j0}$. That is, we estimate the diagonal elements of $\boldsymbol{\Sigma}$ with $\check{\Sigma}_{jj} = \sum_{i=1}^{2} \check{\pi}_i(j)[1-\check{\pi}_i(j)]/n_i$, where $\check{\pi}_i(j)$ is the restricted ML estimate of $\pi_i(j)$ under the *local* null hypothesis H_{0j}, considering only data from variable j. Note that since $\check{\pi}_i(j)$ is consistent under H_{0j}, it is also consistent under the more restrictive parameter space defined by the global null hypothesis $H_0 : \bigcap_{j=1}^{J} H_{0j}$. The advantage of this method is that there is a closed form solution for $\check{\pi}_i(j)$ (Nurminen, 1986) and hence $\max_j |\check{T}_j|$ with $\check{T}_j = (\hat{\delta}_j - \delta_j^0)/\check{\Sigma}_{jj}$ is easy to compute, which helps when inverting the test. However, by viewing each hypothesis in isolation, this method does not provide (restricted) MLEs of joint probabilities $\pi_i(j, j')$ and hence no estimate for the correlation matrix $\boldsymbol{R}$ of the $\check{T}_j$'s. We solve this in two different ways which we call Local1 and Local2:

Local 1: Consider a partition $\{Z_l\}_{l=1}^{L}$ of the index set $Z = \{1, 2, \ldots, J\}$ such that $Z_l \cap Z_{l'} = \emptyset$ for $l \neq l'$ and $\bigcup_{l=1}^{L} Z_l = Z$. The global null hypothesis is then $H_0 = \bigcap_l H_{0l}$ with $H_{0l} = \bigcap_{j \in Z_l} H_{0j}$. By the Bonferroni inequality and with $\check{t}_{\max}$ the observed maximum of the local score statistics, the P-value

$$P_{H_0}\left(\max_j |\check{T}_j| > \check{t}_{\max}\right) = P_{H_0}\left(\bigcup_{j=1}^{J}\{|\check{T}_j| > \check{t}_{\max}\}\right) \leq \sum_{l=1}^{L} P_{H_0}\left(\bigcup_{j \in Z_l}\{|\check{T}_j| > \check{t}_{\max}\}\right).$$

Since the joint null distribution of $\{\check{T}_j\}$ with $j \in Z_l$ only depends on the parameters specified in $H_{0l} \subseteq H_0$, the right hand side equals

$$\sum_{l=1}^{L} P_{H_{0l}}\left(\bigcup_{j \in Z_l}\{|\check{T}_j| > \check{t}_{\max}\}\right) = \sum_{l=1}^{L} P_{H_{0l}}\left(\max_{j \in Z_l} |\check{T}_j| > \check{t}_{\max}\right). \tag{2}$$

For computing these last probabilities in the sum, we only need to estimate the joint null distribution (i.e. null correlation matrix) for those $\check{T}_j$'s with $j \in Z_l$, and hence only need restricted ML estimation of $\pi_i(j, j')$ and $\pi_i(j)$ with $j, j' \in Z_l$. We thus avoid estimation of the full $J \times J$ null correlation matrix by estimating L null correlation matrices of (much) smaller dimensions. SCIs are then obtained by projecting the joint confidence region defined by those $\boldsymbol{\delta}_0$ for which (2) is $\geq \alpha$. This sacrifices some power for computational ease as $\boldsymbol{\delta}_0$ may be in this region but not in the one when using the actual P-value $P_{H_0}\left(\max_j |\check{T}_j| > \check{t}_{\max}\right)$, e.g., when the actual P-value based on the full correlation information equals 0.045 and the right hand side equals 0.055 and $\alpha = 0.05$. On the other hand, we have complete freedom on how we choose the Z_l's and we could group together those variables that are highly associated (as judged, for instance, from the estimated correlation matrix $\hat{\boldsymbol{R}}$ of the adjusted Wald approach) or which naturally form a group in the given context. Note that selecting $Z_l = \{l\},\ l = 1, \ldots, J$ results in the Bonferroni-corrected (local) score intervals, see Section 4.

Local 2: A disadvantage of the Local1 approach above is still the necessity to search over a grid of reasonable $\boldsymbol{\delta}_0$ values (except when using $Z_l = \{l\}$), which becomes impractical as J increases. A straightforward method to estimate the correlation matrix $\boldsymbol{R}$ uses the adjusted sample proportions which are always consistent. This yields the same critical value c as with the adjusted Wald approach in Section 2.1.1 and only requires the computation of a multivariate normal quantile. The jth SCI is then given by all values $\delta_{j0} \in [-1, 1]$ for which $|\check{T}_j| \leq c,\ j = 1, \ldots, J$. This cannot be solved analytically, but fast iterative methods such as interval halving can be used to obtain the lower and upper bounds.

2.2. Likelihood ratio, Pearson and other quadratic forms

Other than taking the maximum as a test statistic, we also considered inverting quadratic forms in the spirit of Hotelling's T^2 for multivariate normal data given by $(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)^t \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)$. When using the null estimate $\tilde{\boldsymbol{\Sigma}}$ instead of $\hat{\boldsymbol{\Sigma}}$, this is the score statistic for H_0 and is algebraically identical to the Pearson Chi-square statistic of form $X^2 = \sum_{i=1}^{2} \sum_{\boldsymbol{a}} (\text{obs} - \text{exp})^2/\text{exp}$, where obs are the observed multinomial counts and exp the expected ones given by $n_i \tilde{\pi}_i(\boldsymbol{a}),\ i = 1, 2$. The likelihood ratio statistic for H_0 has form $G^2 = 2 \sum \text{obs} \log(\text{obs}/\text{exp})$, and both statistics are asymptotically Chi-square with $df = J$. However, similar to inverting Hotelling's T^2 in the two-sample multivariate normal case, inverting X^2 or G^2 turned out to lead to (very) conservative SCIs, often much wider than the Bonferroni-adjusted ones (see Fig. 1), and we do not consider them further here.

Similarly, inverting the maximum $\max_j |Z_j|$ where $\boldsymbol{Z} = \tilde{\boldsymbol{\Sigma}}^{-1/2}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)$, which is asymptotically standard MVN, did not prove competitive. Fig. 1 illustrates the joint confidence regions that result from inverting the Pearson statistic X^2, $\max_j |Z_j|$ and $\max_j |\tilde{T}_j|$ from the previous section, using data from just two variables of the example in Section 5. The projection of the elliptical region corresponding to X^2 (the one for G^2 is even larger) and the parallelogram-shaped region resulting from $\max_j |Z_j|$ yield wider intervals than the projection of the rectangular region resulting from the maximum statistic $\max_j |\tilde{T}_j|$ or any of the other maximum statistics discussed above, such as the Local2 or adjusted Wald approach.

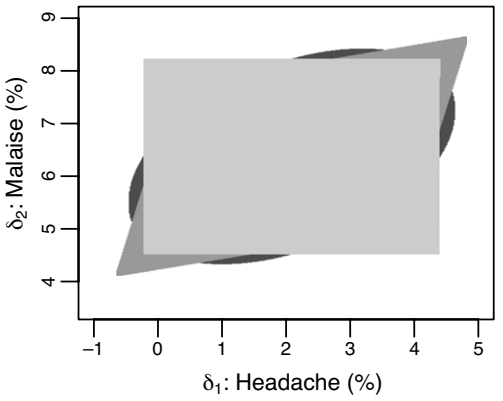

Fig. 1. 2-dimensional simultaneous 95% confidence region for (δ_1, δ_2) based on two variables from the vaccine dataset mentioned in Section 5. The following statistics were inverted to obtain the different regions: X^2 (black), $\max_j |Z_j| = \max |\tilde{\boldsymbol{\Sigma}}^{-1/2}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)|$ (dark-gray) and the maximum of the global score statistics $\max_j |\tilde{T}_j| = \max |\mathrm{diag}(\tilde{\boldsymbol{\Sigma}})^{-1/2}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)|$ (light-gray). The projection of the rectangular-shaped region resulting from $\max_j |\tilde{\boldsymbol{T}}_j|$ yields the shortest intervals.

3. Model-based approaches

In this section we explore other test statistics such as a generalized score statistic under a quasi-likelihood framework that are computationally simpler than the ML approach. A model-based approach models the J marginal probabilities $E[\boldsymbol{y}_{ik}] = \boldsymbol{\pi}_{ik} = (\pi_{ik}(1), \ldots, \pi_{ik}(J))^t$ for the J observations on subject k in group i. Without further subject-specific covariates, one possible model has form

$$\boldsymbol{h}(\boldsymbol{\pi}_{ik}) = \boldsymbol{X}_{ik}\boldsymbol{\beta}, \tag{3}$$

where $\boldsymbol{X}_{ik}$ is a block matrix which equals $\boldsymbol{X}_{1k} = [\boldsymbol{I}_J, \boldsymbol{0}]$ when subject k is in group $i = 1$ and $\boldsymbol{X}_{2k} = [\boldsymbol{0}, \boldsymbol{I}_J]$ when the subject is in group 2, with $\boldsymbol{I}_J$ the identity matrix of dimension J and $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2]^t$ a $2J \times 1$ block vector of unknown parameters. With an identity link $\boldsymbol{h}(\boldsymbol{\pi}_{ik}) = \boldsymbol{\pi}_{ik}$ we get $\boldsymbol{\delta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$ as the vector of the J marginal differences that are our inferential focus.

A popular and computationally fast method to fit this marginal model is via GEE, where one specifies a working correlation matrix $\boldsymbol{R}_i$ for the J variables in each of the two groups. The GEE estimator for $\boldsymbol{\beta}$ is the solution to the quasi-score equation $S(\boldsymbol{\beta}) = \boldsymbol{0}$ which for our setting simplifies considerably (for all derivations and proofs see the Appendix) and yields the vector of marginal sample proportions $\hat{\boldsymbol{\beta}} = [\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2]^t = [\hat{\boldsymbol{\pi}}_1, \hat{\boldsymbol{\pi}}_2]^t$ as the solution, *regardless* of what structure one assumes for the working correlation matrices $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$. Similarly, the robust sandwich-type covariance matrix for $\hat{\boldsymbol{\beta}}$ simplifies to a block diagonal matrix $\boldsymbol{D} = \mathrm{diag}[\boldsymbol{D}_1, \boldsymbol{D}_2]$ with diagonal blocks $\boldsymbol{D}_i = \sum_{k=1}^{n_i}(\boldsymbol{y}_{ik} - \boldsymbol{\beta}_i)(\boldsymbol{y}_{ik} - \boldsymbol{\beta}_i)^t / n_i^2$, and also does not depend on the assumed working correlations.

Another way to fit the marginal model is via ML, which usually becomes computationally demanding as J gets larger because the likelihood refers to the $2(2^J - 1)$ joint probabilities (see Section 2), yet the marginal model is in terms of the $2J$ margins, which are interpreted as restrictions on the multinomial probabilities $\boldsymbol{\pi}(\boldsymbol{a})$ in the likelihood fitting process. However, for the saturated marginal model above ($2J$ parameters for the $2J$ marginal probabilities), the ML approach yields the same solution $\hat{\boldsymbol{\beta}}$ and covariance matrix $\boldsymbol{D}$ as the GEE approach. Under both, GEE and ML fitting, $\hat{\boldsymbol{\beta}}$ is MVN with covariance matrix consistently estimated by $\hat{\boldsymbol{D}} = \mathrm{diag}[\hat{\boldsymbol{D}}_1, \hat{\boldsymbol{D}}_2]$, where $\hat{\boldsymbol{D}}_i$ is $\boldsymbol{D}_i$ evaluated at $\boldsymbol{\beta}_i = \hat{\boldsymbol{\beta}}_i$. Hence, $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2$ is MVN, with estimated covariance matrix $\sum_i \hat{\boldsymbol{D}}_i$, which is precisely the matrix we get when replacing the marginal and joint probabilities in (1) with sample proportions, as mentioned at the beginning of Section 2.1.1. We can then use the general theory of constructing SCIs for the components of $\boldsymbol{\delta}$ in parametric models via the maximum approach as outlined in Hothorn et al. (2008). This leads to the (unadjusted) simultaneous Wald intervals that, as simulation results in Section 4 show, perform poorly.

3.1. Restricted ML and GEE

We expect better performance from estimating (via ML or via GEE) $\boldsymbol{\beta}$ and $\boldsymbol{D}$ under the restriction (null hypothesis) that $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 = \boldsymbol{\delta}_0$, with $\boldsymbol{H} = [\boldsymbol{I}_J, -\boldsymbol{I}_J]$. This yields a covariance matrix estimated under the null. Consequently, the null distribution of our resulting test statistic should be closer to normal and our intervals should have coverage closer to the nominal level. Maximizing the product multinomial likelihood function under these restrictions yields $\tilde{\boldsymbol{\beta}} = [\tilde{\boldsymbol{\pi}}_1, \tilde{\boldsymbol{\pi}}_2]^t = [\tilde{\pi}_1(1), \ldots, \tilde{\pi}_1(J), \tilde{\pi}_2(1), \ldots, \tilde{\pi}_2(J)]^t$, the restricted ML estimates for the marginal proportions introduced in Section 2.1.2. The estimated null covariance matrix is given by $\tilde{\boldsymbol{D}} = \mathrm{diag}[\tilde{\boldsymbol{D}}_1, \tilde{\boldsymbol{D}}_2]$, where $\tilde{\boldsymbol{D}}_i$ is $\boldsymbol{D}_i$ evaluated at $\boldsymbol{\beta}_i = \tilde{\boldsymbol{\beta}}_i$ and $\tilde{\boldsymbol{\beta}}_1 = \boldsymbol{\delta}_0 + \tilde{\boldsymbol{\beta}}_2$. Finally, the estimated covariance matrix of $\hat{\boldsymbol{\delta}} = \boldsymbol{H}\hat{\boldsymbol{\beta}}$ under the null is $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{H}\tilde{\boldsymbol{D}}\boldsymbol{H}^t = \sum_i \tilde{\boldsymbol{D}}_i$, which is the same matrix as in Section 2.1.2, leading to the simultaneous global score intervals introduced there.

Using a computationally much simpler GEE approach to estimate $\boldsymbol{\beta}$ subject to the restriction $\boldsymbol{\beta}_1 = \boldsymbol{\delta}_0 + \boldsymbol{\beta}_2$ yields as the GEE quasi-score equations $S(\boldsymbol{\beta}) - \boldsymbol{H}^t\boldsymbol{\lambda} = \mathbf{0}$, where $\boldsymbol{\lambda}$ is a vector of Lagrange multipliers and $S(\boldsymbol{\beta})$ is given in (5) in the Appendix A. Straightforward manipulations show that the GEE solution for $\boldsymbol{\beta}_2$, $\bar{\boldsymbol{\beta}}_2$, must satisfy

$$\boldsymbol{\beta}_2 = \left[\sum_i n_i V_i^{-1}\right]^{-1} \left(n_1 \boldsymbol{V}_1^{-1}(\hat{\boldsymbol{\pi}}_1 - \boldsymbol{\delta}_0) + n_2 \boldsymbol{V}_2^{-1}\hat{\boldsymbol{\pi}}_2\right), \tag{4}$$

where the working covariance matrix $\boldsymbol{V}_i$ depends on $\boldsymbol{\beta}_i$, $i = 1, 2$. This is a form of weighted average between the shifted (by $\boldsymbol{\delta}_0$) sample proportions $\hat{\boldsymbol{\pi}}_1$ and $\hat{\boldsymbol{\pi}}_2$. To get $\bar{\boldsymbol{\beta}}_2$, one can start with $\boldsymbol{\beta}_2 = \hat{\boldsymbol{\pi}}_2$ and then iterate between updating $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ (i.e., estimating correlation parameters $\boldsymbol{\alpha}$ using $\boldsymbol{\beta}_2$) on the left hand side and updating $\boldsymbol{\beta}_2$ on the right hand side. In our experience, convergence is usually achieved in a couple of iterations for all reasonable $\boldsymbol{\delta}_0$, but problems can occur when components of $\boldsymbol{\beta}_2$ get close to and then step outside their boundary values of 0 or 1. (This happened occasionally when the group sample sizes were small, see simulation results in Section 4.) Note that the solution does depend on the assumed correlation matrices but involves inverting a matrix of dimension $J \times J$ as the most computationally complex step. With $\bar{\boldsymbol{\beta}}_1 = \boldsymbol{\delta}_0 + \bar{\boldsymbol{\beta}}_2$, $\bar{\boldsymbol{\beta}} = [\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2]$ is the restricted GEE estimator. Let $\bar{\boldsymbol{D}} = [\bar{\boldsymbol{D}}_1, \bar{\boldsymbol{D}}_2]$ be its estimated (under the null) covariance matrix, where $\bar{\boldsymbol{D}}_i$ is $\boldsymbol{D}_i$ evaluated at $\boldsymbol{\beta}_i = \bar{\boldsymbol{\beta}}_i$.

3.2. Generalized score statistics

Boos (1992, Eq. (5)) gives a generalized score statistic for testing $H_0 : \boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{\delta}_0$ that uses the estimated covariance matrix of $\boldsymbol{\beta}$ under H_0. For our case, this quadratic form statistic simplifies to $(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)^t \bar{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)$ (see Appendix), where $\bar{\boldsymbol{\Sigma}} = \sum_i \bar{\boldsymbol{D}}_i$. However, since elliptical acceptance regions do not fare well when projected to the axes (see Section 2.2), here we again focus on inverting a maximum statistic given by the maximum of $\bar{\boldsymbol{T}} = \text{diag}[\bar{\boldsymbol{\Sigma}}]^{-1/2}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)$. Since the estimated null correlation matrix $\bar{\boldsymbol{R}} = \text{diag}[\bar{\boldsymbol{\Sigma}}]^{-1/2}\ \bar{\boldsymbol{\Sigma}}\ \text{diag}[\bar{\boldsymbol{\Sigma}}]^{-1/2}$ of $\bar{\boldsymbol{T}}$ depends on $\boldsymbol{\delta}_0$, $\max_j |\bar{T}_j|$ is not pivotal and we again need to search the entire grid of reasonable $\boldsymbol{\delta}_0$ values to find the region where the P-value $P_{H_0}(\max_j |\bar{T}_j| > \bar{t}_{\max})$ is larger than α, with $\bar{t}_{\max}$ the observed maximum. Projecting this region to the J axes yields simultaneous intervals under a GEE approach.

4. Simulation results

We have discussed several strategies for estimating the (null) covariance matrix of the vector $\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0$, either through the adjusted Wald approach ($\hat{\boldsymbol{\Sigma}}$), restricted global ($\tilde{\boldsymbol{\Sigma}}$) or local ML or GEE ($\bar{\boldsymbol{\Sigma}}$) estimation. Each method yields a different maximum statistic and estimated asymptotic distribution, but all are based on the asymptotic normality of $\hat{\boldsymbol{\delta}}$. In this section we evaluate the actual coverage probability and power under various settings for sample sizes and true parameter values. We also want to see if it pays off (in terms of tighter coverage) estimating the covariance matrix under the null (because this is computationally complex) and judge how much we gain by incorporating the correlation information compared to Bonferroni-corrected intervals. For the adjusted Wald approach of Section 2.1.1, Bonferroni-corrected intervals are simply obtained by setting $c = z_{1-\alpha/2J}$, where z_γ is the γ-quantile of the standard normal distribution. There are also Bonferroni corrections for the global score and GEE approaches, but these would only avoid the computation of the multivariate normal integral and not the computationally intensive grid search and so are not considered here. Instead, we compute the straightforward Bonferroni corrected local score intervals obtained by setting $Z_l = \{l\}$, $l = 1, \ldots, J$ for the partition in (2), which becomes $\sum_{j=1}^{J} P_{H_{0j}}(|\check{T}_j| > \check{t}_{\max}) = JP(|W| > \check{t}_{\max})$ with $W \sim N(0, 1)$. Hence, for any $\boldsymbol{\delta}_0$, (2) is $\geq\alpha$ if $|\check{T}_j| \leq z_{1-\alpha/2J}$, $j = 1, \ldots, J$ and no grid search is necessary. Since $\check{T}_j$ is the regular (univariate) score statistic for the difference in proportions, this results in the usual score intervals using a Bonferroni correction, i.e., the j-th interval is given by all $\delta_{j0} \in [-1, 1]$ for which $|\check{T}_j| \leq z_{1-\alpha/2J}$. For the restricted GEE approach we will consider an independence working assumption $\boldsymbol{R}_i = \boldsymbol{I}_J$ or an unstructured correlation where $\boldsymbol{\alpha}_i$ holds all pairwise correlations, but other structures, such as blockwise exchangeable are also possible.

4.1. Unstructured simulations

We first evaluate coverage for the case $J = 4$ by simulating datasets (i.e., multinomial counts) under a fairly unstructured scenario for the marginal probabilities and the dependence structure: We randomly generated marginal probabilities $\pi_i(j)$, $i = 1, 2$, $j = 1, \ldots, 4$ from a uniform distribution on $[0.01, 0.99]$. To simulate the association, we randomly generated log-odds ratios for each of the $\binom{4}{2}$ pairs of binary responses from a $N(1.5, \sigma = 0.8)$ distribution. This implies 25th, 50th and 75th percentiles of the generated pairwise odds ratios of 2.6, 4.5 and 7.7, respectively, inducing a moderate positive dependence between most pair of responses but also allowing for virtually no or a negative dependence. For instance, for some generated datasets pairwise odds ratios were as low as 0.2 or as high as 49. When the generated marginal probabilities and odds ratios are compatible (Qaqish, 2003), we simulated n_i multivariate binary observations in each group by clipping randomly sampled multivariate normal vectors with corresponding mean and covariance matrix (R package binarySimCLF).

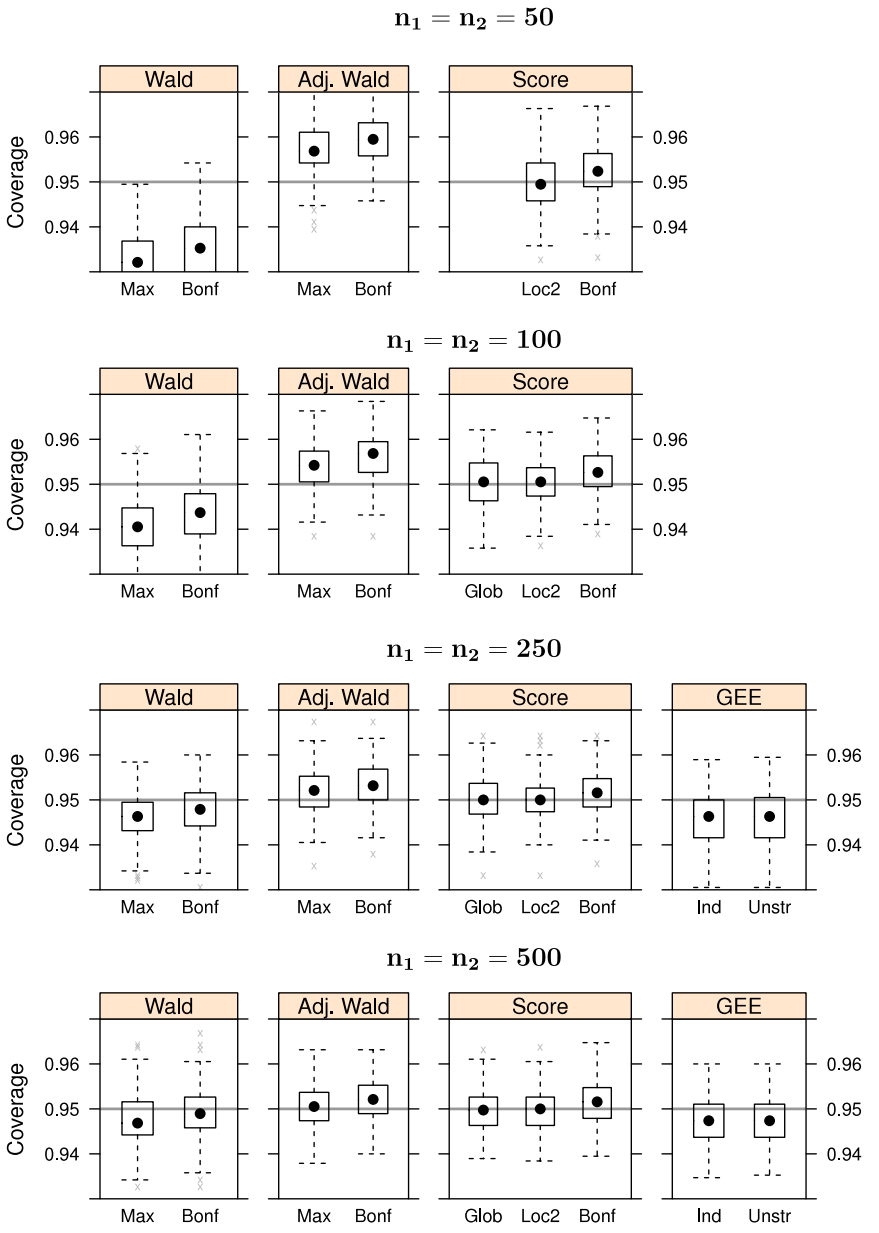

Fig. 2. Coverage probability of $J = 4$ simultaneous confidence intervals over 250 random parameter settings. The first two panels in each row show the coverage probability for the unadjusted and adjusted Wald method (Section 2.1.1) using $\max_j |\hat{T}_j|$ with critical value computed from the multivariate normal distribution with estimated correlation matrix $\hat{\boldsymbol{R}}$ ("Max") or using the Bonferroni adjustment ("Bonf"). The third panel shows the coverage from the global score approach using $\max_j |\tilde{T}_j|$ with estimated correlation matrix $\tilde{\boldsymbol{R}}$ ("Glob", Section 2.1.2) and the local score approach using $\max_j |\check{T}_j|$, with critical value estimated using the correlation matrix $\hat{\boldsymbol{\Sigma}}$ ("Loc2", local2 approach, Section 2.1.3) or the Bonferroni adjustment ("Bonf"). For sample sizes $n_1 = n_2 \geq 250$, the last panel shows the coverage for the restricted GEE approach using $\max_j |\tilde{T}_j|$ under an independence ("Ind") and unstructured ("Unstr") assumption (Section 3.2).

Fig. 2 shows boxplots of the coverage probability for 250 such random parameter settings when $n_1 = n_2 = 50$, 100, 250, or 500. Under each setting, 1900 datasets were generated in order to estimate the coverage to within a margin of error of $\pm 1\%$ for nominal 95.0% coverage. In total, 9 different methods for constructing SCIs were evaluated, see the caption to Fig. 2 for a description. The computationally expensive global score approach (Glob) was not included for the small sample case of $n_1 = n_2 = 50$ as the automated grid search did not converge for some of the 1900 generated datasets. Similarly, the GEE approach did sometimes not converge in the small sample setting and hence is only given for $n_1 = n_2 \geq 250$.

Not surprisingly, based on results in the univariate case, the Wald approach also performs terribly in the multivariate setting and should not be used. The adjusted Wald method is overly conservative in the small sample setting, but performs well for large ($n_i \geq 250$) sample sizes. The Local2 (for all considered n_i) and the global score (for $n_i \geq 100$) methods behave

Table 1
Simulation settings for simulations in Table 2: marginal probabilities $\pi_i(j)$ are in bold on the diagonal, joint probabilities $\pi_i(j, j')$ are below the diagonal. Corresponding pairwise odds ratios that describe the strength of the association are displayed above the diagonal.

Group $i = 1$				
	Y_{11}	Y_{12}	Y_{13}	Y_{14}
Y_{11}	**0.26**	12.4	9.5	11.8
Y_{12}	0.12	**0.17**	7.1	16.0
Y_{13}	0.04	0.03	**0.06**	9.6
Y_{14}	0.05	0.05	0.02	**0.07**
Group $i = 2$				
	Y_{21}	Y_{22}	Y_{23}	Y_{24}
Y_{21}	**0.24**	12.5	11.0	9.6
Y_{22}	0.08	**0.11**	3.9	15.1
Y_{23}	0.04	0.02	**0.06**	4.1
Y_{24}	0.03	0.03	0.01	**0.05**

very well with an average (over the 250 random settings) coverage probability almost identical to the nominal level and a spread of the coverage that is not very large. For the GEE procedure, considered only for $n_i \geq 250$, too many settings resulted in a coverage below the nominal level.

The conservatism introduced by the use of the Bonferroni adjustment is clearly visible in the plots, although it does not seem to be too dramatic. For instance, for the Local2 approach and $n_i = 100$, the minimum, the 25th, 50th, and 75th percentiles and the maximum coverage are 93.6%, 94.7%, 95.0%, 95.4% and 96.2%, respectively, while these values are all about 0.2% points higher with the Bonferroni correction. Not shown in Fig. 2 are the performances of the unadjusted (for multiplicity) Local2 and adjusted Wald methods, which have median coverage of around 82.0% for all settings, well below the nominal level.

4.2. Structured simulations

We next consider simulating correlated binary responses for some given setting for the multinomial probabilities $\{\pi_i(\boldsymbol{a})\}$, $i = 1, 2$. Table 1 shows marginal probabilities $\pi_i(j)$ (on the diagonal) as well as joint probabilities $\pi_i(j, j')$ (on the lower diagonal) and corresponding odds-ratios (on the upper diagonal) of $J = 4$ variables in each of two groups. These values are actually the sample values of the vaccine example introduced in Section 5. Under this scenario, relatively strong associations exist between pairs of variables, with several odds ratios larger than 9 and as high as 16, in both groups. Marginal probabilities range from 25% to 5%. Using these values (and the higher order joint probabilities not shown in Table 1) we reconstruct the corresponding vector of multivariate probabilities $\pi_i(\boldsymbol{a})$ and use it to generated 7600 datasets for each group under various sample sizes. (7600 replications yield a simulation margin of error for the estimated coverage of $\pm 0.5\%$.)

Table 2 shows the estimated coverage for sample sizes ranging from 50 to 1000 per group plus one case with unequal sample sizes ($n_1 = 500$, $n_2 = 250$). These simulations also include the performance of the Local1 approach when variables 2 and 3 are considered to be in the same group. In addition, Table 2 shows the power to detect a significant effect (as judged by the lower bound of the simultaneous confidence interval being larger than 0) for variable 2, for which the true effect used in the simulation is $\delta_2 = 0.171 - 0.107 = 0.064$ and for variable 4, with true $\delta_4 = 0.069 - 0.046 = 0.023$.

For all equal sample size settings in Table 2, the Local2 approach again performs very well and generally beats out all other competitors, for both small and large sample sizes. For instance, for $n_1 = n_2 = 500$, it has coverage probability (95.2%) closest to the nominal level and largest power to detect the effect in both δ_2 and δ_4 (66.9% and 17.2%, respectively). The adjusted Wald and the global score approach also perform acceptable but are generally more conservative, while the Local1 approach does not show much improvement in coverage or power over the Bonferroni version. Overall, for the same test statistic (e.g., adjusted Wald or Local2) we see that using the correlation information results in coverage that is by about 0.3% points closer to nominal compared to the Bonferroni adjustment. The resulting power gain of between 0.3% and 1.1% points seems only moderate. Estimating the correlation under the null (i.e., with the global score approach) did not seem to provide any benefits in terms of coverage or power. Because of its computational complexity, we will not consider it further. Note that the GEE approach, which also uses null estimates, actually performs very well for the particular setting in Table 1 (largest power values while coverage almost exactly nominal), but was shown to lead to unsatisfactory coverage more generally, see Fig. 2. Because of this erratic behavior, we will also not consider the GEE approach further. Finally, for the unequal sample size scenario, the adjusted Wald approach performed slightly better than the Local2 score approach.

In Table 3, coverage and power are shown for various methods when more variables are added to the structure in Table 1 for a total of up to $J = 20$ variables. The probability structure (marginal and pairwise joint probabilities as well as pairwise odds-ratios) underlying these simulations are shown in Tables A1 and A2 in the supplementary materials. They correspond to the sample proportions of a dataset similar to the vaccine example given in Section 5. For instance, for $J = 10$, six variables were added to Table 1 with marginal probabilities equal to $\pi_1(5) = 0.23$, $\pi_1(6) = 0.33$, $\pi_1(7) = 0.20$, $\pi_1(8) = 0.11$, $\pi_1(9) = 0.09$ and $\pi_1(10) = 0.08$ in group 1 and $\pi_2(5) = 0.19$, $\pi_2(6) = 0.06$, $\pi_2(7) = 0.13$, $\pi_2(8) =$

Table 2
Coverage and power (in %) for SCIs under the scenario of Table 1 for various sample sizes. Inverted statistics are the same as in Fig. 2 plus the unadjusted (for multiplicity) versions of the adjusted Wald and Local2 approach. In addition, performance of the Local1 approach was included, using the partition $Z_1 = \{1\}, Z_2 = \{2, 3\}, Z_3 = \{4\}$. The unequal sample size setting corresponds to $n_1 = 500$, $n_2 = 250$.

n_i	Adj. Wald			Score					GEE	
	Max	Bonf	Unadj	Glob	Loc1	Loc2	Bonf	Unadj	Ind	Unstr
Coverage										
50	97.9	98.0	89.9	–	–	96.2	96.8	82.3	–	–
100	96.7	96.9	86.1	96.0	96.0	95.4	95.7	82.2	–	–
250	96.0	96.2	84.9	96.3	95.6	95.3	95.6	83.4	94.9	94.9
500	95.4	95.7	85.0	95.7	95.5	95.2	95.5	84.4	95.2	95.2
1000	95.4	95.7	83.2	95.6	95.6	95.3	95.6	82.9	95.2	95.2
Uneq.	95.3	95.6	83.3	95.8	95.4	94.8	95.2	81.8	94.9	94.5
Power for $j = 2$ $(\delta_2 = 0.171 - 0.107 = 0.064)$										
250	33.2	32.4	53.9	30.9	33.0	34.0	33.1	54.8	35.0	35.0
500	66.2	65.2	82.8	65.1	65.6	66.9	65.8	83.2	67.5	67.6
1000	95.0	94.8	98.4	94.9	94.9	95.1	94.8	98.6	95.3	95.3
Uneq.	47.1	46.0	65.6	42.0	42.8	43.1	42.2	64.8	44.5	49.8
Power for $j = 4$ $(\delta_4 = 0.069 - 0.046 = 0.023)$										
250	6.6	5.9	16.2	6.6	7.6	7.7	7.1	17.8	7.7	7.7
500	16.4	15.5	32.5	16.3	16.5	17.2	16.5	33.7	18.0	18.1
1000	38.0	37.0	58.6	38.0	37.8	38.7	37.8	59.4	38.9	39.0
Uneq.	11.1	10.8	22.7	9.0	9.4	8.2	8.1	21.5	9.5	15.2

Table 3
Coverage (C) and Power (P_{δ_2} and P_{δ_4} for the second and fourth variable) of SCIs when simulating from the first J variables under the scenario given in Table A in the supplementary materials. Power is only shown for $n_1 = n_2 = 500$.

J	n		Adj. Wald			Score		
			Max	Bonf	Unadj	Loc2	Bonf	Unadj
6	100	C:	96.3	96.6	80.1	95.2	95.7	76.1
	250	C:	95.5	96.0	78.3	95.2	95.8	76.8
	500	C:	95.4	95.9	77.7	95.3	95.7	77.1
		P_{δ_2}:	62.8	61.5	83.2	63.2	62.0	83.5
		P_{δ_4}:	13.9	13.2	32.2	14.8	13.9	33.4
10	100	C:	96.4	96.9	71.3	95.5	96.3	65.6
	250	C:	95.9	96.4	67.7	95.4	96.0	65.7
	500	C:	95.9	96.5	68.1	95.7	96.4	67.0
		P_{δ_2}:	55.7	54.0	82.6	56.4	54.3	82.9
		P_{δ_4}:	9.5	8.8	32.7	10.4	9.4	33.9
15	100	C:	97.6	97.8	65.6	96.5	97.0	53.9
	250	C:	96.4	96.9	57.9	95.6	96.2	53.3
	500	C:	95.8	96.2	55.3	95.4	95.7	52.9
		P_{δ_2}:	50.1	48.7	82.2	50.7	48.8	82.5
		P_{δ_4}:	7.6	7.2	32.7	8.2	7.6	33.8
20	100	C:	98.0	98.3	64.1	97.3	97.6	49.4
	250	C:	97.0	97.3	50.5	96.4	96.9	41.1
	500	C:	96.2	96.6	46.2	95.4	95.9	41.6
		P_{δ_2}:	46.2	44.8	83.2	46.7	45.6	83.8
		P_{δ_4}:	6.3	6.0	32.4	6.9	6.5	33.6

$0.11, \pi_2(9) = 0.09$ and $\pi_2(10) = 0.04$ in group 2. The 5-number summary for the pairwise odds ratios in group 1 are equal to $\min = 1.8, Q1 = 4$, median $= 9, Q3 = 13$ and $\max = 32$ and those for group 2 are similar. From Table 3, the Local2 approach again yields coverage close to the nominal value, but this is not always the case for the adjusted Wald method. Throughout, the coverage for the Bonferroni-corrected versions are by about 0.3%–0.6% points larger, translating into a moderate power gain of between 1.1% and 2.1% points for δ_2 and the Local2 method when using the correlation information.

The scenario in Table 3 for $J = 20$ corresponds to a sparse data situation. For instance, about half of the $n_1 = 500$ observed vectors $(Y_{11}, \ldots, Y_{1,20})$ in group 1 are all zeros, i.e. no success was observed for any of the 20 variables. (For the second group, this is even higher at 60%.) Overall, only about 9% of all entries in the data matrix in group 1 (500×20 observations) are successes, and only 6% in group 2. Despite this sparseness, the Local2 approach controls the FWER at the nominal level, which is due to the good asymptotic behavior of the univariate score statistic $\check{T}_j$. For the even more extreme case of $J = 20$ with only $n_1 = n_2 = 100$ or 250 observations in each group, the Local2 approach becomes conservative, with joint coverage of 97.3% and 96.4%, respectively, but still better than the adjusted Wald approach.

Table 4
Marginal sample proportions ($\hat{\pi}_{1j}$ and $\hat{\pi}_{2j}$, in %) for the two groups, their differences ($\hat{\delta}_j$, in % points) and unadjusted, Bonferroni-adjusted and Local2 simultaneous 95% confidence intervals (LB, UB) for the $J = 4$ differences of adverse events in the vaccine trial.

Adverse event	$\hat{\pi}_{1j}$	$\hat{\pi}_{2j}$	$\hat{\delta}_j$	Unadj.		Bonf.		Local2	
				LB	UB	LB	UB	LB	UB
Headache	26.0	23.9	2.1	0.08	4.08	−0.91	5.07	−0.87	5.04
Malaise	17.1	10.7	6.4	4.77	7.96	3.99	8.75	4.01	8.72
Pyrexia	5.6	5.8	−0.2	−1.33	0.83	−1.87	1.36	−1.85	1.35
Arthralgia	6.9	4.6	2.3	1.20	3.37	0.67	3.92	0.69	3.90

We considered other scenarios as well. For instance, we simulated data for $J = 20$ when the variables can be clustered in 4 groups of 5 variables each. We assume variables within a cluster are exchangeable with a common marginal probability and a pairwise odds ratio of 15, but variables from different clusters are independent. The marginal success probabilities for the 4 clusters were set equal to 0.5, 0.4, 0.3 and 0.2 in group 1, and 0.1 less than that in group 2. For $n_1 = n_2 = 500$, the coverage of the Local2 SCIs under this scenario is 95.1%, compared to 95.8% under the Bonferroni adjustment. The power of the Local2 intervals compared to the Bonferroni ones increased by 2.5% (59.5% vs. 57.0%) when $\delta = 0.5 - 0.4 = 0.1$ (i.e., in cluster 1), by 1.9% (63.3% vs. 61.4%) when $\delta = 0.4 - 0.3 = 0.1$ (i.e., in cluster 2), by 1.7% (75.0% vs. 73.3%) when $\delta = 0.3 - 0.2$ and by 1.7% (93.7% vs. 93.0%) when $\delta = 0.2 - 0.1$.

5. Examples

Table 4 shows incidence rates for 4 adverse events (AEs) for an influenza vaccine based on a recent parallel, 2-arm controlled clinical trial with 3600 subjects in both the vaccinated and placebo group. Based on the nature of the compound, investigators hypothesized that the four AEs Headache, Malaise, Pyrexia (fever) and Arthralgia (joint pain) may be associated with the vaccine. The presence or absence of these four AEs were specifically recorded in each patients' diary with the goal to learn about potential differences between the vaccinated and placebo group. In particular, investigators wanted to estimate the size of the differences in the marginal incidence rates.

Remember that Table 1 showed the association structure between the four AEs in each group. For instance, for the vaccinated group, the odds of recording Malaise are 16 times higher when Arthralgia was also recorded. Similarly, the odds of reporting Pyrexia are 9 times higher when Headache was also reported. Table 4 shows Local2 SCIs for estimating the marginal differences, alongside Bonferroni and unadjusted ones. (For SCIs using the other methods of Section 2 and illustration of R code to obtain all SCIs mentioned in this article, please refer to the supplementary materials.) We see that the incidence rate for Malaise is by at least 4.3% points and at most 8.5% points larger in the vaccinated group, and for Arthralgia by at least 1.0% point and at most 3.6% points. Incidence rates for Headache or Pyrexia were not significantly different between the two groups. Note that the FWER associated with these statements is controlled at 5%. Investigators and regulators can use the information on the magnitude of these effects in a risk-benefit analysis.

As a second illustration of our methodology, we consider data from the General Social Survey (GSS, http://sda.berkeley.edu/GSS/). In 2006, part of the survey included questions on the performance of the US government on various issues, such as health care, security, unemployment or protecting the environment. Respondents also indicated if they identify themselves as conservative (group 1, sample size $n_1 = 247$) or liberal (group 2, $n_2 = 214$). Table 5 shows the marginal proportions of $J = 6$ questions on these topics for each group. There are strong associations between some variables. For instance, for both groups, the odds of government success on providing health care for the sick are 12 times higher for respondents who also think the government is successful in providing a decent standard of living for the elderly. There are other variables that do not seem to be highly associated with others, such as the opinion on protecting the environment. Table 5 also shows Local2 and Bonferroni SCIs, controlling the FWER at 5%, alongside unadjusted ones. Perhaps not surprisingly (given a Republican dominated government in 2006), conservatives think the government is significantly more successful than liberals on almost all topics. More interesting is by how much more successful conservatives think the government is on these topics: by at least 13.3% points on unemployment, 10.7% points on providing security, 8.3% points on the environment, 7.4% points on health care (but by no more than 28% points) and 3.7% points on standard of living. Note that by construction, statements such as the above control the FWER at 5%.

6. Summary and discussion

Summing up the simulation results and examples, the Local2 approach emerges as a method that yields close to nominal coverage and shows a moderate power advantage over the straightforward Bonferroni method. It performs very well in both small and large sample scenarios and sparse data situations. In addition, it is quick to implement with the computation of a MVN quantile (e.g., via the R package "mvtnorm", Genz and Bretz, 2009) as the most complex step. Since the computation of the score interval is iterative and needs computer implementation, even for the Bonferroni adjustment, the Local2 approach should always be used over the Bonferroni approach. For large sample sizes (i.e., over 250 in each group) and for unequal sample sizes, the adjusted Wald method is also competitive, with the advantage of a closed form once the MVN quantile is

Table 5

Marginal sample proportions for conservatives and liberals ($\hat{\pi}_{1j}$ and $\hat{\pi}_{2j}$, in %), their differences ($\hat{\delta}_j$, in % points) and unadjusted, Bonferroni-adjusted and Local2 simultaneous 95% confidence intervals (LB, UB) for the $J = 6$ questions on government success in the 2006 GSS. A success is recorded when the respondent indicated that the government is quite or very successful, and no success is recorded when the respondent indicated that the government is quite or very unsuccessful or if the respondent is indifferent.

Government success	$\hat{\pi}_{1j}$	$\hat{\pi}_{2j}$	$\hat{\delta}_j$	Unadj.		Bonf.		Local2	
				LB	UB	LB	UB	LB	UB
Health care	34.8	16.8	18.0	10.1	25.7	7.3	28.3	7.4	28.2
Standard of living	28.7	14.9	13.8	6.3	21.1	3.6	23.7	3.7	23.6
Security threats	61.1	38.3	22.8	13.7	31.5	10.6	34.4	10.7	34.3
Crime	41.7	31.3	10.4	1.6	19.0	−1.5	21.9	−1.4	21.8
Unemployment	42.5	18.2	24.3	16.1	32.2	13.2	34.8	13.3	34.7
Environment	42.1	22.4	19.7	11.2	27.8	8.2	30.6	8.3	30.5

provided. Somewhat surprisingly, the full and quasi-score intervals formed by estimating the variance–covariance matrix under the global null did not seem to improve coverage or power and moreover are computationally very complex. Very conservative performance results from inverting quadratic form test statistics such as Pearson's Chi-squared instead of a maximum statistic.

This article focused on two-sided intervals, and simulation results (not shown) reveal that the non-coverage on either side was fairly balanced at 2.5%. When directional decisions are of interest, for instance upper bounds for the effect on the four marginal differences in the vaccine example, the methodology is easily adjusted. For example, to obtain upper bounds under the Local2 approach, one inverts the test $H_0 : \boldsymbol{\delta} \geq \boldsymbol{\delta}_0$ vs. $H_a : \boldsymbol{\delta} < \boldsymbol{\delta}_0$, using $\min_j \check{T}_j$ as the test statistic. The jth simultaneous upper bound controlling the FWER at level α is then given by the largest value of $\delta_{j0} \in [-1, 1]$ for which $|\check{T}_j| \geq c,\ j = 1, \ldots, J$, where c is the equidistant upper $1-\alpha$ quantile of the MVN distribution with estimated correlation matrix $\hat{\boldsymbol{R}}$. For the adverse event example in Section 5, the four simultaneous upper bounds controlling the FWER at $\alpha = 5\%$ are given (in percent points) by $\delta_1^U = 4.33$, $\delta_2^U = 8.16$, $\delta_3^U = 0.96$ and $\delta_4^U = 3.50$. Compare these to the Bonferroni bounds of $\delta_1^U = 4.37$, $\delta_2^U = 8.19$, $\delta_3^U = 0.98$ and $\delta_4^U = 3.52$.

The methodology also extends to comparisons involving more than two independent groups, for instance, when correlated binary variables are measured for several groups, such as conservatives, moderates and independents in the GSS or a control and several dose groups in the vaccine example. For the latter, confidence intervals for $\pi_i(j) - \pi_0(j)$, where $\pi_0(j)$ denotes the marginal probability in the control group need to be constructed simultaneously for all involved variables $j = 1, \ldots, J$, but also controlling for the multiplicities arising from comparing each dose i to the control. For this, the correlation structure needs to be worked out in a similar manner to (1). Results in the univariate setting ($J = 1$) show the superior performance of the score statistic and also indicate some benefit from incorporating the correlation information (Klingenberg, 2012).

Finally, in this article we only treated the difference of proportions, but results about other effect measures such as relative risks or odds ratios formed with marginal probabilities for comparing the two groups (i.e., using a log or logit link in (3)) are also desirable, as are results when including covariates, such as gender or race in the marginal model for a stratified analysis.

Acknowledgments

We would like to thank the editors and referees for many comments that led to a substantially improved version of the paper.

Appendix A. GEE approach

The GEE estimate for $\boldsymbol{\beta}$ is the solution to the quasi-score equation $S(\boldsymbol{\beta}) = \sum_{i=1}^{2} \sum_{k=1}^{n_i} s_{ik} = 0$ with $s_{ik} = \boldsymbol{D}_{ik}^t V_{ik}^{-1}(\boldsymbol{y}_{ik} - \boldsymbol{\pi}_{ik})$. Here, $\boldsymbol{y}_{ik}$ is the vector of the J correlated binary responses for subject k in group i with $E[\boldsymbol{y}_{ik}] = \boldsymbol{\pi}_{ik}$ and (working) covariance matrix $\boldsymbol{V}_{ik} = \boldsymbol{B}_{ik}^{1/2} \boldsymbol{R}_i \boldsymbol{B}_{ik}^{1/2}$ with $\boldsymbol{B}_{ik} = \text{diag}[\boldsymbol{\pi}_{ik}(1-\boldsymbol{\pi}_{ik})]$ and $\boldsymbol{D}_{ik} = \partial h^{-1}(\boldsymbol{\pi}_{ik})/\partial \boldsymbol{\beta}^t$. In our setting with an identity link ($\boldsymbol{D}_{ik} = \boldsymbol{X}_{ik}$) and the special structure of $\boldsymbol{X}_{ik}$ due to no subject-specific covariates, the estimating equations simplify considerably to

$$S(\boldsymbol{\beta}) = [\boldsymbol{V}_1^{-1}|\boldsymbol{0}]^t n_1(\hat{\boldsymbol{\pi}}_1 - \boldsymbol{\beta}_1) + [\boldsymbol{0}|\boldsymbol{V}_2^{-1}]^t n_2(\hat{\boldsymbol{\pi}}_2 - \boldsymbol{\beta}_2) = \boldsymbol{0}, \tag{5}$$

where $\boldsymbol{V}_i = \boldsymbol{B}_i^{1/2} \boldsymbol{R}_i \boldsymbol{B}_i^{1/2}$ is the working covariance matrix that is the same for all n_i subjects in group i and depends on $\boldsymbol{\beta}_i$ through $\boldsymbol{B}_i = \text{diag}[\boldsymbol{\beta}_i(1-\boldsymbol{\beta}_i)]$. $\boldsymbol{R}_i$ is the working correlation matrix for group i that is defined through correlation parameters $\boldsymbol{\alpha}_i$. The solution to (5) is given by $n_i \boldsymbol{V}_i^{-1} \boldsymbol{\beta}_i = n_i \boldsymbol{V}_i^{-1} \hat{\boldsymbol{\pi}}_i,\ i = 1, 2$, the $\boldsymbol{V}_i$'s drop out and one obtains as the GEE solution $\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\pi}}_i,\ i = 1, 2$, which does not depend on the assumed $\boldsymbol{R}_i$'s.

The sandwich variance–covariance matrix of $\hat{\boldsymbol{\beta}}, \boldsymbol{D} = \text{Var}[\hat{\boldsymbol{\beta}}] = \boldsymbol{L}_0^{-1} \boldsymbol{L} \boldsymbol{L}_0^{-1}$, where $\boldsymbol{L}_0 = \sum_{i=1}^{2} \sum_{k=1}^{n_i} \boldsymbol{D}_{ik}^t \boldsymbol{V}_{ik}^{-1} \boldsymbol{D}_{ik}$ and $\boldsymbol{L} = \sum_{i=1}^{2} \sum_{k=1}^{n_i} \boldsymbol{D}_{ik}^t \boldsymbol{V}_{ik}^{-1} \text{Var}(\boldsymbol{y}_{ik}) \boldsymbol{V}_{ik}^{-1} \boldsymbol{D}_{ik}$ also simplifies considerably. It equals a $2J \times 2J$ block diagonal matrix with the two

$J \times J$ blocks given by $\boldsymbol{D}_i = \sum_{k=1}^{n_i} \mathrm{Var}[\boldsymbol{y}_{ik}]/n_i^2,\ i = 1, 2$, where $\mathrm{Var}[\boldsymbol{y}_{ik}] = (\boldsymbol{y}_{ik} - \boldsymbol{\pi}_{ik})(\boldsymbol{y}_{ik} - \boldsymbol{\pi}_{ik})^t = (\boldsymbol{y}_{ik} - \boldsymbol{\beta}_i)(\boldsymbol{y}_{ik} - \boldsymbol{\beta}_i)^t$ is the *empirical* covariance matrix of $\boldsymbol{y}_{ik}$. Again, we see that $\boldsymbol{D}$ does not depend on the assumed correlation matrices $\boldsymbol{R}_i$.

The generalized score statistic for hypothesis of form $H_0 : \boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{\delta}_0 = \boldsymbol{0}$ equals (Boos, 1992, p. 331, Eq. (5))

$$S(\bar{\boldsymbol{\beta}})^t \bar{\boldsymbol{J}}^{-1} \boldsymbol{H}^t \left[\boldsymbol{H}\bar{\boldsymbol{J}}^{-1}\bar{\boldsymbol{W}}\bar{\boldsymbol{J}}^{-1}\boldsymbol{H}^t\right]^{-1} \boldsymbol{H}\bar{\boldsymbol{J}}^{-1} S(\bar{\boldsymbol{\beta}}),$$

where $S(\bar{\boldsymbol{\beta}})$ is $S(\boldsymbol{\beta})$ evaluated at the restricted GEE estimate $\boldsymbol{\beta} = \bar{\boldsymbol{\beta}}$, the "information" matrix $\bar{\boldsymbol{J}}$ is given by $-\partial S(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^t$ also evaluated at $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{W}}$ equals $\sum_{i=1}^{2}\sum_{k=1}^{n_i} s_{ik} s_{ik}^t$ evaluated at $\bar{\boldsymbol{\beta}}$. Under our model, $\bar{\boldsymbol{J}}$ simplifies to a block-diagonal matrix with blocks $n_i \bar{\boldsymbol{V}}_i^{-1},\ i = 1, 2$, where $\bar{\boldsymbol{V}}_i$ is $\boldsymbol{V}_i$ evaluated at $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\alpha}}_i$, the estimates of the correlation parameters at convergence. Further, $\bar{\boldsymbol{W}}$ simplifies to $\sum_{i=1}^{2}\sum_{k=1}^{n_i} \bar{\boldsymbol{U}}_i^t (\boldsymbol{y}_{ik} - \bar{\boldsymbol{\beta}}_i)(\boldsymbol{y}_{ik} - \bar{\boldsymbol{\beta}}_i)^t \bar{\boldsymbol{U}}_i$ with $\bar{\boldsymbol{U}}_1 = [\bar{\boldsymbol{V}}_1, \boldsymbol{0}]$ and $\bar{\boldsymbol{U}}_2 = [\boldsymbol{0}, \bar{\boldsymbol{V}}_2]$ leading to $\boldsymbol{H}\bar{\boldsymbol{J}}^{-1}\bar{\boldsymbol{W}}\bar{\boldsymbol{J}}^{-1}\boldsymbol{H}^t = \bar{\boldsymbol{\Sigma}} = \sum_{i=1}^{2} \bar{\boldsymbol{D}}_i$, with $\bar{\boldsymbol{D}}_i$ as defined in Section 3.2. Finally, using (5) it is straightforward to show that $S(\bar{\boldsymbol{\beta}})^t \bar{\boldsymbol{J}}^{-1} \boldsymbol{H}^t$ simplifies to $(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)^t$, proving the form of the generalized score statistic as given in Section 3.2.

Appendix B. Supplementary material

Supplementary material describing the underlying correlation structure for the simulations in Table 2 and the use of customized R functions to reproduce all SCIs mentioned in the article can be found online at http://dx.doi.org/10.1016/j.csda.2013.02.016. Some of these rely on R packages "inline" (Sklyar et al., 2012) and "RcppArmadillo" (Francois et al., 2012).

References

Agresti, A., Caffo, B., 2000. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. American Statistician 54, 280–288.

Agresti, A., Klingenberg, B., 2005. Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. Journal of the Royal Statistical Society. Series C (Applied Statistics) 54 (4), 691–706.

Benjamini, Y., Yekutieli, D., 2005. False discovery rate-adjusted multiple confidence intervals for selected parameters. Journal of the American Statistical Association 100, 71–81.

Berry, S., Berry, D., 2004. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. Biometrics 60, 418–426.

Bhapkar, V., Somes, G., 1976. Multiple comparison of matched proportions. Communications in Statistics Series A 5, 17–25.

Boos, D., 1992. On generalized score tests. The American Statistician 46, 327–333.

Chafaï, D., Concordet, D., 2009. Confidence regions for the multinomial parameter with small sample size. Journal of the American Statistical Association 104, 1071–1079.

Davidov, O., Peddada, S., 2011. Order-restricted inference for multivariate binary data with application to toxicology. Journal of the American Statistical Association 106, 1394–1404.

Fitzpatrick, S., Scott, A., 1987. Quick simultaneous confidence intervals for multinomial proportions. Journal of the American Statistical Association 82, 875–878.

Francois, R., Eddelbuettel, D., Bates, D., 2012. RcppArmadillo: Rcpp integration for Armadillo templated linear algebra library. R Package Version 0.3.6.1. URL: http://CRAN.R-project.org/package=RcppArmadillo.

Genz, A., Bretz, F., 2009. Computation of Multivariate Normal and *t* Probabilities. In: Lecture Notes in Statistics, vol. 195.

Goodman, L.A., 1965. On simultaneous confidence intervals for multinomial proportions. Technometrics 7, 247–254.

Hothorn, T., Bretz, F., Westfall, P., 2008. Simultaneous inference in general parametric models. Biometrical Journal 50, 346–363.

Huang, L., Zalkikar, J., Tiwari, R.C., 2011. A likelihood ratio test based method for signal detection with application to FDA's drug safety data. Journal of the American Statistical Association 106, 1230–1241.

Klingenberg, B., 2012. Simultaneous score confidence bounds for risk differences in multiple comparisons to a control. Computational Statistics and Data Analysis 56, 1079–1089.

Lang, J.B., 1996. Maximum likelihood methods for a generalized class of log-linear models. Annals of Statistics 24, 726–752.

Lang, J.B., 2004. Multinomial-Poisson homogeneous models for contingency tables. The Annals of Statistics 32, 340–383.

Lefkopoulou, M., Ryan, L., 1993. Global tests for multiple binary outcomes. Biometrics 49, 975–988.

Lehmacher, W., Wassmer, G., Reitmeir, P., 1991. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. Biometrics 47, 511–521.

Mehrotra, D., Adewale, A., 2012. Flagging clinical adverse experiences: reducing false discoveries without materially compromising power for detecting true signals. Statistics in Medicine 31, 1918–1930.

Mehrotra, D., Heyse, J., 2004. Use of the false discovery rate for evaluating clinical safety data. Statistical Methods in Medical Research 13, 227–238.

Newcombe, R., Nurminen, M., 2011. In defence of score intervals for proportions and their differences. Communications in Statistics—Theory and Methods 40, 1271–1282.

Nurminen, M., 1986. Confidence intervals for the ratio and difference of two binomial proportions. Biometrics 42, 675–676.

Piegorsch, W.W., Richwine, K.A., 2001. Large-sample pairwise comparisons among multinomial proportions with an application to analysis of mutant spectra. Journal of Agricultural, Biological, and Environmental Statistics 6, 305–325.

Pipper, C., Ritz, C., Bisgaard, H., 2012. 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) 61, 315–326.

Pocock, S.J., Geller, N.L., Tsiatis, A.A., 1987. The analysis of multiple endpoints in clinical trials. Biometrics 43, 487–498.

Qaqish, B.F., 2003. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. Biometrika 90, 455–463.

R Core Team, 2012. R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Sison, C.P., Glaz, J., 1995. Simultaneous confidence intervals and sample size determination for multinomial proportions. Journal of the American Statistical Association 90, 366–369.

Sklyar, O., Murdoch, D., Smith, M., Eddelbuettel, D., Francois, R., 2012. inline: Inline C, C++. Fortran function calls from R. R Package Version 0.3.10. URL: http://CRAN.R-project.org/package=inline.

Tamhane, A., Hochberg, Y., 1987. Multiple Comparison Procedures. Wiley & Sons, New York.

Westfall, P., 1985. Simultaneous small–sample multivariate Bernoulli confidence intervals. Biometrics 41, 1001–1013.

Westfall, P.H., Young, S.S., 1989. *P* value adjustments for multiple tests in multivariate binomial models. Journal of the American Statistical Association 84, 780–786.