Appl. Statist. (2005)
54, *Part* 4, *pp.* 691–706

Multivariate tests comparing binomial probabilities, with application to safety studies for drugs

Alan Agresti

University of Florida, Gainesville, USA

and Bernhard Klingenberg

Williams College, Williamstown, USA

[Received October 2003. Final revision August 2004]

Summary. In magazine advertisements for new drugs, it is common to see summary tables that compare the relative frequency of several side-effects for the drug and for a placebo, based on results from placebo-controlled clinical trials. The paper summarizes ways to conduct a global test of equality of the population proportions for the drug and the vector of population proportions for the placebo. For multivariate normal responses, the Hotelling T^2-test is a well-known method for testing equality of a vector of means for two independent samples. The tests in the paper are analogues of this test for vectors of binary responses. The likelihood ratio tests can be computationally intensive or have poor asymptotic performance. Simple quadratic forms comparing the two vectors provide alternative tests. Much better performance results from using a score-type version with a null-estimated covariance matrix than from the sample covariance matrix that applies with an ordinary Wald test. For either type of statistic, asymptotic inference is often inadequate, so we also present alternative, exact permutation tests. Follow-up inferences are also discussed, and our methods are applied to safety data from a phase II clinical trial.

Keywords: Adverse events; Binary data; χ^2-test; Generalized estimating equations; Hotelling test; Marginal homogeneity; Marginal logit model; Random effects

1. Introduction

Table 1 contains summary results of the type that are often found in news magazines (e.g. *Time*) that have full page advertisements promoting a new drug. (In recent years, advertisements of this type have appeared for drugs such as Lamisil by Novartis Pharmaceuticals, Flonase by Glaxo Smith Kline, Clarinex by Schering, Pravachol by Bristol-Myers Squibb, Allegra by Aventis and Botox by Allergan.) Table 1 compares the relative frequency of several undesirable side-effects for a drug and placebo, based on results from placebo-controlled clinical trials. In the pharmaceutical industry, such side-effects are often called *adverse events*, and the studies making such comparisons of a drug with a placebo are called *safety studies*.

The data in Table 1 refer to a safety study for an asthma drug, conducted by Schering-Plough Corp. The adverse events were collected from a double-blind, randomized, phase II clinical trial in which subjects were randomized to one of three treatments: two levels of an active drug and a placebo. Each patient was followed over a period of at least 3 months. The adverse events were reported at scheduled visits to the clinic and were non-solicited reports by the subject to the investigator. The primary objective of the clinical trial was to assess a subject's lung functions as

Address for correspondence: Alan Agresti, Department of Statistics, Box 118545, University of Florida, Gainesville, FL 32611-8545, USA.
E-mail: aa@stat.ufl.edu

 0035–9254/05/54691

Table 1. Summary of incidence of several adverse events in an asthma trial

Adverse event	*Sample %*	
	Drug	*Placebo*
Upper respiratory or cold	40.4	58.5
Muscoloskeletal pain	12.3	23.1
Throat pain	16.4	10.8
Allergic rhinitis	11.6	10.8
Fatigue	10.3	7.7
Diarrhoea	7.5	10.8
Abdominal pain	9.6	4.6
Joint pain	8.2	3.1
Fever	7.5	4.6
Cough	4.1	10.8
Urinary tract infection	6.2	3.1
Sample size	146	65

a response to a treatment *versus* placebo. Subsequently, interest also focused on analysing the evidence of a difference between the occurrence of adverse events in the treated and non-treated groups.

For simplicity of exposition, in Table 1 we combined the results for the two dosage levels of the drug and compared the two drug groups combined with the placebo group. (Section 8 mentions straightforward generalizations for multiple groups.) Of the 211 subjects in the study, 146 were in the drug group and 65 in the placebo group. Table 1 lists the adverse events in order according to their overall frequency in the two groups.

In Table 1, for any given one of the 11 adverse events, a 2×2 table compares the counts on the two possible outcomes for the two groups. We can then use standard inference (e.g. a χ^2-test) to analyse whether the occurrence of that adverse event was significantly different for the two groups. However, how could we conduct a global test to analyse the evidence of a difference between the vector of 11 population proportions for the drug and the vector of 11 population proportions for the placebo? This question was first asked of one of us for similar data from another company a few years ago. In this paper, we survey strategies for answering the question.

1.1. Literature on safety studies and relevant methods

The analysis of adverse event data in clinical trials is an important part of the development, pre- and post-market characterization and safety of pharmaceutical products. Despite that fact, comparative statistical methods for the evaluation of safety outcomes are not as well developed as those for efficacy (O'Neill, 2002). O'Neill (1988) presented a general summary of statistical procedures for analysing safety data.

Lin *et al.* (2001) investigated adverse events in a placebo-controlled clinical study based on proportional hazards and logistic regression models for repeated binary data. The adverse events were handled in a univariate manner, as is the case in almost all the literature on safety studies. A simple way to conduct a global test using the univariate information in Table 1 is with the Bonferroni approach. If P_j is the P-value for the test for the 2×2 table comparing a drug with a placebo for adverse event j, an overall P-value is $11 \min_j(P_j)$ (or 1.0 if this exceeds 1). This

approach is potentially quite conservative, both because of its use of the Bonferroni inequality and because it ignores potential dependence between separate individual inferences. The conservativeness is compounded if we use a small sample discrete method for each individual test (e.g. Fisher's exact test). Less conservative Bonferroni approaches have been developed, such as sequential versions (e.g. Holm (1979)). Westfall and Young (1989) proposed a permutation resampling of the vector responses to find the probability (for each component in the vector) that the minimum P-value of all tests is less than the observed P-value. This gives an adjusted P-value for each component, following a suggestion by Mantel (1980). Their approach is implemented by using Monte Carlo generation of random permutations in the SAS procedure `MULTTEST`, which reports P-values for all individual tests (e.g. based on the marginal χ^2- or Fisher's exact tests) adjusted for correlation and discreteness. This approach does not give a global P-value.

Pocock *et al.* (1987) combined score tests for each individual component to construct a global test for multivariate binary data, extending results from O'Brien (1984). Their test is a special case of more general tests that were proposed by Lefkopoulou and Ryan (1993) that assume that outcomes are uniformly more likely for one group than for another and assume an independence or exchangeable correlation structure among them. Zhang *et al.* (1997) summarized this and related multiple-test approaches for analysing multiple end points in clinical trials with quantitative response variables. For instance, Lehmacher *et al.* (1991) described test procedures that allow, after rejection of the global null hypothesis at level α, a stepwise analysis of differences in subsets of all adverse events or even single adverse events while still maintaining an overall experimentwise error rate of α. More recently, Mehrotra and Heyse (2004) addressed multiplicity by using a less conservative approach of controlling a false discovery rate rather than an experimentwise error rate. In quite a different vein, Berry and Berry (2004) used a three-level hierarchical mixed model to obtain for each adverse event a Bayesian posterior probability that the rate is higher for the treatment. Mehrotra and Heyse (2004) and Berry and Berry (2004) analysed a data set in which only the marginal results are known for the adverse events, so it is not possible to conduct a multivariate analysis.

1.2. The multivariate approaches of this paper

In this paper, we shall consider test statistics that treat the data in a multivariate manner. Chuang-Stein and Mohberg (1993) proposed a related approach, with a multivariate Wald statistic. In Table 1, each group (drug, placebo) has $2^{11} = 2048$ possible response sequences, according to the (yes, no) outcome for the response on each adverse outcome. The percentages in Table 1 refer to the 11 one-dimensional marginal distributions of the 2^{11} contingency table for each group that shows the counts of the possible response sequences. We compare the marginal distributions for the two groups, while using the information in their joint distributions, and we also compare the joint distributions.

For multivariate normal responses, the Hotelling T^2-test is a well-known method for testing equality of a vector of means for two independent samples. (In the two-sample context, it is also called the Mahalanobis test.) We discuss analogues of this test for vectors of binary responses. Section 2 presents a likelihood ratio test comparing the marginal distributions with marginal logit modelling. The test is computationally intensive when each vector has a large number of elements. Section 3 presents a simpler Wald test and a related score-type test. Section 4 discusses tests comparing the joint distributions for the two groups. The emphasis is on permutation tests, since asymptotic tests are not justified even with relatively few side-effects. Section 5 presents analyses based on simpler models, such as random-effects models, that provide structure for the

association between the responses for different adverse events. Section 6 considers the adequacy of the large sample methods when the data are sparse and makes recommendations. Methods of each section are illustrated for the asthma data of the phase II clinical trial. Section 7 describes possible follow-up analyses, and Section 8 briefly discusses extensions to multicategory responses and comparisons of several groups, for which the tests proposed are multivariate versions of likelihood ratio and Pearson tests of independence.

2. Using marginal models for multivariate binomial vectors

For concreteness, in formulating models we refer to Table 1, which has a binary explanatory variable (group) and a multivariate binary response vector. We denote the group by $i=1$ for the drug and $i=2$ for the placebo and we denote the number of binary variables that constitute the multivariate response by c ($c=11$ for Table 1). We assume an independent multinomial distribution for the counts in each subtable of size 2^c, with sample size n_1 for group 1 and n_2 for group 2. For a randomly selected subject assigned $x=i$, let $(y_{i1},\ldots,y_{ic})$ denote the c responses, where $y_{ij}=1$ or $y_{ij}=0$ according to whether side-effect j is present or absent. Let $\pi_i(j)=P(y_{ij}=1)$. Then $\{(\pi_i(j),1-\pi_i(j)), j=1,\ldots,c\}$ are the c one-way marginal distributions for the 2^c cross-classification of responses when $x=i$.

2.1. Simultaneous marginal homogeneity model

This section considers the null hypothesis of equality of the two vectors of binomial parameters $(\pi_1(1),\ldots,\pi_1(c))$ and $(\pi_2(1),\ldots,\pi_2(c))$, i.e., for each side-effect j,

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

We refer to this as the *simultaneous marginal homogeneity* (SMH) hypothesis for the two multivariate distributions. This hypothesis corresponds to the marginal logit model

$$\log\left\{\frac{\pi_i(j)}{1-\pi_i(j)}\right\}=\beta_j, \qquad i=1,2, \quad j=1,\ldots,c. \tag{2}$$

More generally, this and other models that we consider can incorporate explanatory variables in addition to the group.

Model (2) is simple. However, maximum likelihood (ML) fitting is computationally impractical for large c. The models apply to c marginal distributions of the 2^c-table for each group, yet the product multinomial likelihood refers to the multinomial probabilities within those two tables. Note that we cannot fit the model by using only the marginal information in a table such as Table 1; we need the two 2^c joint distributions to incorporate the correlations between responses on different adverse events. See Agresti (2002), pages 464–466, for a brief review of ML methods for fitting marginal logit models.

To maximize the product multinomial likelihood subject to the SMH constraint, one approach iteratively uses Lagrange's method of undetermined multipliers together with the Newton–Raphson method (Aitchison and Silvey, 1958; Haber, 1985). We used an algorithm based on refinements of this method (Lang and Agresti, 1994; Lang, 2004), in which the matrix inverted in the Newton–Raphson step has simpler form. Let $\boldsymbol{\pi}$ denote the vector (with 2×2^c elements) of the two sets of multinomial probabilities. Among the classes of models to which this algorithm applies are the linear model having the matrix form

$$\mathbf{A}\boldsymbol{\pi}=\mathbf{X}\boldsymbol{\beta} \tag{3}$$

and generalized log-linear models of form

$$\mathbf{C}\ \log(\mathbf{A}\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}.$$

In this context, the matrix $\mathbf{A}$ applied to $\boldsymbol{\pi}$ forms the relevant marginal probabilities, and $\boldsymbol{\beta}$ is the vector of the c model parameters. For logit model (2), $\mathbf{C}$ applied to the log-marginal-probabilities forms the marginal logits for the models. An R function (`mph.fit`) for the algorithm applied to such classes of models is available from Professor J. B. Lang (Statistics Department, University of Iowa; e-mail `jblang@stat.uiowa.edu`; details at `www.stat.uiowa.edu/~jblang`). The algorithm becomes more computationally demanding as c increases, but we could use it with $c = 11$ for the example of this paper.

2.2. Testing simultaneous marginal homogeneity

After fitting model (2), likelihood-based methods can test the SMH hypothesis. With large samples, we could use a likelihood ratio or Pearson statistic testing the goodness of fit of logit model (2). Such statistics compare the fit of this model with the saturated model

$$\log\left\{\frac{\pi_i(j)}{1-\pi_i(j)}\right\} = \beta_{ij}, \qquad i = 1, 2, \quad j = 1, \ldots, c. \tag{4}$$

The SMH hypothesis (1) corresponds to $H_0 : \beta_{1j} = \beta_{2j}$, $j = 1, \ldots, c$, in this model.

The likelihood ratio statistic G^2 equals -2 times the logarithm of the ratio of the maximized likelihoods for models (2) and (4). The Pearson statistic compares the 2×2^c observed and fitted counts for model (2), using $X^2 = \Sigma(\text{observed} - \text{fitted})^2/\text{fitted}$. These two statistics have large sample χ^2-distributions with degrees of freedom $\text{df} = c$, the difference in parameter dimensionality of the two models. For these statistics, the resulting null distribution does not assume any particular structure for the joint distribution.

2.3. Drug safety example

In the first step towards a safety analysis, investigators in the phase II trial sought an overall evaluation of the safety profile of the asthma drug. The goodness-of-fit tests of model (2) yield likelihood ratio statistic $G^2 = 16.1$ and Pearson statistic $X^2 = 14.2$, each with $\text{df} = 11$. Neither statistic shows much evidence against the SMH null hypothesis ($P = 0.14$ and $P = 0.22$) for the asthma data. This is valuable information to determine whether proceeding to a larger trial is justified from a safety point of view. It is also relevant for an interim analysis of large, expensive phase III trials, in which an independent data monitoring committee monitors safety and gives recommendations based on their statistical safety analysis. In a different context, the result of such a test might be part of the statistical presentation to federal drug agencies to help to justify a drug approval application.

The joint tables for the asthma data are sparse, having 211 observations in $2 \times 2^{11} = 4096$ cells, so conclusions based on these tests are tentative. The reliability of asymptotics in such cases will be addressed further in Section 6.

3. Wald and score-type tests of simultaneous marginal homogeneity

As c increases, likelihood-based approaches become computationally more difficult. For instance, we could not use the R function that was mentioned earlier for a data set with $c > 11$ variables. Alternative strategies are needed that can also handle large c. The simplest approach to testing SMH is to form a test statistic using solely the marginal sample proportions and their variances and covariances.

In group i, let $\hat{\pi}_i(j)$ denote the sample proportion of subjects who report side-effect j. Let $\mathbf{d} = (d_1, \ldots, d_c)'$ with $d_j = \hat{\pi}_1(j) - \hat{\pi}_2(j)$, $j = 1, \ldots, c$. Appendix A gives the formula for the covariance matrix of $\mathbf{d}$. Let $\hat{\boldsymbol{\Sigma}}$ denote the sample version of this matrix. Then, a Wald statistic for testing the null hypothesis of SMH is

$$W = \mathbf{d}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{d}.$$

This also has an asymptotic null χ^2-distribution with df $= c$ and was used by Chuang-Stein and Mohberg (1993) for comparing adverse events.

In the univariate case ($c = 1$), the Wald statistic is not as reliable a method for comparing two proportions as the Pearson statistic is. For instance, its nominal size tends not to be as close to the actual size. Thus, for any c we prefer an alternative statistic that uses the pooled estimate of the variance and covariance. Appendix A also shows this matrix, which applies under the null hypothesis. Denote the pooled estimate of $\boldsymbol{\Sigma}$ by $\hat{\boldsymbol{\Sigma}}_0$. Let $W_0 = \mathbf{d}'\hat{\boldsymbol{\Sigma}}_0^{-1}\mathbf{d}$. When $c = 1$, this is the Pearson χ^2-statistic, which is the score test. We recommend it over W because of the poor performance in general of Wald inference for proportion data. We shall refer to W_0 as a 'score-type' test, since a full score test for this hypothesis requires estimating the covariances solely under SMH, which is considerably more complex.

For the data that are summarized in Table 1, $W_0 = 19.9$ with df $= 11$ ($P = 0.047$). The evidence against the null hypothesis is somewhat stronger than with the likelihood-based statistics. Of course, there is no guarantee that W_0 performs well for large c or with small n_1 and n_2. Also, Appendix A shows that when $n_1 \neq n_2$ it uses an additional assumption about the second-order marginal distributions. To obtain some feed-back on the validity of the asymptotic P-value, we could construct a P-value by using the bootstrap, repeatedly taking multinomial samples of sizes n_1 and n_2 from the two groups. The multinomial probabilities for the bootstrap are the fitted distribution for the SMH model (2). The bootstrap test P-value is the proportion of generated resamples for which W_0 is at least as large as the sample value. Using 100 000 bootstrap resamples, the bootstrap P-value for the observed value of $W_0 = 19.9$ was 0.045, compared with 0.047 from the asymptotic χ^2-distribution.

When the models are expanded to include explanatory variables, the most straightforward way to obtain parameter estimates in marginal models is the quasi-likelihood approach based on generalized estimating equations (GEEs; Liang and Zeger (1986)). This approach is summarized in Appendix B. Even without explanatory variables, the GEE approach is computationally much simpler than ML for tables with large c. With the binary predictor of group and an unstructured working correlation matrix for the joint distribution of the variables, this corresponds to iterating the weighted least squares approach of Koch *et al.* (1977) (see Miller *et al.* (1993)). The GEE methods are not likelihood based. Thus, tests of hypotheses such as SMH naturally use Wald tests rather than likelihood ratio tests. There has been some work on constructing score-type tests for the GEE approach (e.g. Rotnitzky and Jewell (1990)) which also use empirical covariance estimates to adjust for a misspecified correlation structure.

For the asthma data, the GEE approach assuming an exchangeable correlation structure among the adverse events gives a Wald statistic of 21.7, with df $= 11$ (P-value 0.03). Similar results occurred for the Wald statistic by using other working correlation structures. When applied to the linear model using the identity link function, GEEs compute the empirical covariance of the marginal sample proportions rather than the marginal sample logits. Then, the Wald statistic that is obtained with this approach is the statistic W that was introduced above, which equals 21.1. However, the empirically based standard errors for the GEE approach tend to underestimate the true standard errors (e.g. Firth (1993)), and this is supported by a study

that we conducted that is reported below in Section 6. So, we treat the P-value of 0.03 for this approach with some scepticism.

We do not believe that GEEs with Wald tests are as reliable as the test using the score-type statistic W_0 or the likelihood ratio test of the previous section. This is studied further in Section 6. Its advantages are versatility and readily available software.

4. Tests of identical joint distributions

In some cases, it may be of interest to test the null hypothesis that the entire 2^c joint distributions are identical for the two groups, i.e., for all possible response sequences $(a_1, \ldots, a_c)$,

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

When the null hypothesis is supposed to represent 'no effect', for instance with subjects making the same response whether they take a drug or placebo, then this is a more complete description than SMH of no effect. Although this hypothesis of identical joint distributions (IJDs) is narrower than SMH, in a way it is actually more nearly analogous to the Hotelling approach for normally distributed data. That test assumes a common covariance matrix for the two groups, and hence identical multivariate normal distributions.

The fitted null joint distribution results simply from finding joint sample proportions for the table collapsed over the group, and the fitted counts are these proportions multiplied by the respective sample sizes in the two groups. The likelihood ratio test, which has test statistic $G^2 = 2\,\Sigma$ observed log(observed/fitted), has residual df $= 2^c - 1$. The df-value results from comparing an alternative hypothesis with two independent sets of $2^c - 1$ multinomial probabilities with a null hypothesis with a single set. Although computationally simple, using a χ^2-distribution for this or the related Pearson X^2-statistic is not sensible for even moderate-sized c, because of extreme sparseness and the very large df-value. For instance, for the asthma data on which Table 1 is based, $G^2 = 118.6$ and $X^2 = 31.9$, but these have df $= 2047$.

Instead, we recommend conducting tests of the IJDs hypothesis using the exact permutation distribution under this null structure of exchangeability of distributions. For the sample subjects, consider all $(n_1 + n_2)!/n_1!\,n_2!$ ways of partitioning the sample into n_1 subjects for group 1 and n_2 subjects for group 2. For a chosen test statistic, the P-value is the proportion of these partitions for which the statistic is at least as large as the observed value. This P-value is calculated under the exchangeability assumption for the two groups in terms of their joint distribution, which is the null hypothesis that was mentioned above. With large n_1 or n_2, this permutation approach can be computationally intensive even with a simple test statistic. We can then select a random sample of the possible partitions. For instance, with 5 million random partitions and a true P-value of 0.05, the estimated P-value has a standard error of 0.0001, which is more than sufficient for nearly all purposes.

Even with the modest sample sizes ($n_1 = 146$ and $n_2 = 65$) of the asthma drug safety study, the permutation analysis entails the order of 10^{73} different partitions of the 211 subjects into two groups of these sizes. Thus, we took a random sample of 5 million partitions. Using the permutation distribution, $G^2 = 118.6$ has P-value 0.14 and $X^2 = 31.9$ has P-value 0.29. These P-values provide very similar results to those for the asymptotic tests of the SMH hypothesis using these two statistics.

Likewise, we could generate a P-value under the IJDs hypothesis for a statistic that is designed to detect a particular characteristic for which the two distributions differ. An example is the score-type statistic of the previous section for comparing the marginal proportions. Under the permutation distribution, $W_0 = 19.9$ has P-value equal to 0.041.

5. Tests imbedded in a model for the joint distributions

The main questions of interest for the asthma data refer to the marginal probabilities for the 11 adverse events, for the drug and placebo. The actual form of that joint distribution may be regarded as a nuisance, or at best of secondary interest. Thus, the analyses that are considered in Sections 2 and 3 dealt directly with the marginal distributions and made no attempt to describe the joint distribution of the responses. Alternatively, we can compare the marginal distributions or the joint distributions of the responses while assuming a model for the joint distribution. It is easiest to do this by considering a model for which the SMH hypothesis of Sections 2 and 3 is equivalent to the IJDs hypothesis of Section 4.

This section shows ways to compare the margins while modelling the joint distribution. It also mentions ways potentially to increase the power by considering simpler structure for the marginal inhomogeneity.

5.1. Using random effects to model the dependence

The best-known way to induce an association between the c responses is by using random effects. Let $\pi_{s(i)}(j)$ denote the probability of side-effect j for subject s who is in group i. A logistic–normal random intercept analogue of model (4) is

$$\log\left\{\frac{\pi_{s(i)}(j)}{1-\pi_{s(i)}(j)}\right\} = u_{s(i)} + \beta_{ij}, \qquad i = 1, 2, \quad j = 1, \ldots, c, \tag{5}$$

where the subject-specific random effects $\{u_{s(i)}\}$ are independent from an $N(0, \sigma)$ distribution. Under this structure, SMH and IJDs correspond to the simpler model

$$\log\left\{\frac{\pi_{s(i)}(j)}{1-\pi_{s(i)}(j)}\right\} = u_{s(i)} + \beta_{j}, \qquad i = 1, 2, \quad j = 1, \ldots, c. \tag{6}$$

Since this random-effects model implies a common, non-negative association between pairs of adverse events, it is inappropriate if there is reason to expect negative association between certain pairs of side-effects or associations that vary dramatically in strength.

Assuming this model form, we can test SMH (and IJDs) by the likelihood ratio test comparing models (6) and (5). Again, it has df $= c$. For Table 1, the likelihood ratio statistic equals 22.1 (df $= 11$; P-value 0.023).

5.2. Marginal models with simultaneous model for joint distribution

When many adverse events are measured, it may be that certain associations are negative. Then, there are alternative ways to model the joint distribution. For instance, we could use a log-linear model. This does not require assuming an exchangeability structure for the joint distribution, unless we assume a quasi-symmetric form of log-linear model (which is implied by a random-effects model). The model for the two joint distributions can be specified simultaneously with one for the marginal distributions. We can fit log-linear models simultaneously with compatible marginal models by using methods that were described in Fitzmaurice and Laird (1993) and in Lang and Agresti (1994). Lang's R function that was mentioned above can fit such models. With this approach, however, results of tests of SMH will be similar to results for tests that use a saturated structure for the joint distribution. In standard log-linear models for the joint distribution, the marginal and joint model parameters are orthogonal. In particular, if the marginal

Table 2. Summary of methods for comparing adverse event incidence for drug and placebo groups by testing SMH or IJDs

Method	*Results*
1. Marginal models	
(a) Likelihood ratio test of SMH (e.g. using Lang's R software `mph.fit`)	Likelihood ratio statistic 16.1, df = 11, $P = 0.14$
(b) Score-type test of SMH† (quadratic form using differences and a null covariance matrix)	$W_0 = 19.9$, df = 11, $P = 0.05$
(c) GEE (Wald) test of SMH (quadratic form using differences and the covariance matrix)	$W = 21.1$, df = 11, $P = 0.03$
2. Joint models	
(a) Permutation test of IJDs†	Likelihood ratio statistic 118.6, $P = 0.14$
(b) Likelihood ratio test of SMH and IJDs for random -effects models	Likelihood ratio statistic 22.1, df = 11, $P = 0.02$

†Preferred method for sparse data.

model of SMH holds, the ML estimator of the marginal model parameters is consistent even if the model for the joint distribution is incorrect.

5.3. Structure for the marginal inhomogeneity

Table 2 summarizes the types of analyses that we have applied to the asthma data. Except for the permutation tests, the P-values are based on asymptotics. Since the complete 2×2^{11} table corresponding to Table 1 is sparse, conclusions based on tests having df = 11 must be made cautiously. The asymptotics may not hold well, as we shall discuss in Section 6. More reliable and informative tests use a model-based comparison of the SMH model with a model that provides some structure for the nature of the marginal inhomogeneity. Using a narrower alternative hypothesis provides the potential for increased power and also focuses attention on estimating whatever effects may exist.

One special case of the saturated model (4) that has SMH as a further special case is the logit model

$$\log\left\{\frac{\pi_i(j)}{1-\pi_i(j)}\right\} = \alpha\, I(i=1) + \beta_j, \qquad i = 1, 2, \quad j = 1, \ldots, c. \tag{7}$$

Here, $I(\cdot)$ is an indicator function, and the model permits a shift difference α between the groups for each variable. SMH is the special case $\alpha = 0$. We could use an analogous structure in random-effects model (5). Such alternatives are worthy of attention, for instance, if we expect that each adverse event may be more likely for the drug than for the placebo.

For Table 1, model (7) has ML fit statistics $G^2 = 12.6$ and $X^2 = 8.9$, with df = 10, and $\hat{\alpha} = -0.354$ has se = 0.178. It provides slight evidence of improvement over the SMH model (2), with the change in G^2 equal to 3.5 (df = 1; P-value 0.06).

In the spirit of this model, we could form a simple statistic to summarize results across adverse events that would build power for an alternative by which the probability tends to be higher for one of the groups. For instance, for each subject we could count the total number of adverse events and compare the means for the two groups, using either asymptotic normality of the sample means or assuming a distribution such as the negative binomial distribution or using a nonparametric comparison.

For Table 1, about 80% of the subjects had no more than two adverse events, and the maximum was six. The drug and placebo groups had sample means of 1.34 and 1.48, with standard deviations of 1.33 and 1.34. The two-sample t-test has a two-sided P-value of 0.50. This is also the P-value for the likelihood ratio test comparing negative binomial models with separate means and equal means. The 'exact' Wilcoxon test comparing the two distributions using a conditional test for the 2×7 table cross-classifying the group with the number of adverse events (i.e. conditional on the total number of observations for each adverse event total) had a P-value of 0.44 (using StatXact or procedure `NPAR1WAY` in SAS). The large P-value here partly reflects the substantial discreteness for this conditional test.

Such approaches have the potential for building power, by focusing the effect on a single parameter and single degree of freedom. This can be helpful; for instance, O'Neill (1998) pointed out that pre-market safety databases are often not sufficiently large to have much power for detecting significance for a particular adverse event. However, in practice, adverse events are probably not often uniformly more likely with a drug than a placebo. In Table 1, the sample proportion is higher for the placebo than for the drug in four of the 11 cases, including the case with the largest difference, so it is no surprise that the P-values that are reported in this subsection are not particularly small.

6. Checks of asymptotic tests, and recommendations

A limitation of the ML modelling approach is potential problems due to sparseness of the data. Sparseness can occur in the 2×2^c contingency table if it has many possible adverse events (i.e. large c), or small sample sizes or additional predictors that expand the table even further. In particular, large sample χ^2-tests are more trustworthy when based on small df-values than large df-values.

6.1. Asymptotics for score and Wald statistics

When the asymptotics are questionable for the χ^2-tests that are presented in this paper, it is sensible to use the permutation distribution of the statistic of interest. However, one should realize that the distribution is computed under the IJDs condition, as discussed in Section 4. When we are merely interested in testing SMH, the IJDs condition is narrower than the null hypothesis of interest.

To check the adequacy of the large sample asymptotics, we conducted a simulation study. We used two null joint distributions: the SMH fit and the IJD fit, for the sample distribution that generated Table 1. We used two values of c: $c = 11$, and $c = 5$ with the first five side-effects. We used sample sizes $n_1 = n_2 = 50$ and $n_1 = n_2 = 100$. Since some studies use two or three times as many subjects for the drug as for the placebo, we also considered the unbalanced case ($n_1 = 100$ and $n_2 = 50$), as well as the actual sample sizes for Table 1 ($n_1 = 146$ and $n_2 = 65$).

The theoretical asymptotic distribution for W_0 holds under IJDs, but not under solely SMH, because the covariance matrix assumes second-order homogeneity as well (unless $n_1 = n_2$). Nevertheless, we found that, overall, W_0 performs well for both IJDs and SMH although the data are quite sparse for some choices of c and (n_1, n_2). However, results for the ordinary Wald statistic W were poor. For instance, consider SMH with $c = 11$ and $(n_1, n_2) = (146, 65)$. The simulated mean for W was 13.4 and the variance was 46.7 (compared with nominal χ^2-values of 11 and 22) and, for nominal tail proportion values of 0.10, 0.05 and 0.01, the actual proportions in the tails were 0.23, 0.15 and 0.07. By contrast, for the score statistic W_0, the simulated mean was 10.9, the variance was 21.6 and the tail proportions were 0.097, 0.047 and 0.009. For this

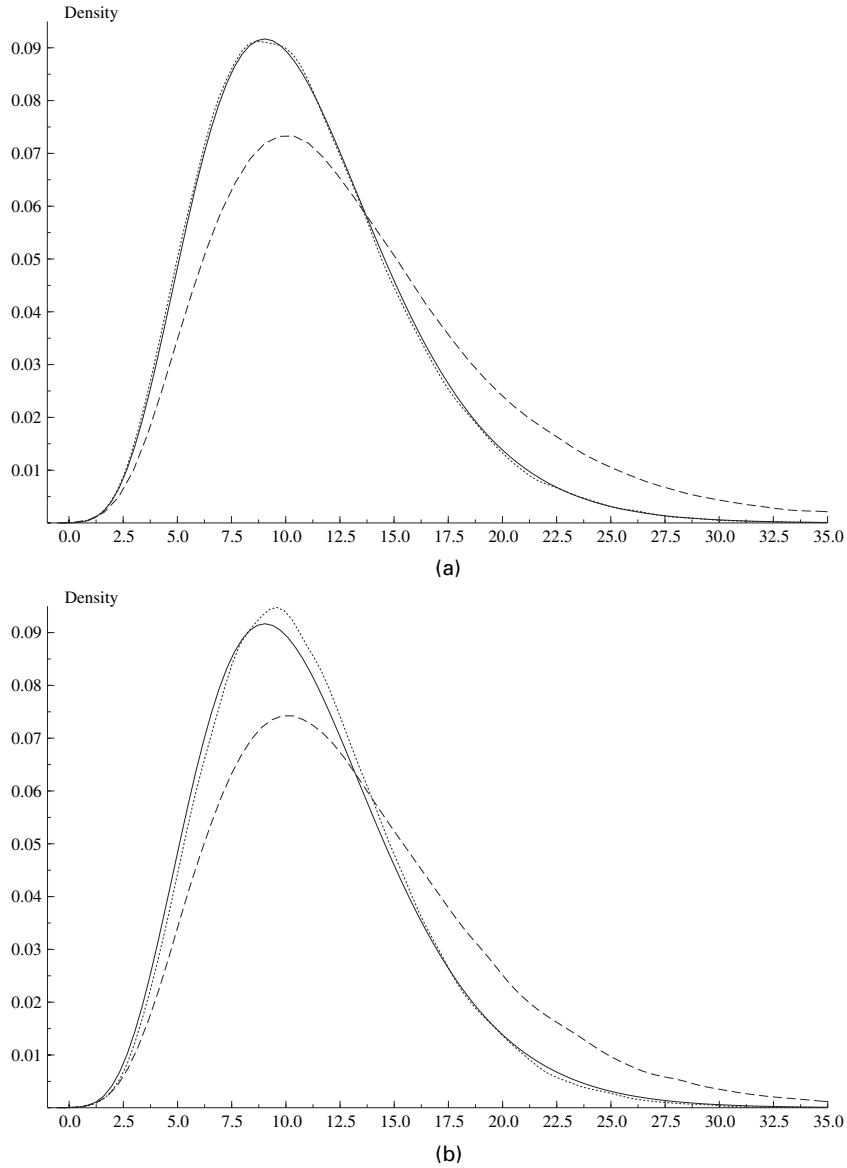

Fig. 1. Estimated probability density functions of the Wald statistic W (– – –) and score-type statistic W_0 (·····) under the assumption of (a) SMH and (b) IJDs (———, reference χ^2-density with df = 11)

case, Fig. 1 shows the simulated density functions of W and W_0 under SMH and IJDs relative to the χ^2-distribution with df $= 11$.

The tests that compare the c marginal distributions have df $= c$, unless we add further structure such as in model (7). For such tests and estimation of corresponding parameters, the sparseness seems to be relevant in terms of the marginal totals of the two possible outcomes for each adverse event, for each group. The marginal models do not have reduced sufficient statistics, but on the basis of what applies to χ^2-statistics in the univariate case it seems sensible to inspect the expected frequencies for the c separate 2×2 marginal tables comparing the two groups on

the binary response. For the sample sizes that were used in the simulation study, for the cases in which the asymptotics performed poorest, the minimum expected frequency was less than 3 and many of the $4c$ expected frequencies were below 5. It is unrealistic to expect a simple sample size guideline to cover all cases well, but a tentative suggestion is to be cautious when using this test when many marginal expected frequencies are smaller than 5.

6.2. Summary recommendations

Refer to the summary of models and tests in Table 2. Overall, we have the following recommendations. To test the IJD hypothesis, use the likelihood ratio or Pearson statistic based on the fitted values for that hypothesis, but use the permutation distribution (randomly sampled, if necessary) to obtain the P-value. To test the SMH hypothesis, use the score-type statistic W_0. We recommend W_0 over the likelihood ratio or Pearson statistic merely because we could conduct simulations to evaluate its asymptotic performance; this is computationally difficult for the ML-based statistics for testing SMH. When some marginal expected frequencies are small to moderate, we can seek corroboration by checking whether similar results apply with a bootstrap for W_0 under the fitted SMH distribution (when it is computationally feasible to obtain that fitted distribution). If results differ in a practical sense, or if many of the marginal expected frequencies are less than about 5, it is safer to use a permutation test of IJDs instead. When $c = 1$, the SMH and IJDs methods are identical, and the likelihood ratio and score-type statistics simplify to the ordinary likelihood ratio and Pearson statistics for testing independence in a 2×2 table.

7. Follow-up comparisons

We presented multivariate methods to assess the evidence of a global difference for two vectors of proportions. When the null hypotheses of SMH or IJDs are rejected, investigators are naturally interested in which specific adverse events or sets of adverse events actually caused the difference. For any given adverse event, a 2×2 table compares the counts on the two possible

Table 3. Follow-up inference for estimating differences of incidence of several adverse events in an asthma trial

Adverse event	*Sample proportion*		*z-statistic*	*Adjusted P-value*	*Bonferroni score confidence interval*
	Drug	*Placebo*			
Upper respiratory or cold	0.404	0.585	−2.43	0.124	(−0.375, 0.030)
Muscoloskeletal pain	0.123	0.231	−1.98	0.391	(−0.293, 0.042)
Throat pain	0.164	0.108	1.07	0.952	(−0.113, 0.186)
Allergic rhinitis	0.116	0.108	0.18	1.000	(−0.157, 0.130)
Fatigue	0.103	0.077	0.59	1.000	(−0.130, 0.137)
Diarrhoea	0.075	0.108	−0.78	1.000	(−0.194, 0.079)
Abdominal pain	0.096	0.046	1.23	0.937	(−0.094, 0.152)
Joint pain	0.082	0.031	1.39	0.890	(−0.084, 0.147)
Fever	0.075	0.046	0.79	1.000	(−0.113, 0.126)
Cough	0.041	0.108	−1.86	0.635	(−0.226, 0.034)
Urinary tract infection	0.062	0.031	0.93	0.997	(−0.103, 0.120)
Sample size	146	65			

outcomes for the two groups. Table 3 shows the signed square root of the Pearson statistic, which is the z-statistic for comparing two independent proportions by using the standard error based on pooling the two samples. Of the 11 z-statistics, only one has absolute value larger than 2, with one other close to 2. The Westfall and Young (1989) adjusted P-values are also shown.

More informatively, we could form simultaneous confidence intervals for a summary measure comparing the drug with the placebo for each adverse event. Table 3 illustrates by showing Bonferroni confidence intervals for the difference of proportions based on inverting the score test (Mee, 1984). This method tends to have actual confidence level nearer the nominal level than the usual Wald interval. Each of these intervals shown in Table 3 has nominal confidence coefficient of 0.99545, so asymptotically the nominal overall level is at least 0.95.

For such follow-up comparisons, it is possible for all to be non-significant, for the significant comparisons to be in a single direction (e.g. always a higher proportion for the drug) or mixed. In the last case, what can we say about the overall safety advantages of one treatment over the other? We could weight the evidence that is provided by the different adverse events, especially if some are regarded as more important than others. Let w_j denote a non-negative weight that is associated with adverse event j. For $\mathbf{w} = (w_1, w_2, \ldots, w_c)'$, $\mathbf{w}'\mathbf{d}$ is a weighted average of the differences. The global score-type statistic W_0 then generalizes to the weighted version

$$\tilde{W}_0 = (\mathbf{w}'\mathbf{d})^2/(\mathbf{w}'\hat{\boldsymbol{\Sigma}}_0\mathbf{w}),$$

with df $= 1$. For instance, investigators considered adverse events 1 and 4 in Table 1 to be more important for the asthma drug. Assigning twice as much weight to these two adverse events, we obtain $\tilde{W}_0 = 1.32$ (P-value 0.25). Such summaries also have the advantage that was mentioned in Section 5.2 of potentially building power from concentrating the effect on a single degree of freedom. Here, this approach did not result in a small P-value, as the placebo had a higher proportion for the first adverse event but the drug did for the fourth adverse event.

We could also incorporate weights in the score statistic itself, without planning to form a weighted summary. We weight the influence of difference j using the weighted difference $\tilde{d}_j = w_j d_j$. The global score-type statistic W_0 then generalizes to the weighted version $\tilde{W}_0 = \tilde{\mathbf{d}}'\tilde{\boldsymbol{\Sigma}}_0^{-1}\tilde{\mathbf{d}}$. It incorporates prior belief about the seriousness of adverse events and the magnitude of their differences between the drug and placebo. Or, as in Berry and Berry (2004), we could take into account the body system, for instance by using weights for adverse events in a common body system that are inversely proportional to the number of adverse events in it. Here, $\tilde{\boldsymbol{\Sigma}}_0$ is constructed from $\hat{\boldsymbol{\Sigma}}_0$ by simply multiplying the jth diagonal element by w_j^2 and the (j,k)th off-diagonal element by $w_j w_k$. The ordinary score-type statistic W_0 is the special case with identical $\{w_j\}$, and this statistic likewise has an asymptotic χ^2 null distribution with df $= c$. If the greater differences between the drug and placebo occur with adverse events considered more serious, this statistic may show greater significance than the ordinary score-type statistic.

8. Extensions

The methods of this paper extend in obvious ways to several groups. To test SMH with g groups and c variables, we can extend the score-type statistic W_0 by forming a vector $\mathbf{d}$ of $c(g-1)$ differences of proportions, comparing a given proportion for each group with the corresponding proportion for an arbitrary base-line group. The variances and covariances of the differences are estimated by using estimates $\{\hat{\pi}(j)\}$ and $\{\hat{\pi}(j,k)\}$ based on pooling the g samples.

The methods also extend in obvious ways to multicategory responses. For comparing g groups simultaneously on c variables, with r_j categories for variable j, the basic likelihood ratio and

score-type sorts of tests have $df = (g-1)(\Sigma_j r_j - c)$. For a single variable, these simplify to the likelihood ratio and Pearson χ^2-tests of homogeneity (or, equivalently, independence) in a two-way $g \times r$ contingency table. With even moderate g and c, asymptotic methods are suspect. A sensible strategy for testing is a permutation test for the various allocations of the subjects to the g groups, computing a relevant sample statistic for each (e.g. the extended W_0-statistic for testing SMH). With covariates, the permutation test is still feasible by using a random sample of the possible permutations, even when some covariates are continuous.

In another context, the SMH hypothesis is a special case of a hypothesis that was studied by Agresti and Liu (1999) in considering survey data in which each subject can pick any number of outcomes for a multiple-category response. See also Loughin and Scherer (1998) for a bootstrap approach for such data. For a related permutation analysis, see Berry and Mielke (2003).

As is generally true, we have seen that different tests and different test statistics for a given hypothesis can lead to quite different P-values. For the asthma data, there was no uniformity relative to the often sacred 0.05-level in terms of whether hypotheses should be rejected. This points out the importance of giving careful thought ahead of time to which is the relevant hypothesis to test (i.e. SMH or IJDs) and which statistic we prefer to summarize the effect. It also points out the ultimate advantage of focusing on the size of the effects rather than mere statistical significance. Confidence intervals based on different methods (e.g. Wald, likelihood ratio or score) can appear relatively similar in practical terms even when P-values of corresponding tests diverge somewhat.

Acknowledgements

This research was supported by grants from the National Institutes of Health and the National Science Foundation. The authors thank Dr Davis Gates at Schering-Plough Corp. for permission to use the data, Dr Joseph Lang for the use of his R function for fitting marginal models and two referees for helpful suggestions to improve the presentation.

Appendix A: Covariance matrices for Wald and score statistics

Let $\mathbf{d} = (d_1, \ldots, d_c)'$ with $d_j = \hat{\pi}_1(j) - \hat{\pi}_2(j)$, $j = 1, \ldots, c$. The vector of differences $\mathbf{d}$ has covariance matrix with elements

$$\mathrm{var}(d_j) = \pi_1(j)\{1-\pi_1(j)\}/n_1 + \pi_2(j)\{1-\pi_2(j)\}/n_2,$$

$$\begin{aligned}\mathrm{cov}(d_j, d_k) &= \mathrm{cov}\{\hat{\pi}_1(j), \hat{\pi}_1(k)\} + \mathrm{cov}\{\hat{\pi}_2(j), \hat{\pi}_2(k)\} \\ &= \{P(y_{1j}=1, y_{1k}=1) - P(y_{1j}=1)\,P(y_{1k}=1)\}/n_1 \\ &\quad + \{P(y_{2j}=1, y_{2k}=1) - P(y_{2j}=1)\,P(y_{2k}=1)\}/n_2.\end{aligned}$$

Under the null hypothesis, the variance is estimated by

$$\hat{\pi}(j)\{1-\hat{\pi}(j)\}\left(\frac{1}{n_1} + \frac{1}{n_2}\right),$$

where the pooled estimate

$$\hat{\pi}(j) = \{n_1\,\hat{\pi}_1(j) + n_2\,\hat{\pi}_2(j)\}/(n_1+n_2).$$

Under the additional assumption that the two groups have the same second-order marginal distributions, the covariance is estimated by

$$\{\hat{\pi}(j,k) - \hat{\pi}(j)\,\hat{\pi}(k)\}\left(\frac{1}{n_1} + \frac{1}{n_2}\right),$$

where $\hat{\pi}(j,k)$ denotes the sample proportion of cases that had both side-effects j and k, after the two samples have been pooled. When $n_1 = n_2$, this estimate is identical to the estimate using only pooled first-order marginal distributions, and we do not need the extra assumption.

Appendix B: A non-technical summary of the generalized estimating equation approach

The GEE approach is a multivariate version of quasi-likelihood, meaning that it specifies only the first two moments rather than a full distribution (Liang and Zeger, 1986). The model applies to the mean of the marginal distribution for each component y_{ij} of the multivariate response (such as model (2)). The method assumes a variance function corresponding to the distribution that it is natural to assume for y_{ij} marginally (such as the binomial distribution) and uses a working guess for the correlation structure among $\{y_{i1}, \ldots, y_{ic}\}$. It does this without assuming a particular multivariate distribution. The estimates are solutions of GEEs. These resemble likelihood equations but are not, since a complete multivariate distribution is not specified (in the univariate case they are likelihood equations under the additional assumptions that the distribution is the member of the exponential family that has the assumed variance function).

The GEE estimates of model parameters are valid even if we misspecify the covariance structure, i.e., when the marginal model is correct, then the GEE model parameter estimators are consistent. Standard errors result from a 'sandwich matrix' adjustment that the GEE method makes using the empirical dependence that the data exhibit. The naïve standard errors based on the working correlation assumption are updated by using the information that the data provide about the actual dependence structure to yield *robust* standard errors that are more appropriate than those based on the guessed working correlation. In theory, choosing the working correlation wisely can pay benefits of improved efficiency of estimation. However, Liang and Zeger (1986) noted that estimators based on treating the responses as independent in the working correlation structure can have surprisingly good efficiency when the actual correlation is weak to moderate.

The GEE approach is appealing because of its computational simplicity, but it has limitations. Since it does not completely specify the joint distribution, there is no likelihood function, and likelihood-based methods are not available. In addition, unless the sample size is quite large, the empirically based standard errors tend to underestimate the true standard errors (Firth, 1993).

References

Agresti, A. (2002) *Categorical Data Analysis*. New York: Wiley.

Agresti, A. and Liu, I.-M. (1999) Modeling a categorical variable allowing arbitrarily many category choices. *Biometrics*, **55**, 936–943.

Aitchison, J. and Silvey, S. D. (1958) Maximum-likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.*, **29**, 813–828.

Berry, K. J. and Mielke, Jr, P. W. (2003) Permutation analysis of data with multiple binary category choices. *Psychol. Rep.*, **92**, 91–98.

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

Chuang-Stein, C. and Mohberg, N. (1993) A unified approach to the analysis of safety data in clinical trials. In *Drug Safety Assessment in Clinical Trials* (ed. G. Sogliero-Gilbert). New York: Dekker.

Firth, D. (1993) Recent developments in quasi-likelihood methods. *Proc. 49th Sess. Int. Statist. Inst.*, 341–358.

Fitzmaurice, G. M. and Laird, N. M. (1993) A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80**, 141–151.

Haber, M. (1985) Maximum likelihood methods for linear and log-linear models in categorical data. *Comput. Statist. Data Anal.*, **3**, 1–10.

Holm, S. (1979) A simple sequential rejective multiple test procedure. *Scand. J. Statist.*, **6**, 65–70.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H. and Lehnen, R. G. (1977) A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics*, **33**, 133–158.

Lang, J. B. (2004) Multinomial-Poisson homogeneous models for contingency tables. *Ann. Statist.*, **32**, 340–383.

Lang, J. B. and Agresti, A. (1994) Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *J. Am. Statist. Ass.*, **89**, 625–632.

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

Lehmacher, W., Wassmer, G. and Reitmeir, P. (1991) Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics*, **47**, 511–521.
Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
Lin, T. C., Hosmane, B. S., Olson, P. J. and Padley, R. J. (2001) Analysis of adverse events in titration studies. *J. Statist. Planng Inf.*, **96**, 129–142.
Loughin, T. M. and Scherer, P. N. (1998) Testing for association in contingency tables with multiple column responses. *Biometrics*, **54**, 630–637.
Mantel, N. (1980) Assessing laboratory evidence for neoplastic activity. *Biometrics*, **36**, 381–399.
Mee, R. W. (1984) Confidence bounds for the difference between two probabilities. *Biometrics*, **40**, 1175–1176.
Mehrotra, D. V. and Heyse, J. F. (2004) Use of the false discovery rate for evaluating clinical safety data. *Statist. Meth. Med. Res.*, **13**, 227–238.
Miller, M. E., Davis, C. S. and Landis, J. R. (1993) The analysis of longitudinal polytomous data: generalized estimating equations and connections with weighted least squares. *Biometrics*, **49**, 1033–1044.
O'Brien, P. C. (1984) Procedures for comparing samples with multiple endpoints. *Biometrics*, **40**, 1079–1087.
O'Neill, R. T. (1988) Assessment of safety. In *Biopharmaceutical Statistics for Drug Development* (ed. K. E. Peace). New York: Dekker.
O'Neill, R. T. (2002) Regulatory perspectives on data monitoring. *Statist. Med.*, **21**, 2831–2842.
Pocock, S. J., Geller, N. L. and Tsiatis, A. A. (1987) The analysis of multiple endpoints in clinical trials. *Biometrics*, **43**, 487–498.
Rotnitzky, A. and Jewell, N. P. (1990) Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, **77**, 485–497.
Westfall, P. H. and Young, S. S. (1989) P value adjustments for multiple tests in multivariate binomial models. *J. Am. Statist. Ass.*, **84**, 780–786.
Zhang, J., Quan, H., Ng, J. and Stepanavage, M. (1997) Some statistical methods for multiple endpoints in clinical trials. *Contr. Clin. Trials*, **18**, 204–221.