STATISTICS IN MEDICINE
Statist. Med. **18**, 2693–2708 (1999)

EXPLAINING HETEROGENEITY IN META-ANALYSIS: A COMPARISON OF METHODS

SIMON G. THOMPSON[1]* AND STEPHEN J. SHARP[2]

[1]*Department of Medical Statistics and Evaluation, Imperial College School of Medicine, Hammersmith Hospital, Du Cane Road, London W12 0NN, U.K.*
[2]*Medical Statistics Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, U.K.*

SUMMARY

Exploring the possible reasons for heterogeneity between studies is an important aspect of conducting a meta-analysis. This paper compares a number of methods which can be used to investigate whether a particular covariate, with a value defined for each study in the meta-analysis, explains any heterogeneity. The main example is from a meta-analysis of randomized trials of serum cholesterol reduction, in which the log-odds ratio for coronary events is related to the average extent of cholesterol reduction achieved in each trial. Different forms of weighted normal errors regression and random effects logistic regression are compared. These analyses quantify the extent to which heterogeneity is explained, as well as the effect of cholesterol reduction on the risk of coronary events. In a second example, the relationship between treatment effect estimates and their precision is examined, in order to assess the evidence for publication bias. We conclude that methods which allow for an additive component of residual heterogeneity should be used. In weighted regression, a restricted maximum likelihood estimator is appropriate, although a number of other estimators are also available. Methods which use the original form of the data explicitly, for example the binomial model for observed proportions rather than assuming normality of the log-odds ratios, are now computationally feasible. Although such methods are preferable in principle, they often give similar results in practice.

1. INTRODUCTION

Meta-analysis aims to compare and possibly combine estimates of effect across related studies. For example, in a meta-analysis of clinical trials, the overall effect of a treatment is often expressed as an estimated odds ratio together with its confidence interval. Methods for providing such an overall estimate are well known, and have been extensively discussed from both classical and Bayesian perspectives.[1,2] Methods which incorporate a between-study component of variance for the treatment effect are based on random effects models;[3] the between-study variance represents the excess variation in observed treatment effects over that expected from the imprecision of results within each study. Another term for such between study variation is heterogeneity.

It is now generally agreed that meta-analysis can and should go further than simply producing overall summaries of effect.[4] In particular, understanding the possible causes of any heterogeneity

* Correspondence to: Simon G. Thompson, Department of Medical Statistics and Evaluation, Imperial College School of Medicine, Hammersmith Hospital, Du Cane Road, London W12 0NN, U.K. E-mail: simon.thompson@.ic.ac.uk

CCC 0277–6715/99/202693–16$17.50
Received December 1997
Accepted January 1999

can increase both the scientific value and clinical relevance of the results from a meta-analysis.[5,6] While this is accepted in principle, there has been little discussion of statistical methods which are appropriate for addressing this issue. The purpose of this paper is to exemplify and compare different methods for investigating the extent to which a particular covariate, with a value defined for each study in a meta-analysis, might explain any heterogeneity. Such analysis is sometimes termed 'meta-regression'. The methods differ in a number of respects, including how they allow for residual heterogeneity, that is heterogeneity which remains unexplained by the covariate. They also extend a method previously proposed by Berkey *et al.*[7]

The relevance of this work is clear from applied medical papers, where a variety of methods are being used with little apparent understanding of their statistical basis. For example, linear regression may be used without it being clear whether it was appropriately weighted,[8] or indeed weighted at all,[9] or whether there was any allowance for residual heterogeneity.[10,11] Similarly, for binary outcome data, logistic regression may be used,[12] or subgroup analyses according to a categorical covariate carried out,[13] without residual heterogeneity being taken into account.

The next section introduces the data for the principal example used in the paper, relating to trials of serum cholesterol reduction. Then possible methods for investigating heterogeneity are described, together with some comments on their different assumptions and limitations. These methods are applied to the cholesterol lowering trials, and the results obtained discussed. A second example follows, in which estimated treatment effects are related to their precision to assess the evidence for publication bias. Finally, the issues raised are considered, recommendations made for choice of method in applied work, and extensions to the methods outlined.

2. THE TRIALS OF SERUM CHOLESTEROL REDUCTION

Using data from the randomized trials of serum cholesterol reduction, the purpose here is to quantify how the average cholesterol reduction achieved in each trial relates to the reduction in the risk of ischaemic heart disease (IHD) events observed. Both fatal IHD and non-fatal myocardial infarction were included as IHD events, and the analysis is based on the 28 trials and the data therein as reviewed by Law *et al.*,[14] omitting trial results that have become available more recently. In these trials, cholesterol had been reduced by a variety of means, namely dietary intervention, drugs, and, in one case, surgery. One motivation for this analysis is that if increased benefit in terms of IHD risk reduction were associated with greater reduction in serum cholesterol, this would lend support to the efficacy of these therapies. Moreover it would allow prediction of the expected IHD risk reduction consequent upon a specified decrease in serum cholesterol.

The data are given in Table I, using the odds ratio of IHD as a summary of the results in each trial. Each odds ratio was estimated as the cross-product of cell counts in the corresponding 2×2 table, with the variance of the log-odds ratio equal to the sum of the reciprocal cell counts, as usual. In the two trials with no events in one group, 0·5 was added to each cell for these calculations.[15] The two active treatment arms in one three-arm trial were combined, and the corresponding cholesterol reductions averaged. The cholesterol reduction was derived as the reduction in the treated group minus that in the control group, averaged over the follow-up period of the trial. This average extent of cholesterol reduction varied widely across the trials, from 0·3 to 1·5 mmol/l. The estimated log-odds ratios of IHD are plotted against the serum cholesterol reduction in Figure 1. In the following analyses, a linear relationship between the log-odds ratio of IHD and cholesterol reduction is investigated.

Table I. Data on IHD events from 28 randomized trials of serum cholesterol reduction

Trial*	Control group		Treated group		Odds ratio	Log-odds ratio	Variance of y_i	Cholesterol reduction (mmol/l)
(i)	Events	No events	Events	No events		(y_i)	(v_i)	(x_i)
1	210	5086	173	5158	0·81	− 0·208	0·0109	0·55
2	85	168	54	190	0·56	− 0·577	0·0415	0·68
3	75	292	54	296	0·71	− 0·342	0·0387	0·85
4	936	1853	676	1546	0·87	− 0·144	0·0037	0·55
5	69	215	42	103	1·27	0·239	0·0527	0·59
6	101	175	73	206	0·61	− 0·488	0·0342	0·84
7	193	1707	157	1749	0·79	− 0·231	0·0127	0·65
8	11	61	6	65	0·51	− 0·670	0·2894	0·85
9	42	1087	36	1113	0·84	− 0·178	0·0534	0·49
10	2	28	2	86	0·33	− 1·122	1·0473	0·68
11	84	1946	56	1995	0·65	− 0·430	0·0308	0·69
12	5	89	1	93	0·19	− 1·653	1·2220	1·35
13	121	4395	131	4410	1·08	0·076	0·0164	0·70
14	65	357	52	372	0·77	− 0·264	0·0401	0·87
15	52	142	45	154	0·80	− 0·226	0·0550	0·95
16	81	148	61	168	0·66	− 0·410	0·0414	1·13
17	24	213	37	184	1·78	0·579	0·0788	0·31
18	11	41	8	20	1·49	0·399	0·2903	0·61
19	50	84	47	83	0·95	− 0·050	0·0652	0·57
20	125	292	82	339	0·57	− 0·571	0·0266	1·43
21	20	1643	62	6520	0·78	− 0·247	0·0669	1·08
22	0	52	2	92	2·84	1·043	2·4299	1·48
23	0	29	1	22	3.93	1·369	2·7450	0·56
24	5	25	3	57	0·26	− 1·335	0·5909	1·06
25	144	871	132	886	0·90	− 0·104	0·0168	0·26
26	24	293	35	276	1·55	0·437	0·0773	0·76
27	4	74	3	76	0·73	− 0·314	0·6100	0·54
28	19	60	7	69	0·32	− 1·138	0·2266	0·68

* Data and original trials' references are provided in Law *et al.*[14]

3. WEIGHTED REGRESSION METHODS FOR INVESTIGATING HETEROGENEITY

First methods that use normal errors regression are considered. The observed log-odds ratio of IHD in each trial (y_i say in trial i, for trial $i = 1 \cdots k$) is assumed to follow a normal distribution. The regression needs to be weighted to take into account the precision of the log-odds ratio estimate in each trial. The cholesterol reduction in each trial is denoted by x_i. All summations below are over $i = 1 \cdots k$; in this example $k = 28$.

In the first model, it is assumed that y_i are independently distributed as

$$y_i \sim \mathrm{N}(\alpha + \beta x_i, v_i) \tag{1}$$

where v_i is the variance of the log-odds ratio within trial i, β represents the change in log-odds ratio of IHD per unit change in cholesterol reduction, and α the log-odds ratio at a cholesterol reduction of zero. Maximum likelihood (ML) estimates of α and β can be obtained by least

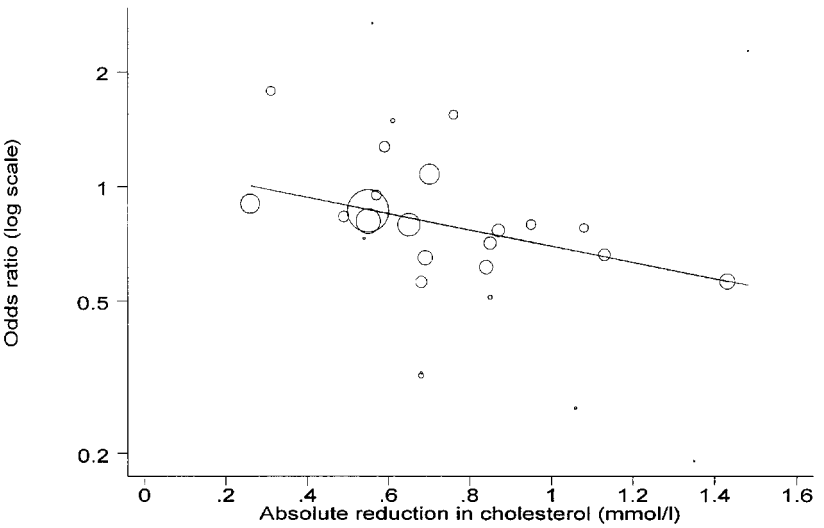

Figure 1. Estimated odds ratios of IHD events in 28 randomized trials of serum cholesterol reduction according to the extent of cholesterol reduction achieved in each trial. The circle corresponding to each trial has area inversely proportional to the variance (v_i) of the log-odds ratio. The superimposed line is obtained by weighted regression using an REML estimate of the residual heterogeneity variance (method (3c))

squares regression of y_i on x_i with weights $w_i = 1/v_i$. If model (1) truly represents the data, then the output from conventional weighted regression programs has to be modified by forcing the mean square error (MSE) to be unity.[16,17] The correct standard errors (SEs) of the regression coefficients are thus obtained by dividing those given by the square root of the reported MSE. In this model, no allowance for residual heterogeneity has been made and the SEs obtained will thus in general be too small.

One method to incorporate residual heterogeneity into the model is to allow a multiplicative factor, greater than 1, to apply to each of the variances v_i. The model then becomes

$$y_i \sim \mathrm{N}(\alpha + \beta x_i, \varphi v_i) \tag{2}$$

where φ is an overdispersion parameter. Here ML estimates of α and β are obtained as before, φ can be estimated as the MSE reported by the weighted regression program, and no modifications to the reported SEs of the regression coefficients should be made. This is conventional weighted regression, where the weights are taken as inversely proportional to the variances v_i rather than necessarily equal to the reciprocal of the variances.

Another method of incorporating residual heterogeneity is to include an additive between-study variance component τ^2. The model is then

$$y_i \sim \mathrm{N}(\alpha + \beta x_i, v_i + \tau^2). \tag{3}$$

ML estimates of α and β are obtained by least squares regression of y_i on x_i with weights $w_i^* = 1/(v_i + \tau^2)$. τ^2 must be explicitly estimated in order to undertake the weighted regression; a number of estimators have been proposed and are described below.

A moment estimator of between-study variance is conventionally used in meta-analysis when no covariate is being considered.[3] Although the extension to the case where there is a covariate is not straightforward, a moment estimator of τ^2 can be derived from the heterogeneity statistic $Q = \sum w_i(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ where $\hat{\alpha}$ and $\hat{\beta}$ are estimated as in (1).[18] In the case of a single covariate

$$\hat{\tau}^2 = \frac{Q - (k-2)}{F(w,x)} \quad \text{if } Q > k-2, \text{ or 0 otherwise} \tag{3a}$$

where

$$F(w,x) = \sum w_i - \frac{\sum w_i^2 \sum w_i x_i^2 - 2\sum w_i^2 x_i \sum w_i x_i + \sum w_i \sum w_i^2 x_i^2}{\sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2}.$$

Then a weighted regression is carried out with weights $w_i^* = 1/(v_i + \hat{\tau}^2)$ to provide new estimates of α and β. This estimator of τ^2 is unbiased if the constraint $\hat{\tau}^2 \geqslant 0$ is removed, and the corresponding formula when there is no covariate[18] reduces to the standard moment estimator.[3]

Other methods which have been proposed for estimating τ^2 require an iterative scheme.[7,19] An ML estimate of τ^2 is provided by solving the iterative equation[20]

$$\tau^2 = \sum w_i^{*2} \{(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 - v_i\} / \sum w_i^{*2} \tag{3b}$$

where $w_i^* = 1/(v_i + \tau^2)$. Starting with $\tau^2 = 0$, a regression using weights w_i^*, then equal to $w_i = 1/v_i$, gives initial estimates of α and β. The right hand side of (3b) is evaluated to yield a new value of τ^2 (subject to the constraint that negative values are set to zero). This then provides modified weights w_i^* for a regression, leading to new estimates of α and β, and thence of τ^2. The procedure continues until convergence. In practice convergence can be slow, and needs to be checked carefully.

The use of restricted maximum likelihood (REML) estimates overcomes the tendency of ML methods to underestimate variances. In this context, an REML estimate of τ^2 is obtained by modifying the right hand side of (3b) using a factor $k/(k-p)$, allowing for the p parameters estimated in the regression. In this case, where α and β are estimated and $p = 2$, the equation becomes

$$\tau^2 = \sum w_i^{*2} \{[k/(k-2)]\,(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 - v_i\} / \sum w_i^{*2}. \tag{3c}$$

An iterative solution is sought as before.

These same ML and REML estimates have been proposed in another context,[19] but using different iterative formulae. They are mentioned in this form by Berkey *et al.*,[7] who instead concentrate on an empirical Bayes estimate of τ^2 obtained by replacing w_i^{*2} by w_i^* in (3c):[21]

$$\tau^2 = \sum w_i^{*} \{[k/(k-2)]\,(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 - v_i\} / \sum w_i^{*}. \tag{3d}$$

When the value of τ^2 has been estimated, either non-iteratively (method (3a)) or iteratively (methods (3b)–(3d)), the weights w_i^* used in the regression are equal to the (estimated) reciprocal variances. The reported SEs should therefore be adjusted using the reported MSE, as for model (1).

The methods (3a) to (3d) suffer from the disadvantage that, while the SEs of $\hat{\alpha}$ and $\hat{\beta}$ take into account the estimate of τ^2, they consider τ^2 as known and equal to its estimated value. When the number of studies is limited, as often is the case in meta-analysis, the estimate of τ^2 is imprecise.

Taking account of this imprecision, which would be expected to increase the SEs of $\hat{\alpha}$ and $\hat{\beta}$, can be achieved in a full Bayes analysis. This can be done using the BUGS implementation of Markov chain Monte Carlo density estimation,[2,22] writing model (3) equivalently as

$$\begin{aligned} y_i &\sim \mathrm{N}(\xi_i, \psi_i) \\ \xi_i &= \alpha + \beta x_i \\ \psi_i &= v_i + \tau^2. \end{aligned} \tag{3e}$$

Priors are needed for the parameters α, β and τ^2. In our analyses, non-informative priors suitable for log-odds ratios were used, namely N(0, 100) for α, N(0, 10000) for β, and an inverse gamma(0·001, 0·001) for τ^2 as advocated elsewhere.[2] A 'burn-in' of 5000 iterations was used, and the posterior distributions of the parameters derived from the following 5000 iterations.[22] The posterior medians and SDs of these distributions provide the equivalent of the estimates and SEs from classical analysis methods. In addition 95 per cent credible intervals are available which do not rely on asymptotic SEs. This method is referred to as 'full Bayes', in contrast to the empirical Bayes method (3d); however it simply uses non-informative priors to approximate a likelihood-based analysis.

All these methods based on weighted regression, numbered (1) to (3e), suffer some theoretical disadvantages. First, it has been assumed that the log-odds ratios have normal distributions. This may be inadequate when, as in the cholesterol trials example (Table I), results of some studies are based on small numbers of events. In particular, when there is a zero count in a cell of the 2×2 table, the odds ratio and variance estimates are not finite. Here 0·5 has been added to all the cells of the 2×2 table in which there is a zero cell, as suggested by others,[15] to circumvent this problem. None of the methods takes account of the fact that the v_i are estimated from the data rather than known. This again will be of particular concern for small studies. These problems are overcome by the logistic regression methods that follow (methods (4) to (6b)). Also accounting for the fact that the residual heterogeneity parameter τ^2 has been estimated is only addressed in the full Bayes analysis (method (6b)).

4. LOGISTIC REGRESSION METHODS FOR INVESTIGATING HETEROGENEITY

In this section, models which use the binomial structure of the data directly are described. Let π_{ij} denote the true risk of IHD events in the jth group ($j = 0$ control, $j = 1$ treated) of trial i, z_j an indicator variable for treatment group (0 for control, 1 for treated), and x_{ij} the cholesterol reduction achieved in the jth group of trial i (relative to the control group). The data provide the number of IHD events (y_{ij}) and number of subjects (n_{ij}) in the jth group of trial i; the form of the data used is shown in Table II.

Conventional logistic regression uses the model

$$\mathrm{logit}(\pi_{ij}) = \gamma_i + \alpha z_j + \beta x_{ij}. \tag{4}$$

Here the parameters α and β are interpreted as before, while the inclusion of fixed parameters γ_i ($i = 1, \ldots, k$) provides for an analysis stratified by study, preserving the comparison of randomized groups as is appropriate.[23] This model gives the log-odds ratio in trial i as $\mathrm{logit}(\pi_{i1}) - \mathrm{logit}(\pi_{i0}) = \alpha + \beta(x_{i1} - x_{i0})$. We express the cholesterol reduction relative to the control group, so that $x_{i0} = 0$, only for convenience. Standard logistic regression programs will provide ML estimates of the parameters of the model, together with asymptotic SEs. However,

Table II. Data for the first two trials of Table I, arranged in a format suitable for logistic regression

Trial (i)	Group (j)	z_j	Events (y_{ij})	Total subjects (n_{ij})	Cholesterol reduction (mmol/l) (x_{ij})
1	Control	0	210	5296	0
1	Treatment	1	173	5331	0·55
2	Control	0	85	253	0
2	Treatment	1	54	244	0·68

there is no allowance for residual heterogeneity (usually termed overdispersion in the context of generalized linear model such as this).

A simple 'correction' for overdispersion is to multiply the reported SEs by a scale factor adjustment.[24] Here it is assumed that

$$\mathrm{var}(y_{ij}) = \varphi n_{ij}\pi_{ij}(1 - \pi_{ij}). \tag{5}$$

The parameter φ can be estimated as the Pearson χ^2 statistic divided by the residual degrees of freedom of the model. The SEs are multiplied by $\sqrt{\hat{\varphi}}$. However the reliability of this adjustment is in doubt when there are small expected (that is, fitted) values. Moreover, the imprecision in estimating the scale factor is not taken into account.

Multi-level models[25] are one way of allowing for an additive component of between trial variability. An appropriate model can be written

$$\mathrm{logit}(\pi_{ij}) = \gamma_i + \alpha z_j + \beta x_{ij} + \alpha_i z_j \tag{6a}$$

where α_i are random parameters with mean zero expressing how the log-odds ratio in trial i deviates from the overall average α. It is assumed that the α_i are distributed as $N(0, \tau^2)$; τ^2 is estimated rather than the individual α_i. Using restricted iterative generalized least squares (RIGLS) under predictive quasi-likelihood (PQL) with second-order approximations,[26] the software MLwiN (formerly MLn)[27] provides approximate REML estimates of the parameters, together with asymptotic SEs. However the SEs of the estimates of the fixed parameters α and β do not take into account the imprecision in estimating τ^2.[28] This problem is overcome by the full Bayes method which follows.

A Bayesian analysis of the above random effects logistic regression model can be obtained using BUGS.[2,22] Here

$$\begin{aligned} y_{ij} &\sim \mathrm{Binomial}(\pi_{ij}, n_{ij}) \\ \mathrm{logit}(\pi_{i0}) &= \gamma_i \\ \mathrm{logit}(\pi_{i1}) &= \gamma_i + \delta_i \\ \delta_i &= \alpha + \beta(x_{i1} - x_{i0}) + \varepsilon_i \\ \varepsilon_i &\sim \mathrm{Normal}(0, \tau^2) \end{aligned} \tag{6b}$$

with x_{ij} defined as in Table II, so that $x_{i0} = 0$. This is the same model as (6a), but written in a form for direct implementation in BUGS. In the analyses which follow, the same priors as in method (3e) were used, with additionally N(0, 100) priors for γ_i. For this analysis, 5000 iterations in BUGS took about 4 minutes on a Pentium 120 processor.

5. APPLICATION OF METHODS TO TRIALS OF SERUM CHOLESTEROL REDUCTION

The above methods were applied to the cholesterol trials data. The results are presented in Table III, numbered by models (1) to (6b) as in Sections 3 and 4.

The first weighted regression did not take any account of residual heterogeneity, and so the SEs were the smallest of any of those obtained by the weighted regression methods. In the second regression, the SEs were increased by a multiplicative factor of 1·21 (the square root of the MSE which was 1·46). In the third set of methods, the SEs were also generally larger, but by different extents for the estimates of α and β, and the estimates themselves were also altered. The ML estimate of τ^2 was zero, and so this analysis gave the same results as method (1). However the bias in the ML estimate is indicated by a positive, albeit small, REML estimate of τ^2, while the moment and empirical Bayes estimates were substantially larger. The full Bayes estimate was close to the REML estimate. The varying estimates of α and β, and their SEs, given by the different methods reflect the estimates of τ^2 obtained.

In none of these weighted regression analyses was there convincing evidence that α was non-zero. Hence the data are compatible with an odds ratio of unity at zero cholesterol reduction. The estimate of β was convincingly negative, so that the IHD risk reduction indeed increases according to extent of cholesterol reduction. The estimate of τ^2, in comparison to that when the covariate is omitted, allows the proportion of the heterogeneity variance explained by the covariate to be calculated. From the results of the REML analysis (method (3c))

Table III. Estimates of the linear regression relationship between log-odds ratio of IHD and extent of serum cholesterol reduction (mmol/l) in 28 randomized trials, obtained by different methods (see Figure 1)

Method*	Residual heterogeneity	Esimates (SEs)		Heterogeneity	
		Intercept (α)	Slope (β)	Multiplicative (φ)	Additive (τ^2)
Weighted regression					
(1)	None	0·121 (0·097)	− 0·475 (0·138)	1	0
(2)	Multiplicative	0·121 (0·117)	− 0·475 (0·167)	1·46	–
(3a)	Additive (MM)	0·160 (0·137)	− 0·521 (0·180)	–	0·017
(3b)	Additive (ML)	0·121 (0·097)	− 0·475 (0·138)	–	0
(3c)	Additive (REML)	0·135 (0·112)	− 0·492 (0·153)	–	0·005
(3d)	Additive (EB)	0·177 (0·156)	− 0·541 (0·203)	–	0·029
(3e)	Additive (FB)	0·148 (0·130)	− 0·508 (0·175)	–	0·007
Logistic regression					
(4)	None	0·121 (0·097)	− 0·476 (0·137)	1	0
(5)	Multiplicative	0·121 (0·120)	− 0·476 (0·171)	1·55	–
(6a)	Additive (MLM)	0·148 (0·126)	− 0·509 (0·167)	–	0·011
(6b)	Additive (FB)	0·117 (0·122)	− 0·469 (0·162)	–	0·007

* Number of model or equation in text, – not applicable, MM method of moments, ML maximum likelihood, REML restricted maximum likelihood, EB empirical Bayes, FB full Bayes, MLM multi-level model

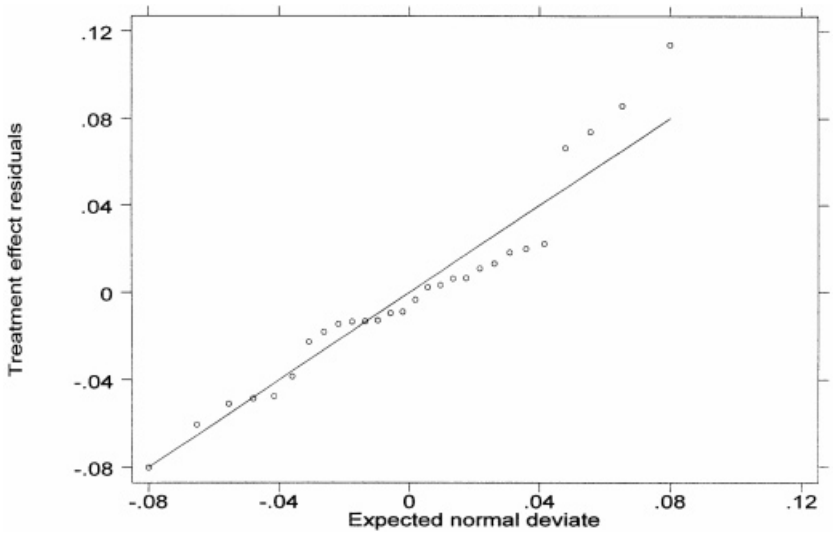

Figure 2. Normal plot of the treatment effect residuals (α_i) from a multi-level model analysis (method (6a)) for the 28 trials of serum cholesterol reduction

for the cholesterol reduction covariate, this proportion was 84 per cent. From this analysis, each 1 mmol/l cholesterol reduction was estimated to reduce the odds ratio of IHD by 39 per cent, that is $1 - \exp(-0{\cdot}492)$; this relation is depicted in Figure 1. For a 1 mmol/l absolute cholesterol reduction, the 95 per cent range of true odds ratios for different studies is estimated as $\exp(0{\cdot}135 - 0{\cdot}492 \pm 2\,\hat{\tau})$, that is from 0·61 to 0·80.

In the logistic regression analyses, the conventional and scale factor adjusted methods yielded almost identical results compared to the first two weighted regression methods. Thus the use of normal approximations for the log-odds ratios was adequate in this example. The multi-level model (6a) corresponds to the REML method (3c), but gave somewhat bigger estimates of τ^2. Figure 2 shows the normal plot for the random treatment effect parameters α_i, obtained as shrunken residuals from the multi-level model;[25] these appear to satisfy at least approximate normality. In the full Bayes analysis, the SEs of $\hat{\alpha}$ and $\hat{\beta}$ are similar to those in the multi-level model analysis despite the smaller estimate of τ^2 obtained. This reflects the full Bayes analysis allowing for the imprecision in $\hat{\tau}^2$. The 95 per cent credible intervals for α and β in the Bayesian analysis were $-0{\cdot}115$ to 0·382 and $-0{\cdot}813$ to $-0{\cdot}167$, respectively, showing little difference from those that would be derived under the assumption of asymptotic normality. The 95 per cent credible interval for τ^2 was 0·0006 to 0·061, reflecting therefore very considerable uncertainty in the estimate. The differences between the estimates for τ^2 obtained by different methods in Table III should thus be viewed with this perspective.

The full Bayes analysis (method (6b)) estimated the 95 per cent range of true odds ratios per 1 mmol/l absolute cholesterol reduction for different studies as 0·59 to 0·83. This is slightly wider than the range given above as obtained from the REML analysis (3c) because the estimate of τ^2 was larger in the Bayesian analysis. As a check of robustness for the choice of prior for τ^2, we replaced the inverse gamma (0·001, 0·001) prior for τ^2 by first a positive half Normal (0, 10) and then a positive half Normal (0, 100) prior for τ. The estimates of τ^2 were somewhat greater, 0·015

and 0·013, respectively, and $\hat{\beta}$ and its SE correspondingly greater, $-0{\cdot}534$ (0·177) and $-0{\cdot}504$ (0·191), respectively. This exemplifies that there is very little information about τ^2 in the data set and the resulting estimate is very imprecise. The consequence is that choice of prior, although all intended to be non-informative, makes some difference to the results obtained.

In this example there is only a moderate degree of heterogeneity between trials, either before or after including the extent of cholesterol reduction as a covariate in the analysis. Hence the differences in results between those methods which do not encompass residual heterogeneity and those which do is not extreme, although the latter yield somewhat larger SEs as expected. If the residual heterogeneity had been more substantial, the differences would have been more marked; this is shown in the second example that follows in the next section. Although there are some small numbers of events amongst the trial data in Table I, the overall analysis is dominated by the results of the larger trials. Hence the assumption of normality of the log-odds ratios seems adequate in this example. The results obtained from the two methods assuming multiplicative residual heterogeneity are similar to each other, as are (in general) the results from those which assume additive residual heterogeneity. One advantage of the multi-level model and full Bayes analyses is that it is easy to inspect the distribution of the random treatment effects, for example, using a normal plot as in Figure 2, thus investigating an important assumption of the model.

Any of the above analyses could have been further pursued by restricting α to be zero, although this would require a different formula for the moment estimator (3a). This would then have focused on the effect of cholesterol reduction *per se*, with the assumption that zero reduction in cholesterol in the trials was associated with no change in the risk of IHD.[23] With α unrestricted, the analysis focuses more on the particular regimens used to lower cholesterol so that, even with a zero cholesterol reduction, there could still be a benefit or hazard of the intervention. Alternatively, the assumption of linearity of the effect of cholesterol reduction on the log-odds ratio of IHD could have been investigated, for example, by including a quadratic term. One of the trials was in fact a multi-arm trial, with a different cholesterol reduction achieved in each treated group. Only the logistic regression methods (4) to (6b) could have been extended to incorporate this facet of the original data; here each line of data (Table II) refers to one group within a trial, rather than one trial as in Table I, and so the extension is straightforward.[23] The use of post-randomization data, necessary to calculate the cholesterol reduction covariate, rather than simply baseline values or trial characteristics, is akin to evaluating how changes in a surrogate marker relate to changes in clinical outcomes.[29]

6. APPLICATION OF METHODS TO TRIALS OF SCLEROTHERAPY

Pagliaro *et al.*[30] performed a meta-analysis of 19 randomized trials assessing the effectiveness of endoscopic sclerotherapy for the prevention of bleeding and death in patients with cirrhosis and oesophagogastric varices. We also considered these data in investigating whether there was a relationship between underlying risk and treatment benefit.[31] Here the various weighted and logistic regression methods described in Sections 3 and 4 are used to assess the evidence of publication bias for the bleeding outcome in these trials, by investigating the linear relationship between treatment effect estimate and precision (as given by the standard error of the estimate). This provides a regression analysis for a funnel plot,[32] and is an objective way of assessing the evidence for publication bias rather than having to rely on unreliable subjective assessment of funnel plot asymmetry.[33] When there is no heterogeneity, the analysis is equivalent to that proposed by Egger *et al.*[11] who performed an unweighted regression of standard normal deviate

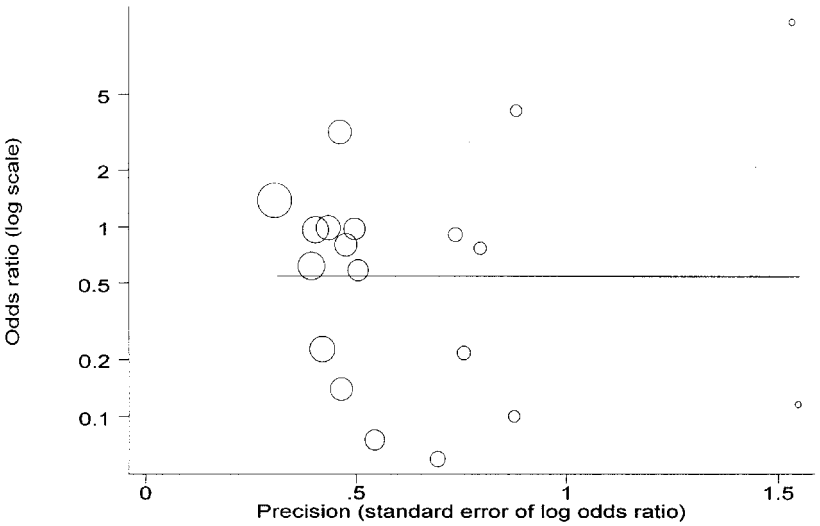

Figure 3. Estimated odds ratios of bleeding in 19 randomized trials of sclerotherapy according to precision (standard error of log-odds ratio). The circle corresponding to each trial has area inversely proportional to the variance (v_i) of the log-odds ratio. The superimposed line is obtained by weighted regression using an REML estimate of the residual heterogeneity variance (method (3c))

(treatment effect estimate divided by its SE) on the reciprocal of the SE, equivalent to a regression on a Galbraith plot.[34] The slope in the analysis below is equivalent to the intercept in Egger's analysis; a non-zero value provides evidence for publication or other small study bias. The following analysis however extends Egger's by appropriately allowing for the possibility of residual heterogeneity.

The original data for the 19 sclerotherapy trials is provided by Thompson *et al.*,[31] the number of subjects in the different trials ranged from 29 to 281, and the number of bleeding outcomes from 3 to 54. There was evidence of substantial heterogeneity between trials in the odds ratios of bleeding comparing sclerotherapy to control.[31] In the regressions we use log-odds ratios, as in the previous example. In two trials with zero events in one randomized group, 0·5 was added to each of the cells of the 2×2 table as before in order to calculate odds ratio estimates for the weighted regression methods and SEs. A plot of log-odds ratios against precision is shown in Figure 3.

The results of the different analyses are shown in Table IV. Methods (1) and (4), which do not allow for heterogeneity, gave some evidence of publication bias, since the slope estimate was negative and larger than its SE. However, the multiplicative heterogeneity variance factors in methods (2) and (5) were much greater in this example than in the cholesterol trials example, so the SEs were more substantially increased as compared to methods (1) and (4). Any evidence for publication bias is removed after allowing for residual heterogeneity. In contrast to the previous example, the direct use of the binomial structure of the data (methods (4) and (5)) here gave results which were noticeably different from those which assumed normality of the log-odds ratios (methods (1) and (2)), because there were more trials with small numbers of events and fewer large trials.

The results using an additive component of variance (methods (3) and (6)) gave generally similar results to each other, but in this example rather different from those when using

Table IV. Estimates of the linear regression relationship between log-odds ratio of bleeding and precision (standard error of log odds ratio) in 19 randomized trials of sclerotherapy, obtained by different methods (see Figure 3)

Method*	Residual heterogeneity	Esimates (SEs) Intercept (α)	Slope (β)	Heterogeneity Multiplicative (φ)	Additive (τ^2)
Weighted regression					
(1)	None	0·056 (0·341)	− 1·113 (0·656)	1	0
(2)	Multiplicative	0·056 (0·734)	− 1·113 (1·411)	4·62	–
(3a)	Additive (MM)	− 0·587 (0·665)	− 0·040 (1·037)	–	0·994
(3b)	Additive (ML)	− 0·594 (0·674)	− 0·028 (1·048)	–	1·037
(3c)	Additive (REML)	− 0·621 (0·712)	0·016 (1·093)	–	1·225
(3d)	Additive (EB)	− 0·639 (0·741)	0·046 (1·128)	–	1·378
(3e)	Additive (FB)	− 0·582 (0·738)	− 0·034 (1·121)	–	1·210
Logistic regression					
(4)	None	− 0·153 (0·320)	− 0·818 (0·599)	1	0
(5)	Multiplicative	− 0·153 (0·767)	− 0·818 (1·435)	5·74	–
(6a)	Additive (MLM)	− 0·642 (0·744)	− 0·122 (1·137)	–	1·367
(6b)	Additive (FB)	− 0·527 (0·745)	− 0·192 (1·061)	–	1·578

* Number of model or equation in text, – not applicable, MM method of moments, ML maximum likelihood, REML restricted maximum likelihood, EB empirical Bayes, FB full Bayes, MLM multi-level model

multiplicative heterogeneity (methods (2) and (5)). There was some difference in the slope estimate when using the binomial structure of the data (method (6)) rather than summary estimates (method (3)). The estimates of τ^2 were fairly similar from the different methods, the ML and MM estimates being the lowest, and the empirical and full Bayes estimates being the largest. The fact that the full Bayes analyses allow for the imprecision in the estimate of τ^2 is illustrated clearly in this example; the estimate of τ^2 in the full Bayes analysis using summary statistics (method (3e)) is less than that in the REML analysis (method (3c)) but the SEs of $\hat{\alpha}$ and $\hat{\beta}$ are greater.

In this example there is only any evidence of publication bias, as expressed by a relationship between log-odds ratio estimate and precision, when residual heterogeneity is ignored. This shows the importance of allowing for residual heterogeneity in such analyses. Two further technical issues are not addressed here. The first is that the estimated SE of a log-odds ratio and the estimated log-odds ratio itself may be artefactually correlated because they are both derived from the same 2×2 table;[7] the second is that, since the estimated SE is imprecise, its use as a covariate in a regression analysis may lead to underestimation of the true regression relationship, through regression dilution bias.[35,36]

7. DISCUSSION

Analyses which investigate whether particular variables may explain some of the heterogeneity of results in meta-analysis are becoming more common. They are prompted both by the desire for a full scientific understanding of the results of meta-analyses, and by the increasingly comprehensive information on the studies which are available. Such methods can also be applicable in the analysis of multi-centre trials.[20] Thus it has become relevant to discuss the appropriate statistical

techniques for carrying out such analyses, especially as a variety of different methods are being used in practice. The paper has addressed the simplest situation, where one trial-level covariate is being investigated.

All analyses have assumed that it is a prerequisite to take into account the precision of the estimated effects within each trial. On statistical grounds, the inverse variance of the estimated effects in each trial has been used directly rather than the number of events in each trial, as has been done in some applied papers.[8] Most fundamentally this paper shows that it is also important to take into account the possibility of residual heterogeneity, that is, heterogeneity not explained by the covariate. Ignoring such residual heterogeneity will underestimate the SEs of the regression coefficients, and thus overstate the importance of the covariate. The use of appropriate SEs would also be important, for example, in calculating a prediction interval for the treatment effect around the estimated regression line.

The rationale for using a multiplicative factor for variance inflation is weak. The idea that the variance of the estimated effect within each study should be multiplied by some constant has little intuitive appeal, and leads to the same dominance of large studies over small studies that has been critized in the context of fixed effect meta-analysis.[37] Thus, despite the fact that such analyses are easy to carry out, and might therefore be used as a quick and approximate way of assessing the impact of residual heterogeneity on the results, we do not recommended them in practice. The use of an additive component of variance to represent heterogeneity between studies is more intuitively appealing, and of course is the usual way of representing heterogeneity in meta-analysis without covariates[3] as well as in many other situations.[25]

The choice of method for estimating the additive component of variance τ^2 is less straightforward. As has been shown in the results, it is the often value of τ^2 that primarily influences the parameter estimates and SEs obtained. In meta-analysis without covariates, a moment estimator of τ^2 is most often used in practice because it is simple to calculate, being non-iterative. The calculations become more difficult when a covariate is introduced, as in (3a), and indeed involve matrix manipulation if more than one covariate were considered.[18] Moreover, ML methods, which are asymptotically efficient, are preferable, but the downward bias of the estimate of τ^2 suggests that an REML estimate is more appropriate here as in other situations.[25] The empirical Bayes estimate, although supported by simulation studies for one particular meta-analysis,[7] gave substantially larger estimates than the other methods, including the full Bayes method. For the methods using the binary data directly, the multi-level REML estimates as implemented in MLwiN[27] may sometimes underestimate the residual heterogeneity variance if the data are sparse.[26] Conversely there is a possibility that the use of a strictly positive prior for τ^2 in the full Bayes method may produce inflated estimates when τ^2 is close to zero. The full Bayes analysis, like some analyses proposed for meta-analysis without covariates,[20,38] has the advantage that the SEs of the regression estimates take into account the imprecision in estimating τ^2. However the effect of this in practice often appears to be small, as has been observed in other circumstances.[20] Therefore a recommendation to use an REML estimate of τ^2 may be most appropriate in practice. A Stata program to fit models (3a) to (3d) is now available.[39] Similar stand-alone programs are also available from Belmont Research.[40]

In the first example considered in this paper, the methods making direct use of the binary data structure produced almost identical results to those using summary estimates of log-odds ratios. Although some small trials with few events were included, their impact was very limited because of the dominating presence of some large trials, as depicted in Figure 1. In the second example, all the trials were relatively small; in this case there was a noticeable difference between these

methods, although this was not substantial enough to change the interpretation. Thus it would only be in cases when all the trials were small that taking the binary nature of the data into account might be important, or possibly where small trials were at the extremes of the range of the covariate, these then being highly influential in the analysis. Thus in general, despite the current availability of specialist software for multi-level logistic regression, its use is in practice likely to give similar results to those from methods based on summary data. The use of individual data would of course become necessary in the developing area where individual-level rather than trial covariates were being investigated,[41] and both the multi-level method (6a) and full Bayes method (6b) naturally extend to this case.

In this paper it has been assumed that the covariate being investigated has been specified in advance of data inspection. In such a case the question of whether the covariate explains any of the observed heterogeneity can be answered from a simple hypothesis testing perspective. However, in practice there may be many such covariates to choose from, or the ones selected may have been chosen with knowledge of the results of some of the studies. While all the methods we have described extend to the inclusion of multiple covariates, the interpretation needs to be more cautious to take into account the dangers of *post hoc* data-dredging.[5] In the extreme case, any set of $(k-1)$ non-linearly dependent trial-level covariates will 'explain' all the heterogeneity between the results of k trials. In practice, near-collinearity of categorical variables describing trial characteristics can also be a problem.[10]

In applied papers, investigation of heterogeneity often proceeds by division of the set of trials into subgroups according to some characteristic. The methods proposed here are applicable to binary or categorical covariates, as well as continuous covariates such as those presented in the examples. In general, because of the imprecision of the estimated residual heterogeneity variance, it may be preferable to estimate a single variance, rather than, for example, allowing this variance to vary according to subgroups as defined by the categorical covariate.

Results have been shown for analyses where effect sizes are on the log-odds ratio scale. The methods which employ summary data could be used for any scale of measuring outcome that yields approximate normality. These include mean differences for continuous data, or log relative risks or absolute risk differences for binary outcomes. It is also possible to use 'smoothed' estimates of the variances of log-odds ratios, to reduce the correlation between the estimates and their variances,[7] but we have not pursued this in this paper. In computer software, a standard Normal distribution has been used to derive statistical tests and confidence intervals from the estimated SEs obtained by these methods.[39] Others have advocated a t-distribution; based on simulations for one data set a t_{k-p-4} distribution was suggested.[7] However, the issue as to which is the better approximation is currently unresolved, and is likely to depend on the particular characteristics of the data set being analysed. This problem can be avoided in the full Bayes analyses since 95 per cent credible intervals can be directly obtained from the posterior distributions simulated.

Extensions of methods (6a) and (6b) to individual-level data for means of continuous outcomes is straightforward, but their development for use on binary outcome data on scales other than the log-odds ratio is problematic. The Bayesian approach (6b) however has the advantage that it can be extended to allow for distributions other than Normal for the random effects; for example, t-distributions can be used to encompass heavy tailed distributions.[42] The methods presented in this paper should not however be used for the more complex issue of whether the treatment effects across a set of trials depend on underlying risk, where for example underlying risk is assessed from

the observed risk in the control group in each trial, because of the problems induced by regression to the mean.[31,43,44]

The associations between trial-level covariates and estimates of treatment effect are between-trial associations, and thus observational in nature. Hence they do not necessarily have the causal interpretation that can be ascribed to treatment comparisons within randomized trials. In particular, the relationship between treatment effects and average covariates at the trial level does not have the same interpretation as a relationship derived from data on individuals.[45] Moreover, if the covariate is measured imprecisely, the strength of the regression relationship will be underestimated; in some contexts it will be important to allow for this regression dilution bias.[29] Any evidence for causality in such associations is not direct, and alternative explanations in terms of confounding have always to be considered.

ACKNOWLEDGEMENTS

We thank Catherine Berkey, Richard Tweedie, Brad Biggerstaff, William DuMouchel, Jesse Berlin and one anonymous referee for their advice on the various estimators of heterogeneity variance discussed in this paper and for other constructive comments.

REFERENCES

1. Fleiss, J. L. 'The statistical basis of meta-analysis', *Statistical Methods in Medical Research*, **2**, 121–145 (1993).
2. Smith, T. C., Spiegelhalter, D. J. and Thomas, A. 'Bayesian approaches to meta-analysis: a comparative study', *Statistics in Medicine*, **14**, 2685–2699 (1995).
3. DerSimonian, R. and Laird, N. 'Meta-analysis in clinical trials', *Controlled Clinical Trials*, **7**, 177–188 (1986).
4. Rubin, D. 'A new perspective', in Wachter, K. W. and Straf, M. L. (eds), *The Future of Meta-analysis*, Russell Sage Foundation, New York, 1990, Chapter 14, pp. 155–165.
5. Thompson, S. G. 'Why sources of heterogeneity in meta-analysis should be investigated', *British Medical Journal*, **309**, 1351–1355 (1994).
6. Berlin J. A. 'Benefits of heterogeneity in meta-analysis of data from epidemiologic studies', *American Journal of Epidemiology*, **142**, 383–387 (1995).
7. Berkey, C. S., Hoaglin, D. C., Mosteller, F. and Colditz, G. A. 'A random-effects regression model for meta-analysis', *Statistics in Medicine*, **14**, 395–411 (1995).
8. Boersma, E., Maas, A. C. P., Deckers, J. W. and Simoons, M. L. 'Early thrombolytic treatment in acute myocardial infarction: reappraisal of the golden hour', *Lancet*, **348**, 771–775 (1996).
9. Bucher, H. C., Cook, R. J., Guyatt, G. H., Lang, J. D., Cook, D. J., Hatala, R. and Hunt, D. L. 'Effects of dietary calcium supplementation on blood pressure: a meta-analysis of randomized controlled trials', *Journal of American Medical Association*, **275**, 1016–1022 (1996).
10. Berlin, J. A. and Antman, E. M. 'Advantages and limitations of meta-analytic regressions of clinical trials data', *Online Journal of Current Clinical Trials*, **3**, document 134 (1994).
11. Egger, M., Davey Smith, G., Schneider, M. and Minder, C. 'Bias in meta-analysis detected by a simple, graphical test', *British Medical Journal*, **315**, 629–634 (1997).
12. Christensen, E. and Gluud, C. 'Glucocorticoids are ineffective in alcoholic hepatitis: a meta-analysis adjusting for confounding variables', *Gut*, **37**, 113–118 (1995).
13. Lefering, R. and Neugebauer, E. A. M. 'Steroid controversy in sepsis and septic shock: a meta-analysis', *Critical Care Medicine*, **23**, 1294–1303 (1995).
14. Law, M. R., Wald, N. J. and Thompson, S. G. 'By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease?', *British Medical Journal*, **308**, 367–373 (1994).
15. Cox, D. R. and Snell, E. J. *Analysis of Binary Data*, 2nd edn, Chapman and Hall, London, 1989.
16. Greenland, S. 'Quantitative methods in the review of epidemiologic literature', *Epidemiologic Reviews* **9**, 1–30 (1987).

17. Chene, G. and Thompson, S. G. 'Methods for summarizing the risk associations of quantitative variables in epidemiologic studies in a consistent form', *American Journal of Epidemiology*, **144**, 610–621 (1996).
18. DuMouchel, W. H. and Harris, J. E. 'Bayes methods for combining the results of cancer studies in humans and other species', *Journal of the American Statistical Association*, **78**, 293–308 (1983).
19. Pocock, S. J., Cook, D. G. and Beresford, S. A. A. 'Regression of area mortality rates on explanatory variables: what weighting is appropriate?', *Applied Statistics* **30**, 286–295 (1981).
20. Hardy, R. and Thompson, S. G. 'A likelihood approach to meta-analysis with random effects', *Statistics in Medicine*, **15**, 619–629 (1996).
21. Morris, C. N. 'Parametric empirical Bayes inference: theory and applications', *Journal of the American Statistical Association*, **78**, 47–55 (1983).
22. Gilks, W. R., Thomas, A. and Spiegelhalter, D. J. 'A language and program for complex Bayesian modelling', *Statistician*, **43**, 169–177 (1994).
23. Thompson, S. G. 'Controversies in meta-analysis: the case of the trials of serum cholesterol reduction', *Statistical Methods in Medical Research*, **2**, 173–192 (1993).
24. McCullagh, P. J. and Nelder, J. A. *Generalized Linear Models*, 2nd edn, Chapman and Hall, London, 1989.
25. Goldstein, H. *Multilevel Statistical Models*, 2nd edn, Edward Arnold, London, 1995.
26. Goldstein H. and Rasbash, J. 'Improved approximations for multilevel models with binary responses', *Journal of the Royal Statistical Society*, *Series A*, **159**, 505–513 (1996).
27. Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G. and Healy, M. J. R. *A User's Guide to MLwiN*, Institute of Education, London, 1998.
28. Kenwood, M. G. and Rogers, J. H. 'Small sample inference for fixed effects from restricted maximum likelihood', *Biometrics*, **53**, 983–997 (1997).
29. Daniels, M. J. and Hughes, M. D. 'Meta-analysis for the evaluation of potential surrogate markers', *Statistics in Medicine*, **16**, 1965–1982 (1997).
30. Pagliaro, L., D'Amico, G., Sorensen, T. I. A., Lebrec, D., Burroughs, A. K., Morabito, A., Tiné F., Politi, F. and Traina, M. 'Prevention of first bleeding in cirrhosis – a meta-analysis of randomized trials of nonsurgical treatment', *Annals of Internal Medicine*, **117**, 59–70 (1992).
31. Thompson S. G., Smith, T. C. and Sharp, S. J. 'Investigating underlying risk as a source of heterogeneity in meta-analysis', *Statistics in Medicine*, **16**, 2741–2758 (1997).
32. Egger, M. and Davey Smith, G. 'Misleading meta-analysis. Lessons from "an effective, safe, simple" intervention that wasn't', *British Medical Journal*, **310**, 752–754 (1995).
33. Villar, J. Piaggio, G., Carroli, G. and Donner, A. 'Factors affecting the comparability of meta-analyses and largest trials' results in perinatology', *Journal of Clinical Epidemiology*, **50**, 997–1002 (1997).
34. Galbraith, R. F. 'A note on the graphical presentation of estimated odds ratios from several clinical trials', *Statistics in Medicine*, **7**, 889–894 (1988).
35. Irwig, L., Macaskill, P., Berry, G. and Glasziou, P. 'Bias in meta-analysis detected by a simple, graphical test (letter)', *British Medical Journal*, **316**, 470 (1998).
36. Sterne, J. A. C. and Egger, M. 'Methods to detect bias in meta-analysis, and their application to placebo-controlled trials of homeopathy', in press.
37. Thompson, S. G. and Pocock, S. J. 'Can meta-analyses be trusted', *Lancet* **338**, 1127–1130 (1991).
38. Biggerstaff, B. J. and Tweedie, R. L. 'Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis', *Statistics in Medicine*, **16**, 753–768 (1997).
39. Sharp, S. J. 'Meta-analysis regression', *Stata Technical Bulletin*', **42**, 16–22 (1998).
40. DuMouchel, W., Fram, D., Jin, Z., Normand, S., Snow, B., Taylor, S. and Tweedie, R. 'Software for exploration and modeling of meta-analyses (abstract)', *Controlled Clinical Trials*, **18**, 181S (1997).
41. Stewart, L. A. and Clarke, M. J. 'Practical methodology of meta-analysis (overviews) using updated individual patient data', *Statistics in Medicine*, **14**, 2057–2079 (1995).
42. Best, N. G., Spiegelhalter, D. J., Thomas, A. and Brayne, C. E. G. 'Bayesian analysis of realistically complex models', *Journal of the Royal Statistical Society*, *Series A*, **159**, 323–342 (1996).
43. McIntosh, M. 'The population risk as an explanatory variable in research synthesis of clinical trials', *Statistics in Medicine*, **15**, 1713–1728 (1996).
44. Walter, S. D. 'Variation in baseline risk as an explanation of heterogeneity in meta-analysis', *Statistics in Medicine*, **16**, 2883–2900 (1997).
45. Morgenstern, H. 'Uses of ecological analysis in epidemiologic research', *American Journal of Public Health*, **72**, 127–130 (1982).

