DOI: 10.1111/j.1541-0420.2006.00525.x

Multivariate Extensions of McNemar's Test

Bernhard Klingenberg

Department of Mathematics and Statistics, Williams College, Williamstown, MA 01267, U.S.A.
email: bklingen@williams.edu

and

Alan Agresti

Department of Statistics, University of Florida, Gainesville, FL 32611-8545, U.S.A.
email: aa@stat.ufl.edu

Summary. This article considers global tests of differences between paired vectors of binomial probabilities, based on data from two dependent multivariate binary samples. Difference is defined as either an inhomogeneity in the marginal distributions or asymmetry in the joint distribution. For detecting the first type of difference, we propose a multivariate extension of McNemar's test and show that it is a generalized score test under a GEE approach. Univariate features such as the relationship between the Wald and score test and the dropout of pairs with the same response carry over to the multivariate case and the test does not depend on the working correlation assumption among the components of the multivariate response. For Q1 sparse or imbalanced data, such as occurs when the number of variables is large or the proportions are close to zero, the test is best implemented using a bootstrap, and if this is computationally too complex, a permutation distribution. We apply the test to safety data for a drug, in which two doses are evaluated by comparing multiple responses by the same subjects to each one of them.

Key words: Adverse events; Correlated binary data; Drug safety; Generalized score test, Marginal homogeneity; Matched pairs; Quality of life; Toxicity.

1. Introduction

Safety, toxicity, or quality-of-life assessments become important issues in early developmental stages of pharmaceutical products. For instance, the analysis of adverse event (AE) data from clinical trials is crucial in testing (and subsequent marketing) the safety of a drug. Sponsoring and regulatory agencies involved in this process continually assess whether clinical trials need to be adjusted or terminated because of safety or other concerns. Usually, not many subjects are available in early stages and several responses have to be explored jointly, many with small incidence rates. Furthermore, studies with a primary endpoint of safety, toxicity, or quality of life often measure this multivariate response at two or more occasions, employing crossover or longitudinal designs and leading to paired or repeated multivariate data.

Safety concerns were the motivation behind a secondary analysis of several AEs recorded in a small crossover clinical trial about the efficiency of a new antidepressive drug. Investigators aimed to test whether significant differences existed in incidents of AEs between varying doses of the active ingredient. A sample of 28 healthy volunteers were first given a low dose of 50 mg and then two higher doses of 200 mg and 500 mg, with a sufficient washout period in between. A placebo treatment was also mixed in either at the beginning, the end, or at any of the intermittent stages of a subject's increasing dose sequence. All doses were judged to be acceptable from previous safety studies involving a single dose. In this article, we compare the incidence of AEs under either of the two higher doses to the incidence of AEs under either the placebo or the low dose, which was a particular contrast of interest to the investigators. Extensions of the methodology to handle several doses at once are mentioned at the end of this article.

Table 1 displays summary data for the four most common AEs that were observed under these two grouped dose levels, which we label low and high. For each individual AE, the McNemar test is a well known procedure for comparing the paired incidence rates. However, for several, possibly correlated AEs, how can we conduct a multivariate extension of that test?

Generalizations of McNemar's test to the case of two *independent* samples of paired univariate binary responses were discussed by Feuer and Kessler (1989) and for binary crossover data by Becker and Balagtas (1993). Agresti and Klingenberg (2005) developed strategies for comparing two independent *multivariate* binary vectors for a global, comparative evaluation of marginal incidence rates in two groups. They proposed likelihood ratio (LR) and score-type tests, supplemented with exact permutation approaches in cases of sparse data. Adjustments to the regular McNemar statistic in case of *dependent*

Table 1

Summary of proportions for four adverse events recorded at each of two dose levels of a drug tested on 28 *patients, and* 98.75% *score confidence intervals for the true differences. Results of a global score-type test are indicated in the last row*

	Drug dose		
Adverse event	Low	High	Score CI
Headache	4/28	4/28	[−0.23, 0.23]
Somnolence	2/28	4/28	[−0.16, 0.30]
Ecchymosis	4/28	1/28	[−0.35, 0.13]
Sore throat	1/28	3/28	[−0.16, 0.30]
SMH: $W_0 = 5.05$ (bootstrap P-value = 0.35)			

samples of clustered binary data were proposed by Eliasziw and Donner (1991) and Obuchowski (1998). Pesarin (2001) developed methods for comparing two dependent *vectors* of sample proportions. He combined exact results from multiple univariate tests through a nonparametric combination function. This approach calculates an exact P-value for the McNemar statistic for each of the 2 × 2 tables formed with an individual AE. A test for the global hypothesis of no difference is then obtained by suitably combining the (dependent) individual P-values. He illustrates with an example involving two AEs.

For the univariate case, the McNemar test is the score statistic. In Section 2 of this article we develop an extension of that test to the multivariate case, using multivariate methods rather than combining univariate results. Several properties carry over to the multivariate case, such as the relationship between the Wald and score test for paired binary and multicategory responses (Ireland, Ku, and Kullback, 1969) and the dropout of pairs with the same response sequence in each sample. We also present connections to generalized score tests under a GEE approach and show that the statistic is invariant to the working correlation assumption among the multiple binary responses. Ordinary LR and score tests are also discussed, but these are intractable when the number of AEs is large. All these tests focus on the equality of the marginal proportions of each AE at the two doses and have asymptotic chi-squared distributions. We establish guidelines for the asymptotic behavior, but in general recommend a bootstrap approach. Section 3 considers a second, narrower hypothesis that specifies symmetry in the joint distribution of the multivariate response with respect to the two doses. The corresponding test is best implemented using a permutation distribution rather than an asymptotic one. For sparse data, which occur when the number of AEs is large and the bootstrap approach becomes computationally infeasible, we also recommend conducting the test about the first-order margins under this stronger set of restrictions. The final section briefly discusses extensions to multiple responses or multiple categories.

Our presentation is framed entirely in terms of a safety analysis comparing two doses of a drug. However, the methods and results apply to any setting in which paired or repeated multivariate binary observations are obtained, such as in testing neurotoxicity of a substance as was recently discussed by Han et al. (2004) or as an alternative to summary score approaches in assessing quality of life (Ribaudo and Thompson, 2002).

Table 2

Safety profiles (1 = *present,* 0 = *absent*) *of four adverse events* (*AE*) *observed under two doses of a drug on* 28 *subjects*

Low dose				High dose				
AE 1	AE 2	AE 3	AE 4	AE 1	AE 2	AE 3	AE 4	Count
1	1	0	0	1	1	0	0	1
1	0	0	0	1	1	0	0	1
1	0	0	0	0	1	0	0	1
1	0	0	0	0	0	0	1	1
0	1	1	0	1	0	0	0	1
0	0	1	1	0	0	0	0	1
0	0	1	0	0	0	0	1	1
0	0	1	0	0	0	0	0	1
0	0	0	0	1	0	0	0	1
0	0	0	0	0	1	0	0	1
0	0	0	0	0	0	1	0	1
0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	16

Profiles with a count of zero (e.g., those not observed) are not shown.

2. Tests of Simultaneous Marginal Homogeneity

We consider paired multivariate binary data, with c binary variables indicating the incidence of c AEs observed under two conditions (e.g., doses). For dose i, let $y_{ij} = 1$ if a subject experiences AE j and $y_{ij} = 0$ otherwise, $j = 1, \ldots, c$, $i = 1, 2$. Let $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)' = (y_{11}, \ldots, y_{1c}, y_{21}, \ldots, y_{2c})'$ denote the $2c$-dimensional binary responses for a randomly selected subject, where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ refer to the responses at the low and high dose, respectively. A 2^{2c} contingency table summarizes all possible outcomes for $\boldsymbol{y}$, which are often referred to as safety profiles. Table 2 shows the nonempty cells of this table for the data summarized in Table 1. We see that 16 of the 28 subjects did not experience any AEs.

We assume a multinomial distribution for the counts in this 2^{2c} table, with sample size equal to n. Let $\pi_i(j) = P(y_{ij} = 1)$ denote the first-order marginal probability of observing AE j at dose i. Similarly, let $\pi_i(j, k) = P(y_{ij} = 1,\ y_{ik} = 1)$ and $\pi(j, k) = P(y_{1j} = 1,\ y_{2k} = 1)$. This section considers the null hypothesis

$$H_0 : \pi_1(j) = \pi_2(j), \quad j = 1, 2, \ldots, c. \tag{1}$$

We say that the 2^{2c} table cross classifying all safety profiles satisfies *simultaneous marginal homogeneity* if this holds, and we denote it by SMH. Q2

2.1 *Wald and Score-Type Tests of SMH*

Let $\boldsymbol{d} = (d_1, \ldots, d_c)'$, where $d_j = \hat{\pi}_1(j) - \hat{\pi}_2(j)$ is the jth difference between the marginal sample proportions. The vector of differences $\boldsymbol{d}$ has covariance matrix Σ with

$$\mathrm{Var}(d_j) = \left[\pi_1(j) + \pi_2(j) - 2\pi(j, j) - \{\pi_1(j) - \pi_2(j)\}^2\right]\big/ n$$

$$\mathrm{Cov}(d_j, d_k) = [\pi_1(j, k) + \pi_2(j, k) - \{\pi(j, k) + \pi(k, j)\} - \{\pi_1(j) - \pi_2(j)\}\{\pi_1(k) - \pi_2(k)\}]/n.$$

A Wald statistic is obtained by replacing the unknown proportions in Σ with the corresponding sample proportions $\hat{\pi}_i(j), \hat{\pi}_i(j, k)$, and $\hat{\pi}(j, k), i = 1, 2,\ j, k = 1, \ldots, c$ and calculating the quadratic form $W = \boldsymbol{d}'\hat{\Sigma}^{-1}\boldsymbol{d}$.

The multinomial assumption for the counts in the 2^{2c} table implies asymptotic normality of the marginal sample proportions $\{\hat{\pi}_i(j)\}$ and $\boldsymbol{d}$. The above formula for the variance makes it clear that in order to avoid a degenerate distribution, (a) we need $0 < \pi_i(j) < 1$ for all j and at least one of $i = 1$ or 2, and (b) we cannot have $\pi_1(j) = \pi_2(j) = \pi(j, j)$. The former is not restrictive in practice, as an AE that was never (or always) observed under either dose would hold no information about marginal inhomogeneity. Under these conditions and SMH the test statistic W has an asymptotic chi-squared distribution with d.f. = c.

For a variety of basic univariate contingency table analyses, the chi-squared approximation for a Wald statistic has been shown to be inadequate unless n is very large. Thus, for any n and c we prefer an alternative statistic that uses the pooled estimate of the covariance matrix under the null hypothesis of SMH. The pooled estimate for the common proportion $\pi_0(j)$ in each dose group is given by

$$\hat{\pi}_0(j) = \{\hat{\pi}_1(j) + \hat{\pi}_2(j)\}/2. \tag{2}$$

The variance of d_j and the covariance of d_j and d_k then simplify to

$$\begin{aligned}\mathrm{Var}_0(d_j) &= [\pi_1(j) + \pi_2(j) - 2\pi(j, j)]/n \\ &= 2\{\pi_0(j) - \pi(j, j)\}/n \\ \mathrm{Cov}_0(d_j, d_k) &= [\pi_1(j, k) + \pi_2(j, k) - \{\pi(j, k) + \pi(k, j)\}]/n.\end{aligned}$$

Using corresponding sample proportions, denote the estimate of the variance–covariance matrix of $\boldsymbol{d}$ under these expressions by $\hat{\Sigma}_0$. The quadratic form statistic is then $W_0 = \boldsymbol{d}'\hat{\Sigma}_0^{-1}\boldsymbol{d}$ and we can relax condition (a) to allow for the case $\pi_1(j) = 0$, $\pi_2(j) = 1$ or vice versa. We refer to W_0 as a "score-type" statistic, because a full score test for this hypothesis requires estimating the covariances solely under SMH, which is considerably more complex and discussed in Section 2.3. In the univariate case ($c = 1$) with just a single AE, W_0 reduces to the well-known McNemar statistic for binary matched pairs.

There is a surprisingly simple relationship between W and W_0. The two estimates of the covariance matrix are linked through $\hat{\Sigma} = \hat{\Sigma}_0 - \boldsymbol{d}\boldsymbol{d}'/n$, which leads to the relationship

$$W = W_0/(1 - W_0/n) \tag{3}$$

between the Wald and score-type statistic. Ireland et al. (1969) showed the same type of result for paired multicategorical responses in the univariate case, for which Stuart (1955) had proposed a score-type statistic. Although (3) implies that W and W_0 converge to the same asymptotic χ^2_c distribution, Sections 2.4 and 2.5 show that the convergence is much faster for the score-type statistic.

2.2 *A GEE Approach to Testing SMH*

In this section, we show that W_0 is a generalized score statistic for testing SMH using the marginal modeling approach based on solving GEEs (Zeger and Liang, 1986). The GEE approach specifies a model $E[\boldsymbol{y}] = \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$ for the vector of marginal probabilities and postulates a working correlation matrix for $\boldsymbol{y}$ rather than a model for its joint distribution. A GEE estimator of a parameter vector $\boldsymbol{\beta}$ is the solution to the score equation

$$S(\boldsymbol{\beta}) = (\partial\boldsymbol{\pi}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}')V^{-1}(\boldsymbol{\beta})\sum_{k=1}^{n}\{\boldsymbol{y}_k - \boldsymbol{\pi}(\boldsymbol{\beta})\} = 0, \tag{4}$$

where $V(\boldsymbol{\beta}) = D^{\frac{1}{2}}(\boldsymbol{\beta})RD^{\frac{1}{2}}(\boldsymbol{\beta})$ is a working covariance matrix for the $2c$ marginal responses, $D(\boldsymbol{\beta})$ is a $2c \times 2c$ diagonal matrix with $\mathrm{Var}(y_{ij})$ as diagonal elements, and R is the working correlation matrix. Due to the lack of subject-specific covariates, $V(\boldsymbol{\beta})$ is the same for all observations $\{\boldsymbol{y}_k\}$, which simplifies expressions considerably.

The model $\boldsymbol{\pi}(\boldsymbol{\beta}) = \boldsymbol{\beta}$ with a separate parameter for each marginal probability has GEE estimate $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\pi}}$, the marginal sample proportions. The SMH hypothesis (1) corresponds to the restriction $H\boldsymbol{\beta} = 0$, where $H = (I_c \,|\, {-I_c})$ is a block matrix consisting of two identity matrices of dimension c. Let $L = (I_c \,|\, I_c)$ be another such block matrix. Then, the solution to (4) under SMH is given by the fixed point equation

$$\hat{\boldsymbol{\beta}}_0 = L'\big(LV^{-1}(\hat{\boldsymbol{\beta}}_0)L'\big)^{-1}LV^{-1}(\hat{\boldsymbol{\beta}}_0)\hat{\boldsymbol{\pi}}. \tag{5}$$

This solution is a weighted average of the marginal sample proportions. Interestingly, under independence or exchangeable working correlation assumptions, the solution can be obtained explicitly as $\hat{\boldsymbol{\beta}}_0 = L'L\hat{\boldsymbol{\pi}}/2 = \hat{\boldsymbol{\pi}}_0$, the pooled marginal sample proportions (2). For other working correlation matrices (such as an unstructured one), the SMH solutions must be obtained iteratively by alternating between updating the right- and left-hand side of (5), taking into account the assumed structure of R. In general, these solutions differ from the pooled marginal sample proportions.

Under a GEE approach, generalized score tests for model parameters have been proposed by Rotnitzky and Jewell (1990) and Boos (1992). For testing our hypothesis $H\boldsymbol{\beta} = 0$ in $\boldsymbol{\pi}(\boldsymbol{\beta}) = \boldsymbol{\beta}$, a statistic discussed by Boos (1992, p. 331) has the form

$$T_{GS} = S(\hat{\boldsymbol{\beta}}_0)'I^{-1}(\hat{\boldsymbol{\beta}}_0)H'(H\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_0)H')^{-1}HI^{-1}(\hat{\boldsymbol{\beta}}_0)S(\hat{\boldsymbol{\beta}}_0),$$

where $I(\boldsymbol{\beta}) = nV^{-1}(\boldsymbol{\beta})$ is the expected information matrix (see technical report available at `www.tibs.org/biometrics`) and $\widehat{\mathrm{Cov}}(\boldsymbol{\beta}) = \sum_{k=1}^{n}\{\boldsymbol{y}_k - \boldsymbol{\pi}(\boldsymbol{\beta})\}\{\boldsymbol{y}_k - \boldsymbol{\pi}(\boldsymbol{\beta})\}'/n^2$.

In the context of testing SMH, T_{GS} simplifies considerably to

$$n^2\boldsymbol{d}'\left(H\sum_{k=1}^{n}\boldsymbol{y}_k\boldsymbol{y}_k'\,H'\right)^{-1}\boldsymbol{d},$$

which is precisely our statistic W_0 written in matrix notation. It does not depend on R and hence is invariant to the choice of the working correlation structure among the $2c$ responses. Furthermore, on moving H inside the sum, we see that for subjects with safety profiles $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)'$ that have $\boldsymbol{y}_1 = \boldsymbol{y}_2$, $H\boldsymbol{y}\boldsymbol{y}'H' = 0$ because $H\boldsymbol{y} = 0$. Hence, as is the case for McNemar's test in the univariate case, subjects with the same profile under the two doses do not contribute to the test statistic. (W_0 calculated with the 11 subjects with different profiles $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ equals W_0 calculated with all 28 subjects.)

Similarly, the generalized Wald statistic

$$T_{GWII} = (H\hat{\boldsymbol{\beta}})'(HI^{-1}(\hat{\boldsymbol{\beta}})\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})I^{-1}(\hat{\boldsymbol{\beta}})H')^{-1}H\hat{\boldsymbol{\beta}}$$

given in Boos (1992, p. 329) reduces in our context of testing SMH to

$$n^2\boldsymbol{d}'\left(H\sum_{k=1}^{n}(\boldsymbol{y}_k-\hat{\boldsymbol{\pi}})(\boldsymbol{y}_k-\hat{\boldsymbol{\pi}})'\,H'\right)^{-1}\boldsymbol{d}.$$

This is our Wald statistic W in matrix form, and it is also seen to be invariant to the working correlation assumption on the $2c$ responses. However, subjects with identical safety profiles under the two doses do contribute to it.

2.3 *Likelihood Ratio and Ordinary Score Test of SMH*

Alternative tests of SMH can be based on the maximum likelihood (ML) estimates of the cell probabilities under SMH. For instance, one could construct the LR statistic. To obtain the maximized likelihood under SMH, one must maximize a multinomial likelihood with $2^{2c}-1$ joint probabilities, subject to equality constraints relating two sets of c marginal probabilities. One approach is to use the Lagrange method of undetermined multipliers together with the Newton–Raphson method as implemented by Lang and Agresti (1994). An R-function ("mph.fit") for the algorithm is available from Prof. J. B. Lang (Statistics Dept., Univ. of Iowa, e-mail: `jblang@stat.uiowa.edu`, details at `www.stat.uiowa.edu/∼jblang`). However, this approach becomes computationally difficult as c increases. (e.g., we were not able to use the "mph.fit" software for $c > 4$ AEs in our data set.) The LR statistic G^2 equals -2 times the log of the ratio of the maximized likelihoods under SMH and under the unrestricted case.

A corresponding Pearson statistic compares the 2^{2c} observed and fitted counts for the SMH model, using the usual $X^2 = \sum(\text{observed} - \text{fitted})^2/\text{fitted}$. This is the actual score statistic for testing SMH. Again, like the LR test, it is computationally infeasible with current software unless c is small. The LR and score statistics also have large-sample chi-squared distributions with d.f. = c.

2.4 *SMH for the Drug Safety Data*

The generalized score statistic, and hence W_0, is available in SAS (PROC GENMOD with the GEE implementation; sample SAS and Ox code are available from the first author's website `www.williams.edu/∼bklingen`), as is the Wald statistic W. For our data with $c = 4$ AEs for an antidepressive drug administered to $n = 28$ subjects, there are $2^{2\times 4} = 256$ possible safety profiles, according to the (yes, no) outcome for each AE at the two dose levels. The sample proportions in Table 1 refer to the eight first-order marginal probabilities $\{\pi_i(j),\ j = 1, \ldots, 4,\ i = 1,\ 2\}$ of the underlying $2^{2\times 4}$ contingency table. For these data, $W = 6.17$ and $W_0 = 5.05$ with d.f. = 4. These have asymptotic P-values of 0.19 and 0.28 for testing SMH. However, we are skeptical that the chi-squared limiting distribution is valid with such sparse data in which n is small and the sample marginal proportions are near 0.

To get some feedback about whether asymptotic results are sensible, we generated a P-value for these statistics using the bootstrap method. We repeatedly took multinomial samples of size n, using as the multinomial probabilities the ML estimates obtained under the SMH hypothesis. The P-value is then the proportion of generated samples for which the test statistic is at least as large as the observed value. Results are identical for the Wald statistic W and the score-type statistic W_0, because of the one-to-one relationship between them. For 500,000 resamples, the bootstrap P-value equals 0.35. Figure 1 presents a density estimate of the bootstrap distribution for the two statistics and compares it with their asymptotic chi-squared distribution. The plot shows that using asymptotic results would be misleading with such sparse data, although the tail behavior is much closer to its asymptotic distribution for W_0 than for W. We will check more closely the asymptotic behavior of W and W_0 in the next subsection.

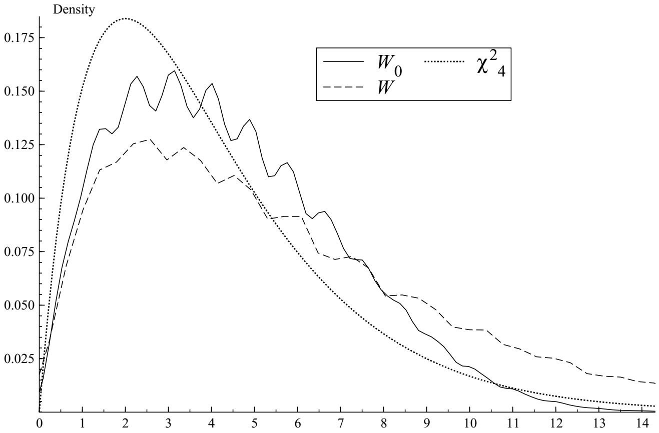

Figure 1. Bootstrap distribution of the score-type (W_0) and Wald (W) statistic for testing SMH for data in Table 1 (500,000 resamples). The asymptotic chi-squared reference distribution with d.f. = 4 is also shown.

For the ML fit of SMH, the likelihood-based statistics are $G^2 = 3.73$ and $X^2 = 2.42$. Again, the asymptotic P-values of 0.44 and 0.66 need to be treated with skepticism because of the sparseness. These statistics are computationally too complex to implement in the bootstrap or to simulate for an asymptotic evaluation.

2.5 *Asymptotic Behavior of W and W*$_0$

This section reports simulation results in order to study adequacy of asymptotic chi-squared distributions for the Wald and score-type statistics for a relatively small c. As in the previous section, we use the ML estimates of the multivariate probabilities under SMH for our sparse and imbalanced safety data set to generate samples. To explore the asymptotic behavior, we generated multinomial samples of size $n = 20$ to $n = 200$ with the first two ($c = 2$) and all four ($c = 4$) AEs of Table 1. The results in Table 3, based on 100,000 simulations for each case, enable us to compare the mean and variance of W and W_0 to the nominal values of c and $2c$ and to compare the actual proportions in the tails of their sampling distributions to nominal values of 0.10, 0.05, and 0.01. We see that for sample sizes less than 100, neither statistic is well approximated by a χ^2 distribution and consequently we recommend the bootstrap under such circumstances.

Based on what applies for the quality of the approximation of McNemar's test to a chi-squared in the univariate case, it seems sensible to inspect the sum n^* of the two off-diagonal elements in each of the c 2×2 tables that cross-classify an AE at the two dose levels. The normal approximation to the null

Table 3

Results of a simulation study for W and W_0 with 100,000 simulated data sets at each combination of c and n. n^* are the sum of the off-diagonal elements in each marginal 2×2 table cross-classifying incidence of an AE at the two dose levels

						$P(W, W_0 > \chi^2_{c,\alpha})$		
c	n	$n^* < 10$	Stat.	Mean	Var.	$\alpha = 0.100$	$\alpha = 0.050$	$\alpha = 0.010$
4	20	100%	W	6.21	16.20	0.303	0.191	0.063
			W_0	4.41	4.73	0.068	0.010	0.000
	50	75%	W	4.98	15.19	0.184	0.117	0.043
			W_0	4.32	8.65	0.126	0.065	0.011
	100	6%	W	4.47	11.8	0.141	0.083	0.026
			W_0	4.18	8.93	0.115	0.061	0.014
	150	0%	W	4.29	11.23	0.125	0.070	0.019
			W_0	4.11	8.67	0.108	0.056	0.013
	200	0%	W	4.21	9.72	0.118	0.065	0.017
			W_0	4.08	8.53	0.106	0.055	0.011
2	20	100%	W	2.58	3.75	0.125	0.055	0.006
			W_0	2.17	2.00	0.055	0.015	0.000
	50	87%	W	2.28	4.71	0.130	0.071	0.014
			W_0	2.11	3.47	0.103	0.050	0.005
	100	11%	W	2.14	4.52	0.114	0.059	0.013
			W_0	2.05	3.84	0.102	0.050	0.009
	150	0%	W	2.09	4.34	0.110	0.057	0.013
			W_0	2.03	3.90	0.102	0.050	0.009
	200	0%	W	2.07	4.27	0.107	0.055	0.012
			W_0	2.03	3.93	0.100	0.050	0.010

Note: For $n = 20$, 16% of the simulated data sets had no observation at both doses for at least one AE. For all other n, this percentage was less than 1%. These data sets were eliminated from the summary statistics.

binomial distribution of this sum, which leads to the univariate McNemar statistic, works well if $n^* \geq 10$. The chi-squared approximation for W_0 seems to hold adequately at $n = 150$ and at both c values, when none of the c marginal tables has $n^* < 10$. Table 3 also shows that a much larger n is required for W to be approximately chi-squared. This evidence is one reason for our strong preference for using W_0 instead of W. For our data set, none of the $c = 4$ n^*'s was larger than 10 and hence using the asymptotic χ^2_4 distribution is not justified, as we saw in the previous section.

3. A Stronger Symmetry Hypothesis for the Multivariate Response

The previous section compared the one-dimensional marginal distributions for each AE. This would normally be the main focus. However, in some cases it might be of interest to compare the entire c-dimensional distributions of all AEs jointly under the two doses. For instance, even if differences between marginal probabilities are insignificant, the elevated joint incidence of some of them under the higher dose (possibly due to higher correlations of AEs under it) might pose a serious safety concern. The null hypothesis is then that

$$P(y_{11} = a_1, \ldots, y_{1c} = a_c) = P(y_{21} = a_1, \ldots, y_{2c} = a_c) \qquad (6)$$

for all possible safety profiles $(a_1, \ldots, a_c)$ for the c AEs. That is, the 2^c joint distribution of all AEs is identical under the two doses and there is symmetry in a subject's safety profile. This is a more complete description of "no dose effect," one that is implied by the situation in which each subject makes the same responses for the c AEs regardless of the dose. Note that SMH is a special case.

3.1 *Permutation Test for Sparse Data*

Although one could construct large-sample tests of this hypothesis, they would have extremely limited scope when c is large, because of sparseness of the data relative to a large d.f. value. However, it is straightforward to construct a permutation test, which applies with any c and n. Since (6) implies that the two c-dimensional distributions are exchangeable for the two doses, we consider all 2^n possible ways of interchanging $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ in a subjects' observed safety profile $\boldsymbol{y}$. This generates an exact distribution for a test statistic of interest. In practice, when n is moderate or large, this is computationally infeasible. One can then merely randomly generate a very large number of the permutations to obtain a suitably precise estimate of the exact P-value.

Regarding a test statistic for this narrow hypothesis, again the LR statistic is computationally unattractive, as it involves (under the null) maximizing a likelihood over $2^{2c} - 1$ parameters, subject to equality constraints that replace $2 \times (2^c - 1)$ parameters by $2^c - 1$ parameters. A statistic that is computationally simpler treats the data in the form of n strata, one for each subject. The table for subject i is a 2×2^c table, listing all possible safety profiles for the c AEs across columns, with row 1 for the low and row 2 for the high dose. Each table has one observation in each row. Now, suppose that given a particular subject, the safety profiles are equally likely at the two doses (i.e., there is conditional independence between the safety profile and dose, given subject). This implies that the safety profiles are also equally likely to occur under each of the two doses in the 2×2^c marginal table collapsed over subjects. But this is exactly condition (6). Hence, one can test the hypothesis by testing for conditional independence in the stratified table, using a generalized Mantel–Haenszel statistic for a multicategory response with unordered categories (Landis, Heyman, and Koch, 1978). Standard software (such as SAS's PROC FREQ) provide this statistic. In general, its large-sample distribution would have d.f. $= 2^c - 1$, but in practice many profiles would not be observed, and for results to be valid regardless of the values of c and n one should use the permutation distribution. See Darroch (1981) for a discussion of the use of such statistics for the various hypotheses that can be considered for repeated measurement on a categorical response.

For the safety data in Table 2, only eight different safety profiles (out of 16 possible) were observed under the low or high dose. Conditional on observing safety profiles of these kinds, the generalized Mantel–Haenszel statistic equals 8.74, with d.f. = 7 (P-value = 0.29). (For SAS or Ox code, see the first author's website.) As with the score-type statistic W_0, subjects with the same profile under the two doses do not contribute to the test statistic. Again, because of sparseness and the relatively large d.f. value, we treat this P-value with skepticism. A more valid P-value results from randomly selecting permutations from the 2^{28} possible ones. Using a random sample of 5 million of them (which takes less than 7 minutes in Ox), we obtained an estimated P-value of 0.21.

3.2 *Permutation Small-Sample Testing of SMH*

In Section 2, we recommended the bootstrap to obtain P-values for the SMH hypothesis under small to moderate sample sizes. However, with a larger number of AE, fitting the multinomial model under SMH can be computationally impractical. For these cases, we need another approach for testing SMH under the common case of small sample and/or sparse data situations. This section proposes the permutation test under the more restrictive null hypothesis (6) for testing SMH. That is, one uses the procedure of the previous subsection, but with W or W_0 as the test statistic which are designed to detect a shift in the marginal distribution. Because of the relationship (3) between W and W_0, they each have the same permutation P-value.

For the data in Table 1, using 5 million permutations, we get an estimated P-value of 0.32 (compared to 0.35 from the bootstrap analysis) for the score-type statistic $W_0 = 5.05$. One needs to realize that this P-value is generated under stronger conditions (6) that imply SMH. The effect of this is shown in Figure 2, which compares the permutation and bootstrap distributions of W_0 for data sets generated under SMH. The range of the permutation distribution is less than that of the bootstrap distribution because it does not consider as extreme reconfigurations of the data. For instance, the largest marginal count of an AE obtainable in any permutation of the original data is 6, while this might be higher when generating data using the fitted multinomial distribution. Under the latter, some generated tables showed marginal counts of more than 6 for one or two AE under the high dose, coupled with a low count of less than 3 under the low dose. These tables led to (highly significant) W_0 values of 13 and larger, explaining the fatter tail of the bootstrap distribution.

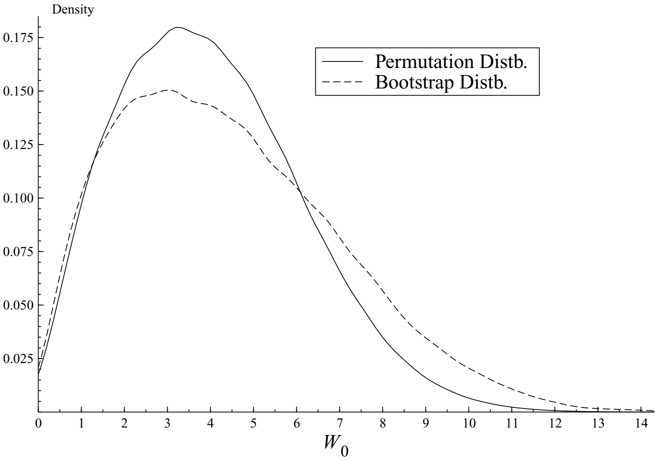

Figure 2. Comparison of permutation and bootstrap distribution of W_0 for data sets generated under SMH. We considered 3000 permutations for each of 20,000 generated data sets.

The permutation distribution's thinner tail leads to elevated type I error rates for testing SMH under it, compared to the reference bootstrap distribution. (Note that we recommend the permutation approach only for situations where the bootstrap is computationally impractical). If simpler models for the joint probability structure can be chosen (cf. the next section), more appropriate exact tests for SMH could be constructed. This is a topic for future research.

3.3 *Modeling the Marginal Inhomogeneity*

An alternative approach of comparing AE incidence rates, particularly in the presence of additional covariates, is through generalized linear mixed models (GLMMs). This model class assumes random effects specific to each subject. SMH can be tested by comparing the maximized log likelihood under a restricted model that satisfies SMH to a more general model. However, GLMMs focus on subject-specific probabilities, while interest here lies in the comparison of marginal probabilities. More importantly, GLMMs postulate a nonnegative, exchangeable correlation structure among AEs sharing a common random effect. This is inappropriate for our safety data set, for which the pairwise sample correlations range between −0.17 and 0.71.

Regardless of the modeling approach, if one expects probabilities for AEs to be higher in one group, one could attempt to build power by focusing a single-degree-of-freedom test on this common effect. For example, the marginal model $\pi_i(j) = \beta_j + \alpha \mathrm{I}(i = 2)$ (or its subject-specific counterpart) is sensible if we wish to build power for detecting an increased incidence of AEs at the higher dose. Estimation of such a marginal model can proceed via ML (using, e.g., Lang's R-function and possibly restricting higher-order interactions) or GEE. The SMH hypothesis corresponds to $H_0 : \alpha = 0$. For the data analyzed in this article, both approaches yield nonsignificant estimates: $\hat{\alpha} = 0.011\,(0.029)$ with LR test P-value 0.70 for the ML approach and $\hat{\alpha} = 0.035\,(0.027)$ with generalized score test P-value 0.42 under the GEE approach with an unstructured correlation matrix. However, these tests are not very sensitive for these data because we do not expect all incidence rates to be uniformly larger under the higher dose.

4. Follow-Up and Extensions

In the study used here, the low dose of 50 mg was always given before the higher doses of 200 and 500 mg. Thus, proper care must be taken to ensure that any difference between the two groups is not an artifact of the design (e.g., blinding to the dose sequence and no period or carry-over effects). Also, of course, in practice one would want to follow-up the test by an assessment comparing the individual AEs.

One way to do this is with a confidence interval comparing the two proportions for each side effect. To ensure actual confidence level being relatively near the nominal level, we recommend using the score confidence interval (Tango, 1998), which works well even when n is relatively small and the data have relatively few outcomes of the event of interest (Newcombe, 1998; Agresti and Min, 2005). Table 1 presents Bonferroni score intervals for the individual AEs, with asymptotic simultaneous confidence level of at least 0.95 (i.e., each one is a 98.75% score confidence interval). For these data, neither the global tests nor the individual tests show a significant difference of incidence rates between the two dose levels.

The methods of this article extend in obvious ways to several repeated measures. To test SMH with T repeated measures on c variables, one can extend the score-type statistic W_0 by forming a vector $\boldsymbol{d}$ of $c(T-1)$ differences of proportions, comparing a given proportion for each dose to the corresponding proportion for an arbitrary baseline dose. The methods also extend in obvious ways to multicategory responses. For instance, it is increasingly common to classify an AE by its severity, using categories not present, as mild and severe. For ordered categories and corresponding scores, one could then base the statistic on a vector of $c(T-1)$ differences of means.

In either the binary or ordinal case, one could form a statistic that is sensitive to a linear trend for some or all of the dimensions (e.g., to help detect an increasing trend in the proportion of times an AE occurs as the dose level increases), rather than generally utilize the $T-1$ differences. This gives the potential for building power relative to the general statistics which have relatively large d.f. values.

For a nominal-scale comparison of T repeated measures simultaneously on c variables, with r_j categories for variable j, the score-type test has d.f. $= (T-1)(\sum_j r_j - c)$. For a single variable, these simplify to the extension of the Stuart (1955) test to a r^T contingency table. With even moderate T and c, asymptotic methods are suspect. A sensible strategy for testing is a permutation test for the $(T!)^n$ allocations of the subjects' sub-vectors of responses to the T times, computing the extended W_0 statistic for each (or, for a random sample of them).

Acknowledgements

This research was supported by grants from the National Institutes of Health and the National Science Foundation. The authors thank Dr Christy Chuang-Stein for permission to use the data, Dr Joseph Lang for permission to use his software mph.fit, Dr Jurgen Doornik for the free matrix program Ox, and the reviewers for comments that led to an improvement of the article.

References

Agresti, A. and Klingenberg, B. (2005). Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. *Applied Statistics* **54,** 691–816.

Agresti, A. and Min, Y. (2005). Simple improved confidence intervals for comparing matched proportions. *Statistics in Medicine* **24,** 729–740.

Becker, M. P. and Balagtas, C. C. (1993). Marginal modeling of binary cross-over data. *Biometrics* **49,** 997–1009.

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

Darroch, J. N. (1981). The Mantel-Haenszel test and tests of marginal symmetry; fixed-effects and mixed models for a categorical response. *International Statistical Review* **49,** 285–307.

Eliasziw, M. and Donner, A. (1991). Application of the McNemar test to non-independent matched pair data. *Statistics in Medicine* **10,** 1981–1991.

Feuer, E. J. and Kessler, L. J. (1989). Test statistic and sample size for a two-sample McNemar test. *Biometrics* **45,** 629–636.

Han, K. E., Catalano, P. J., Senchaudhuri, P., and Mehta, C. (2004). Exact analysis of dose response for multiple correlated binary outcomes. *Biometrics* **60,** 216–224.

Ireland, C. T., Ku, H. H., and Kullback, S. (1969). Symmetry and marginal homogeneity of an r × r contingency table. *Journal of the American Statistical Association* **64,** 1323–1341.

Landis, J. R., Heyman, E. R., and Koch, G. G. (1978). Average partial association in three-way contingency tables: A review and discussion of alternative tests. *International Statistical Review* **46,** 237–254.

Lang, J. B. and Agresti, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association* **89,** 625–632.

Newcombe, R. (1998). Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine* **17,** 2635–2650.

Obuchowski, N. (1998). On the comparison of correlated proportions for clustered data. *Statistics in Medicine* **17,** 1495–1507.

Pesarin, F. (2001). *Multivariate Permutation Tests with Applications in Biostatistics.* Chichester: John Wiley & Sons.

Ribaudo, H. J. and Thompson, S. G. (2002). The analysis of repeated multivariate binary quality of life data: A hierarchical model approach. *Statistical Methods in Medical Research* **11,** 69–83.

Rotnitzky, A. and Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for clustered correlated data. *Biometrika* **77,** 485–497.

Stuart, A. (1955). A test for homogeneity of the marginal distributions of a two-way classification. *Biometrika* **42,** 412–416.

Tango, T. (1998). Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine* **17,** 891–908.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42,** 121–130.

Received February 2005. *Revised November* 2005.
Accepted November 2005.

Queries

Q1 Author: Please check the sentence “For sparse or imbalanced data. . . ” for clarity.

Q2 Author: Please spell out ‘SMH’.