STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:3417–3432

A multilevel model framework for meta-analysis of clinical trials with binary outcomes[‡]

Rebecca M. Turner[1,*,†], Rumana Z. Omar[2], Min Yang[3], Harvey Goldstein[3] and Simon G. Thompson[4]

[1] *MRC Clinical Trials Unit, 222 Euston Road, London NW1 2DA, U.K.*
[2] *Department of Epidemiology and Public Health, Imperial College School of Medicine, Du Cane Road, London W12 ONN, U.K.*
[3] *Mathematical Sciences Group, Institute of Education, 20 Bedford Way, London WC1H OAL, U.K.*
[4] *MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 2SR, U.K.*

SUMMARY

In this paper we explore the potential of multilevel models for meta-analysis of trials with binary outcomes for both summary data, such as log-odds ratios, and individual patient data. Conventional fixed effect and random effects models are put into a multilevel model framework, which provides maximum likelihood or restricted maximum likelihood estimation. To exemplify the methods, we use the results from 22 trials to prevent respiratory tract infections; we also make comparisons with a second example data set comprising fewer trials. Within summary data methods, confidence intervals for the overall treatment effect and for the between-trial variance may be derived from likelihood based methods or a parametric bootstrap as well as from Wald methods; the bootstrap intervals are preferred because they relax the assumptions required by the other two methods. When modelling individual patient data, a bias corrected bootstrap may be used to provide unbiased estimation and correctly located confidence intervals; this method is particularly valuable for the between-trial variance. The trial effects may be modelled as either fixed or random within individual data models, and we discuss the corresponding assumptions and implications. If random trial effects are used, the covariance between these and the random treatment effects should be included; the resulting model is equivalent to a bivariate approach to meta-analysis. Having implemented these techniques, the flexibility of multilevel modelling may be exploited in facilitating extensions to standard meta-analysis methods. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

Meta-analysis of a set of clinical trials involves a series of choices. The decisions at each stage are similar whether the meta-analyst has only summary data from published results or full individual

[*] Correspondence to: Rebecca M. Turner, MRC Clinical Trials Unit, 222 Euston Road, London, NW1 2DA, U.K.
[†] E-mail: rebecca.turner@ctu.mrc.ac.uk
[‡] Presented at the International Society for Clinical Biostatistics, Nineteenth International Meeting, Dundee, U.K., August 1998.

Contract/grant sponsor: Economic and Social Research Council

patient data. The first choice is between fixed effect and random effects models [1], and in either case the method of estimation must be selected from a number of alternatives [2]. If fitting a random effects model, more decisions arise: how to allow for uncertainty in estimation of the between-trial variance when constructing a confidence interval for the treatment effect [3, 4]; how to obtain confidence intervals for the between-trial variance [3]; how to incorporate trial-level covariates [5–8], and how to investigate sources of between-trial heterogeneity [9–12]. Having selected appropriate methods from those available, the meta-analyst may require extensions to these to deal with more complex data sets, for example survival data [13] or multivariate outcomes [14]. Many approaches have been proposed to address particular issues, but these may not always extend easily to other situations. Bayesian methods [6, 15] serve as one possible approach to the issues mentioned above, and facilitate many extensions besides. Here we discuss multilevel modelling in a frequentist framework which provides a unified approach to meta-analysis, and which may be carried out within widely available software.

Multilevel modelling is now an accepted statistical analysis tool for hierarchical data [16]. In this paper we put existing methods for meta-analysis of two-arm clinical trials with binary outcomes into the general framework of multilevel modelling. At present, the majority of meta-analyses are performed with access only to published treatment effects expressed as summary data, such as log-odds ratios, an approach referred to as meta-analysis of the literature [17]. Meta-analyses of individual patient data are however likely to become more common in the future. We therefore demonstrate the use of multilevel models in meta-analysis of both summary data and of individual binary data. The flexibility of the multilevel model framework may then be exploited in providing extensions to standard methods [18].

The meta-analysis data sets used to exemplify the multilevel modelling methods are described briefly in Section 2. Methods for meta-analysis of summary data are discussed in Section 3, as applied to the main example data set. Section 4 examines the application of individual data methods to the same example. A second example data set is of a different size and structure to the first; the results from meta-analysis of these data are presented for comparison in Section 5. When performing a meta-analysis using individual patient data, the trial effects on the outcome may be regarded as either fixed or random. Since the latter approach raises different issues, we consider this separately in Section 6.

2. EXAMPLES

The main data set used in exemplifying the methods consists of 22 trials performed to investigate the effect of selective decontamination of the digestive tract on the risk of respiratory tract infection; patients in intensive care units were randomized to receive treatment by a combination of non-absorbable antibiotics or to receive no treatment [6]. The numbers of patients on each treatment in each outcome category are available (Table I), as well as the summary log-odds ratios with their variances. In order to examine the modelling issues associated with relatively small meta-analyses and to provide some comparison between data sets, we also use a second example comprising fewer trials. This example involves nine clinical trials examining the effect of taking diuretics during pregnancy on the risk of pre-eclampsia (Table II) [19]. A third example is used in Section 6 to enable comparison with the bivariate approach proposed by van Houwelingen *et al.* [20], where the raw data are tabulated; this involves 25 trials for the treatment of upper gastrointestinal bleeding by a histamine H_2 antagonist.

Table I. Respiratory tract infections in treated and control groups of 22 trials for selective decontamination of the digestive tract.

Trial	Infections/total		Odds ratio	Log OR	Variance (log OR)
	Treated	Control			
1	7/47	25/54	0.20	−1.59	0.24
2	4/38	24/41	0.08	−2.48	0.38
3	20/96	37/95	0.41	−0.89	0.11
4	1/14	11/17	0.04	−3.17	1.33
5	10/48	26/49	0.23	−1.46	0.21
6	2/101	13/84	0.11	−2.20	0.60
7	12/161	38/170	0.28	−1.27	0.12
8	1/28	29/60	0.04	−3.23	1.10
9	1/19	9/20	0.07	−2.69	1.26
10	22/49	44/47	0.06	−2.89	0.44
11	25/162	30/160	0.79	−0.23	0.09
12	31/200	40/185	0.66	−0.41	0.07
13	9/39	10/41	0.93	−0.07	0.28
14	22/193	40/185	0.47	−0.76	0.08
15*	0/45	4/46	0.10	−2.27	2.27
16	31/131	60/140	0.41	−0.88	0.07
17	4/75	12/75	0.30	−1.22	0.36
18	31/220	42/225	0.71	−0.34	0.07
19	7/55	26/57	0.17	−1.75	0.23
20	3/91	17/92	0.15	−1.89	0.42
21*	14/25	23/23	0.03	−3.62	2.20
22	3/65	6/68	0.50	−0.69	0.53

* 0.5 added to all cells of the 2 × 2 table in calculation of the log odds ratio and its variance, to avoid degeneracy.

Table II. Cases of pre-eclampsia in treated and control groups of nine diuretics trials.

Trial	Cases of pre-eclampsia/total		Odds ratio	Log OR	Variance (log OR)
	Treated	Control			
1	14/131	14/136	1.04	0.04	0.16
2	21/385	17/134	0.40	−0.92	0.12
3	14/57	24/48	0.33	−1.12	0.18
4	6/38	18/40	0.23	−1.47	0.30
5	12/1011	35/760	0.25	−1.39	0.11
6	138/1370	175/1336	0.74	−0.30	0.01
7	15/506	20/524	0.77	−0.26	0.12
8	6/108	2/103	2.97	1.09	0.69
9	65/153	40/102	1.14	0.14	0.07

3. SUMMARY DATA METHODS

3.1. Standard methods

A fixed effect model for meta-analysis assumes the true treatment effects to be homogeneous across trials, and accordingly estimates the common treatment effect θ by a weighted average of the

Table III. Estimates from meta-analysis of respiratory tract infections data, using summary data methods and individual data methods with fixed trial effects.

	Log OR (θ)	(95% CI)	Between-trial variance (τ^2)	(95% CI)
Summary data methods				
Conventional				
Fixed effect	-0.94	$(-1.12, -0.77)$	0	–
Random effects	-1.27	$(-1.61, -0.92)$	0.36	–
Multilevel modelling				
ML: (Wald)	-1.29	$(-1.65, -0.92)$	0.42	(0.01, 0.83)
(Profile likelihood)		$(-1.73, -0.92)$		(0.12, 1.19)
(Bootstrap)		$(-1.63, -0.92)$		(0.06, 0.86)
REML: (Wald)	-1.30	$(-1.67, -0.93)$	0.47	(0.02, 0.91)
(Bootstrap)		$(-1.66, -0.93)$		(0.10, 1.01)
Individual data methods with fixed trial effects				
Fixed effect	-1.06	$(-1.24, -0.89)$	0	–
Random effects				
ML: (Wald)	-1.43	$(-1.80, -1.07)$	0.46	(0.08, 0.84)
REML: (Wald)	-1.49	$(-1.90, -1.07)$	0.64	(0.14, 1.14)
(Bootstrap with bias correction)	-1.66	$(-2.26, -1.05)$	0.71	[0.00, 1.13)

trial-specific estimates, with weights equal to the reciprocals of their within-trial variances [2]. The traditional random effects model [21] assumes the true treatment effects to vary randomly between trials. This model includes a between-trial component of variance τ^2 which is usually estimated non-iteratively by a method of moments estimator. Assuming normality of the observed and true treatment effects, the random effects model can be written

$$\begin{aligned} y_i &\sim \mathrm{N}(\theta + v_i, \sigma_i^2) \\ v_i &\sim \mathrm{N}(0, \tau^2) \end{aligned} \tag{1}$$

where σ_i^2 is the variance of the observed treatment effect y_i in the ith trial, usually assumed to be known. Under the assumption of normality for the y_i, a confidence interval may be calculated for the average treatment effect θ. A commonly used measure of treatment effect in binary event data is the log-odds ratio; the normality assumption required is more easily satisfied for this than for alternative measures such as risk difference. The fixed effect model ($\tau^2 = 0$) and the random effects model with a moment estimator of τ^2 have been fitted to the summary log-odds ratios from the respiratory tract infections data set; the results are given in Table III.

There is evidence of heterogeneity ($\chi^2_{21} = 60.1, P < 0.001$) across trials, so the random effects estimate of θ is not equal to the fixed effect estimate, and has a wider confidence interval. The extent and direction of the difference between the fixed effect and random effects estimates is explained by the likely presence of publication bias ($P < 0.001$ from a regression asymmetry test [22]). The smallest trials in the infections data set tend to have the more extreme estimates of treatment effect (Table I), and smaller trials are given greater relative weight in calculation of the random effects estimate compared with the fixed effect estimate [2]. Some of the approaches for addressing publication bias are cited in Section 7; this issue is separate from the aims of the current paper and so is not addressed here.

3.2. *Multilevel modelling methods*

The random effects method for summary data may be expressed as a multilevel model [23]. In this approach, the model specified by (1) is regarded as a random effects regression model, with the observed log-odds ratios considered as a continuous outcome. Maximum likelihood (ML) and restricted maximum likelihood (REML) estimates may then be obtained easily using multilevel modelling software. We use the MLn/MLwiN software [24,25] to fit this model to the respiratory tract infections data; ML and REML estimates are found via the iteratively generalized least squares (IGLS) and restricted iteratively generalized least squares (RIGLS) algorithms, respectively [16].

The three random effects approaches of ML, REML and conventional non-iterative method of moments estimation [21] all lead to similar estimates of θ (Table III). ML estimates of variance components such as τ^2 do not take into account the use of the same data in estimation of the fixed effects, so are biased downwards in general; in REML estimation a modified likelihood is used to produce unbiased estimates [26]. The ML estimate is about 10 per cent smaller than the REML estimate in our example, and the moment estimate of τ^2 is smaller than both ML and REML estimates. The moment estimator is also unbiased for τ^2 and differences between moment and REML estimates are generally small [21]. The ML and REML estimates presented here agree with those calculated using the corresponding iterative formulae directly [5].

3.3. *Calculation of confidence intervals*

When using the multilevel models framework, it is possible to calculate likelihood based and bootstrap confidence intervals for both θ and τ^2, as well as Wald intervals using asymptotic standard errors. In the calculation of Wald confidence intervals for θ, τ^2 is assumed to be known rather than estimated and the resulting intervals are too narrow [4]. When obtaining confidence intervals for τ^2, the normality assumption required by the Wald approach is almost certainly invalid unless the number of trials is very large [16].

Likelihood based intervals for θ are preferable to Wald intervals [3] since they allow for the imprecision in estimation of τ^2; intervals based on likelihood are also preferable when the log-likelihood of a parameter is highly skewed or irregularly shaped [27]. However, the likelihood approach does require normality if the range of values obtained is to be interpreted as a confidence interval in a strict sense [28], so problems remain in constructing confidence intervals for τ^2. The likelihood based method requires computation of the profile log-likelihood $l(\lambda)$ for a range of values of the parameter λ of interest; this can be implemented within the multilevel models framework. Likelihood ratio test statistics computed from REML log-likelihood may not be strictly valid so require modification before use in testing of fixed effects [29]; REML profile likelihood confidence intervals may not therefore be constructed in the usual way for θ. However, REML likelihood ratio test statistics may be used directly for testing of variance components such as τ^2 [30]. Since the method for constructing REML likelihood based intervals for fixed effects is not currently implemented within most multilevel software, we present profile likelihood confidence intervals based on ML only.

The parametric bootstrap [31] can be employed to provide confidence intervals for θ and τ^2. This method requires no normality assumption for the estimate about which the confidence interval is constructed and is therefore useful when sample sizes are small, in particular for τ^2. A series of data sets are simulated under the distributional assumptions of the initial model and a bootstrap set of parameter estimates is generated from each. We may then base confidence intervals on the bootstrap replications of each parameter of interest, using the smoothed percentiles of the bootstrap distributions. Parametric bootstrapping is directly available in the MLwiN software [25]. The bootstrap confidence intervals presented for θ and τ^2 are based on smoothed bootstrap percentiles from 1000 replications.

Differences between the widths of the three Wald confidence intervals for θ using conventional random effects, ML and REML estimation reflect the differences in the estimates of τ^2. The method of profile likelihood produced the widest confidence intervals for both θ and τ^2. The likelihood based interval for θ is somewhat wider than the ML Wald interval, and slightly asymmetric about the estimate; the ML and REML bootstrap intervals are slightly narrower than the corresponding Wald intervals. The increase in the width of the confidence interval from using likelihood rather than Wald methods is more substantial for τ^2 than for θ. The bootstrap intervals for τ^2 are of similar width to the Wald intervals, but both bootstrap and likelihood intervals cover ranges of values further from zero than those covered by the two Wald intervals. The Wald method has not performed badly for θ in this data set, but for τ^2 the forced symmetry about the estimate is inappropriate. Bootstrap intervals are preferred over likelihood intervals because the bootstrap method relaxes the normality assumption required for interpretation of likelihood support intervals as approximate confidence intervals. Since this assumption is less sound for τ^2 than for θ, a greater discrepancy is seen between bootstrap and likelihood intervals for the former.

4. INDIVIDUAL DATA METHODS

4.1. Description of model

One approach to meta-analysis of individual binary data is to fit logistic regression models, with fixed trial effects allowing the log-odds to vary across the n trials. A fixed effect analysis may be performed by means of a model containing as explanatory variables treatment group together with n dummy variables representing trial effects. To carry out a random effects meta-analysis comparable with the DerSimonian and Laird model [21], we require in addition terms v_i representing the deviation of each trial's true treatment effect (log-odds ratio) from the average:

$$\begin{aligned} \text{logit}(\pi_{ij}) &= (\theta + v_i)x_{ij} + \phi_i \\ v_i &\sim \text{N}(0, \tau^2) \end{aligned} \qquad (2)$$

where π_{ij} is the true response probability for the jth individual in the ith trial, $x_{ij} = 0/1$ indicates their control/treatment group, and the set of fixed parameters ϕ_i indicates their trial membership.

An alternative approach in which trial effects on the log-odds scale are regarded as random rather than fixed is discussed in Section 6, but the appropriateness of either approach should be considered carefully. When regarding trial effects as fixed we estimate a nuisance parameter for every trial included, so reducing the information available for estimating any one model parameter; inconsistent estimates could result if sample sizes in the trials are small [32]. It has also been argued that modelling trial effects as random is inappropriate since the true treatment effects from a set of trials may not be assumed drawn at random from an underlying population. An alternative approach of conditioning on the marginals of the 2×2 tables [20] is not currently easy to implement within standard software.

4.2. Estimation of parameters

A fixed effect meta-analysis model for individual data requires only standard logistic regression, while a model incorporating random treatment effects requires specialist software. Within the MLn/MLwiN software [24, 25], the iterative estimation procedure for a random effects logistic regression model involves use of either marginal quasi-likelihood [33] (MQL) or penalized quasi-likelihood [34] (PQL), and either first-order or second-order Taylor expansion approximations. PQL produces improved estimates of variance components in mixed models, in general, while model convergence is more easily achieved with MQL [16]. However, even the optimal estimation procedure of PQL with second-order approximations may give downwardly biased estimates of the between-trial variance when the number of trials is small [35, 36] or when the probabilities of events are extreme.

To remove bias from the estimate of τ^2, the bias corrected parametric bootstrap procedure may be used [37, 38]. As a first step, a set of bootstrap replications is generated to provide estimates of the bias present in the initial estimators. Since these estimates are themselves biased [39], the bootstrap bias estimation process is repeated, using as a basis the new bias corrected estimates. The second process produces a better estimate of bias than the first, on average; the initial estimates are again corrected using the second bias estimates, and the bootstrap is applied again. Iteration between bias estimation and bias correction continues until convergence [37], at which time the parameter estimates provided are asymptotically consistent and unbiased [38]. The bias corrected parametric bootstrap procedure is directly available in the MLwiN software [25]. It is however computationally intensive and should therefore be used only in the final analyses to provide unbiased estimates.

The published summary outcome data from the respiratory tract infections example may be converted to an individual binary data format, represented as a series of zeros and ones. We thus fitted model (2) to the infections example, using ML and REML estimation; the results are presented in Table III. Identical results may be obtained by fitting model (2) to the grouped binomial data, since no individual level covariates are included. We applied the bias correction procedure within the individual data model with random treatment effects and REML estimation, using 15 sets of 800 replications.

The bias corrected estimate of τ^2 is larger than the standard REML estimate (Table III), showing that the latter was indeed downwardly biased. The ML and REML estimates of τ^2 obtained from individual data methods are larger than the corresponding estimates resulting from summary data methods. When using summary data methods, the quantities σ_i^2 (model (1)) are assumed to be known and consequently different estimates of between-trial variance τ^2 may be expected. The fixed effect and random effects estimates of θ from individual data methods all differ noticeably from the corresponding summary data estimates; each of the latter indicates a smaller treatment effect than its counterpart. The summary data methods appear to perform badly in this data set, possibly because of the correction for zero cells required [40], or the extreme response probabilities in some trials. The individual data model (2) assumes normality for the random treatment effects v_i. To gain some insight regarding the validity of this we examine the normal plot Figure 1(a) of empirical Bayes estimates [16] of v_i. We interpret Figure 1(a) with caution since the predicted residuals shown are estimated with differing precision and are therefore subject to varying degrees of shrinkage towards the mean [41]. The plot however shows no strong evidence against normality.

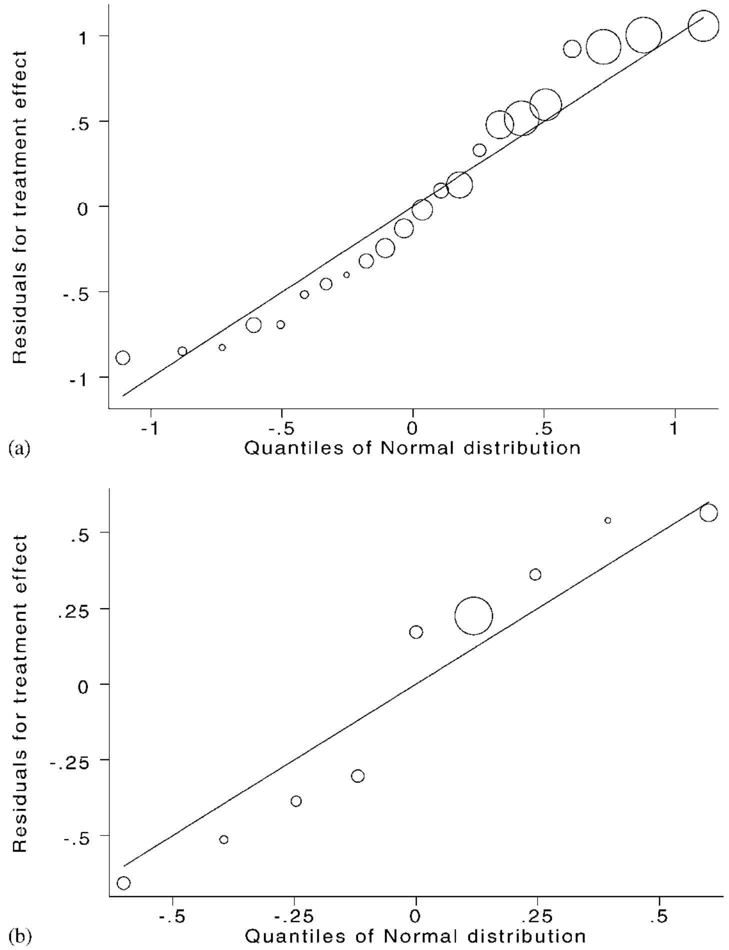

Figure 1. Between-trial residuals for treatment effect (*a*) from the respiratory tract infections data; (*b*) from the pre-eclampsia data. Area of circle inversely proportional to variance of log-odds ratio estimate.

4.3. Calculation of confidence intervals

It is not yet possible using MLn/MLwiN to obtain reliable log-likelihood values for multilevel models fitted to binary data [16], so profile likelihood based intervals are not presented when applying individual data methods. We can however construct 95 per cent confidence intervals for both θ and τ^2 using quantiles of the bootstrap distributions obtained from the parametric bootstrap with bias correction. The intervals are computed using a procedure suggested by Kuk [38] involving a scaling parameter to adjust the empirical quantiles. The bias corrected bootstrap interval for τ^2 is wider than the corresponding REML Wald confidence interval (Table III), and the proportional increase in width is greater for θ. As explained in Section 3.3, Wald intervals for θ may be too narrow and those for τ^2 are likely to be invalid. The discrepancy between the bootstrap and Wald intervals is greater here than when using summary data methods. The summary

Table IV. Estimates from meta-analysis of pre-eclampsia data, using summary data methods and individual data methods with fixed trial effects.

	Log OR (θ)	(95% CI)	Between-trial variance (τ^2)	(95% CI)
Summary data methods				
Conventional				
Fixed effect	−0.40	(−0.58, −0.22)	0	–
Random effects	−0.52	(−0.91, −0.13)	0.23	–
Multilevel modelling				
ML: (Wald)	−0.52	(−0.93, −0.11)	0.24	[0.00, 0.70)
(Profile likelihood)		(−0.98, −0.05)		(0.03, 1.13)
(Bootstrap)		(−0.89, −0.11)		(0.00, 0.62)
REML: (Wald)	−0.52	(−0.98, −0.08)	0.30	[0.00, 0.71)
(Bootstrap)		(−0.92, −0.07)		(0.00, 0.83)
Individual data methods with fixed trial effects				
Fixed effect	−0.41	(−0.59, −0.23)	0	–
Random effects				
ML: (Wald)	−0.54	(−0.91, −0.17)	0.18	[0.00, 0.42)
REML: (Wald)	−0.56	(−1.01, −0.11)	0.32	[0.00, 0.70)
(Bootstrap with bias correction)	−0.66	(−1.39, 0.06)	0.37	(0.00, 0.63)

data confidence intervals for θ are narrower than the corresponding intervals from individual data methods, reflecting the respective estimates of τ^2. For τ^2 also, the bootstrap interval from summary data is narrower than the bias corrected bootstrap interval from individual binary data. In this data set, the summary data results do not fully represent the uncertainty surrounding the estimates.

5. RESULTS FOR PRE-ECLAMPSIA DATA

Here we apply the methods discussed above to the pre-eclampsia data set (Table II), which comprises fewer trials (nine rather than 22); the results are given in Table IV. The difference between the conventional random effects and fixed effect estimates of θ is smaller in this data set than in the respiratory tract infections example; here the test for heterogeneity gives $\chi^2_8 = 27.3$, with $P < 0.001$. The regression asymmetry test [22] gives no evidence ($P = 0.41$) of publication bias for the pre-eclampsia example. When using summary data methods, the differences between ML, REML and conventional random effects estimates of θ and of τ^2 are similar to those found in analysis of the respiratory tract infections data. The three estimates of θ are very similar here and, as in the infections data, the method of moments estimate of τ^2 is lower than the REML estimate.

Within summary data analyses, the method of profile likelihood again produces the widest confidence intervals for both θ and τ^2. The bootstrap confidence intervals for θ are again narrower than the corresponding Wald intervals, but the differences in width are greater here than in the respiratory tract infections data set. The difference in width between the likelihood based and bootstrap intervals for τ^2 is also greater in this data set. The shape of the likelihood curve for τ^2 is less symmetric when the number of trials is smaller, so more caution is necessary when interpreting likelihood support intervals as confidence intervals. In the pre-eclampsia data, both ML and REML Wald confidence intervals for τ^2 had negative lower limits and so required truncation at zero. Given the necessity for truncation in Wald intervals and the interpretation difficulties involved with the likelihood interval, the bootstrap confidence interval for τ^2 should be regarded as the most reliable summary data interval in the pre-eclampsia data set.

In analysis of the individual binary data, the bias corrected estimate of τ^2 is again larger than the standard REML estimate. In this data set, the bias corrected bootstrap confidence interval for τ^2 is narrower than the corresponding REML Wald interval. Since both Wald intervals required truncation at zero, as in the analysis of summary data, the bias corrected bootstrap interval is preferred. Differences between summary data and individual binary data estimates of θ and of τ^2 are generally smaller here than in the infections example. The better performance of summary data methods in the pre-eclampsia example may be explained by the larger trial sizes and the lack of extreme probabilities of events; also this data set required no correction for zero cells. There may be some evidence against the assumption of normality for the random treatment effects, as shown in Figure 1(*b*).

6. INDIVIDUAL DATA METHODS WITH RANDOM TRIAL EFFECTS

6.1. Description of model

In the previous section, fixed trial effects were used to allow the log-odds to vary across trials in the meta-analysis of individual binary data. An alternative approach is to fit random trial effects, thereby assuming the log-odds to be drawn from a random distribution [20]. The random effects meta-analysis model with random trial effects includes the effects u_i of trial on the log-odds as well as the effects v_i of trial on treatment effect:

$$\begin{gathered}\text{logit}(\pi_{ij}) = \alpha + u_i + (\theta + v_i)x_{ij} \\ u_i \sim \text{N}(0, \sigma^2), \quad v_i \sim \text{N}(0, \tau^2), \quad \text{cov}(u_i, v_i) = \rho\sigma\tau\end{gathered} \tag{3}$$

As before, we initially code $x_{ij} = 0/1$ to indicate control/treatment group. It is important to include the covariance between the u_i and the v_i. When modelling random treatment and trial effects, we are implicitly modelling the variance-covariance matrix associated with the bivariate log-odds parameter. If $\text{cov}(u_i, v_i)$ is assumed to be zero, the between-trial variance of the log-odds across control groups is modelled by σ^2, while that across intervention groups is modelled by $\sigma^2 + \tau^2$ (Table V). The variation across trials for control groups is thereby forced to be less than or equal to the variation across trials for intervention groups; this assumption may well be inappropriate. Furthermore, the covariance between control group and intervention group log-odds is assumed to be equal to the between-trial variance of the log-odds in the control groups (Table V). When $\text{cov}(u_i, v_i)$ is estimated rather than assumed to be zero, the variance-covariance matrix of the bivariate log-odds parameter estimates is modelled freely by combinations of the three parameters σ, τ and ρ (Table V).

Using second-order PQL methods in MLwiN for the respiratory tract infections example, the estimated variance-covariance matrix changes markedly when $\text{cov}(u_i, v_i)$ is fitted (Table VI). With the constraints removed, the variation across trials for control groups is estimated as substantially larger than the variation across trials for intervention groups. This example illustrates the dangers of the zero $\text{cov}(u_i, v_i)$ model, in which estimates of θ and τ^2 may be based on invalid assumptions. The zero and non-zero $\text{cov}(u_i, v_i)$ models give different estimates of θ and τ^2.

Table V. Variance-covariance matrices for the bivariate log-odds parameter, corresponding to codings of 0/1 or ±1/2 combined with zero or non-zero covariance terms.

	$x = 0/1$		$x = \pm 1/2$	
	Treatment A	Treatment B	Treatment A	Treatment B
Zero covariance				
Treatment A (or control)	σ^2	σ^2	$\sigma^2 + \tau^2/4$	$\sigma^2 - \tau^2/4$
Treatment B (or active)	σ^2	$\sigma^2 + \tau^2$	$\sigma^2 - \tau^2/4$	$\sigma^2 + \tau^2/4$
Non-zero covariance				
Treatment A (or control)	σ^2	$\sigma^2 + \rho\sigma\tau$	$\sigma^2 - \rho\sigma\tau + \tau^2/4$	$\sigma^2 - \tau^2/4$
Treatment B (or active)	$\sigma^2 + \rho\sigma\tau$	$\sigma^2 + 2\rho\sigma\tau + \tau^2$	$\sigma^2 - \tau^2/4$	$\sigma^2 + \rho\sigma\tau + \tau^2/4$

Table VI. Estimates from individual data random effects meta-analysis of respiratory tract infections data, with random trial effects.

	Zero covariance, with 0/1 coding	Zero covariance, with ±1/2 coding	Non-zero covariance
Log OR	−1.40	−1.40	−1.36
Variance of log OR	0.48	0.56	0.56
Bivariate log-odds	$(-2.05, -0.65)$	$(-2.03, -0.63)$	$(-1.97, -0.61)$
Variance of bivariate log-odds	$\begin{pmatrix} 1.376 & 1.376 \\ 1.376 & 1.858 \end{pmatrix}$	$\begin{pmatrix} 1.301 & 1.021 \\ 1.021 & 1.301 \end{pmatrix}$	$\begin{pmatrix} 1.753 & 1.022 \\ 1.022 & 0.848 \end{pmatrix}$

6.2. Bivariate approach

Estimation of model (3) involves use of the full likelihood for binomial data, and the set of control group and intervention group log-odds are assumed to follow a bivariate normal distribution. This model is a multilevel representation of the bivariate approach proposed by van Houwelingen *et al.* [20], in which maximum likelihood estimates were obtained using exact likelihood and two approximate procedures (of which the second was advocated). Here we use penalized quasi-likelihood and second-order Taylor approximations within the MLn/MLwiN software. To demonstrate the similarity between results from our approach and the exact likelihood approach, we present estimates (Table VII) from the upper gastrointestinal bleeding data set used by van Houwelingen *et al.* The estimates obtained from multilevel modelling and exact likelihood are very similar; the differences are of like magnitude to those observed between the exact likelihood approach and the recommended approximate approach [20].

6.3. Coding of treatment covariate

Treatment covariates are usually in practice coded as 0/1. The coding of ±1/2 may however be advantageous when fitting a random effects meta-analysis model with random trial effects in data sets with few degrees of freedom, where estimation of a covariance between two random effects is problematic or impossible. When assuming $\text{cov}(u_i, v_i)$ to be zero and using ±1/2 coding, the variance of the log-odds in control group patients is modelled as equal to that in intervention group patients, and the covariance between control group and intervention group log-odds is modelled

Table VII. Estimates from bivariate random effects meta-analysis of upper gastrointestinal bleeding data, comparing a multilevel approach to those of van Houwelingen (VH) *et al.* [20].

	Multilevel	VH exact	VH approximate
Log OR	−0.17	−0.17	−0.17
Variance of log OR	0.11	0.12	0.11
Bivariate log-odds	$(-1.36, -1.19)$	$(-1.35, -1.19)$	$(-1.34, -1.17)$
Variance of bivariate log-odds	$\begin{pmatrix} 0.123 & 0.068 \\ 0.068 & 0.127 \end{pmatrix}$	$\begin{pmatrix} 0.126 & 0.067 \\ 0.067 & 0.129 \end{pmatrix}$	$\begin{pmatrix} 0.122 & 0.067 \\ 0.067 & 0.122 \end{pmatrix}$

separately (Table V). In the respiratory tract infections example, it can be seen that the difference between the estimated variance-covariance matrices from models with zero and non-zero $\text{cov}(u_i, v_i)$ is smaller when the $\pm 1/2$ coding is used (Table VI). In the pre-eclampsia data it was not possible to achieve convergence when $\text{cov}(u_i, v_i)$ was included, owing to the small number of trials. In such situations, the most appropriate model has to be simplified, and a zero $\text{cov}(u_i, v_i)$ model with $\pm 1/2$ coding may be preferable to a zero $\text{cov}(u_i, v_i)$ model with the usual 0/1 coding.

6.4. Model checking

In addition to the assumption of normality for the random treatment effects v_i as required by model (2), model (3) assumes the random trial effects u_i to be normally distributed. In Figure 2(*a*) we see some evidence against this latter assumption for the infections example; two trials have particularly large residuals. When non-normality occurs, the individual data model (2) with fixed trial effects would be more appropriate than model (3). Figure 2(*b*) gives us some insight into the relationship between the u_i and the v_i; we see that more negative treatment effects were observed in trials with higher than average risk of respiratory tract infection. This further demonstrates the need for including the (u_i, v_i) covariance in model (3), and represents a relationship between underlying risk and the extent of treatment benefit [10, 42].

7. DISCUSSION

In this paper we have focused on the application of a multilevel model framework to meta-analysis of binary outcomes, but corresponding methods exist for other outcomes. Multilevel modelling techniques for meta-analysis of continuous outcomes are now well developed; these include methods for combining individual patient data and summary data [18]. The models presented for meta-analysis of binary outcomes can be adapted for application to ordinal outcomes, again for either summary data or individual patient data. It is also possible to perform meta-analysis of survival outcomes using multilevel models [16]. Models for random effects meta-analysis of multiple correlated continuous outcomes were described recently [14]. Maximum likelihood or REML estimates for such models could be obtained within the multilevel modelling framework [43], with the possibility of analysing multiple non-continuous outcomes or even mixed outcomes, for example, one binary and one continuous outcome.

Multilevel modelling methods extend naturally to include both individual-level and trial-level covariates. The meta-regression approach described by Berkey *et al.* [5] may be implemented,

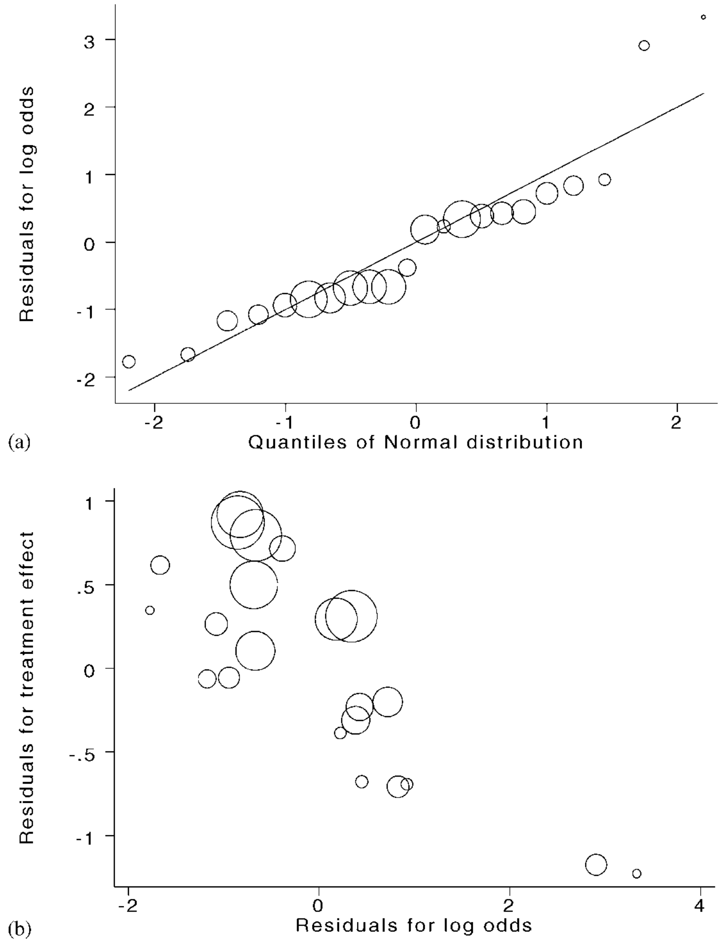

Figure 2. Between-trial residuals (*a*) for log-odds of respiratory tract infection; (*b*) for treatment effect plotted against residuals for log-odds. Area of circle inversely proportional to variance of log-odds ratio estimate.

allowing the investigation of sources of between-trial heterogeneity with the aid of trial-level covariates [44]. In this paper we describe models with only two levels, corresponding to individuals within trials. The three-level model is an obvious extension of these. Three-level models would be a valuable tool for meta-analysis of multi-centre trials, meta-analysis of cluster randomized trials or meta-analysis of trials with repeated measures. A variation on the three-level model is the cross-classified multilevel model; this would be of use if, for example, the trials under analysis had drawn subjects from multiple centres, while some centres had also contributed subjects to several trials. In such situations the between-trial variation is crossed with centres; this can be modelled using the MLwiN software [25].

The methods described in the paper should be possible within any specialized multilevel software, together with the extensions mentioned above, although the bias corrected parametric bootstrap may not yet be widely available. Non-Bayesian approaches to estimation of hierarchical models for meta-analysis have been proposed previously [7, 8, 45]; in this paper we have concentrated on methods which are easily implemented in multilevel software. The models for both summary data and individual data have been based around the log-odds ratio scale throughout; analyses may in principle be carried out on other scales such as log relative risk or absolute risk difference, though this would be computationally less straightforward.

Bayesian hierarchical modelling provides an alternative framework for meta-analysis. Carlin [46] considered a semi-Bayesian random effects model for summary data analysis, which uses numerical integration to allow for the uncertainty in estimation of τ^2 but regards the σ_i^2 as known quantities. Smith *et al.* [6] later described a fully Bayesian approach, in which Gibbs sampling is used to perform random effects meta-analysis of individual binary data. This method enables flexibility in modelling, for example, allowing the inclusion of trial- and individual-level covariates. Pauler and Wakefield [15] give a thorough description of the Bayesian framework for meta-analysis and its implementation. Classical multilevel modelling and Bayesian hierarchical modelling offer similar possibilities to the meta-analyst; one advantage of the latter is its natural ability to utilize information from previous studies and thus improve precision in estimation of τ^2 [47].

Our main example, the respiratory tract infections data set, showed evidence of publication bias but we have not considered correction for this within the multilevel modelling techniques presented. Models proposed to correct for publication bias in estimates of treatment effect generally assign to each study a weight which is a function of the selection probability for that study [48]. Such models require assumptions about the specific form taken by the selection probabilities, and may involve rather arbitrary decisions to which robustness is lacking [49]. Copas [50] has recently recommended a sensitivity approach to the problem of publication bias, as an alternative to obtaining corrected estimates. The proposed method involves examination of the extent to which the estimation of θ depends on parameters describing the selection probabilities. This procedure yields a range of plausible estimates of θ rather than a single corrected estimate.

The multilevel models approach to meta-analysis encompasses standard methods, while its flexibility offers a wealth of extensions. In this paper we have demonstrated the essential techniques for meta-analysis of binary data using multilevel modelling. Having once performed these methods in a multilevel software package, the implementation of many of the extensions mentioned is straightforward. Further work is required, however, to investigate the full potential of multilevel models for more complex extensions such as, for example, mixed multivariate outcomes.

ACKNOWLEDGEMENTS

This work was funded by the Economic and Social Research Council under the programme for the Analysis of Large and Complex Datasets. We are grateful for the helpful comments of two referees.

REFERENCES

1. Thompson SG. Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. *Statistical Methods in Medical Research* 1993; **2**:173–192.
2. Thompson SG. Meta-analysis of clinical trials. In *Encyclopedia of Biostatistics*, Armitage P, Colton T (eds). Wiley: New York, 1998; 2570–2579.
3. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* 1996; **15**:619–629.
4. Biggerstaff BJ, Tweedie RL. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine* 1997; **16**:753–768.
5. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Statistics in Medicine* 1995; **14**:395–411.

6. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* 1995; **14**:2685–2699.
7. Stram DO. Meta-analysis of published data using a linear mixed-effects model. *Biometrics* 1996; **52**:536–544.
8. Breslow N, Leroux B, Platt R. Approximate hierarchical modelling of discrete data in epidemiology. *Statistical Methods in Medical Research* 1998; **7**:49–62.
9. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal* 1994; **309**:1351–1355.
10. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine* 1997; **16**:2741–2758.
11. Walter SD. Variation in baseline risk as an explanation of heterogeneity in meta-analysis. *Statistics in Medicine* 1997; **16**:2883–2900.
12. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine* 1998; **17**: 841–856.
13. Matsuyama Y, Sakamoto J, Ohashi Y. A Bayesian hierarchical survival model for the institutional effects in a multi-centre cancer clinical trial. *Statistics in Medicine* 1998; **17**:1893–1908.
14. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine* 1998; **17**:2537–2550.
15. Pauler DK, Wakefield JC. Modeling and implementation issues in Bayesian meta-analysis. In *Meta-Analysis in Medicine and Health Policy*, Stangl DK Berry DA, (eds). Marcel Dekker: New York, 2000.
16. Goldstein H. *Multilevel Statistical Models*. Edward Arnold: London, 1995.
17. Stewart LA, Parmar MKB. Meta-analysis of the literature or of individual patient data: is there a difference? *Lancet* 1993; **341**:418–422.
18. Goldstein H, Yang M, Omar RZ, Turner RM, Thompson SG. Meta-analysis using multilevel models with an application to the study of class size effects. *Journal of the Royal Statistical Society*, *Series C* 2000 (in press).
19. Thompson SG, Pocock SJ. Can meta-analyses be trusted? *Lancet* 1991; **338**:1127–1130.
20. van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. *Statistics in Medicine* 1993; **12**:2273–2284.
21. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
22. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* 1997; **315**:629–634.
23. Lambert PC, Abrams KR. Meta-analysis using multilevel models. *Multilevel Modelling Newsletter* 1995; **7**:17–19.
24. Woodhouse G, Rasbash J, Goldstein H, Yang M, Plewis I. *Multilevel Modelling Applications*: *A Guide for Users of MLn*. Institute of Education: London, 1996.
25. Goldstein H, Rasbash J, Plewis I, Draper D, Browne W, Yang M, Woodhouse G, Healy MJR. *A User's Guide to MLwiN*. Institute of Education: London, 1998.
26. Brown HK, Kempton RA, The application of REML in clinical trials. *Statistics in Medicine* 1994; **13**:1601–1617.
27. Cox DR, Oakes D. *Analysis of Survival Data*. Chapman and Hall: London, 1984; 35–36.
28. Clayton D, Hills M. *Statistical Models in Epidemiology*. Oxford University Press: Oxford, 1993; 89–91.
29. Welham SJ, Thompson R. Likelihood ratio tests for fixed model terms using residual maximum likelihood. *Journal of the Royal Statistical Society*, *Series B* 1997; **59**:701–714.
30. Morrell CH. Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics* 1998; **54**:1560–1568.
31. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman and Hall: New York, 1993.
32. Breslow NE, Day NE. *Statistical Methods in Cancer Research. Volume 1 - The Analysis of Case-control Studies*. IARC: Lyon, 1980; 249.
33. Goldstein H. Nonlinear multilevel models with an application to discrete response data. *Biometrika* 1991; **78**:45–51.
34. Breslow NE, Clayton DG. Approximate inference in generalised linear mixed models. *Journal of the American Statistical Association* 1993; **88**:9–25.
35. Goldstein H, Rasbash J. Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society*, *Series A* 1996; **159**:505–513.
36. Rodriquez G, Goldman N. An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society*, *Series A* 1995; **158**:73–89.
37. Goldstein H. Consistent estimators for multilevel generalised linear models using an iterated bootstrap. *Multilevel Modelling Newsletter* 1996; **8**:3–6.
38. Kuk AYC. Asymptotically unbiased estimation in generalised linear models with random effects. *Journal of the Royal Statistical Society*, *Series B* 1995; **57**:395–407.
39. Davison AC, Hinkley DV. *Bootstrap Methods and their Application*. Cambridge University Press: Cambridge, 1997; 103–105.
40. Cox DR, Snell EJ. *Analysis of Binary Data*. Chapman and Hall: London, 1989; 31–32.
41. Langford IH, Lewis T. Outliers in multilevel data. *Journal of the Royal Statistical Society*, *Series A* 1998; **161**: 121–160.

42. McIntosh M. The population risk as an explanatory variable in research synthesis of clinical trials. *Statistics in Medicine* 1996; **15**:1713–1728.
43. Beacon HJ, Thompson SG. Multi-level models for repeated measurement data: application to quality of life data in clinical trials. *Statistics in Medicine* 1996; **15**:2717–2732.
44. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine* 1999; **18**:2693–2708.
45. Aitkin M. Meta-analysis by random-effect modelling in generalized linear models. *Statistics in Medicine* 1999; **18**:2343–2351.
46. Carlin JB. Meta-analysis for 2 × 2 tables: a Bayesian approach. *Statistics in Medicine* 1992; **11**:141–158.
47. Higgins JPT, Whitehead A. Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine* 1996; **15**:2733–2749.
48. Vevea JL, Hedges LV. A general linear model for estimating effect size in the presence of publication bias. *Psychometrika* 1995; **60**:419–435.
49. Little RJA. A note about models for selectivity bias. *Econometrica* 1985; **53**:1469–1474.
50. Copas J. What works?: selectivity models and meta analysis. *Journal of the Royal Statistical Society*, *Series A* 1999; **162**:95–109.