Journal of the
Royal Statistical Society
SERIES A
Statistics
in Society

J. R. Statist. Soc. A (2009)
172, *Part* 1, *pp.* 119–136

Models for potentially biased evidence in meta-analysis using empirically based priors

N. J. Welton and A. E. Ades,

University of Bristol, UK

J. B. Carlin,

Murdoch Children's Research Institute and University of Melbourne, Australia

D. G. Altman

Centre for Statistics in Medicine, Oxford, UK

and J. A. C. Sterne

University of Bristol, UK

[Received April 2007. Revised February 2008]

Summary. We present models for the combined analysis of evidence from randomized controlled trials categorized as being at either low or high risk of bias due to a flaw in their conduct. We formulate a bias model that incorporates between-study and between-meta-analysis heterogeneity in bias, and uncertainty in overall mean bias. We obtain algebraic expressions for the posterior distribution of the bias-adjusted treatment effect, which provide limiting values for the information that can be obtained from studies at high risk of bias. The parameters of the bias model can be estimated from collections of previously published meta-analyses. We explore alternative models for such data, and alternative methods for introducing prior information on the bias parameters into a new meta-analysis. Results from an illustrative example show that the bias-adjusted treatment effect estimates are sensitive to the way in which the meta-epidemiological data are modelled, but that using point estimates for bias parameters provides an adequate approximation to using a full joint prior distribution. A sensitivity analysis shows that the gain in precision from including studies at high risk of bias is likely to be low, however numerous or large their size, and that little is gained by incorporating such studies, unless the information from studies at low risk of bias is limited. We discuss approaches that might increase the value of including studies at high risk of bias, and the acceptability of the methods in the evaluation of health care interventions.

Keywords: Bayesian methods; Bias; Health technology assessment; Markov chain Monte Carlo methods; Randomized controlled trials

1. Introduction

Various studies have provided empirical evidence that specific flaws in the conduct of randomized controlled trials may bias estimates of treatment effects (Gluud, 2006). In particular, there is good evidence that failure to conceal randomized allocation at the time of patient recruitment, and the lack of double blinding, are associated with exaggeration of treatment effect estimates (Schulz *et al.*, 1995; Moher *et al.*, 1998; Egger *et al.*, 2003; McAuley *et al.*, 2000; Kjaergard *et al.*,

Address for correspondence: N. J. Welton, Academic Unit of Primary Health Care, Department of Community Based Medicine, University of Bristol, Cotham House, Cotham Hill, Bristol, BS6 6JL, UK.
E-mail: Nicky.Welton@bristol.ac.uk

 0964–1998/09/172119

2001; Wood *et al.*, 2008). Trial characteristics, such as the adequacy of allocation concealment or blinding, are usually treated as binary indicators of a high or low risk of bias. Trials with flaws in their conduct are commonly reported in the medical literature and often represent a substantial proportion of the evidence that is included in systematic reviews (Egger *et al.*, 2003). Although multiple analyses including or omitting the high risk evidence may be conducted and reported, meta-analyses are increasingly used in decision analysis, where a single 'best' estimate must be reported. Meta-analysts are then faced with a choice about whether one should take a 'best available evidence' approach by restricting attention to the trials at low risk of bias, or an 'all available evidence' approach, in which all trials are included.

The all available evidence approach is in the spirit of organizations such as the UK National Institute of Clinical Excellence (NICE), where the focus is on a decision analysis that reflects the body of evidence that is available at the current moment in time. However, such an approach raises methodological issues in the formulation of a model to account for potential bias. If trials at high risk of bias are to be included in a meta-analysis, then questions are also raised about appropriate inclusion–exclusion criteria in the process of systematic review and extraction of data.

Proposed methods for inclusion of potentially biased evidence have focused either on *down-weighting* studies with high risk of bias in the synthesis of the evidence (Begg and Pilote, 1991; Li and Begg, 1994; Larose and Dey, 1997; Prevost *et al.*, 2000; Spiegelhalter and Best, 2003) or on detailed modelling of study-specific biases that is based on characteristics of individual studies, which are then used to adjust observed treatment effects study by study *before* synthesis of the evidence (Eddy *et al.*, 1992; Wolpert and Mengersen, 2004; Greenland, 2005). Here we consider an alternative framework in which we adjust for expected bias as well as downweighting studies at high risk of bias, within a Bayesian paradigm. Meta-epidemiological studies (in which a collection of meta-analyses provides evidence on the association of study characteristics with treatment effect estimates) are used to provide empirically based prior information on the degree of bias that can be expected from studies at high risk of bias, the heterogeneity in bias between studies within a particular meta-analysis and the additional heterogeneity in mean bias between meta-analyses.

The paper is organized as follows. We first define our bias model for combining trials at low and high risk of bias, and obtain results in algebraic form, providing insights on the informational content of trials at high risk of bias. We then show how the parameters of the bias model can be estimated from meta-epidemiological data (Schulz *et al.*, 1995) and investigate various extensions. Next we consider various ways in which the outputs from the meta-epidemiological analysis can be used to introduce prior information on bias parameters in a new meta-analysis. We apply the model to an example meta-analysis of Clozapine *versus* neuroleptic medication for treatment of schizophrenia, to draw specific conclusions on the relative value of trials with adequate and inadequate allocation concealment in that area. We present sensitivity analyses investigating how the final estimate of the treatment effect and its precision depend on bias parameter inputs. We end with a discussion of the various modelling assumptions that are being made, how acceptable such an approach is likely to be to a national decision maker such as the NICE and what further work needs to be done before such models can be confidently used in practice, and finally we discuss our methods in the context of other approaches.

2. Model for combining adequately and inadequately conducted trials in a single meta-analysis

2.1. Bias model

In a given single meta-analysis m, suppose that studies are classified as being either at low

(L-studies) or high (H-studies) risk of bias due to a specific flaw in their conduct, such that there are $n_{\mathrm{L},m}$ L-studies and $n_{\mathrm{H},m}$ H-studies. Each study i provides a summary treatment effect estimate $y_{i,m}$, with standard error $\sigma_{i,m}$ (where $i = 1, \ldots, n_{\mathrm{L},m}$ indexes the L-studies and $i = (n_{\mathrm{L},m} + 1), \ldots, (n_{\mathrm{L},m} + n_{\mathrm{H},m})$ the H-studies, and m indexes the meta-analysis of current interest). We assume that the L-studies provide an unbiased estimate of a (fixed) true treatment effect of interest, which is denoted d_m, and the H-studies estimate this same treatment effect with a study-specific bias, $\beta_{i,m}$. To obtain analytical results, we assume that the treatment effect is measured on a continuous scale, e.g. log-odds-ratios for binary outcomes, with an, at least approximately, normal likelihood, so we have the basic models

$$y_{i,m} \sim N(d_m, \sigma^2_{i,m}) \qquad i = 1, \ldots, n_{\mathrm{L},m}, \tag{1}$$

$$y_{i,m} \sim N(d_m + \beta_{i,m}, \sigma^2_{i,m}) \qquad i = (n_{\mathrm{L},m} + 1), \ldots, (n_{\mathrm{L},m} + n_{\mathrm{H},m}). \tag{2}$$

The sampling variances $\sigma^2_{i,m}$ are considered to be known, on the basis that they are usually well estimated from the data in each meta-analysis. The fixed underlying treatment effect is given a flat normal prior distribution, $d_m \sim N(0, 100^2)$.

We put a hierarchical model on the study-specific biases that captures the nature of the empirical evidence that is available to inform these parameters:

$$\beta_{i,m} \sim N(b_m, \kappa^2) \qquad i = (n_{\mathrm{L},m} + 1), \ldots, (n_{\mathrm{L},m} + n_{\mathrm{H},m}), \tag{3}$$

$$b_m \sim N(b_0, \varphi^2), \tag{4}$$

$$b_0 \sim N(B_0, V_0). \tag{5}$$

H-studies in meta-analysis m have overall meta-analysis-specific mean bias b_m, and between-study within-meta-analysis variance κ^2. We make the (rather strong—see Section 7) assumption that mean bias b_m in meta-analysis m is exchangeable with the mean bias from other meta-analyses, with common mean bias across all relevant meta-analyses b_0 and between-meta-analysis variance in mean bias φ^2. The mean bias b_0 itself is uncertain with expectation B_0 and variance V_0. Note that V_0 represents *uncertainty*, which can be reduced by further information, whereas κ^2 and φ^2 are measures of intrinsic *variation*. In Section 3 we show how estimates of κ^2, φ^2, B_0 and V_0 can be obtained from meta-epidemiological data.

This formulation is sufficiently simple to allow us to obtain some analytical results on the posterior distribution for the true treatment effect d_m, while still capturing the heterogeneity and uncertainties that are inherent in the evidence base. This approach provides insights into the key determinants of the posterior for d_m and facilitates examination of the sensitivity of the results to various model inputs. However, when we consider extensions to this basic model, we use Markov chain Monte Carlo (MCMC) simulation to obtain results.

2.2. Posterior distribution for d_m in a single meta-analysis m

In what follows we drop the subscript m for compactness. Following Gelman *et al.* (2003) (section 15.3), we can view our hierarchical prior structure as additional data and write this hierarchical linear model in the form of a single likelihood:

$$y|X, \gamma, \Sigma \sim N(X\gamma, \Sigma)$$

where

$$y = \begin{pmatrix} \left.\begin{matrix} y_1 \\ y_2 \\ \vdots \\ y_{n_L} \end{matrix}\right\} n_L \\ \left.\begin{matrix} y_{(n_L+1)} \\ y_{(n_L+2)} \\ \vdots \\ y_{(n_L+n_H)} \end{matrix}\right\} n_H \\ \left.\begin{matrix} 0 \\ 0 \\ \vdots \\ 0 \end{matrix}\right\} n_H \\ 0 \\ B_0 \end{pmatrix}, \qquad X = \begin{pmatrix} \left.\begin{matrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 & 0 & 0 \end{matrix}\right\} n_L \\ \left.\begin{matrix} 1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 & 0 & 0 \end{matrix}\right\} n_H \\ \left.\begin{matrix} 0 & 1 & 0 & \cdots & 0 & -1 & 0 \\ 0 & 0 & 1 & \cdots & 0 & -1 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 & 0 \end{matrix}\right\} n_H \\ \begin{matrix} 0 & 0 & 0 & \cdots & 0 & 1 & -1 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{matrix} \end{pmatrix},$$

$$\gamma = \begin{pmatrix} d \\ \left.\begin{matrix} \beta_{(n_L+1)} \\ \vdots \\ \beta_{(n_L+n_H)} \end{matrix}\right\} n_H \\ b \\ b_0 \end{pmatrix}, \qquad \Sigma = \operatorname{diag}\begin{pmatrix} \left.\begin{matrix} \sigma_1^2 \\ \vdots \\ \sigma_{n_L}^2 \end{matrix}\right\} n_L \\ \left.\begin{matrix} \sigma_{(n_L+1)}^2 \\ \vdots \\ \sigma_{(n_L+n_H)}^2 \end{matrix}\right\} n_H \\ \left.\begin{matrix} \kappa^2 \\ \vdots \\ \kappa^2 \end{matrix}\right\} n_H \\ \varphi^2 \\ V_0 \end{pmatrix}.$$

The first $n_L + n_H$ rows of data vector y and matrix X simply pick out the relevant likelihoods, conditional on $\{d, \beta_i, \sigma_i^2\}$, for the observed low and high risk studies (equations (1) and (2)). The following n_H rows represent the hierarchical model for bias between studies, within meta-analysis (equation (3)). The 'observed' data are set equal to 0. This gives the correct mean, $E[\beta_i - b] = 0$, while reflecting the between-study, within-meta-analysis variance κ^2. The next row represents between meta-analysis variation (equation (4)). Again the 'observed' data are set equal to 0 to give the correct mean, $E[b - b_0] = 0$, while reflecting the between meta-analysis variation φ^2. The final row represents the prior for mean bias across meta-analyses (equation (5)). The observed data are set to B_0 to give the correct mean, $E[b_0] = B_0$, while reflecting the uncertainty in this, V_0.

The variance matrix Σ is assumed known, and so the posterior for parameters γ can be obtained by weighted least squares regression (Lindley and Smith, 1972; Gelman *et al.*, 2003):

$$\gamma | y, X, \Sigma \sim N\{(X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}y, (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}\}.$$

For our model, the marginal posterior for d (the first element of γ) can be found in closed form:

$$E[d|\{y_i,\sigma_i^2\}]=\frac{\sum_{i=1}^{n_L} y_i/\sigma_i^2+(1/w)\sum_{i=n_L+1}^{n_L+n_H}(y_i-B_0)/(\sigma_i^2+\kappa^2)}{\sum_{i=1}^{n_L} 1/\sigma_i^2+(1/w)\sum_{i=n_L+1}^{n_L+n_H}1/(\sigma_i^2+\kappa^2)},$$

$$\mathrm{var}(d|\{y_i,\sigma_i^2\})=\left\{\sum_{i=1}^{n_L} 1/\sigma_i^2+(1/w)\sum_{i=n_L+1}^{n_L+n_H}1/(\sigma_i^2+\kappa^2)\right\}^{-1}, \quad (6)$$

where

$$w=1+\sum_{i=n_L+1}^{n_L+n_H}(V_0+\varphi^2)/(\sigma_i^2+\kappa^2).$$

The posterior mean treatment effect is a weighted average of the L- and H-studies. The L-studies are taken at face value with inverse variance weights. Individual H-studies are weighted by the inverse of the variance of the estimate plus between-study within-meta-analysis variance κ^2. However, this is also multiplied by the inverse of a second weighting factor w, which increases with the ratio of uncertainty in the meta-analysis mean bias $V_0+\varphi^2$ to between-study, within-meta-analysis uncertainty in estimated treatment effects $\sigma_i^2+\kappa^2$. So, if the uncertainty in the meta-analysis mean bias is large compared with the estimation uncertainty and between-study heterogeneity, then H-studies are given less weight. In other words, if we do not know the bias adjustment to make, then an H-study cannot tell us much about the treatment effect however large it is, or however many such studies there are.

Sterne *et al.* (2008) have described some properties of the posterior distribution (equations (6)). In particular, the informational content of H-studies is limited. If there is a single H-study, then even if that trial is very large it is still downweighted by $V_0+\varphi^2+\kappa^2$. Even if there are infinitely many H-studies, the posterior variance still depends on the uncertainty in the meta-analysis mean bias $V_0+\varphi^2$. Note that, for this model to revert to a standard fixed effect meta-analysis that treats all high and low risk evidence at face value, we need to assume that all of $V_0=\varphi^2=\kappa^2=B_0=0$.

Both to adjust for and to downweight H-studies properly, we therefore need information on variance parameters V_0, φ^2 and κ^2, as well as on mean bias b_0. Whereas V_0, κ^2 and b_0 can (in theory) be estimated from a single meta-analysis, φ^2 cannot. However, all parameters can be estimated from meta-epidemiological studies (Sterne *et al.*, 2008).

3. Meta-epidemiological analysis to estimate inputs to bias model

3.1. Models for meta-epidemiological analysis

We now show how to estimate B_0, V_0, φ^2 and κ^2 from data. Schulz *et al.* (1995) analysed a data set consisting of 33 meta-analyses, $m=1,\ldots,33$, where each study was characterized according to whether concealment of allocation was adequate or inadequate. There were 250 trials in all: 79 adequately concealed (L-studies) and 171 inadequately concealed (H-studies).

Sterne *et al.* (2008) have presented an analysis of the Schulz data (model M1) that was based on a standard model for Bayesian meta-analysis (Smith *et al.*, 1995). They assumed a fixed treatment effect in which d_m denotes the true treatment effect in all trials that are included in meta-analysis m, regardless of allocation concealment. The outcome $r_{a,i,m}$ for arm a of trial i in meta-analysis m is assumed to have a binomial likelihood (for given denominator $n_{a,i,m}$):

$$r_{a,i,m}\sim\mathrm{binomial}(p_{a,i,m},n_{a,i,m}).$$

The probability of success, $p_{a,i,m}$, is modelled by a logistic regression:

$$\text{logit}(p_{a,i,m}) = \begin{cases} \mu_{i,m} & a \text{ the control arm,} \\ \mu_{i,m} + d_m + \beta_{i,m} X_{i,m} & a \text{ the treatment arm,} \end{cases} \quad (7)$$

where $\mu_{i,m}$ is the log-odds of success in the control arm, d_m is the treatment effect, $X_{i,m}$ is an indicator of allocation concealment ($X_{i,m} = 1$, inadequate; $X_{i,m} = 0$, adequate) and $\beta_{i,m}$ is the bias in treatment effect in study i of meta-analysis m. This model simply generalizes the approximate normal likelihood model for $y_{i,m}$ that was used in the previous section, to allow for the binomial variation and to estimate the base risk as well as the odds ratio. The model for bias is exactly the same as set out in equations (1)–(4); however, in the meta-epidemiological analysis we put a flat prior on the mean bias over meta-analyses b_0. The resulting posterior (with mean B_0 and variance V_0) then forms the prior for b_0 in equation (5) for a future meta-analysis.

$N(0,100^2)$ priors were given to $\mu_{i,m}$, d_m and mean bias b_0. The variance parameters were given uniform(0,10) priors on the standard deviation scale. We used uniform priors for standard deviations as suggested by Gelman (2006) and chose the range (0,10) to be large on a log-odds scale. Results were fairly robust to changes in these priors. Posterior summaries from this model were obtained by using MCMC simulation implemented in the WinBUGS 1.4.1 software (Spiegelhalter *et al.*, 2000). In all the results that are presented, two chains were run until convergence (25000 iterations) according to the Brooks–Gelman–Rubin diagnostic tool (Brooks and Gelman, 1998). These 'burn-in' simulations were then discarded and a further 50000 iterations run on which all inference is based. Reparameterizing so that the hierarchical models are centred on zero may improve the speed of convergence.

We extend model M1 in three ways. We may expect that κ^2 varies between meta-analyses, and so a natural extension is to estimate a hierarchical gamma distribution for between-study, within-meta-analysis, precisions, $1/\kappa_m^2 \sim \text{gamma}(\eta\lambda, \lambda)$, with flat gamma priors on the common parameters η and λ. Model M2 extends model M1 to incorporate this random-effects distribution for $1/\kappa_m^2$. Posterior median values for η and λ can then inform a gamma prior for $1/\kappa_m^2$ in a new meta-analysis.

Model M3 extends M1 by incorporating meta-analysis-specific random-effects distributions for the treatment effects, so the fixed treatment effect d_m is replaced by $\delta_{i,m} \sim N(d_m, \tau_m^2)$, where the mean effects in each meta-analysis, d_m, are given flat normal priors and τ_m are given uniform(0,10) priors.

Finally, model M4 incorporates both meta-analysis-specific random-effects distributions for the treatment effect and a random-effects distribution for $1/\kappa_m^2$.

The WinBUGS code for all the models that are presented here can be downloaded from `http://www.bristol.ac.uk/cobm/research/mpes`.

3.2. Results of the meta-epidemiological analyses

For model M1 the mean bias b_0 had a posterior mean of -0.47 with posterior standard deviation 0.095 (Table 1). These outputs provide estimates of the parameters in equation (5), i.e. $\hat{B}_0 = -0.47$ and $\hat{V}_0 = 0.095^2$. We summarize the variance parameters with their posterior median values. The posterior median of the between-study within-meta-analysis standard deviation κ was 0.49, and the between-meta-analysis standard deviation φ was 0.26. In Section 5 we compare various approaches to using the posteriors from the Schulz analysis to inform priors for use in a new meta-analysis where studies are characterized according to whether randomization allocation concealment was adequate or inadequate.

Table 1 also shows posterior summaries for models M2–M4. Estimated mean bias $\hat{B}_0$ is reasonably robust to the choice of model. However, the estimates for both between-meta-

Table 1. Posterior summaries from models M1–M4 described in Section 3 for the meta-epidemiological analysis applied to the Schulz data†

Parameter	*Mean*	*Standard deviation*	*Median*	*95% credible interval*
Model M1: Schulz analysis (fixed treatment effect; κ^2 fixed)				
b_0	−0.47	0.095	−0.47	(−0.65, −0.28)
φ	0.26	0.131	0.26	(0.02, 0.52)
κ	0.50	0.062	0.49	(0.38, 0.62)
Model M2: Schulz analysis (fixed treatment effect; $1/\kappa_m^2 \sim gamma(\eta\lambda, \lambda)$)				
b_0	−0.45	0.094	−0.45	(−0.64, −0.27)
η	7.28	3.892	6.26	(3.17, 17.27)
λ	1.38	3.138	0.41	(0.06, 10.04)
φ	0.24	0.131	0.24	(0.01, 0.51)
κ_m	0.50	0.879	0.44	(0.19, 1.13)
Model M3: Schulz analysis (random treatment effects, κ^2 fixed)				
b_0	−0.46	0.108	−0.47	(−0.66, −0.25)
φ	0.15	0.106	0.13	(0.01, 0.39)
κ	0.11	0.085	0.10	(0.00, 0.30)
Model M4: Schulz analysis (random treatment effects; $1/\kappa_m^2 \sim gamma(\eta\lambda, \lambda)$)				
b_0	−0.44	0.119	−0.43	(−0.68, −0.21)
η	22.27	11.20	19.83	(7.24, 51.23)
λ	3.21	5.083	1.26	(0.06, 18.30)
φ	0.14	0.110	0.12	(0.00, 0.40)
κ_m	0.24	0.104	0.23	(0.13, 0.43)

†Where there is a random-effect distribution for between-study, within meta-analysis, precision in bias, the predictive distribution, on the standard deviation scale, in a new meta-analysis, κ_m, is presented.

analysis variation in bias $\hat{\varphi}^2$ and between-study, within-meta-analysis variation $\hat{\kappa}^2$ are lower if a random-effects model is used for treatment effect. This has the potential to impact strongly on the presumed informational content of studies at high risk of bias. We therefore need to consider the choice of models for the meta-epidemiological data carefully before forming priors for a new meta-analysis. It is also important that the same model is used for a new meta-analysis as was used to form prior inputs.

4. Fixed or random treatment effects: model fit and selection

In the bias model (equations (3)–(5)) we are already assuming random effects; therefore it seems reasonable to use the adequately concealed evidence alone to decide whether to use a fixed or random-effects model for the true treatment effect. Table 2 shows the posterior mean of the residual deviance, $\bar{D}_{\text{res}}$, effective number of parameters, p_D, and the deviance information criterion DIC (Spiegelhalter *et al.*, 2002) for both fixed and random-effects treatment models for the Schulz data, on the basis of the adequately concealed studies alone. DIC is the sum of the residual deviance $\bar{D}$ and the effective number of parameters, p_D, and provides a measure of model fit that penalizes model complexity. Because of the non-linearity between the likelihood and the model parameters, we calculate p_D at the posterior mean of the fitted values rather than at the posterior mean of the parameters (Welton and Ades, 2005). We see that, although the random treatment effects model has a better fit (lower $\bar{D}_{\text{res}}$), it also has substantially more

Table 2. Posterior mean residual deviance $\bar{D}_{\text{res}}$, effective number of parameters, p_D, and deviance information criterion DIC for the meta-epidemiological analysis†

Model	$\bar{D}_{\text{res}}$	p_D	*DIC*
Adequately concealed evidence from Schulz data			
Fixed treatment effects	179.7	115.0	294.7
Random treatment effects	156.5	145.8	302.3
Adequately and inadequately concealed evidence from Schulz data			
M1, fixed treatment effect; κ fixed	538.1	373.7	911.8
M2, fixed treatment effect; κ_m random	535.8	370.2	906.0
M3, random treatment effect; κ fixed	500.6	401.6	902.2
M4, random treatment effect; κ_m random	497.3	405.6	902.9

†Results are presented for models M1–M4 described in Section 3, applied to adequately concealed evidence only from the Schulz data and adequately and inadequately concealed evidence from the Schulz data.

parameters (higher p_D) and consequently has higher DIC. The adequately concealed evidence alone therefore suggests that a fixed treatment effect model is the more parsimonious. However, even if there is no evidence of heterogeneity in treatment effects, we may still not feel that it is appropriate to assume a fixed effects model—as this assumption will be forced to hold across all meta-analyses, including a new meta-analysis.

Table 2 also shows model fit based on both adequately and inadequately concealed trials, for the four different models M1–M4. It is interesting to see that, although model M2 (random effects for $1/\kappa_m^2$) has a similar $\bar{D}_{\text{res}}$, the effective number of parameters is actually reduced, leading to a lower DIC in the fixed treatment effect model. This seems counterintuitive—the more complicated model has lower complexity, p_D. However, the random-effects model for $1/\kappa_m^2$ has the consequence that overall there is a higher degree of shrinkage in the bias parameters—leading to a lower effective number of parameters overall.

On the basis of the evidence, there is little to choose between the models. We would tentatively propose M4, the random treatment effect model and random effects for $1/\kappa_m^2$, on the grounds that a fixed treatment effect model is too restrictive for general use, and that it is unlikely that between-study variation in bias is the same in different meta-analyses. However, we must accept that the degree to which the inadequately concealed evidence is downweighted could be sensitive to model choice. Whichever model we decide on for the meta-epidemiological analysis, the same model should be used in the analysis of a new meta-analysis.

5. Bias-adjusted treatment effect estimates in a new meta-analysis: sensitivity to model choice and format of priors

5.1. Sensitivity to model choice

We now show how evidence-based priors for bias parameters that are obtained from the analysis of the Schulz data can be introduced into a new meta-analysis. We use as an example a meta-analysis of studies comparing Clozapine *versus* neuroleptic medication for treatment of schizophrenia (Wahlbeck *et al.*, 1998). Allocation concealment in each study was classified as either inadequate or unclear, or adequate. The data were in the form of binomial counts, and we used a logistic regression model as presented in equation (7), for a single meta-analysis m, and the bias model as given by equations (1)–(5).

Table 3. Posterior summaries for the pooled treatment effect (log-odds-ratio) in the Clozapine meta-analysis†

Bias model	*Posterior mean (95% credible interval) for the following treatment effect models:*	
	Fixed effects	*Random effects*
Face value		
Adequately concealed studies	−0.321 (−0.836, 0.193)	−0.065 (−1.682, 2.840)
Inadequately concealed studies	−0.884 (−1.129, −0.641)	−0.533 (−1.031, 0.130)
Adequately and inadequately concealed studies combined	−0.781 (−1.002, −0.562)	−0.452 (−0.883, 0.081)
	Model M1	*Model M3*
Bias adjustment, κ fixed		
Non-parametric prior sampled from joint posterior from Schulz	−0.244 (−0.656, 0.152)	−0.145 (−0.630, 0.438)
Parametric priors		
κ, φ constant	−0.249 (−0.663, 0.162)	−0.149 (−0.613, 0.430)
κ, φ stochastic	Bivariate normal, −0.241 (−0.656, 0.165)	Independent gammas, −0.133 (−0.609, 0.448)
	Model M2	*Model M4*
Bias adjustment, κ_m random		
Non-parametric prior sampled from joint posterior from Schulz	−0.259 (−0.664, 0.143)	−0.150 (−0.644, 0.450)
Parameteric priors		
η, λ, φ constant	−0.256 (−0.658, 0.144)	−0.144 (−0.625, 0.439)
η, λ, φ stochastic	η, λ gamma; φ normal, −0.260 (−0.669, 0.147)	η, λ, φ gamma, −0.142 (−0.634, 0.437)

†Results where all evidence is taken at face value are shown separately for inadequately and adequately concealed trials, as well as when both types of study are combined. Results are also presented for bias adjustment models M1–M4, for different methods for introducing prior information on the bias parameters.

We begin with estimates of pooled treatment effect based on separate meta-analyses of the trials at high and low risk of bias (Table 3). This allows us to investigate the extent of bias in the high risk studies, before fitting more complex models that adjust for bias. If only the adequately concealed trials are included, the pooled (fixed effect) log-odds-ratio has posterior mean −0.321 (95% credible interval (CI) −0.84 to 0.19), compared with −0.884 (95% CI −1.13 to −0.64) for the inadequately concealed studies, suggesting that these studies are subject to bias. Note that the difference in posterior mean pooled log-odds-ratio between the adequately and inadequately concealed studies is −0.563, which is consistent with the estimated 95% CI for mean bias from the analysis of the Schulz data (95% CI −0.65 to −0.28) (Table 1). If all the evidence is taken at face value, then the posterior mean pooled log-odds-ratio is −0.781 (95% CI −1.13 to −0.64), which lies between the two but closer to the mean of the inadequately concealed studies which are larger and more numerous. Different estimates are obtained with random-effects models, because the larger studies, which showed the strongest effects, have relatively less weight (Fig. 1). However, the combined effect taking all evidence at face value is again between the pooled effects of the adequately and inadequately concealed studies.

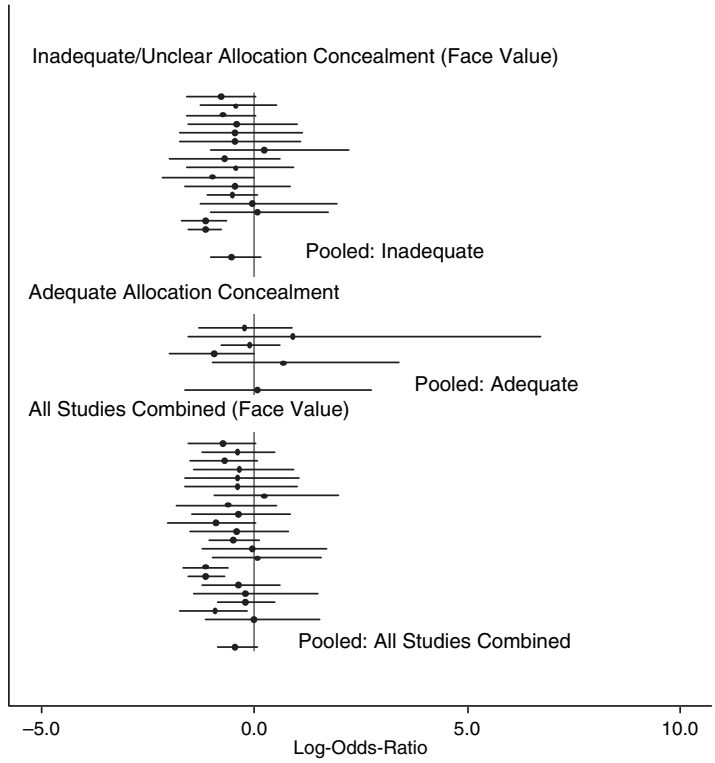

Fig. 1. Posterior mean study-specific and pooled treatment effects (on a log-odds-ratio scale), with 95% CIs, for random-effects models separately for inadequately concealed (taken at face value) and adequately concealed trials: in addition results are shown when both types of study are combined and taken at face value

We might expect the bias-adjusted posterior mean pooled log-odds-ratio to lie (for the fixed treatment effect model) between -0.321, the face value estimate from the adequately concealed studies, and $-0.884+0.47=-0.414$, the face value estimate from the inadequately concealed studies adjusted for mean bias. Perhaps counterintuitively, the posterior mean bias-adjusted pooled estimate is -0.244 (95% CI -0.66 to 0.15) (Table 3). However, the bias model is hierarchical in nature and, just like a random treatment effects model, gives relatively less weight to the large studies, which in this example are also the studies showing the greatest treatment effect. The combined effect of the random-effects model for bias and the bias adjustment is to shift the combined treatment effect estimate substantially in the direction of no effect. Each of the bias-adjusted models gives little evidence of a treatment effect, in contrast with the analyses that take the inadequately concealed evidence at face value (Table 3).

Extending the analysis to a random treatment effects model leads to a greater posterior standard deviation, and reduced treatment effect sizes. This is in part because the larger studies (which have relatively more weight in a fixed effect analysis) show the bigger treatment effects in this example (Fig. 1), but also because the prior inputs for the variance parameters are sensitive to the use of a random- or fixed effects model (Table 1). Arguing as above, we might expect that the bias-adjusted posterior mean log-odds-ratio lies between -0.065 and $-0.533+0.47=-0.063$, when in fact the pooled bias-adjusted estimate is -0.145 (95% CI -0.63 to 0.44) (Table 3). For the random-effects model the treatment effect is stronger than expected by intuition. Again this

is a result of the hierarchical model for bias, which leads to a reduction in estimated between-study heterogeneity in treatment effect in the bias-adjusted models. As a consequence, more weight is given to the large studies with strong treatment effects than for models where the evidence is taken at face value.

5.2. Sensitivity to format of priors

We next consider in more detail how to introduce priors based on the analysis of the Schulz data into a new meta-analysis. In the results that were presented above (Table 3), we sampled from the joint posterior distribution from the Schulz analysis to form a joint prior for b_0, κ and φ (for κ fixed; models M1 and M3) or b_0, η, λ and φ (for κ_m random; models M2 and M4). We achieved this by saving the MCMC chains (thinned every 25 iterations) from the Schulz analysis (resulting in 2000 records) and then sampled with replacement from these 2000 records to provide a non-parametric joint prior.

Meta-analysts may not have access to either a full relevant meta-epidemiological data set or the MCMC chains from a resulting meta-epidemiological analysis. There are several possible parametric approximations that can be made. We could simply enter κ and φ (for κ fixed; models M1 and M3) or η, λ and φ (for κ_m random; models M2 and M4) as constants in equations (3)–(5), taking their posterior median values from the Schulz output. Alternatively, we could treat κ and φ, or η, λ and φ, as stochastic, and attempt to approximate their joint distributions parametrically, on the basis of the output from the Schulz analysis. Inevitably, there will be a wide range of modelling options here (the approximations that are described in Table 3 represent

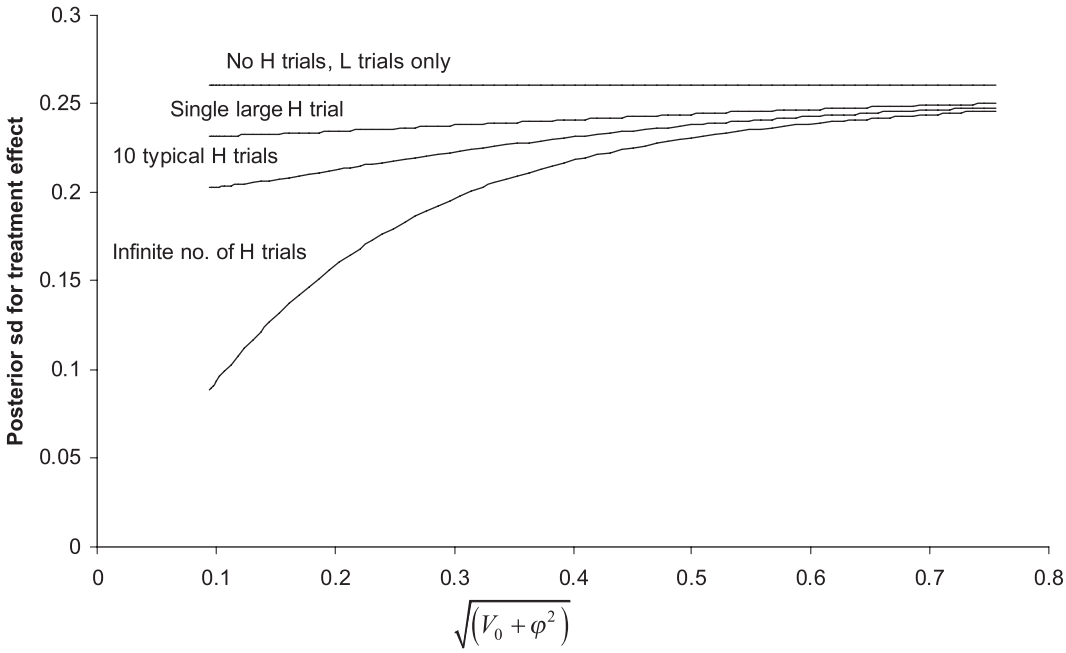

Fig. 2. Posterior standard deviation of fixed treatment effect d_m, plotted against prior standard deviation for the meta-analysis-specific mean bias $\sqrt{(V_0 + \varphi^2)}$, for four scenarios (where L denotes low risk of bias and H high risk of bias), L-trials only, a single infinitely sized H-trial, 10 H-trials with typical (from Schulz data) variance of $\sigma_t^2 = 0.7$ and an infinite number of H-trials: in all cases the L-evidence from the Clozapine example is used and $\hat{\kappa}^2$ is set equal to 0.25, the posterior mean from the fixed treatment effect model for the Schulz data; the posterior CI for $\sqrt{(V_0 + \varphi^2)}$ from the fixed treatment effect models based on the Schulz data is (0.17, 0.73)

just one such choice), raising the further question about whether the MCMC outputs have been adequately captured in the approximation.

These different methods for forming a prior are broadly comparable (Table 3), with both stochastic and constant prior bias parameters giving estimated bias-adjusted treatment effects that are close to those obtained by using the full joint posterior from Schulz *et al.* (1995) as a prior. For the stochastic models, slightly different results were found with different models, but no consistent pattern could be seen.

6. Contribution of evidence at high risk of bias: sensitivity to parameters of the bias model

It is clear from this analysis that the bias-adjusted estimated treatment effect and its posterior uncertainty will be sensitive to the inputs to the bias model from the meta-epidemiological modelling. The algebraic solution (equation (6)) to the basic bias model allows us to carry out sensitivity analysis to prior inputs in more detail, either by evaluating posterior summaries directly for given input values or by looking at derivatives with respect to given inputs. For example, Fig. 2 shows how the posterior standard deviation for the treatment effect increases as the prior uncertainty (on the standard deviation scale) in meta-analysis-specific mean bias increases (for $\hat{\kappa}^2$ set equal to 0.25, the posterior mean from the fixed treatment effect model for the Schulz data). The higher the value of $\sqrt{(V_0 + \varphi^2)}$, the more the evidence at high risk of bias is downweighted. The posterior standard deviation of the treatment effect decreases as the number

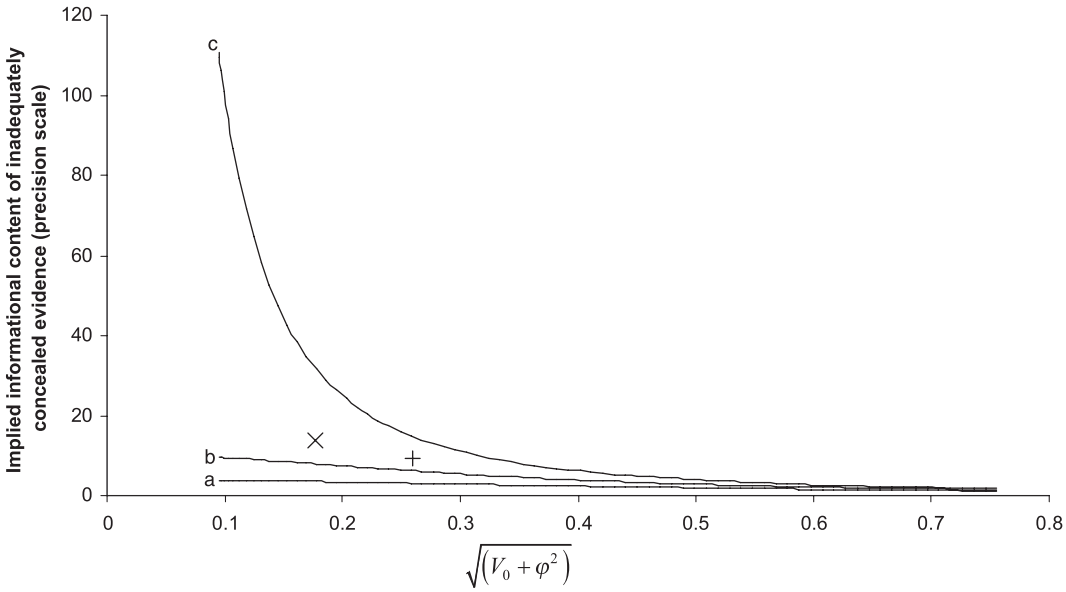

Fig. 3. Implied informational content of the evidence at high risk of bias on a precision scale (precision including H-trials minus precision with L-trials only), plotted against prior standard deviation of the meta-analysis-specific mean bias, $\sqrt{(V_0 + \varphi^2)}$: results are presented for a single infinitely sized H-trial (curve a), 10 H-trials with typical (from the Schulz data) variance of $\sigma_l^2 = 0.7$ (curve b) and an infinite number of H-trials (curve c); in all cases the L-evidence from the Clozapine example is used and $\hat{\kappa}^2$ is set equal to 0.25, the posterior mean from fixed treatment effect model for the Schulz data; numerical results are superimposed for fixed and random treatment effect models with random κ_m^2 for the Clozapine meta-analysis; the posterior 95% CI for $\sqrt{(V_0 + \varphi^2)}$ from the fixed treatment effect models based on the Schulz data is (0.17, 0.73) (×, Clozapine, random-effects model; +, Clozapine, fixed effects model)

of studies at high risk of bias increases from a single (even if very large) study, to 10 typically sized studies, to infinitely many studies (Fig. 2). The effect on posterior uncertainty depends on the value of $\sqrt{(V_0+\varphi^2)}$. On the basis of the fixed treatment effect model for the Schulz data, the posterior mean for $\sqrt{(V_0+\varphi^2)}$ was 0.52 with a CI of (0.17, 0.73). When $\sqrt{(V_0+\varphi^2)}=0.52$, then the posterior standard deviation falls from 0.260 with no studies at high risk of bias to 0.245 with a single large study at high risk of bias, to 0.239 for 10 typical studies at high risk of bias and to 0.233 for an infinite number of studies at high risk of bias. The reduction in posterior uncertainty that is attributable to the use of data at high risk of bias is therefore relatively minor in this case.

It is also instructive to look at the *gain* in precision (which is defined as 1/variance) due to incorporation of the evidence at high risk of bias. This is simply the precision with the H-trials minus precision with no H-trials. This is plotted for various scenarios in Fig. 3. For example if $\sqrt{(V_0+\varphi^2)}$ takes values over the 95% CI range (0.17, 0.73), the gain in precision from including 10 typical H-trials (with precision $1/0.7=1.43$) is equivalent to between 5.7 ($=8.1/1.43$) and 1.1 ($=1.6/1.43$) typical trials taken at face value. We see that the gain in precision from including the evidence at high risk of bias is limited, unless $\sqrt{(V_0+\varphi^2)}$ is low (near the lower end of its CI limits in this analysis). Fig. 3 also shows numerical results of fitting fixed and random treatment effect models with κ_m^2 random to the Clozapine meta-analysis, where there are 16 trials at high risk of bias. The main feature of the random-effects analysis is that the posterior mean for $\sqrt{(V_0+\varphi^2)}$ is at the lower end of the CI, and the gain in precision from including the trials at high risk of bias is greater. Thus it would appear (at least for the Clozapine example) that a fixed effect model might lead to more downweighting of the evidence at high risk of bias than a random-effects model.

Fig. 4 shows how this relationship, for 10 typically sized studies at high risk of bias, changes

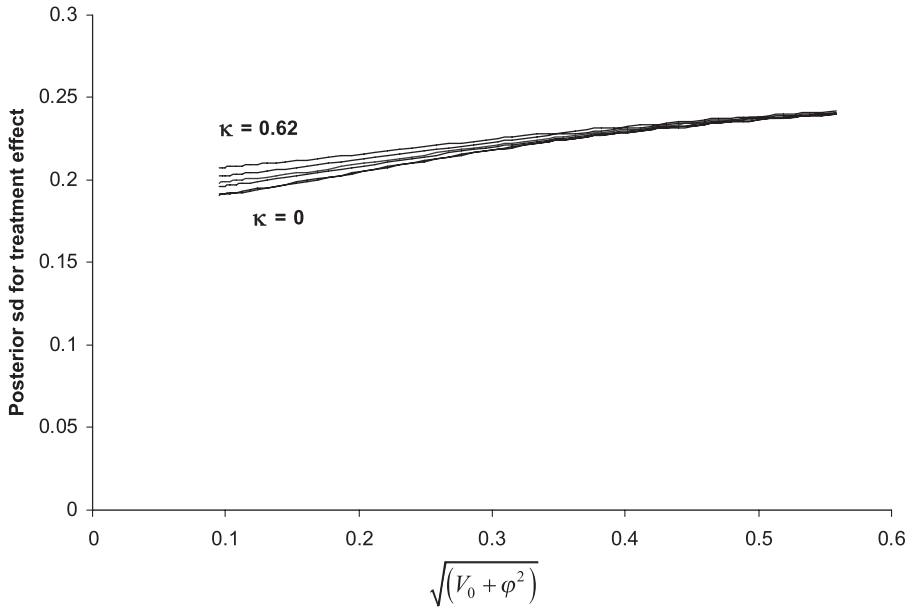

Fig. 4. Posterior standard deviation of treatment effect d_m, plotted against prior standard deviation for the meta-analysis-specific mean bias $\sqrt{(V_0+\varphi^2)}$ for six values of κ, representing the 2.5%, 50% and 95% percentiles of the posterior from the fixed ($\kappa=0.38$, 0.49, 0.62) and random- ($\kappa=0$, 0.1, 0.3) effects models for the Schulz data: in all cases the evidence at low risk of bias from the Clozapine example is used, with 10 H-trials with typical (from the Schulz data) variance of $\sigma_i^2=0.7$

with κ. Increasing the between-study heterogeneity in bias κ leads to a modest increase in posterior standard deviation in treatment effect. For example, when $\sqrt{(V_0+\varphi^2)}=0.52$ (the mean from a fixed treatment effects model), then the posterior standard deviation falls from 0.240 to 0.238 as κ falls from 0.62 to 0. A random-effects model leads to lower values of κ, and consequently more weight given to the evidence at high risk of bias. However, Figs 2 and 4 show that the posterior standard deviation in treatment effect is much more sensitive to changes in uncertainty in meta-analysis-specific mean bias, $\sqrt{(V_0+\varphi^2)}$, than to κ. This can be shown to hold in general, for the fixed effect model, by considering the respective derivatives in equation (6).

7. Discussion

Any meta-analysis containing studies at high and low risk of bias should compare results, for estimated treatment effect and corresponding precision, both including and excluding the high risk evidence (taken at face value). However, there is usually too little information within a meta-analysis to allow precise estimation of the differences between treatment effects in low and high risk studies (Sterne *et al.*, 2002). Therefore, regardless of whether treatment effect estimates including or excluding the high risk evidence appear consistent, we should still have more faith in the low risk evidence. Taking the high risk studies at face value will underestimate uncertainty in treatment effects, which may have implications for any resulting decision analysis. The bias adjustment models that are presented here attempt to capture this uncertainty, as well as systematic differences between treatment effect estimates in low and high risk studies.

There has been considerable recent discussion of methods for addressing bias in observational or randomized studies. Most of the proposals that have been published so far are reweighting schemes, which accord evidence with a high risk of bias a lower weight (Begg and Pilote, 1991; Li and Begg, 1994; Larose and Dey, 1997; Prevost *et al.*, 2000; Spiegelhalter and Best, 2003). Eddy *et al.* (1992) pointed out that this mitigates the bias but does not eliminate it. Our approach can be distinguished from this previous work in two ways. First, our proposed model for bias incorporates variation in the extent of mean bias both between studies, within meta-analyses and between meta-analyses. In addition, we include an estimated overall mean bias term and the corresponding uncertainty in this estimate. This not only leads to downweighting of potentially biased evidence but also, at least in principle, to an unbiased pooled estimate. Second, we base the parameters of our bias model on empirical evidence from collections of previously published meta-analyses, because single meta-analyses typically provide only limited information on the extent of bias (Egger *et al.*, 2003; Sterne *et al.*, 2002). This, of course, entails the strong assumption that the mean bias in a new meta-analysis is exchangeable with the mean biases in the meta-analyses included in previous empirical (meta-epidemiological) studies. For example, the meta-analyses that were included in the study of Schulz *et al.* (1995) are mostly from maternity and child care studies, and we must doubt whether the mean bias in studies on drugs for schizophrenia (the Clozapine example meta-analysis) is exchangeable with the mean biases in this collection of meta-analyses.

Our example focused on the problem of bias due to inadequate allocation concealment, which has been the subject of several empirical investigations (although in some ways this was just a vehicle for exploring the statistical methods and the implications of the modelling assumptions). Our sensitivity analyses suggest that studies with inadequate concealment of allocation contribute very little gain in precision on the treatment effect, owing to a relatively high degree of uncertainty in the meta-analysis-specific mean bias $\sqrt{(V_0+\varphi^2)}$. In cases where the precision in treatment effect based on adequately concealed evidence alone is low, owing to little adequately

concealed evidence being available and/or because a random treatment effects model is being used, then the inadequately concealed evidence may be useful. Otherwise we would tentatively suggest that, on the basis of the evidence that is available, there is little value in incorporating evidence from inadequately concealed trials.

We have clarified some of the technical issues in introducing information from previous meta-epidemiological studies into a new meta-analysis. Ideally, the joint posterior distribution from the meta-epidemiological analysis should be used to form the prior for a new meta-analysis. This would require either access to the complete meta-epidemiological database (which is unlikely to be practical in general) or MCMC outputs from such an analysis, which we would advocate making generally available. We found that plugging in posterior median values for the bias variance parameters produces results which, in this example, give an adequate approximation to the full joint distribution of bias model parameters. However, this will need to be explored more fully in a wider range of examples—in particular when the joint distribution is asymmetrical and/or when parameters are highly correlated.

Recently, a much larger database of meta-analyses has been developed (Wood *et al.*, 2008), which could be used to form priors for bias in new meta-analyses. Wood *et al.* (2008) confirmed that treatment effects were dependent on allocation concealment and also found evidence for an effect of lack of blinding. Interestingly, the evidence for bias was much stronger when outcomes were subjectively assessed than when objectively assessed or all-cause mortality outcomes were used. This suggests that if the meta-epidemiological data can be more specifically tailored to the new meta-analysis, perhaps focusing on studies in a similar area of medicine, with the same outcome measures, it might be possible to reduce the *variety* of different meta-analyses, and thus to generate an evidence base in which estimates of φ^2, the between-meta-analysis variation, were lower. This approach might also address the difficult, but critical, exchangeability assumption, i.e. that the mean bias in studies at high risk of bias in the meta-analysis in question is exchangeable with the mean biases in the meta-analyses in the epidemiological database. The reduction in the size of the evidence base would, however, increase V_0, the uncertainty in the expectation of mean bias. It is likely, therefore, that reducing either V_0 or φ^2 can only be done at the expense of the other, and this suggests that there are clear limits in the amount of information that can be provided by studies at high risk of bias, however carefully one tailors the evidence base for priors.

Meta-epidemiological studies are observational in nature and may therefore be affected by confounding. Confounding might occur because studies at high risk of bias due to a particular flaw in their conduct are more likely than other studies to have further flaws in their conduct. This possibility was addressed by Wood *et al.* (2008), who found the effect of inadequate allocation concealment to be modestly attenuated when controlled for absence of blinding, and vice versa. The potential for such confounding implies that inadequate allocation concealment may not be the *cause* of the bias that was identified in Table 1; however, inadequate allocation concealment might still be considered a proxy for studies at a higher risk of bias.

Siersma *et al.* (2007) discussed statistical models that allow for such confounding. Potential confounders can be adjusted for in the meta-epidemiological analysis by adding regression terms to equation (7). It is also straightforward to incorporate interaction terms to equation (7) that allow the average bias to vary according to characteristics such as whether the outcome variable was subjectively or objectively assessed. In a new meta-analysis, we can also add covariate terms to equations (1) and (2), e.g. patient subgroups for whom there is a differential treatment effect, which may be of interest to clinicians.

The models that are presented here all assume statistical independence of trials and meta-analyses. Such assumptions will be violated if some trials are included in more than one meta-

analysis. Most published meta-epidemiological analyses have dealt with this issue informally, by including meta-analyses that address different clinical questions and by including only one meta-analysis per systematic review. Wood *et al.* (2008) combined data from three meta-epidemiological studies. They dealt with overlap by indexing all trials by using PubMed or similar identifiers, which were used to identify those meta-analyses containing overlapping trials. Meta-analyses were removed to ensure that there was minimal overlap in the data set that was analysed. An alternative would be to use modelling techniques that can allow for the correlation that duplication implies, such as cross-classified random effects (Patterson and Thompson, 1971) that have been modelled by using MCMC simulation (Browne *et al.*, 2001).

Here we have restricted attention to randomized controlled trials. Extension of these methods to include evidence from observational studies would be of interest and would in principle be possible. However, there are substantial practical difficulties in assembling empirical meta-epidemiological evidence on the magnitude of and variability in biases in estimates of treatment effects from observational studies, compared with randomized controlled trials. Although there have been several evaluations of differences between intervention effects in randomized controlled trials and observational studies (Deeks *et al.*, 2003), we are not aware of published estimates of the variance of these differences. Much empirical research is based on convenience samples of published meta-analyses (MacLehose *et al.*, 2000; Concato *et al.*, 2000), which may not be representative. Therefore, exchangeability assumptions may be more difficult to justify when combining evidence from observational studies and randomized controlled trials.

We end by reflecting on the implications of our results for the potential future use of the bias adjustment approach in health care intervention assessment. In the context of a national decision maker such as the NICE, any decision that is taken needs to be supported by evidence that is accepted by, among others, groups of patients and the pharmaceutical industry. It is therefore crucial that all modelling assumptions are transparent and reproducible, and not open to interpretation. There seem to be at least three obstacles that would prevent these methods from being adopted for routine use at present.

First, a critical assumption is that the new meta-analysis can be considered exchangeable with those in the existing evidence base. The results of Wood *et al.* (2008) suggest that we would certainly need to consider the variability in bias between clinical areas and according to outcome. However, there are likely to be other possible mechanisms for bias, which may leave the exchangeability assumption open to debate. Second, this paper has focused on a single source of bias, coded as a binary variable. Further work would be required to generalize this to account for multiple sources of bias and to deal with confounding, as discussed above. Although such models for multiple sources of bias are technically feasible (Eddy *et al.*, 1992; Greenland, 2005; Siersma *et al.*, 2007), reliance on exchangeability assumptions and sensitivity to the model for the meta-epidemiological data will be increased. A third obstacle to the acceptance of these methods is the sensitivity of the estimated treatment effect to whether a fixed or random-effects treatment model is employed (Table 3). A choice of a fixed effect model may lead to greater downweighting of evidence at high risk of bias, whereas a random-effects model may give the evidence that is at high risk of bias relatively more weight.

The possibility of using carefully tailored subsets of previous evidence, and the issue of multiple sources of bias, can, of course, be examined in more detail as more comprehensive meta-epidemiological databases are established. These analyses will no doubt add to our knowledge of bias mechanisms, lead to refinements in the methods and most importantly provide a better empirical basis for bias adjustment. However, it is all too easy to imagine that a company whose drug is not recommended might dispute a methodology that downweights its evidence of

efficacy, especially if a slight change in modelling assumptions or in the meta-epidemiological data gives a more favourable result.

In summary, an 'all available evidence' approach to health intervention assessment trades off increased precision at the expense of an increased risk of bias. The models that are presented here provide a methodology that downweights and adjusts for potential bias. However, at our present state of knowledge, the models that are proposed here cannot yet be confidently used for health intervention assessment, unless the decision on which intervention to adopt can survive a thorough and wide-ranging analysis of sensitivity to model inputs.

Acknowledgements

The authors thank Professor Simon Thompson (Medical Research Council Biostatistics Unit, Cambridge) and others who commented on an earlier version of this paper, which was presented to the workshop 'Explaining the results of a complex probabilistic modelling exercise: conflict, consistency and sensitivity analysis', which was sponsored by the Medical Research Council Population Health Sciences Research Network, in Cambridge, September 7th–8th, 2006. We also thank two referees for their very helpful comments on an earlier draft of this paper.

References

Begg, C. B. and Pilote, L. (1991) A model for incorporating historical controls into a meta-analysis. *Biometrics*, **47**, 899–906.

Brooks, S. P. and Gelman, A. (1998) Alternative methods for monitoring convergence of iterative simulations. *J. Computnl Graph. Statist.*, **7**, 434–455.

Browne, W. J., Goldstein, H. and Rasbash, J. (2001) Multiple membership classification (MMMC) models. *Statist. Modllng*, **1**, 103–124.

Concato, J., Shah, N. and Horwitz, R. I. (2000) Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New Engl. J. Med.*, **342**, 1887–1892.

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., Petticrew, M. and Altman, D. G. (2003) Evaluating non-randomised intervention studies. *Hlth Technol. Assessmnt*, **7**, no. 27.

Eddy, D. M., Hasselblad, V. and Shachter, R. (1992) *Meta-analysis by the Confidence Profile Method: the Statistical Synthesis of Evidence*. London: Academic Press.

Egger, M., Juni, P., Bartlett, C., Holenstein, F. and Sterne, J. A. C. (2003) How important are comprehensive literature searches and the assessment of trial quality in systematic reviews?: empirical study. *Hlth Technol. Assessmnt*, **7**, no. 1.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayes. Anal.*, **1**, 515–534.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003) *Bayesian Data Analysis*, 2nd edn. Boca Raton: Chapman and Hall–CRC.

Gluud, L. L. (2006) Bias in clinical intervention research. *Am. J. Epidem.*, **163**, 493–501.

Greenland, S. (2005) Multiple-bias modelling for analysis of observational data (with discussion). *J. R. Statist. Soc.* A, **168**, 267–306.

Kjaergard, L. L., Villumsen, J. and Gluud, C. (2001) Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Ann. Intern. Med.*, **135**, 982–989.

Larose, D. T. and Dey, D. K. (1997) Grouped random effects models for Bayesian meta-analysis. *Statist. Med.*, **16**, 1817–1829.

Li, Z. and Begg, C. B. (1994) Random effects models for combining results from controlled and uncontrolled studies in meta-analysis. *J. Am. Statist. Ass.*, **89**, 1523–1527.

Lindley, D. V. and Smith, A. F. M. (1972) Bayes estimates for the linear model (with discussion). *J. R. Statist. Soc.* B, **34**, 1–41.

MacLehose, R. R., Reeves, B. C., Harvey, I. M., Sheldon, T. A., Russell, I. T. and Black, A. M. S. (2000) A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Hlth Technol. Assessmnt*, **4**, no. 34.

McAuley, L., Pham, B., Tugwell, P. and Moher, D. (2000) Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *Lancet*, **356**, 1228–1231.

Moher, D., Pham, B., Jones, A., Cook, D. J., Jadad, A. R., Moher, M., Tugwell, P. and Klassen, T. P. (1998) Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet*, **352**, 609–613.

Patterson, H. D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Prevost, T. C. Abrams, K. R. and Jones, D. R. (2000) Hierarchical models in generalised synthesis of evidence: an example based on studies of breast cancer screening. *Statist. Med.*, **19**, 3359–3376.

Schulz, K. F., Chalmers, I., Hayes, R. J. and Altman, D. G. (1995) Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *J. Am. Med. Ass.*, **273**, 408–412.

Siersma, V., Als-Nielsen, B., Chen, W., Hilden, J., Gluud, L. L. and Gluud, C. (2007) Multivariable modelling for meta-epidemiological assessment of the association between trial quality and treatment effects estimated in randomized clinical trials. *Statist. Med.*, **26**, 2745–2758.

Smith, T. C., Spiegelhalter, D. J. and Thomas, A. (1995) Bayesian approaches to random-effects meta-analysis: a comparative study. *Statist. Med.*, **14**, 2685–2699.

Spiegelhalter, D. J. and Best, N. G. (2003) Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statist. Med.*, **22**, 3687–3709.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.

Spiegelhalter, D. J., Thomas, A. and Best, N. (2000) *WinBUGS User Manual*. Cambridge: Medical Research Council Biostatistics Unit.

Sterne, J. A. C., Jüni, P., Schulz, K. F., Altman, D. G., Bartlett, C. and Egger, M. (2002) Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Statist. Med.*, **21**, 1513–1524.

Sterne, J. A. C., Welton, N. J., Ades, A. E., Altman, D. G. and Carlin, J. B. (2008) Incorporation of potentially biased evidence in systematic reviews and meta-analyses. Submitted to *Int. J. Epidem*.

Wahlbeck, K., Cheine, M. V. and Essali, A. (1998) Clozapine versus typical neuroleptic medication for schizophrenia. *Cochr. Database Syst. Rev.*, **1**.

Welton, N. J. and Ades, A. E. (2005) A model of toxoplasmosis incidence in the UK: evidence synthesis and consistency of evidence. *Appl. Statist.*, **54**, 385–404.

Wolpert, R. L. and Mengersen, K. L. (2004) Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. *Statist. Sci.*, **19**, 450–471.

Wood, L., Egger, M., Gluud, L. L., Schulz, K., Jüni, P., Altman, D., Gluud, C., Martin, R. M., Wood, A. J. G. and Sterne, J. A. C. (2008) Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *Br. Med. J.*, **336**, 601–605.