STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:1115–1139

TUTORIAL IN BIOSTATISTICS
Strategies for comparing treatments on a binary response with multi-centre data

Alan Agresti*,† and Jonathan Hartzel

Department of Statistics, University of Florida, Gainesville, Florida 32611-8545, U.S.A.

SUMMARY

This paper surveys methods for comparing treatments on a binary response when observations occur for several strata. A common application is multi-centre clinical trials, in which the strata refer to a sample of centres or sites of some type. Questions of interest include how one should summarize the difference between the treatments, how one should make inferential comparisons, how one should investigate whether treatment-by-centre interaction exists, how one should describe effects when interaction exists, whether one should treat centres and centre-specific treatment effects as fixed or random, and whether centres that have either 0 successes or 0 failures should contribute to the analysis. This article discusses these matters in the context of various strategies for analysing such data, in particular focusing on special problems presented by sparse data. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

For motivation regarding the questions addressed in this paper, we begin with an example. Table I shows results of a clinical trial conducted at eight centres. The purpose was to compare two cream preparations, an active drug and a control, with respect to their success or failure in curing an infection [1]. This table illustrates a common situation in many pharmaceutical and biomedical applications – comparison of two treatments on a binary response ('success' or 'failure') when observations occur for several strata. The strata are often medical centres or clinics, or they may be levels of a control variable, such as age or severity of the condition being treated, or combinations of levels of several control variables, or they may be different studies of the same sort evaluated in a meta analysis [2–10].

Table I exhibits a potential difficulty that often occurs with multi-centre clinical trials or stratification using several control variables: the sample sizes for the treatments in many of the clinics are modest, and the corresponding cell counts are relatively small. Indeed, for the control group

*Correspondence to: Alan Agresti, Department of Statistics, University of Florida, Gainesville, Florida 32611-8545, U.S.A.
†E-mail: AA@STAT.UFL.EDU

Contract/grant sponsor: NIH
Contract/grant sponsor: NSF

Received December 1998
Accepted June 1999

Table I. Clinical trial relating treatment to response for eight centres.

Centre	Treatment	Response		Total	Per cent 'success'
		Success	Failure		
1	Drug	11	25	36	30.6
	Control	10	27	37	27.0
2	Drug	16	4	20	80.0
	Control	22	10	32	68.8
3	Drug	14	5	19	73.7
	Control	7	12	19	36.8
4	Drug	2	14	16	12.5
	Control	1	16	17	5.9
5	Drug	6	11	17	35.3
	Control	0	12	12	0.0
6	Drug	1	10	11	9.1
	Control	0	10	10	0.0
7	Drug	1	4	5	20.0
	Control	1	8	9	11.1
8	Drug	4	2	6	66.7
	Control	6	1	7	85.7
Total	Drug	55	75	130	42.3
	Control	47	96	143	32.9

Source of data: Beitler and Landis [1].

in two centres, all observations are failures. Ordinary maximum likelihood (ML) estimation can provide badly biased (even infinite) estimates of some parameters in such cases, and in certain asymptotic frameworks it can even be inconsistent. Bias also occurs, however, from combining strata to increase the stratum-specific sample sizes.

Among the questions of interest for data of this sort are the following: (i) How should one summarize, descriptively, the difference between the treatments? (ii) How should one make inferential comparisons of the treatments? (iii) How should one investigate whether there is treatment-by-centre interaction? (iv) If such interaction exists, how should one describe the effect heterogeneity? (v) Should centres be treated as fixed or random, and could that choice affect any results in a substantive way? (vi) Should centres with 0 successes or with 0 failures contribute to descriptive and inferential analyses? (vii) Should one combine centres or add small constants to empty cells in descriptive and inferential analyses, for instance to use information that otherwise is discarded in the statistical analysis?

In considering these questions, this article discusses strategies for analysing data of the form of Table I. Section 2 presents some possible models for the data and corresponding summaries of the effects. Section 3 presents ways of estimating those effects, and Section 4 illustrates the models for Table I. Section 5 discusses inferential analyses for the models. Section 6 studies the effects of severe sparseness on the analyses, using a data set that is even more sparse than Table I. Section 7 compares the strategies, makes some recommendations, and mentions extensions, alternative approaches, and open questions.

Possible analyses result from all combinations of several factors, including: (i) the choice of link function relating response probabilities to predictors in the model; (ii) whether the model

permits interaction; (iii) whether the model treats centres as random or fixed; (iv) whether inference uses a small-sample analysis or an asymptotic one with the number of centres fixed or an asymptotic one with the number of centres growing with the sample size; (v) whether one uses a Bayes or a frequentist approach or some non-likelihood-based method such as generalized estimating equations (GEE). Here, we consider only the frequentist approach and binary responses. Other papers have presented related discussion of some of these issues in the contexts of Bayesian approaches [4, 6, 11, 12] and continuous responses [13–15]. Also, we do not consider other issues of importance in actual clinical trials, such as adequacy of sample size and selection of centres.

This paper does not claim any new or surprising results, and although it is called a 'tutorial', we fully expect that many readers will have strong opinions about the appropriateness of certain methods. We hope, however, that a unified discussion of various strategies may be helpful for many biostatisticians and quantitatively-oriented medical researchers and perhaps even stimulate research on alternative approaches.

2. MODELS AND SUMMARIES OF EFFECTS

For data in the form of Table I, let X denote treatment, let Y denote the response variable, and let Z denote the stratification factor. Let $X=1$ denote the drug and $X=2$ denote the control (or placebo), and let $Y=1$ denote 'success' and $Y=2$ denote 'failure'. Let $\pi_{ik}=P(Y=1|X=i,Z=k)$, for $i=1,2,k=1,\ldots,K$. Let n_{ijk} denote the cell count for treatment i and response outcome j in stratum k. In this article we often refer to Z using the generic term 'centre', although as mentioned above it might refer to different studies or combinations of levels of control variables.

2.1. Models assuming a lack of interaction

A simple model for Table I, although usually only plausible to a rough approximation, has additive treatment and centre effects on some scale. For instance, with the logit link function (that is, log of the odds) $\text{logit}(\pi_{ik})=\log[\pi_{ik}/(1-\pi_{ik})]$, this is

$$\text{logit}(\pi_{1k})=\alpha_k+\beta/2, \quad \text{logit}(\pi_{2k})=\alpha_k-\beta/2, \ k=1,\ldots,K \tag{1}$$

That is, β is the difference between the logit for drug and the logit for control. One could include an overall intercept in this model and then use a constraint such as $\sum_k \alpha_k=0$ or $\alpha_1=0$, but we use parameterization (1) to discuss more easily (later in the paper) the effects of strata with 0 successes or with 0 failures.

This model assumes a lack of treatment-by-centre interaction. For the logit scale, β refers to a log-odds ratio, so a lack of interaction implies that the true odds ratio e^{β} between X and Y is the same in all centres. Usually primary interest focuses on estimating the treatment effect β rather than the centre effects $\{\alpha_k\}$.

When additivity exists, it need not be on the logit scale. In addition, many practitioners have difficulty interpreting the odds ratio. One could use the same predictor form with an alternative link function such as the probit or log-log or complementary log-log, although these can also be difficult to interpret. Simpler interpretations occur with the log link, by which

$$\log(\pi_{1k})=\alpha_k+\phi/2, \ \log(\pi_{2k})=\alpha_k-\phi/2 \tag{2}$$

With this model, $\exp(\phi) = \pi_{1k}/\pi_{2k}$ is a ratio of success rates, analogous to a relative risk in each centre. (Here, we use notation ϕ rather than β to reflect the effect having a different meaning than in model (1); likewise, the intercept also refers to a different scale, but we use common α_k notation for simplicity since this parameter is not the main focus of interest.)

Model (2) has the structural disadvantage of constraining $\alpha_k \pm \phi/2$ to be negative, so that π_{ik} falls between 0 and 1. Iterative methods for fitting the model may either ignore this, perhaps yielding estimates of some π_{ik} above the permissible $[0, 1]$ range, or may fail to converge if estimates at some stage violate this restriction; normally this does not happen when $\{\pi_{ik}\}$ are not near 1. This model approximates the logit model when $\{\pi_{ik}\}$ are close to 0, but it has interpretations for ratios of probabilities rather than ratios of odds. Model (2) refers to a ratio of success rates, and unlike other models considered in this subsection, when it holds it no longer applies if one interchanges the labelling of 'success' and 'failure' categories.

Simple interpretations also occur with the identity link, by which

$$\pi_{1k} = \alpha_k + \delta/2, \;\; \pi_{2k} = \alpha_k - \delta/2 \tag{3}$$

For this model, the probability of success is $\pi_{1k} - \pi_{2k} = \delta$ higher for drug than control in each centre. This model has the severe constraint that $\alpha_k \pm \delta/2$ must fall in $[0, 1]$. Iterative methods often fail for it. It is unlikely to fit well when any π_{ik} are near 0 or 1 as well as somewhat removed from those boundary values, since smaller values of $\pi_{1k} - \pi_{2k}$ typically occur near the parameter space boundary. Thus, the model has less scope than the ones with logit and log links. Even so, unless the model fits very poorly, an advantage of summarizing the effect by δ is its ease of interpretation by non-statisticians.

In summarizing association for a set of centres by a single measure such as the odds ratio or relative risk, it is preferable to use the measure that is more nearly constant across those centres. In practice, however, for sparse data it is usually difficult to establish superiority of one link function over others, especially when all $\{\pi_{ik}\}$ are close to 0. This article discusses all three of these link functions but pays greatest attention to the logit, which is the most popular one in practice.

2.2. *Random effects models*

The standard ML approach for fitting models such as (1) treats $\{\alpha_k\}$ as fixed effects. In many applications, such as multi-centre clinical trials and meta analyses, the strata are themselves a sample. When this is true and one would like inferences to apply more generally than to the strata sampled, a random effects approach may be more natural. In practice, the sample of strata are rarely randomly selected. However, Grizzle [16] expressed the belief of many statisticians when he argued that 'Although the clinics are not randomly chosen, the assumption of random clinic effect will result in tests and confidence intervals that better capture the variability inherent in the system more realistically than when clinic effects are considered fixed'. This approach seems reasonable to us for many applications of this type.

For the logit link, a logit-normal random effects model [17] with the same form as (1)

$$\text{logit}(\pi_{1k}) = a_k + \beta/2, \quad \text{logit}(\pi_{2k}) = a_k - \beta/2 \tag{4}$$

assumes that $\{a_k\}$ are independent from a $N(\alpha, \sigma)$ distribution. The parameter σ, itself unknown, summarizes centre heterogeneity in the success probabilities. This model also makes the strong assumption that the treatment effect β is constant over strata.

For binary data, random effects models are most commonly used with logit or probit link functions. A structural defect exists with the log and identity links in treating $\{a_k\}$ as normally distributed; for any parameter values with $\sigma > 0$, with positive probability a particular realization of the random effect corresponds to π_{ik} outside $[0, 1]$.

2.3. Treatment-by-centre interaction

Even if a model that is additive in centre and treatment effects fits sample data adequately, it is usually unrealistic to expect the *true* association to be identical (or essentially identical) in each stratum. This subsection considers models that permit interaction. With a fixed-effects approach, the model

$$\text{logit}(\pi_{1k}) = \alpha_k + \beta_k/2, \quad \text{logit}(\pi_{2k}) = \alpha_k - \beta_k/2 \tag{5}$$

has odds ratio e^{β_k} in centre k. It is saturated (residual d.f. $= 0$), having $2K$ parameters for the $2K$ binomial probabilities. The ML estimate of β_k is the sample log-odds ratio in stratum k, $\hat{\beta}_k = \log(n_{11k}n_{22k}/n_{12k}n_{21k})$.

Usually, such as in meta analyses, one would want to extend such a model to determine explanations for the variability in associations among the strata. When the strata have a natural ordering with scores $\{z_k\}$, an unsaturated model (d.f. $= K - 2$) results from assuming a linear trend in the log-odds ratios; that is, by replacing β_k in model (5) by $\beta + z_k\lambda$. Often other explanatory variables are available for modelling the odds ratio [18–20]. Then, one could construct a model of form

$$\beta_k = \mathbf{z}_k'\boldsymbol{\lambda}$$

describing the centre-specific log-odds ratios, where $\mathbf{z}_k$ is a column vector of explanatory variables and $\boldsymbol{\lambda}$ is a column vector of parameters. A related model adds a random effect term for each centre to reflect unexplained variability [21].

For the random effects approach without other explanatory variables, an additional parameter can represent variability in the true effects. The logit-normal model is

$$\text{logit}(\pi_{1k}) = a_k + b_k/2, \quad \text{logit}(\pi_{2k}) = a_k - b_k/2 \tag{6}$$

where $\{a_k\}$ are independent from $N(\alpha, \sigma_a)$, $\{b_k\}$ are independent from $N(\beta, \sigma_b)$, and $\{a_k\}$ are independent of $\{b_k\}$. Here, β is the expected value of centre-specific log-odds ratios, and σ_b describes their variability. An equivalent model form is $\text{logit}(\pi_{ik}) = a_k + \beta x_i + b_{ik}$, where x_i is a treatment dummy variable ($x_1 = 1, x_2 = 0$) and b_{1k} and b_{2k} are independent $N(0, \sigma)$, where σ^2 corresponds to $\sigma_b^2/2$ in parameterization (6). Note that one should not formulate the model as $\text{logit}(\pi_{ik}) = a_k + b_k x_i$, since the model then imposes greater variability on the logit for the first treatment unless one permits (a_k, b_k) to be correlated.

Analogous random effects models apply with alternative link functions. Again, the models with identity or log link are structurally improper when either variance component is positive. This suggests a caution, as results reported for software using a particular estimation method may depend on whether the parameter constraints are recognized. In our experience, the identity link often has convergence problems. Good initial estimates of $(\alpha, \beta, \sigma_a, \sigma_b)$ can be helpful, such as using values suggested by fixed effects modelling. In some applications it is also sensible to let (a_k, b_k) be correlated, by treating it as a bivariate normal random effect [22]. With the identity link, for instance, centres with a_k close to 0 may tend to have values of b_k relatively close to

0. We do not discuss such models here, as such modelling is better supported with moderate to large K, and our examples have relatively small K with sparse data. With such examples, some will think it bold or foolhardy of us to use even relatively simple random effects models!

3. MODEL FITTING AND ESTIMATING EFFECTS

We now discuss model fitting and parameter estimation. Unless stated otherwise, the discussion refers to the logit models.

3.1. Model fitting

It is straightforward to fit the fixed effects models with standard software. Possibilities include software for binary responses such as PROC LOGISTIC in SAS, or software for generalized linear models such as PROC GENMOD in SAS and the *glm* function in S-plus.

Random effects models for binary data are more difficult to fit. One must integrate the joint mass function of the responses with respect to the random effects distributions to obtain the likelihood function [23], which is a function of β and the other parameters of those distributions. With the logit interaction model (6), for instance, the likelihood function equals

$$\ell(\alpha, \beta, \sigma_a, \sigma_b) = \Pi_k \Pi_i \left[\int_{a_k} \int_{b_k} \pi_{ik}^{n_{i1k}} (1 - \pi_{ik})^{n_{i2k}} \, \mathrm{d}G(b_k) \, \mathrm{d}F(a_k) \right]$$

where F is a $\mathrm{N}(\alpha, \sigma_a)$ CDF, G is a $\mathrm{N}(\beta, \sigma_b)$ CDF, $\pi_{1k} = \exp(a_k + b_k/2)/[1 + \exp(a_k + b_k/2)]$, and $\pi_{2k} = \exp(a_k - b_k/2)/[1 + \exp(a_k - b_k/2)]$. One can approximate the likelihood function using numerical integration methods, such as Gauss–Hermite quadrature. The approximation improves as the number of quadrature points q increases, more points being needed as the variance components increase in size. Performance is enhanced by an adaptive version of quadrature that transforms the variable of integration so that the integrand is sampled in an appropriate region [24, 25]. Having approximated the likelihood, one can use standard maximization methods such as Newton–Raphson to obtain the estimates. As a by-product, the observed information matrix, based on the curvature (second derivatives) of the log-likelihood at the ML estimates, is inverted to provide an estimated asymptotic covariance matrix.

Other approximations for integrating out the random effects lead to related approximations of the likelihood function and the ML estimates. Most of these utilize linearizations of the model. A Laplace approximation yields penalized quasi-likelihood (PQL) estimates [26], and a related generalization includes an extra scale parameter [27]. These approximations can behave poorly when variance components are large or when distributions are far from normal, such as Bernoulli or binomial with small indices at each setting of predictors [25, 26, 28]. When feasible, it is better to use adaptive Gauss–Hermite quadrature with sufficiently large q, the determination of 'sufficiently large' being based on monitoring the convergence of estimates and standard errors as q increases. Other promising ML approximations use Monte Carlo approximation methods [28, 29], for which the approximation error is estimable and decreases as the number of simulations increases. One can also use Markov chain Monte Carlo methods with an approximating Bayes model that uses flat prior distributions for the other parameters [30], although the danger exists of improper posterior distributions [31–33].

Most major software packages are not equipped to fit generalized linear models with random effects. Version 7 of SAS includes PROC NLMIXED, which can provide a good approximation to ML using adaptive Gauss–Hermite quadrature. The linearization approximations [26, 27] are available in earlier versions with a SAS macro, called GLIMMIX, that uses iterative calling of PROC MIXED. Most other specialized programs for hierarchical models with random effects likewise use various normal approximations to the working response in the mixed logit model.

3.2. The sparse asymptotic framework

In many applications, such as when the strata are centres, asymptotic arguments for increasing the sample size most naturally refer to increasing simultaneously the number of strata, K. A disadvantage then of the usual large-sample methods with the fixed effects logit models is that they are based on $n \to \infty$ with a *fixed* number of parameters (for example, K fixed), whereas the more appropriate 'sparse asymptotic' framework has $K \to \infty$ as $n \to \infty$. For sparse asymptotics, consistency of ordinary ML estimators breaks down for the odds ratio, relative risk, and difference of proportions [34]. An extreme case (Anderson [35], p. 244) occurs with matched-pairs data (two observations for each k), in which case the ordinary ML estimator of β in model (1) converges in probability to 2β.

The sparse asymptotic framework does not cause special problems for the random effects approach. After integrating out the random effects, the likelihood function depends only on the remaining parameters (for example, α, β and σ in model (4)), so the parameter space does not increase as K does. In particular, if the random effects model holds, the ordinary ML estimator of β is consistent. In practice, however, if n and K have only moderate size, as in Table I, inferences about the size of the variance components may be very imprecise.

In the logit fixed effects model (1), the conditional likelihood approach provides an alternative way of guaranteeing a consistent estimator of β. With it, one eliminates $\{\alpha_k\}$ in constructing the likelihood function by conditioning on their sufficient statistics [36]. Software is available for this approach, such as LogXact [37]. It has the advantage of not requiring a distributional assumption about the random effects yet still being valid for sparse asymptotics. A disadvantage of conditional ML is that the fitting procedure does not provide predicted values for $\{\alpha_k\}$ or an estimate of their variability. Also, this approach is applicable only with the logit link (that is, only the canonical link of a generalized linear model provides reduced sufficient statistics).

3.3. Mantel–Haenszel type estimators of common effects

An alternative estimator of β in the no interaction model (1) is the Mantel–Haenszel (M–H) estimator [38]

$$\hat{\beta}_{\mathrm{MH}} = \log\left(\frac{\sum_k n_{11k} n_{22k}/n_{++k}}{\sum_k n_{12k} n_{21k}/n_{++k}}\right) \tag{7}$$

Like the conditional ML estimator, it is consistent both in sparse-stratum (K increases with n) or large-stratum (K fixed but n increases) asymptotics. It has the advantage over conditional ML of simplicity. It suffers no efficiency loss when $\beta = 0$ and usually little otherwise.

Mantel–Haenszel type estimators are also available for the relative risk and the difference of proportions. As noted in Section 2, models for these parameters have severe parameter restrictions.

Even if the model holds only approximately, however, a summary measure of this type is useful for communication with scientists who are unfamiliar with odds ratios. The M–H type estimator of a common log relative risk [39, 40] (that is, ϕ in model (2)) is

$$\hat{\phi}_{\mathrm{MH}} = \log\left(\frac{\sum_k n_{11k}n_{2+k}/n_{++k}}{\sum_k n_{21k}n_{1+k}/n_{++k}}\right) \tag{8}$$

whereas the M–H type estimator of a common difference of proportions [34] (that is, δ in model (3)) is

$$\hat{\delta}_{\mathrm{MH}} = \frac{\sum_k (n_{11k}n_{2+k}/n_{++k} - n_{21k}n_{1+k}/n_{++k})}{\sum_k n_{1+k}n_{2+k}/n_{++k}} \tag{9}$$

If a no interaction model fits adequately but the data are highly sparse, the corresponding M–H estimator may even be preferred to the ML estimator, because of the bias that exists in sparse asymptotics [34] for the ML estimator. The conditional ML approach does not apply to the log and identity link functions and the random effects model has structural problems (for example, probabilities outside the [0, 1] interval), so these estimates are particularly useful for these link functions.

Given their good performance under sparse asymptotics and their ease of computation, one might consider always using M–H instead of ML estimators. However, for large-stratum asymptotics (fixed K), M–H estimators lose some efficiency compared to ML, and the efficiency loss can be considerable for $\hat{\phi}_{\mathrm{MH}}$ and $\hat{\delta}_{\mathrm{MH}}$ in some cases [34]. Moreover, software for ML estimation is widely available for fixed effects analyses and becoming more so for random effects analyses. Thus, if the data have moderate to large samples in each stratum, it is better to use the model-based ML estimators.

3.4. Centre estimates

In most applications, main interest focuses on the treatment effect and its variability among centres. However, centre estimates also result from the fixed effects or random effects ML approaches. With the random effects approach, the expected values of $\{a_k\}$ given the data are analogues of best linear unbiased predictors (BLUP) for mixed models with normal responses. These expected values themselves depend on unknown parameters, so one obtains the predicted values by plugging in the ML estimates of those parameters. Ordinary standard errors of these predictors, like those of empirical Bayes estimators, do not take into account that the variance component is estimated rather than known; hence, they tend to be too small, and adjustments are available [41, 42]. Adjustments are also available to help account for the bias in estimating the variance components [43], which can be considerable, but we shall not address that issue here.

For fixed effects logit models, the sufficient statistic for α_k is n_{+1k}, conditional on the binomial sample sizes in that stratum. By contrast, for the random effects models estimates of centre effects 'borrow from the whole', and the estimate of a_k can be considerably affected by results in other strata. As the sample size grows in stratum k, however, the influence of other strata decreases.

3.5. Logit model: Allowing interaction

For the random effects model (6) that permits interaction, the complexity of model fitting is compounded by estimating two variance components. When the data are sparse but do contain sufficient

information to provide estimates of $(\alpha, \beta, \sigma_a, \sigma_b)$, the estimated average effects $(\hat{\alpha}, \hat{\beta})$ are more reliable than the estimated variability $(\hat{\sigma}_a, \hat{\sigma}_b)$ of effects, especially when K is not especially large. When $\hat{\sigma}_b > 0$, the standard error of $\hat{\beta}$ is typically larger than with the model (4) of homogeneous odds ratios (that is, the special case in which $\sigma_b = 0$), because of the extra variance component due to treating the treatment effect as random rather than fixed.

Liu and Pierce [44] proposed an alternative way of estimating (β, σ_b) for the model (6) that assumes the log-odds ratios are a $N(\beta, \sigma_b)$ random sample. They first eliminated $\{a_k\}$ by a conditioning argument, focusing solely on the variability in association, and then provided a simple solution based on an approximation to the likelihood function using Laplace's method. They suggested that their method is primarily intended for cases in which cell counts are relatively large and the variability σ_b is not great, say, $\sigma_b < 1$. See Raghunathan and Ii [45] and Liang and Self [46] for related work.

4. MODEL FITTING FOR TABLE I

We now apply these methods to Table I. For these data the sample success rates vary markedly among centres both for the control and drug treatments, but in all except the last centre that rate is higher for drug. Normally in using models with random centre and possibly random treatment effects, one would prefer to have more than $K = 8$ centres; keeping in mind the difficulty particularly of getting good variance component estimates with such a small value of K, we use these data to illustrate the models. Table II shows the use of SAS (PROC NLMIXED and PROC GENMOD) for ML fitting of logit models to Table I. Alternative link functions utilize similar statements. For the random effects interaction model, for instance, the code `pi=exp(a+b*treat)` requests the log link model and `pi=a+b*treat` requests the identity link model. In the NLMIXED code in Table II for the no interaction model with random centre effects, the 'predict' option requests the logit estimates of $a_k \pm \beta/2$ for the eight centres and stores them in the data set OUT1.

Table III summarizes results of estimating the treatment effect β using various logit models. The parameter β is the common log-odds ratio for the no interaction models and the expected value of the log-odds ratio for the interaction model with random treatment effects. For the random effects model (6) permitting interaction, the estimated standard deviation of the log-odds ratios is relatively small, $\hat{\sigma}_b = 0.15$ (standard error $= 1.1$). For all approaches, estimates of the common log-odds ratio or its expected value are similar. In each case the estimated value of about 0.75 equals about 2.5 standard errors; this corresponds to an estimated common odds ratio of about $e^{0.75} = 2.1$ and a 95 per cent confidence interval for the common odds ratio of about (1.2, 3.8). There is considerable evidence of a drug effect, but with such a small sample one cannot determine whether that effect is weak or moderate.

For the interaction model, since $\hat{\sigma}_b$ is small, the random effects model provides a considerable smoothing of the sample odds ratios. Table IV shows the eight sample odds ratios and their random effects model estimates, computed by exponentiating the estimated expected log-odds ratios given the sample data. The smoothed estimates show considerably less variability and do not have the same ordering as the sample values. For instance, the smoothed estimate is greater for centre 3 than for centre 6 even though the sample value is infinite for the latter, partly reflecting the greater shrinkage that occurs when sample sizes are smaller. When $\hat{\sigma}_b = 0$, the interaction model provides the same fit as the no interaction model, so the model estimated odds ratios are identical in each centre.

Table II. Example of SAS code for using GENMOD to fit fixed effects logit model and NLMIXED to fit random effects logit models to Table I.

```
data binomial;
input center treat y n @@ ; * y successes out of n trials;
if treat=1 then treat=.5; else treat=-.5;
cards;
1 1 11 36    1 0 10 37    2 1 16 20    2 0 22 32
3 1 14 19    3 0  7 19    4 1  2 16    4 0  1 17
5 1  6 17    5 0  0 12    6 1  1 11    6 0  0 10
7 1  1  5    7 0  1  9    8 1  4  6    8 0  6  7
;
run;

proc genmod data=binomial; * fixed effects, no interaction model;
  class center;
  model y/n=treat center / dist=bin link=logit noint;
run;

proc nlmixed data=binomial qpoints=15; * random effects, no interaction;
  parms alpha=-1 beta=1 sig=1; * initial values for parameter estimates;
  pi=exp(a + beta*treat)/(1+exp(a + beta*treat)); * logistic formula for prob;
  model y ~ binomial(n, pi);
  random a ~ normal(alpha, sig*sig) subject=center;
  predict a + beta*treat out=OUT1;
run;

proc nlmixed data=binomial qpoints=15; * random effects, interaction;
  parms alpha=-1 beta=1 sig_a=1 sig_b=1; * initial values;
  pi=exp(a + b*treat)/(1+exp(a + b*treat));
  model y ~ binomial(n, pi);
  random a b ~ normal([alpha,beta], [sig_a*sig_a,0,sig_b*sig_b]) subject=center;
run;
```

Table III also summarizes estimates for other descriptive measures, with ML results obtained using GENMOD and NLMIXED in SAS. As noted before, the restricted parameter space for the log and identity links can provide problems. Having good starting values increases the chance of proper convergence. We used starting values near the estimates obtained with the SAS GLIMMIX macro.

For the random effects interaction model with the log link, the ML estimated standard deviation of the log relative risks equals 0. Hence, the fitted relative risks are the same in each centre, the estimate of 1.27 being identical to that for the random effects no interaction model. For this sample the association is more nearly constant for the relative risk than the odds ratio.

For the random effects interaction model with the identity link, we were unable to obtain convergence with NLMIXED. Using GLIMMIX, Littell *et al.* [47] reported an estimated mean of 0.120 (standard error = 0.051) and an estimated standard deviation of 0.098 for the clinic-specific differences of proportions, but we could not obtain these results even with GLIMMIX. A weighted least squares estimate of the clinic-specific difference of proportions [8] is 0.131 (standard error = 0.052) with an estimated standard deviation of 0.075.

Table III. Estimated treatment effect and standard error, and results of likelihood ratio (LR) and Wald tests of hypothesis of no treatment effect, for Table I.

Measure (Equation number)	Interaction	Centre	Method	Estimate	Standard error	Wald statistic	LR statistic	*P*-value
Odds ratio (1)	No	Fixed	ML	0.777 (2.2)	0.307	6.4	6.7	0.01
			Cond. ML	0.756 (2.1)	0.303	6.2		
			M–H	0.758 (2.1)	0.304	6.2		
(4)		Random	ML	0.739 (2.1)	0.300	6.0	6.3	0.01
(6)	Yes	Random	ML	0.746 (2.1)	0.325	5.3	4.6	0.03
Relative risk (2)	No	Fixed	ML	0.247 (1.3)	0.126	3.8	3.9	0.05
			M–H	0.354 (1.4)	0.142	6.2		
		Random	ML	0.241 (1.3)	0.126	3.6	3.8	0.05
	Yes	Random	ML	0.241 (1.3)	0.126	3.6	3.7	0.08
Difference of prop. (3)	No	Fixed	ML	0.137	0.055	6.2	6.6	0.01
			M–H	0.130	0.050	6.7		
		Random	ML	0.148	0.055	7.2	7.6	0.01

Odds ratio and relative risk estimates appear in parentheses next to their log estimates. Wald and LR test statistics have approximate null chi-squared distributions with d.f. = 1; *P*-value refers to LR statistic.

Table IV. Estimated centre-specific odds ratio and relative risk for Table I, based on sample and on predictions for random effects interaction models.

Centre	Odds ratio		Relative risk	
	Sample	Model	Sample	Model
1	1.19	2.02	1.13	1.27
2	1.82	2.09	1.16	1.27
3	4.80	2.19	2.00	1.27
4	2.29	2.11	2.13	1.27
5	∞	2.18	∞	1.27
6	∞	2.12	∞	1.27
7	2.00	2.11	1.80	1.27
8	0.33	2.06	0.78	1.27

5. INFERENCE ABOUT EFFECTS

5.1. Inference for logit models

For the fixed and random effects logit models, standard methods yield inferences about the treatment effect. For instance, the likelihood-ratio test statistic is minus twice the difference in maximized log-likelihoods between model (1) or (4) with $\beta = 0$ and the model with unrestricted β. It has a null chi-squared distribution with d.f. = 1, as does the Wald statistic, which is the squared ratio of the estimate to its standard error. The standard error is obtained from the inverse information matrix. The simple Wald form of 95 per cent confidence interval for the common odds ratio is obtained by exponentiating the endpoints of $\hat{\beta} \pm 1.96$(standard error). Better, one could construct a profile likelihood confidence interval (for example, for the fixed effects solution using the LRCI option in PROC GENMOD) or an interval based on inverting a score test [48].

With the fixed effects approach, for highly sparse data it is preferable to conduct inference using the conditional likelihood. For model (1), this likelihood depends only on β. A 95 per cent large-sample likelihood-based confidence interval consists of all β values for which minus two times the log-likelihood falls within 3.84 (the 95th percentile of the χ^2_1 distribution) of the maximum. Tests and confidence intervals with this approach are available with LogXact.

5.2. Mantel–Haenszel inference

For model (1), the Mantel–Haenszel estimator (7) of a common log-odds ratio has a standard error estimate [49] that is valid for both large-stratum and sparse-stratum asymptotics. The variance estimate equals

$$\widehat{\text{var}}(\hat{\beta}_{\text{MH}}) = \frac{\sum_k (n_{11k}+n_{22k})(n_{11k}n_{22k})/n_{++k}^2}{2(\sum_k n_{11k}n_{22k}/n_{++k})^2} + \frac{\sum_k (n_{12k}+n_{21k})(n_{12k}n_{21k})/n_{++k}^2}{2(\sum_k n_{12k}n_{21k}/n_{++k})^2} + \frac{\sum_k [(n_{11k}+n_{22k})(n_{12k}n_{21k}) + (n_{12k}+n_{21k})(n_{11k}n_{22k})]/n_{++k}^2}{2(\sum_k n_{11k}n_{22k}/n_{++k})(\sum_k n_{12k}n_{21k}/n_{++k})}$$

One can use this to form a confidence interval for the common log-odds ratio, exponentiating endpoints to obtain the interval for the odds ratio. Like the conditional ML approach, this is preferred over ordinary intervals for the fixed-effects logit model (1) when the data are highly sparse.

Similarly, estimated variances for both types of asymptotics are available for the estimator of a common difference of proportions (9) and the estimator of a common log relative risk (8). Let $R_k = n_{11k}n_{2+k}/n_{++k}$ and $S_k = n_{21k}n_{1+k}/n_{++k}$. For the log relative risk, the estimated variance is [34]

$$\widehat{\text{var}}(\hat{\phi}_{\text{MH}}) = \frac{\sum_k (n_{1+k}n_{2+k}n_{+1k} - n_{11k}n_{21k}n_{++k})/n_{++k}^2)}{(\sum_k R_k)(\sum_k S_k)} \tag{10}$$

For the difference in proportions, the estimated variance is [50]

$$\widehat{\text{var}}(\hat{\delta}_{\text{MH}}) = \frac{\hat{\delta}_{\text{MH}}(\sum_k P_k) + (\sum_k Q_k)}{(\sum_k n_{1+k}n_{2+k}/n_{++k})^2} \tag{11}$$

where

$$P_k = [n_{1+k}^2 n_{21k} - n_{2+k}^2 n_{11k} + n_{1+k}n_{2+k}(n_{2+k} - n_{1+k})/2]/n_{++k}^2$$

and

$$Q_k = [n_{11k}n_{22k} + n_{21k}n_{12k}]/2n_{++k}$$

A disadvantage of these inferences is their restriction to the no interaction models and their treatment effects. Since the variance formulae assume a common treatment effect for each centre, they should not be used when substantial heterogeneity exists.

5.3. Tests of no interaction

A test of no interaction for the fixed effects logit model is equivalently a goodness-of-fit test of model (1) and a test for equality of the K true odds ratios. When the data are not sparse and K is fixed, one can use ordinary likelihood-ratio and Pearson chi-squared statistics for this purpose, with

d.f. $= (K-1)$. The likelihood-ratio statistic refers to the likelihood-ratio test comparing the model (1) to the saturated model. An alternative chi-squared test provided by some software for this case is the Breslow–Day test [36, 51], which is based on heterogeneity in the K sample log-odds ratios. For this situation with K fixed, an exact conditional test [52] of equality of odds ratios is available in StatXact [53].

When the stratum-specific sample sizes are small and K is large, none of these tests has much power. It may be possible to increase power by checking for a particular type of interaction, such as a linear trend in the log-odds ratios when the strata have a natural ordering. For the fixed effects approach, a benefit of using the simpler model when the degree of interaction is not significant is that the common odds ratio estimator can be a better estimator of the true stratum-specific odds ratios than the separate sample values (for example, having smaller total mean squared error) even when those true odds ratios are not identical, for the usual reasons of model parsimony.

For the random effects approach, one can test for a lack of interaction by testing that $\sigma_b = 0$ in model (6). The score test with an arbitrary mixture distribution for the random effect leads to an asymptotically normal statistic [46]. Under the null, the likelihood-ratio statistic equals 0 (that is, because $\hat{\sigma}_b = 0$) or approximately a χ^2_1 variate, each with probability about 0.5; thus, the usual chi-squared right-tail probability is halved to get the P-value. However, for random effects models, one might question the entire enterprise of conducting tests of no interaction. Typically the likelihood reveals that values of $\sigma_b > 0$ are consistent with the data, and when $\hat{\sigma}_b > 0$ the confidence interval for β with the interaction model is somewhat wider than with the no interaction model, better reflecting the actual heterogeneity that ordinarily occurs in practice.

5.4. Summarizing effects when interaction exists

When significant interaction exists, with the fixed effects approach the saturated model provides an odds ratio estimate for each stratum. Alternatively, a covariate may be apparent such that odds ratios are more nearly constant after adjusting for that covariate. For instance, there may be one or two centres that are considerably different from the others in some way. With the random effects approach (6), it is natural to describe the interaction by $(\hat{\beta}, \hat{\sigma}_b)$, providing an estimate of an average log-odds ratio and the variability about that average. With the random effects model, one can also obtain approximate BLUP estimates of the log-odds ratios $\{b_k\}$. These provide a smoothing of the sample log-odds ratio estimates from the fixed-effects saturated model. As the sample size increases within a particular stratum, the random effects estimate becomes more similar to the sample value for that stratum.

Similar remarks regarding interaction apply for analyses involving the difference of proportions and relative risk. For example, for fixed K, large-sample d.f. $= K - 1$ chi-squared tests exist of whether the difference of proportions is the same for all strata [8, 48]. A corresponding test holds for sparse data with K large [3]. When interaction exists with a fixed effects model with parameter θ_k in stratum k, an alternative [54–56] to simply reporting the stratum-specific estimates is to estimate $\sum_k \rho_k \theta_k$, where ρ_k is the population proportion classified in stratum k (or if this is unknown, simply $\rho_k = 1/K$).

For the random effects interaction model with identity link, alternative estimates exist of the mean and variance of the stratum-specific differences of proportions [1, 8]. These approaches weight the sample estimate from each stratum inversely proportional to its estimated variance. A modified approach uses an alternative weighting scheme to reduce bias [57]. An analogous random effects analysis exists for the relative risk [9].

5.5. *Goodness-of-fit*

We mentioned above the goodness-of-fit test for the fixed effects models. These treat K as fixed. For sparse data with large K, these tests lack power and may be poorly approximated by chi-squared distributions. Model checking is very challenging with highly sparse data.

The random effects model (4) assuming no interaction also satisfies the fixed effects structure (1). So, lack of fit in the ordinary goodness-of-fit test for model (1) also implies lack of fit in the random effects model. When the random effects model holds, its fit behaves asymptotically like that of the fixed effects model, for fixed K with $n \to \infty$. Similarly, for the models permitting interaction with fixed K, the fixed and random effects estimates are asymptotically equivalent. It is not obvious how to check the fit of such models for sparse asymptotics in which $K \to \infty$. The usual goodness-of-fit statistics are then approximately normal [58], and there is some evidence that the jack-knife can work well in estimating asymptotic variances of such statistics [59]. We are unaware, however, of any checks on this yet for models of the type discussed in this article.

5.6. *Inferential results for Table I*

Table III also shows standard errors for the various estimators and the results of Wald and likelihood-ratio tests of no effect. Substantive results are similar with all link functions, with evidence of a better success rate with drug than with control, although the model-based inferences with the relative risk provide slightly less evidence of association. The estimated effect can be described by a stratum-specific odds ratio of about 2.1, relative risk of about 1.3, or difference of proportions of about 0.14. For each measure the data do not contradict the models that assume a lack of interaction; for instance, the interaction models provide similar summary estimates and standard errors. Also, the traditional goodness-of-fit statistics do not show lack of fit when applied to the fixed effects versions of the no interaction models. The Pearson statistic equals 8.0 for the logit link, 9.9 for the log link, and 9.9 for the identity link, each with d.f. $= 7$.

6. EFFECTS OF SEVERE SPARSENESS

This section summarizes some special considerations and results when the data are severely sparse, such as effects of centres containing certain patterns of empty cells and effects of modifying the data such as by adding constants to empty cells or combining centres. Table V is an example of such data [60]. This table was shown to the first author a few years back by an attendee of a short course on categorical data analysis. It shows results for five centres of a clinical trial designed to compare an active drug to placebo in treating toenail fungal infections. Again, success rates vary markedly among centres, but note that the binomial sample sizes are very small. Here, two centres have no successes and one centre has only one success. Although one cannot expect to conduct precise inference with such small n and K and although normally K would be much larger than 5 in the application of random effects models (especially to estimate variance components), these data are useful for illustrating effects of such severe sparseness.

Here, a reasonable asymptotic framework is the sparse one whereby K increases proportionally to n. When n is small, it is difficult to detect when heterogeneity truly exists among strata in the treatment effects. Thus, our remarks are directed primarily toward models such as (1) and (4), that is, we assume that reality is reasonably well described by the fixed effects or random effects model with homogeneous odds ratios.

Table V. Clinical trial relating treatment to response for five centres.

Centre	Treatment	Response		Total	Per cent 'success'
		Success	Failure		
1	Active drug	0	5	5	0.0
	Placebo	0	9	9	0.0
2	Active drug	1	12	13	7.7
	Placebo	0	10	10	0.0
3	Active drug	0	7	7	0.0
	Placebo	0	5	5	0.0
4	Active drug	6	3	9	66.7
	Placebo	2	6	8	25.0
5	Active drug	5	9	14	35.7
	Placebo	2	12	14	14.3
Total	Active drug	12	36	48	25.0
	Placebo	4	42	46	8.7

Source: Agresti [60], p. 193.

For severely sparse data, the strata sample sizes are very small and using ordinary ML with the fixed effects model may provide seriously biased estimates. If that approach is used, it is safest to do so using conditional ML estimation.

6.1. *Extreme cases: centres with 0 successes or 0 failures*

For stratum k, let $s_k = n_{11k} + n_{21k}$ denote the number of successes and let $f_k = n_{12k} + n_{22k}$ denote the number of failures. First, we study the effects on the analyses of strata that have either $s_k = 0$ or $f_k = 0$, such as centres 1 and 3 of Table V.

Consider fixed effects models relating to the odds ratio. Then, ML estimates exist only in the extended sense that $\hat{\alpha}_k = -\infty$ when $s_k = 0$ and $\hat{\alpha}_k = \infty$ when $f_k = 0$. The likelihood approaches its maximum in the limit as these estimates grow unboundedly in the appropriate direction and $\hat{\beta}$ and $\{\hat{\alpha}_k\}$ for strata with $\min(s_k, f_k) > 0$ take the finite values the ML estimates assume after deleting the offending strata from the data set. Although $\hat{\alpha}_k$ is infinite when $\min(s_k, f_k) = 0$, in practice it is common for software to be fooled by the very flat log-likelihood and converge, reporting large centre estimates. The reported standard errors for such strata are huge, since they are based on inverting a matrix that summarizes the curvature of the log-likelihood at convergence.

In any case, for logit model (1), centres with $s_k = 0$ or $f_k = 0$ have no effect on $\hat{\beta}$. Similarly, the conditional likelihood approach to fitting model (1) ignores strata with $s_k = 0$ or $f_k = 0$, as does the M–H estimate and its standard error. When one conditions on row and column totals, the observed counts in the stratum are the only ones possible, and the distribution is degenerate; for instance, conditionally, the count in the first cell equals the observed value with probability 1, and the variance of the distribution of that count is 0. Similarly, the M–H test statistic [38] for testing that the stratified treatment effect is null, which was originally derived for such conditional distributions, is unaffected by such tables.

Next, consider the random effects approach. Since it borrows from the whole, one obtains a finite estimate of a_k even when $s_k = 0$ or $f_k = 0$. Strata with $s_k = 0$ or $f_k = 0$ are relevant for the random effects model (4) also in terms of estimating the variance σ^2 of the centre estimates.

Deleting such a stratum usually has a decreasing effect on $\hat{\sigma}$, since the remaining strata show less variability in their overall success rates. Certainly, one would want to utilize data from all the centres if one were interested in estimating centre variability or individual centre effects $\{a_k\}$. Normally such tables have little effect on $\hat{\beta}$ or inference about β, an exception being mentioned below. Similar comments apply to random effects models for the relative risk or difference of proportions.

For inference with the relative risk or the difference in proportions, we next study analyses based on M–H estimators, for which effects of 0 column totals in strata are clear from the relevant formulae. The relative risk estimator (8) is unaffected by strata with $s_k = 0$, but strata with $f_k = 0$ provide a shrinkage toward 1.0. This is sensible, since when the two sample proportions of success fall within a small $\varepsilon > 0$ of 0, the sample relative risk can be any non-negative value, but when the two sample proportions are within ε of 1, the sample relative risk must fall very close to 1. Similarly, strata with $s_k = 0$ make no contribution to the estimated variance (10), and strata with $f_k = 0$ contribute to the denominator alone, thus providing a shrinkage in the variance estimate. For testing, strata with $s_k = 0$ make no contribution to the ratio of estimate to standard error, and provide no information about whether this type of effect exists.

For the M–H estimator (9) of the difference of proportions, strata with $s_k = 0$ or $f_k = 0$ make no contribution to the numerator but do contribute to the denominator. Thus, including such strata has the effect of shrinking the estimated difference of proportions toward 0 compared to the estimate that excludes them. This is expected, since such strata have a sample difference of proportions of 0. There is a compensating shrinkage effect on the standard error, and the ratio of estimate to standard error is unaffected by such strata. Thus, these strata also provide no information about whether this type of effect exists, although they do contribute toward estimating the size of the effect and hence provide evidence about whether interaction exists.

6.2. *Analyses of Table V*

Keeping in mind the highly tentative nature of any random effects modelling with such a small K, we summarize in Tables VI and VII various logit model analyses of Table V. The first row of Table VI reports estimates of the log-odds ratio β and their standard errors, for the ML fixed effects approach, the ML random effects approaches, and the M–H approach. Results are similar for all approaches, with the estimated common log-odds ratio of 1.5 (odds ratio of about 4.5) being about 2.2 standard errors.

The two centres with no successes can provide no information about the log-odds ratio treatment effect β as estimated by the fixed effects model or the M–H method. Very similar results occur with the random effects approach for the reduced data set deleting centres 1 and 3, as shown in the second line of Table VI.

For the no interaction models, the first row of Table VII reports ML estimates of $\{\alpha_k\}$ for the fixed effects model (1) and approximate BLUP estimates of $\{a_k\}$ for the random effects model (4). Because $s_1 = s_3 = 0, \hat{\alpha}_1 = \hat{\alpha}_3 = -\infty$ for model (1). Software may provide misleading indications in such situations, and a danger sign is when standard errors are enormous compared to the estimates, reflecting the very flat log-likelihood. The values in Table VII are those reported by PROC GENMOD in SAS (Version 7). PROC LOGISTIC provides $\hat{\alpha}_1 = -15.0$ (standard error $= 312.8$) and $\hat{\alpha}_3 = -15.3$ (standard error $= 339.7$) but warns that the ML estimates may not exist. The other centre estimates are the same for both procedures and the same as one obtains by deleting centres 1 and 3 from the data set (see row 2 of Table VII).

Table VI. Estimated treatment log-odds ratio (standard error in parentheses) for various logit models with Table V.

Data	Logit model (equation number)				
	Fixed, no interaction (1)	Random, no interaction (4)	Random non-parametric, no interaction	Random interaction (6)	Mantel–Haenszel (7)
Table V unadjusted	1.55 (0.70)	1.52 (0.70)	1.53 (0.69)	1.52 (0.70)	1.55 (0.71)
Delete centres 1,3	1.55 (0.70)	1.48 (0.70)	1.51 (0.69)	1.48 (0.70)	1.55 (0.71)
Combine centres 1–3	1.56 (0.70)	1.54 (0.70)	1.53 (0.69)	1.54 (0.70)	1.56 (0.70)
Add 0.000001 all cells	1.55 (0.70)	1.52 (0.70)	1.53 (0.69)	1.52 (0.70)	1.55 (0.71)
Add 0.05 all cells	1.48 (0.68)	1.45 (0.67)	1.46 (0.67)	1.45 (0.67)	1.48 (0.68)

Corresponding odds ratio estimates vary between $e^{1.45} = 4.3$ and $e^{1.56} = 4.8$.

As noted before, naive standard errors of estimates of random effects ignore the fact that the variance of those random effects is itself estimated. (Moreover, one is naive to expect to estimate well a variance component when K and n are as small as in the examples of this article!) Booth and Hobert [42] proposed a method for calculating standard errors based on the conditional mean squared error of prediction (CMSEP), given the data. This method incorporates a positive correction for the variability of the parameter estimates as well as an estimate of the bias incurred by using an estimate for the unknown conditional variance. Although this bias is often larger than the variance correction and thus non-ignorable, it is computationally difficult to calculate. Morris [41] proposed an analytic correction which can work well for the logistic mixed model [61]. Table VII reports the standard errors for the random effects centre estimates provided by NLMIXED, using the PREDICT option, which are based on a Laplace approximation to the CMSEP.

An ML estimate $\hat{\alpha}_k = -\infty$ is not very appealing when one truly believes that $\pi_{ik} > 0$. Because of the normality assumption, the random effects estimate of a_k also uses information from other centres and is finite. For centre 1, for instance, the estimate $\hat{a}_1 = -1.07$ provides an estimated success probability of $\exp(-1.07)/[1 + \exp(-1.07) = 0.255$ for placebo, even though that group had no successes at that centre. The estimated standard deviation of the centre effects is $\hat{\sigma} = 1.8$. Although centres with $\min(s_k, f_k) = 0$ provide no information about the treatment effect, deleting them from the analysis will tend to decrease $\hat{\sigma}$. In this case, $\hat{\sigma}$ decreases to 1.1.

The no interaction models, whether fixed effects or random effects, showed moderate evidence of a treatment effect. The random effects model permitting interaction has identical results (see Table VI), since the ML estimate of the standard deviation of the log-odds ratio is 0. This also happens when deleting centres 1 and 3 or when combining centres 1–3.

6.3. Extreme cases: centres with one observation per treatment

Simplified forms of the various estimates and standard errors occur for matched pairs data in which each row of each stratum contains a single observation. This is an extreme form of sparseness

Table VII. Estimated centre effects (standard error in parentheses) for no interaction models with Table V.

Data	Fixed effects model (1)					Random effects model (4)				
	α_1	α_2	α_3	α_4	α_5	a_1	a_2	a_3	a_4	a_5
Table V unadjusted	−28.0 (2.1×10^5)	−4.2 (1.2)	−27.9 (1.9×10^5)	−1.0 (0.7)	−2.0 (0.7)	−1.1 (1.4)	−0.5 (1.2)	−1.5 (1.4)	2.3 (1.2)	1.4 (1.2)
Delete centres 1,3		−4.2 (1.2)		−1.0 (0.7)	−2.0 (0.7)		−1.2 (0.9)		1.1 (0.8)	0.2 (0.8)
Combine centres 1–3	−4.9 (1.2)			−1.0 (0.7)	−2.0 (0.7)	−1.8 (1.1)			1.5 (1.0)	0.5 (1.0)
Add 0.000001 all cells	−16.6 (707.1)	−4.2 (1.2)	−16.8 (707.1)	−1.0 (0.7)	−2.0 (0.7)	−1.1 (1.4)	−0.5 (1.2)	−1.2 (1.4)	2.3 (1.2)	1.4 (1.2)
Add 0.05 all cells	−5.7 (3.2)	−4.1 (1.1)	−5.9 (3.2)	−0.9 (0.6)	−2.0 (0.6)	−0.9 (1.2)	−0.6 (1.0)	−1.0 (1.2)	2.1 (1.1)	1.2 (1.0)

Fixed effects estimates obtained using PROC GENMOD in SAS.

in which $n = 2K$. An important application is in cross-over studies, in which stratum k provides subject k's response for each treatment.

Let $a = \sum_k n_{11k}n_{21k}$ denote the number of pairs where both observations are successes, $b = \sum_k n_{11k}n_{22k}$ the number where the first is a success and the second is a failure, $c = \sum_k n_{12k}n_{21k}$ the number where the first is a failure and the second is a success, and $d = \sum_k n_{12k}n_{22k}$ the number where both are failures. Then, the M–H log-odds ratio estimate simplifies to

$$\hat{\beta}_{\mathrm{MH}} = \log(b/c), \quad \widehat{\mathrm{var}}(\hat{\beta}_{\mathrm{MH}}) = b^{-1} + c^{-1}$$

which is identical to the conditional ML estimate. Also, the M–H type of log relative risk estimate is

$$\hat{\phi}_{\mathrm{MH}} = \log[(a+b)/(a+c)], \quad \widehat{\mathrm{var}}(\hat{\phi}_{\mathrm{MH}}) = (b+c)/(a+b)(a+c)$$

and the M–H type of difference of proportions estimate is

$$\hat{\delta}_{\mathrm{MH}} = (b-c)/K, \quad \widehat{\mathrm{var}}(\hat{\delta}_{\mathrm{MH}}) = [(b+c) - (b-c)^2/K]/K^2$$

where $K = a + b + c + d$.

For this degree of sparseness, it is inappropriate to use ordinary ML estimators of these parameters based on models such as (1), as such estimators are inconsistent. The random effects version (4) is adequate, since the number of parameters in the marginal likelihood stays constant as K increases. In fact, suppose the association between the two responses is non-negative, in the sense that $\log(ad/bc) \geqslant 0$; then, for any parametric random effects model that is consistent with the data, the estimate of the log-odds ratio β is identical [62] to the M–H and conditional ML estimates, namely $\log(b/c)$.

6.4. Effects of adding constants or combining centres

When finite ML estimates do not exist, one approach in fixed effects models for contingency table analysis is to add a small positive constant to each cell (or to the empty cells), thus ensuring that all resulting estimates are finite. When that constant is small, however, the resulting value of $\hat{\beta}$ for model (1) and its standard error are usually almost identical to what one obtains by ignoring

strata for which $s_k = 0$ or $f_k = 0$. Table VI illustrates, showing the effect of adding 0.000001 to each cell and adding one observation to the data set (1/20 to each cell) for the sparse data of Table V. The treatment effects and goodness-of-fit are stable, as the addition of any such constant less than 0.001 to each cell yields $\hat{\beta} = 1.55$ (ASE = 0.70) and a G^2 goodness-of-fit statistic equal to 0.50.

Although this process also provides finite centre estimates for strata with $\min(s_k, f_k) = 0$, the estimates for these strata depend strongly on the constant chosen. Table VII illustrates, again showing the effect for added constants of 0.000001 and 0.05. The *ad hoc* nature of this approach is a severe disadvantage. Random effects and Bayesian approaches seem more suited to smoothing effects of zeros, and do not require adding arbitrary constants. Thus, we do not advocate adding constants in order to artificially include data from certain centres in the analysis.

An alternative strategy in multi-centre analyses combines centres of a similar type. Then, if each resulting partial table has responses with both outcomes, the ordinary descriptions and inferences use all the data. This, however, can affect somewhat the interpretations and conclusions made from those inferences. An extreme form of combining centres results from adding together all K tables and performing inference and description for that marginal X-by-Y 2×2 table. Although apparently sometimes done in practice, this can be dangerous, as Simpson's paradox [60] illustrates.

It seems reasonable to combine two centres if the descriptive measure of interest is similar for each and similar to what one gets by combining them. Sufficient conditions exist for when this happens. For instance, suppose the relative risk or the difference of proportions is identical for two centres. Then, the value of that measure takes the same value when the centres are combined [63] if the sample size ratio n_{1+k}/n_{2+k} is the same for each centre. For the odds ratio, collapsibility is more complex, sufficient conditions being the conditional independence of Z with either X or Y for those two strata. These conditions have limited relevance here, however, since when $\min(s_k, f_k) = 0$, there is no information about the size of the relative risk or odds ratio for that centre. Thus, it seems dangerous to combine that centre with others unless there are good reasons to believe that those centres are very similar and could be expected to share similar values of the measure of interest. For the difference of proportions, it is unnecessary in any case to combine a centre having $\min(s_k, f_k) = 0$ with other centres, since it makes a contribution as it is to the summary difference of proportions (although, as noted above, it provides no information about the significance of that difference).

For Table V, perhaps centres 1 and 3 are similar to centre 2, since the success rate is also very low for that centre. Table VI also shows the results of combining these three centres and re-fitting the models to this table and the tables for the other two centres. Here, the effect is negligible. In summary, with frequentist approaches there seems to be no loss of information regarding the significance of treatment effects by simply deleting centres having $\min(s_k, f_k) = 0$, although it is useful to include them for random effects analyses designed to estimate centre variability.

6.5. *Assumptions in models*

For severely sparse data, effects of model misspecification can be especially worrisome. Rarely would it be possible to check assumptions about homogeneity of effects or about a form of distribution for random effects. In estimating parameters in random effects models with sparse data, one might be concerned about how much those estimates may depend on the assumption for the random effects distribution. One way to check this assumption is to compare results to those obtained with a distribution-free approach for the random effects distribution, [64, 65] which

estimates that distribution using a finite number of mass points and probabilities. This approach is available with a GLIM macro [66]. Results for examples here are very similar to those assuming a normal random effect. Table VI illustrates for the no interaction model applied to Table V, providing results supplied by that GLIM macro.

At a minimum, it seems sensible to conduct some analyses designed to investigate sensitivity to assumptions and the influence of changes in the model and slight changes in the data. A *model sensitivity* study checks whether conclusions about the treatment effect are similar for a variety of plausible models. A *case sensitivity* analysis checks the effect on estimates and test statistics of deleting or adding a single observation or changing a single observation from success to failure or vice versa, checking this separately for each cell in the contingency table.

We illustrate the case sensitivity analysis for the no interaction random effects model with Table V. Checking the influence of each observation by deleting it from the data set, the estimated mean log-odds ratios vary from 1.42 to 1.87 with standard error ranging from 0.69 to 0.77, compared to the values of 1.52 and 0.696 for the observed data; the ratios of estimates to standard errors range from 2.02 to 2.42, compared to the observed $1.519/0.696 = 2.18$. The two smallest estimates and ratios result from deleting a success for the active drug in centre 4 or 5. When we instead add a single observation, the estimated mean log-odds ratios range from 1.16 to 1.61 with standard errors ranging from 0.63 to 0.70, while the ratios of estimates to standard errors range from 1.84 to 2.36. The five smallest estimates and ratios result from adding a success in the placebo group, in turn for each centre. After changing a single observation from success to failure or vice versa, the estimated mean log-odds ratios vary from 1.15 to 1.90 with standard errors ranging from 0.63 to 0.77; the ratios of estimates to standard errors range from 1.81 to 2.47. These results indicate the very tentative nature of any conclusions about the significance of the results in Table V.

As mentioned, it can be difficult to estimate well the variance components or standard errors of those components or the random effects. To check whether certain ones seem plausible, one might use the jack-knife or else treat the fitted model as if it were the true one and conduct a parametric bootstrap for independent binomials of the given row sizes satisfying that model [59]. This may be useful also to provide alternative confidence intervals. There is no guarantee that bootstrap methods will work well for highly sparse data, but a dramatically different result can suggest potential problems with the standard error estimate and corresponding Wald interval estimates.

6.6. Dependence of results on method of fitting

When using Gauss–Hermite quadrature to approximate the likelihood function in obtaining ML estimates for random effects models, the resulting quality of the approximations for the ML estimates can depend strongly on the number of quadrature points used. This is especially true when the data are sparse or the variance components are large. We recommend that the number of quadrature points be increased until the change in parameter estimates and standard errors is negligible. In our experience the standard errors and variance component estimates usually require a greater number of quadrature points for convergence than the treatment parameter estimates.

The number of quadrature points can be greatly decreased by centring the quadrature nodes at the mode of the function being integrated and scaling them by the curvature at the mode [24, 25]. Using this approach with the no interaction model for Table I, we needed only 9 quadrature points to obtain convergence (to four decimal places) in the parameter estimates and about 13 for the

standard errors, as opposed to about 200 quadrature points (about 270 for the standard errors) using the standard Gauss–Hermite nodes and weights. With the centred nodes approach one must take care when calculating predicted centre effects and interaction effects, since the functions being approximated may not be unimodal.

By default, PROC NLMIXED in SAS selects the number of quadrature points. Starting with one quadrature point, the log-likelihood is evaluated at the parameter starting values. The number of quadrature points is then increased and the log-likelihood re-evaluated until the difference between two successive evaluations is less than some user-controlled epsilon. That necessary number of quadrature points at the initial values is then used in all successive cycles in determining the parameter values that maximize the likelihood function. In our experience this often leads to only five or six points and can be inadequate for standard error calculations or predictions. Users can avoid the default method by using the QPOINTS= option. We also recommend expressing the variance components as products of standard deviations in the RANDOM statement of NLMIXED. Estimation of the standard deviation often avoids convergence problems when the estimated variance component is close to zero.

7. SUMMARY COMMENTS AND RECOMMENDATIONS

7.1. Similarities and differences in substantive results

For the examples in this paper, we reached similar conclusions about the treatment effect whether we used fixed effects or random effects models. Our experience with a variety of examples indicates that the fixed effects model and the random effects model assuming no interaction tend to provide similar results about the common treatment effect. Those results are also similar to the ones for the mean of the treatment effects for the random effects interaction model when the variance component estimate for the treatment effects equals 0 or close to 0. The latter model may provide a much wider confidence interval for the average effect when that variance component estimate is substantial. To illustrate, we alter Table I slightly, changing three of the failures to successes for drug in centre 3 and three of the successes to failures for drug in centre 8. Then the ML estimates are $\hat{\beta} = 0.759$ with SE $= 0.305$ for fixed effects model (1) and $\hat{\beta} = 0.722$ with SE $= 0.299$ for the random effects model (4) without interaction, but $\hat{\beta} = 0.767$ with SE $= 0.623$ for the random effects model (6) permitting interaction. For the latter model, $\hat{\sigma}_b = 1.37$, compared to 0.15 for the actual data.

By contrast to the usual similarity of estimates of overall treatment effects with fixed effects and random effects models, the two model types can provide quite different estimates of individual centre or treatment effects. For instance, when all observations in a centre fall in the same outcome category, the random effects models smooth the centre effects considerably from the infinite values obtained with the fixed effects models.

An interesting question is to study the types of sparse data configurations or highly unbalanced data sets that can result in the two types of analyses giving substantively different treatment estimates or inferences about the treatment effect. As an extreme example (mentioned to us by Dr C. McCulloch in a personal communication), suppose that some strata have observations only for treatment 1 and the other strata have observations only for treatment 2. The fixed effects approach has insufficient information to estimate the treatment effect, since there are K binomial observations but $K + 1$ parameters for the no interaction model. By contrast, with the random effects approach

data of this form provide information about the treatment effect and about a mean and standard deviation of the random effects distribution, at least if one can regard the strata of each of these two types (having observations with only treatment 1 or having observations with only treatment 2) as a random sample from that distribution.

7.2. Strategies for choice of model and analysis

In selecting a method, a key determinant is the intended scope of inferences. If the strata truly are a sample of all possible strata and one would like to make inferences that apply more generally than to only the strata sampled, then the random effects approach is more natural. Data from multi-centre clinical trials and meta analyses are usually of that type, although the samples are usually not random. However, many share the view quoted earlier of Grizzle [16] that a random effects approach still better reflects all the actual sources of variability. If the strata sampled are the only ones of interest, such as when the strata are levels of control variables such as gender and race, the fixed effects approach is natural. Even when the strata are not a sample, however, the random effects estimators can be beneficial because of their smoothing effects. For instance, when there is significant interaction, the random effects estimates of stratum-specific log-odds ratios might be preferred to the separate sample values, especially when some of those sample values are infinite. See Senn [14] for a more sceptical view noting potential problems with using random effects approaches.

The choice of a fixed effects or random effects analysis can be a complex one having many considerations. [14, 16]. Among statistical considerations, for random effects modelling one should preferably have many more centres than the 8 in Table I and the 5 in Table V, yet the combining of information that occurs with random effects modelling is often very appealing. Among non-statistical considerations, a 'centre' is often quite arbitrary and not as well defined as a 'subject', yet we develop treatments not just for the subjects who attended the centres used in the study [14]. A referee has pointed out that one could consider 'fixed' and 'random' as but two labels for a continuum of sampling models that includes, for instance, systematic cases that are more representative than a random sample in certain senses and illustrative cases that are less so. Further development of such a framework of types of effects would be an interesting topic for further research.

Next, whatever one's choice of fixed or random effects model, one must decide whether to include interaction terms in the model. With many strata or highly sparse data, the power of tests of the hypothesis of no interaction may be weak. The safest approach is then to use the interaction model; otherwise, if one uses the simpler model but interaction truly exists, the standard error of the estimated treatment effect may be unrealistically low. Fixed effects and random effects no interaction models will tend to report smaller standard errors for the treatment effect than the interaction model, since the latter model permits an extra component of variance. Even when $\hat{\sigma}_b = 0$, the likelihood function often reveals that values of σ_b quite far from 0 are also plausible; thus, it is safest to use the interaction model. One may pay a penalty for doing so, having an increased standard error, but this simply reflects scepticism about the homogeneity model and the desire for inferences to apply more generally than for only the centres sampled.

With the random effects approach, one must also consider the validity of a normal assumption for the random effects. When the primary interest is in the treatment effect, the choice of distribution for the random effect should not be crucial [67], as a wide variety of mixing distributions lead to similar marginal distributions (averaged over the random effects). For model (4), for instance, if the normal distribution can induce an intracluster correlation approximately equal to the intracluster

correlation for the actual mixing distribution, then there is little bias in estimation of β or in the standard error estimates [67]. When the actual distribution is highly skewed, some bias [65, 67] may occur in estimating α.

The above remarks refer to the treatment effect. When estimation of centre effects are the focus, it is of interest to study the degree to which the estimates could depend on the choice of distribution. For fixed K, asymptotically this does not seem to be a problem. For instance, when the additive model form (1) holds, for any finite set of centre effects, as $\{n_{ik}\}$ increase the random effect estimators of treatment and centre parameters behave like the fixed effect estimators; in particular, both sets converge to the true values. In practice this manifests itself by the random effects estimates being very similar to the fixed effects estimates when the stratum-specific sample sizes are large.

An interesting open question is to study the effect of misspecification of the random effects distribution for the sparse asymptotic framework in which K grows with n. It is then too much to ask for consistency of centre estimates, but does one obtain consistency of estimation of the treatment effect and the variance components? One way to check the effect of the normality assumption is to compare results to those obtained with a non-parametric approach [64]. An advantage of the normal choice, other than convenience, is that it extends naturally to multivariate random effects that may have some correlation structure.

We have seen that centres with 0 successes or 0 failures can be disregarded in terms of deciding whether a treatment effect exists. They are needed, however, for estimating the variance component of centre effects in the random effects model, and for estimating the size of the effect in fixed and random effects models for the difference of proportions.

7.3. Extensions and alternative methods

Our emphasis has been on binary data with two groups, but the models and issues discussed generalize to multinomial data and several groups. For instance, for an ordinal response, one can use a proportional odds model with centre and treatment effects, with the centre effects being treated either as fixed or random. Recent work has focused on ways of fitting such models with random effects [68, 69], and one can use NLMIXED to fit the proportional odds model and related models (such as with probit link) based on the Gauss–Hermite quadrature approximation of the likelihood function.

Finally, this paper has focused on frequentist approaches. Alternative approaches include Bayes and empirical Bayes methods [4, 6, 11, 12, 70, 71]. The random effects model has much in common with empirical Bayes, in that it assumes a distribution for a set of parameters and uses the data to estimate parameters of that distribution.

ACKNOWLEDGEMENTS

This work was partially supported by grants from NIH and NSF. We appreciate helpful comments from Brent A. Coull, Ranjini Natarajan, Ramon Littell, Brett Presnell, Russell Wolfinger, and three referees.

REFERENCES

1. Beitler PJ, Landis JR. A mixed-effects model for categorical data (correction **42**: 1009). *Biometrics* 1985; **41**:991–1000.
2. Draper D, Gaver Jr, DP, Goel PK, Greenhouse JB, Hedges LV, Morris CN, Tucker JR, Waternaux CMA, Berlin JAR. *Combining Information*: *Statistical Issues and Opportunities for Research*. National Academy Press: 1992.

3. Lipsitz SR, Dear KBG, Laird NM, Molenberghs G. Tests for homogeneity of the risk difference when data are sparse. *Biometrics* 1998; **54**:148–160.
4. Berry SM. Understanding and testing for heterogeneity across 2×2 tables: application to meta-analysis. *Statistics in Medicine* 1998; **17**:2353–2369.
5. Givens GH, Smith DD, Tweedie RL. Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science* 1997; **12**:221–250.
6. Carlin JB. Meta-analysis for 2×2 tables: a Bayesian approach. *Statistics in Medicine* 1992; **11**:141–159.
7. Normand S-LT. Meta analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine* 1999; **18**:321–359.
8. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
9. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Statistics in Medicine* 1995; **14**:395–411.
10. Emerson JD. Combining estimates of the odds ratio: the state of the art. *Statistical Methods in Medical Research* 1994; **3**:157–178.
11. Skene AM, Wakefield JC. Hierarchical models for multicentre binary response studies. *Statistics in Medicine* 1990; **9**:919–929.
12. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* 1995; **14**:2685–2699.
13. Gallo PP. Practical issues in linear models analyses in multicentre clinical trials. *Biopharmaceutical Report of the Biopharmaceutical Section of the American Statistical Association* 1998; **6**:1–9.
14. Senn S. Some controversies in planning and analysing multi-centre trials. *Statistics in Medicine* 1998; **17**:1753–1765.
15. Jones B, Teather D, Wang J, Lewis JA. A comparison of various estimators of a treatment difference for a multi-centre clinical trial. *Statistics in Medicine* 1998; **17**:1767–1777.
16. Grizzle JE. Letter to the editor. *Controlled Clinical Trials* 1987; **8**:392–393.
17. Pierce DA, Sands BR. Extra-Bernoulli variation in regression of binary data. Oregon State University, Department of Statistics Technical Report 46, 1975.
18. Breslow N. Regression analysis of the log odds ratio: a method for retrospective studies. *Biometrics* 1976; **32**:409–416.
19. Hanfelt JJ, Liang K-Y. Inference for odds ratio regression models with sparse dependent data. *Biometrics* 1998; **54**:136–147.
20. Davis LJ. Generalization of the Mantel-Haenszel estimator to nonconstant odds ratios. *Biometrics* 1985; **41**:487–495.
21. Platt R, Leroux B, Breslow N. Generalized linear mixed models for meta-analysis. *Statistics in Medicine* 1999; **18**:643–654.
22. Coull BA, Agresti A. Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics* 2000; **56**:162–168.
23. Anderson DA, Aitkin M. Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society*, *Series B*, *Methodological* 1985; **47**:203–210.
24. Liu Q, Pierce DA. A note on Gauss–Hermite quadrature. *Biometrika* 1994; **81**:624–629.
25. Wolfinger RD. Towards practical application of generalized linear mixed models. Proceedings of 13th International Workshop on Statistical Modeling. Marx B, Friedl H. (eds), 1998:388–395.
26. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; **88**:9–25.
27. Wolfinger R, O'Connell M. Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* 1993; **48**:233–243.
28. McCulloch C. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* 1997; **92**:162–170.
29. Booth JG, Hobert JP. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society*, *Series B* 1999; **61**:265–285.
30. Zeger SL, Karim MR. Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* 1991; **86**:79–86.
31. Hobert J, Casella G. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* 1996; **91**:1461–1473.
32. Natarajan R, McCulloch CE. A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika* 1995; **82**:639–643.
33. Ghosh M, Ghosh A, Chen M, Agresti A. Noninformative priors for one parameter item response models, *Journal of Statistical Planning and Inference* 2000: in press.
34. Greenland S, Robins JM. Estimation of a common effect parameter from sparse follow-up data (correction **45**: 1323–1324). *Biometrics* 1985; **41**:55–68.
35. Andersen EB. Discrete *Statistical Models with Social Science Applications.* North-Holland/Elsevier: New York, 1980.
36. Breslow N, Day NE. *Statistical Methods in Cancer Research, Vol I: The Analysis of Case-Control Studies.* IARC: Lyon, 1980.

37. LogXact. *Logistic Regression Software Featuring Exact Methods.* Cytel Software: Cambridge, MA, 1993.
38. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 1959; **22**:719–748.
39. Tarone RE. On summary estimators of relative risk. *Journal of Chronic Diseases* 1981; **34**:463–468.
40. Nurminen M. Asymptotic efficiency of general noniterative estimators of common relative risk. *Biometrika* 1981; **68**:525–530.
41. Morris CN. Parametric empirical Bayes inference: theory and applications (correction pp. 55–65). *Journal of the American Statistical Association* 1983; **78**:47–55.
42. Booth JG, Hobert JP. Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association* 1998; **93**:262–272.
43. Breslow NE, Lin X. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* 1995; **82**:81–91.
44. Liu Q, Pierce DA. Heterogeneity in Mantel-Haenszel-type models. *Biometrika* 1993; **80**:543–556.
45. Raghunathan TE, Ii Y. Analysis of binary data from a multicentre clinical trial. *Biometrika* 1993; **80**:127–139.
46. Liang K-Y, Self SG. Tests for homogeneity of odds ratios when the data are sparse. *Biometrika* 1985; **72**:353–358.
47. Littell RC, Milliken GA, Stroup WW, Wolfinger RD. *SAS System for Mixed Models*. SAS Institute Inc.: Cary, NC, 1996.
48. Gart JJ, Nam J. Approximate interval estimation of the difference in binomial parameters: Correction for skewness and extension to multiple tables (correction **47**:357; **47**:979). *Biometrics* 1990; **46**:637–643.
49. Robins J, Breslow N, Greenland S. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* 1986; **42**:311–323.
50. Sato T. Comments on 'Estimation of a common effect parameter from sparse follow-up data' (**41**:55–68). *Biometrics* 1989; **45**:1323–1324.
51. Tarone RE. On heterogeneity tests based on efficient scores. *Biometrika* 1985; **72**:91–95.
52. Zelen M. The analysis of several 2×2 contingency tables. *Biometrika* 1971; **58**:129–137.
53. StatXact. *A Statistical Package for Exact Nonparametric Inference* (*version 4.0*). Cytel Software: Cambridge, MA, 1998.
54. Cochran WG. Some methods for strengthening the common chi-square tests. *Biometrics* 1954; **10**: 417–451.
55. Gart JJ. On the combination of relative risks. *Biometrics* 1962; **18**:601–610.
56. Radhakrishna S. Combination of results from several 2×2 contingency tables. *Biometrics* 1965; **21**: 86–98.
57. Emerson JD, Hoaglin DC, Mosteller F. A modified random-effect procedure for combining risk difference in sets of 2×2 tables from clinical trials. *Journal of the Italian Statistical Society* 1993; **2**:269–290.
58. Morris C. Central limit theorems for multinomial sums. *Annals of Statistics* 1975; **3**:165–188.
59. Simonoff JS. Jackknifing and bootstrapping goodness-of-fit statistics in sparse multinomials. *Journal of the American Statistical Association* 1986; **81**:1005–1011.
60. Agresti A. *An Introduction to Categorical Data Analysis*. Wiley: New York, 1996.
61. Greenland S. Second-stage least squares versus penalized quasi-likelihood for fitting hierarchical models. *Statistics in Medicine* 1997; **16**:515–526.
62. Neuhaus JM, Kalbfleisch JD, Hauck WW. Conditions for consistent estimation in mixed-effects models for binary matched-pairs data. *Canadian Journal of Statistics* 1994; **22**:139–148.
63. Shapiro SH. Collapsing contingency tables — A geometric approach. *American Statistician* 1982; **36**: 43–46.
64. Aitkin M. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* 1999; **55**:117–128.
65. Heckman J, Singer B. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 1984; **52**:271–320.
66. Aitkin M, Francis B. Fitting overdispersed generalized linear models by non-parametric maximum likelihood. *GLIM Newsletter* 1995; **25**:37–45.
67. Neuhaus JM, Hauck WW, Kalbfleisch JD. The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* 1992; **79**:755–762.
68. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics* 1994; **50**: 933–944.
69. Tutz G, Hennevogl W. Random effects in ordinal regression models. *Computational Statistics & Data Analysis* 1996; **22**:537–557.
70. Liao JG. A hierarchical Bayesian model for combining multiple 2×2 tables using conditional likelihoods. *Biometrics* 1999; **55**:21–26.
71. Louis TA. Using empirical Bayes methods in biopharmaceutical research (Discussion pp. 828–829). *Statistics in Medicine* 1991; **10**:811–827.

