Assessing publication bias in meta-analyses in the presence of between-study heterogeneity

Author(s): Jaime L. Peters, Alex J. Sutton, David R. Jones, Keith R. Abrams, Lesley Rushton and Santiago G. Moreno

Source: *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, JULY 2010, Vol. 173, No. 3 (JULY 2010), pp. 575-591

Published by: Wiley for the Royal Statistical Society

Stable URL: http://www.jstor.com/stable/40666276

J. R. Statist. Soc. A (2010)
173, *Part* 3, *pp.* 575–591

Assessing publication bias in meta-analyses in the presence of between-study heterogeneity

Jaime L. Peters, Alex J. Sutton, David R. Jones and Keith R. Abrams

University of Leicester, UK

Lesley Rushton

Imperial College London, UK

and Santiago G. Moreno

University of Leicester, UK

[Received February 2009. Revised September 2009]

Summary. Between-study heterogeneity and publication bias are common features of a meta-analysis that can be present simultaneously. When both are suspected, consideration must be made of each in the assessment of the other. We consider extended funnel plot tests for detecting publication bias, and selection modelling and trim-and-fill methods to adjust for publication bias in the presence of between-study heterogeneity. These methods are applied to two example data sets. Results indicate that ignoring between-study heterogeneity when assessing publication bias can be misleading, but that methods to test or adjust for publication bias in the presence of heterogeneity may not be powerful when the meta-analysis is not large. It is therefore unrealistic to expect to disentangle the effects of publication bias and heterogeneity reliably in all except the largest meta-analyses.

Keywords: Funnel plots; Heterogeneity; Meta-analysis; Metaregression; Publication bias

1. Introduction

Meta-analysis is a commonly used tool for the synthesis and evaluation of evidence on a particular area of interest. Regardless of the area of application, evidence that between-study heterogeneity (the variability in the true underlying effects between studies (Higgins and Thompson, 2002) and publication bias (the propensity for research which is statistically significant to be published, and hence included in a meta-analysis, is higher than for research which is not) are present in a meta-analysis is not uncommon (Engels *et al.*, 2000; Song *et al.*, 2000; Sutton *et al.*, 2000b), and both can exist simultaneously. In this paper we consider two types of between-study heterogeneity: explainable and residual. *Explainable* heterogeneity is heterogeneity that can be explained by an observed (study level) covariate(s), whereas *residual* heterogeneity refers to variation that cannot be explained (and is often accommodated in meta-analysis models by the inclusion of random effects). Undetected publication bias can lead to erroneous conclusions being drawn from a meta-analysis. Between-study heterogeneity can also lead to problems for the meta-analyst if incorrectly interpreted, but investigation into its possible causes can lead to

Address for correspondence: Jaime L. Peters, Peninsula Medical School, Noy Scott House, Barrack Road, Exeter, EX2 5DW, UK.
E-mail: jaime.peters@pms.ac.uk

 0964–1998/10/173575

a better understanding of the research question(s) of interest through identifying factors which explain the variability.

There are various conditions under which the assessment of publication bias by using published methods has been found to be inappropriate. One main limitation is where there is between-study heterogeneity. When such heterogeneity is unexplainable, methods to assess visually (e.g. funnel plots (Light and Pillemar, 1984)), to test (e.g. the rank correlation test (Begg and Mazumdar, 1994), or regression tests (Egger *et al.*, 1997; Harbord *et al.*, 2006; Macaskill *et al.*, 2001; Peters *et al.*, 2006; Rücker *et al.*, 2008)) or to adjust for publication bias (e.g. trim and fill (Duval and Tweedie, 2000a, b)) are known to be particularly poor (Harbord *et al.*, 2006; Peters *et al.*, 2006; Schwarzer *et al.*, 2002; Terrin *et al.*, 2005) owing to the extra (often unmodelled) variability in the analysis, especially when there are few studies in the meta-analysis. Nevertheless, researchers continue to use these methods in scenarios where unexplainable heterogeneity is present even though this may lead to unreliable conclusions (Ioannidis and Trikalinos, 2007).

Further, even when (a proportion) of between-study heterogeneity can be explained by (study level) covariates, such predictable effects are often ignored when making an assessment of publication bias. For example in the meta-analysis of Lewis *et al.* (2005), to investigate an association between MTHFR 677C->T polymorphism and coronary heart disease significant between-study heterogeneity was noted to have 'largely disappeared after stratification by geographical region'. Despite this, Lewis *et al.* (2005) assessed evidence of publication bias in the meta-analysis by using Egger's regression test (Egger *et al.*, 1997), ignoring the effects of geographical region in the analysis. In such situations, the conclusions from that assessment of publication bias may be misleading because the observed, and potentially explainable, between-study heterogeneity may be 'confounding' the observed relationship between effect size and some measure of study precision (e.g. standard error or sample size) that is used to assess publication bias. This is explained in more detail below.

The aim of this paper is to present and appraise critically ways to assess publication bias when explainable and/or residual between-study heterogeneity is observed in a meta-analysis. To date, little guidance has been available for how to proceed in such situations. We use two published meta-analysis examples (Mapstone *et al.*, 2003; Raudenbush, 1984) to illustrate some of the methods and techniques that are discussed herein. These examples are described in Section 3. In Section 4 we focus our attention on methods to *test* for publication bias under scenarios where explainable and/or residual heterogeneity is present. We then move on to describe and illustrate methods that are used to *adjust* for publication bias and how this may be done in the presence of between-study heterogeneity (Section 5). We present our discussion which includes recommendations for use in practice in Section 6. First, a brief introduction to meta-analysis and publication bias methods is given.

2. Meta-analysis and publication bias methods

2.1. Meta-analysis methods

A fixed effects meta-analysis is a variance-weighted average of the effect estimates from the studies included, where all studies are assumed to be estimating the same underlying effect, so that any differences that are observed between estimates are due to sampling error alone. The pooled estimate μ from a fixed effects inverse variance weighted meta-analysis is given by

$$y_i = \mu + \varepsilon_i, \qquad \varepsilon_i \sim N(0, v_i), \tag{1}$$

where y_i is the effect estimate and v_i is the associated variance from study i.

In many meta-analyses, studies may be carried out in different locations or with different populations, leading to excess heterogeneity being observed. In such cases, the fixed effect meta-analysis model is not appropriate and a random-effects inverse variance meta-analysis can be used. Assuming the above definitions of y_i and v_i, the pooled estimate from a random-effects meta-analysis is given by

$$y_i = \theta_i + \varepsilon_i, \qquad \theta_i \sim N(\mu, \tau^2), \quad \varepsilon_i \sim N(0, v_i), \tag{2}$$

where τ^2 is the estimate of the between-study heterogeneity in the meta-analysis, so that, when $\tau^2 = 0$, the random-effects meta-analysis model is equivalent to the fixed effects model (Sutton *et al.*, 2000a).

2.2. Heterogeneity, subgroup analyses and meta-regression

An alternative measure of between-study heterogeneity is the inconsistency measure I^2 (Higgins and Thompson, 2002). I^2 is the percentage of total variation across studies that is due to between-study heterogeneity rather than chance (Higgins and Thompson, 2002). It is given as

$$I^2 = (H^2 - 1)/H^2 \tag{3}$$

where $H^2 = Q/(k-1)$. Q is the Q-statistic from the homogeneity hypothesis test that all studies are estimating the same underlying effect:

$$Q = \sum_{i=1}^{k} w_i (y_i - \mu)^2, \qquad w_i = 1/v_i, \tag{4}$$

(Sutton *et al.*, 2000a) and k is the number of studies in the meta-analysis.

In a random-effects model investigation of the source(s) of the heterogeneity is not undertaken. Subgroup analyses and meta-regression are common techniques for exploring potential sources of heterogeneity but should only be considered as exploratory analyses since any associations may occur by chance or due to confounding, rather than reflecting a true association. Subgroup analyses are only useful to investigate heterogeneity between categorical covariates (e.g. animal species or strain). Metaregression, in contrast, involves the regression of effect size on categorical and/or continuous study level covariates that are believed to explain some of the between-study heterogeneity (e.g. the year of publication).

2.3. Funnel plots for publication bias

The funnel plot is one of the simplest methods used to investigate publication bias (Sterne *et al.*, 2005). It is a scatter plot of a measure of study size against a measure of effect size. If no bias is present it should appear funnel shaped, since effect sizes should be evenly distributed (i.e. symmetric) around the underlying true effect size with more variability in the smaller studies than in the larger studies owing to the greater influence of sampling error. If gaps in the lower extremities of the funnel are observed, causing the plot to appear asymmetrical, publication bias may be suspected (Light and Pillemar, 1984). Funnel plot asymmetry helps to detect small study effects (the tendency for smaller studies to show effects further from the null than larger studies (Sterne *et al.*, 2000)) which may be due to publication bias. In a funnel plot it is assumed that all effect estimates in the meta-analysis are estimating the same underlying effect (i.e. fixed effect), or are sampled from a symmetrical, common distribution of effects (i.e. random effects) (Light and Pillemar, 1984). If, however, there is a covariate effect in the meta-analysis, such as geographical region (as in the meta-analysis of Lewis *et al.* (2005) that was described in

Section 1), this assumption does not hold and so the appearance of the funnel plot may be distorted by the covariate effect, making it susceptible to misinterpretation. Many methods to test or adjust for publication bias are based on the premises of the funnel plot, so they will also encounter such problems. However, publication bias is only one reason why a funnel plot may appear asymmetric; poor methodological quality, between-study heterogeneity and chance are among others (see Sterne *et al.* (2000) for further details).

2.4. Regression tests for publication bias

Various regression tests for publication bias, based on the funnel plot, have been proposed (Egger *et al.*, 1997; Harbord *et al.*, 2006; Macaskill *et al.*, 2001; Peters *et al.*, 2006) to test for an association between the effect estimates and their standard errors (in particular, studies with more extreme effect estimates will have larger standard errors). The most commonly cited regression test is the weighted regression model that was published by Egger *et al.* (1997) and is given by

$$y_i = \alpha + \beta \, \mathrm{se}_i + \varepsilon_i \tag{5}$$

which is weighted by $1/\mathrm{se}_i^2$, with $\varepsilon_i \sim N(0, \mathrm{se}_i^{2*}\varphi)$. y_i is the effect estimate as defined above, $\mathrm{se}_i = \sqrt{v_i}$ is the standard error from study i and φ is the unknown multiplicative overdispersion parameter that is estimated in the model. This parameter accounts for residual heterogeneity, which is assumed larger among the smaller studies owing to the greater influence of sampling error (see Moreno *et al.* (2009) for more discussion of this dispersion parameter). There are some statistical concerns regarding the use of the Egger test with dichotomous data (Deeks *et al.*, 2005; Higgins and Green, 2008; Macaskill *et al.*, 2001). Briefly, there is a mathematical relationship between the odds ratio OR and its standard error, which can lead to inflated type I errors with the Egger test (Deeks *et al.*, 2005); also the independent variable, se_i in equation (6), is assumed known but is in fact estimated. Various alternatives have been published to overcome these concerns with the use of Egger's test for ORs (see Higgins and Green (2008) for a summary). As a comparison with the Egger regression test, we use the Peters regression test (Peters *et al.*, 2006), which is a modification of the test that was proposed by Macaskill *et al.* (2001). In simulation studies, the Peters test (given in equation (6)) was found to have more appealing properties than the Macaskill and Egger tests (Peters *et al.*, 2006; Rücker *et al.*, 2008):

$$y_i = \alpha + \frac{\beta}{\mathrm{size}_i} + \varepsilon_i. \tag{6}$$

This regression model is weighted by

$$\left(\frac{1}{a_i + b_i} + \frac{1}{c_i + d_i} \right)^{-1},$$

with $\varepsilon_i \sim N(0, \mathrm{se}_i^{2*}\varphi)$. size_i is the total sample size for study i, and a_i, b_i, c_i and d_i are the cell values from the usual 2×2 tables for calculation of the ORs (e.g. $\mathrm{OR} = a_i d_i / b_i c_i$).

All the regression tests that have so far been evaluated in the literature for detecting publication bias are fixed effects models with multiplicative overdispersion error. It is therefore not surprising that regression tests have been shown to have limited performance in detecting publication bias in the presence of between-study heterogeneity (Harbord *et al.*, 2006; Peters *et al.*, 2006). Random-effect versions of the Egger, Macaskill and Peters regression models (to

account for residual between-study heterogeneity) and extended regression models to account for explainable between-study heterogeneity have been evaluated in a simulated study (Peters *et al.*, 2009). The possibility of extending these regression models to account for heterogeneity has previously been noted by others (Sterne *et al.*, 2000).

The results of the simulation study suggest that only the extended fixed effects version of the Peters test was found to have some power (albeit limited) to detect publication bias, but this was in the presence of explainable heterogeneity when there was no residual between-study heterogeneity (Peters *et al.*, 2009). In the most likely scenarios for meta-analyses, where both explainable and residual between-study heterogeneity exist, the extended tests exhibited poor performance in the simulation study (Peters *et al.*, 2009). Nevertheless, there was some evidence that the extended tests may offer some ability to disentangle publication bias and explainable and residual between-study heterogeneity for large meta-analyses, such as those in the psychology or education literature. Below are the extended fixed effects Egger and Peters tests where group_i represents the covariates which are applied later in the paper for comparison. More details on the above-mentioned simulation study can be found in Peters *et al.* (2009).

The fixed effects extended Egger test is

$$y_i = \alpha + \beta \text{se}_i + \gamma \text{group}_i + \varepsilon_i \tag{7}$$

weighted by $1/\text{se}_i^2$, with $\varepsilon_i \sim N(0, \text{se}_i^{2*}\varphi)$.

The fixed effects extended Peters test is

$$y_i = \alpha + \frac{\beta}{\text{size}_i} + \gamma \text{group}_i + \varepsilon_i \tag{8}$$

weighted by

$$\left(\frac{1}{a_i + b_i} + \frac{1}{c_i + d_i}\right)^{-1},$$

with $\varepsilon_i \sim N(0, \text{se}_i^{2*}\varphi)$.

2.5. Trim-and-fill method

The trim-and-fill model is an iterative non-parametric method to adjust for publication bias based on the one-sided asymmetry of a funnel plot (Duval and Tweedie, 2000a, b). As opposed to the above regression models, the trim-and-fill method assumes that studies with the most extreme effect sizes are suppressed (i.e. publication bias is a factor of effect size only). Specifically, it is assumed that extreme estimates on the left-hand side of a funnel plot are suppressed, and the number of studies on the right-hand side of the funnel that have no counterpart on the left-hand side, z, are estimated. These z-studies are then 'trimmed' from the right-hand side of the plot and the pooled estimate from the studies of this now symmetrical plot is calculated. The z-studies are replaced along with their left-hand side counterparts and a pooled effect is estimated. This process is repeated until the pooled estimates that are calculated at subsequent iterations become stable. See Duval and Tweedie (2000a, b) for more information. Recent research suggests that using either fixed or random-effects modelling in the trimming stage of the algorithm followed by a random-effect model to pool the fitted data set outperforms other alternative model combinations (Peters *et al.*, 2007).

2.6. Selection modelling

Selection models include weight functions that allow estimation of a pooled effect given

that the studies in the meta-analysis are a selection of the totality of relevant studies, i.e. that publication bias is present (Hedges and Vevea, 2005). Not only is the estimation of the pooled effect modelled, but the selection process is also, and so a pooled effect, given publication bias, is estimated. The modelling of the pooled effect can incorporate study level covariates and so this method can simultaneously model explainable between-study heterogeneity and publication bias. Details on the selection model that is used in this paper are given in Section 5.2.

3. Examples

3.1. Meta-analysis of animal toxicology experiments

The first example is a meta-analysis of animal toxicology experiments by Mapstone *et al.* (2003). Relative risks RR from animal experiments investigating fluid resuscitation and the risk of death from haemorrhage were synthesized. Although possible sources of between-study heterogeneity were explored, Mapstone *et al.* (2003) did not assess publication bias in their original analysis. Data from 43 animal experiments using three different species (rats ($n = 35$), pigs ($n = 7$) and sheep ($n = 1$)) are combined by using a random-effects metaregression giving a pooled RR of 0.88 (and 95% confidence interval (CI) of 0.73, 1.07). The I^2-value for this meta-analysis is 61% (95% CI: 45%, 72%), suggesting moderately large between-study heterogeneity (Higgins *et al.*, 2003). Further analyses by Mapstone *et al.* (2003) considered the influence of study level covariates (in particular the method that was used to induce haemorrhage in the animals, follow-up time of the study and volume of fluid infused) which explained some of the heterogeneity between studies (Mapstone *et al.*, 2003). Consequently, Mapstone *et al.* (2003) reported RRs stratified by the method that was used to induce haemorrhage and adjusted (using metaregression) for fluid used, follow-up time and species of animal. They reported that there was little difference in the RRs with and without adjustment for species, but differences between rat strains were not considered by them. Our reanalysis of these data suggests that there is some evidence of a strain effect in the rat experiments (see Fig. 1 where study-specific estimates are stratified by species and strain). It is, however, plausible that the differences between the observed RRs from the rat strains are not necessarily reflecting true differences in the effect of fluid resuscitation on mortality through haemorrhage between rat strains; rat strain may be a proxy for some other factor, e.g. experimenter effect, environment effect or some other effect that is related to the laboratory used or the group of researchers.

Although, as described above, several possible sources of between-study heterogeneity were explored by Mapstone *et al.* (2003), for illustration in this paper, only the species and strain differences are considered and all other heterogeneity is considered unexplainable.

3.2. Meta-analysis of teacher expectancy

The second example is a meta-analysis of studies that were carried out to determine the influence of teacher's expectations on the measured intelligence of their students (Raudenbush, 1984). In each study, a group of students (the experimental group) is expected to do well in an intelligence test. This experimental group and a control group of students (who were not expected to do as well) are given the intelligence test before and after the experimental group has been identified to teachers as likely to do well. The standardized mean differences (Sutton *et al.*, 2000a) in test scores before and after are calculated and used as the outcome measure in each study. 19 estimates from 18 studies are included in this meta-analysis. A fixed effects inverse variance pooled estimate of the standardized mean differences (0.06 (95% CI: −0.01, 0.13)) suggests that students

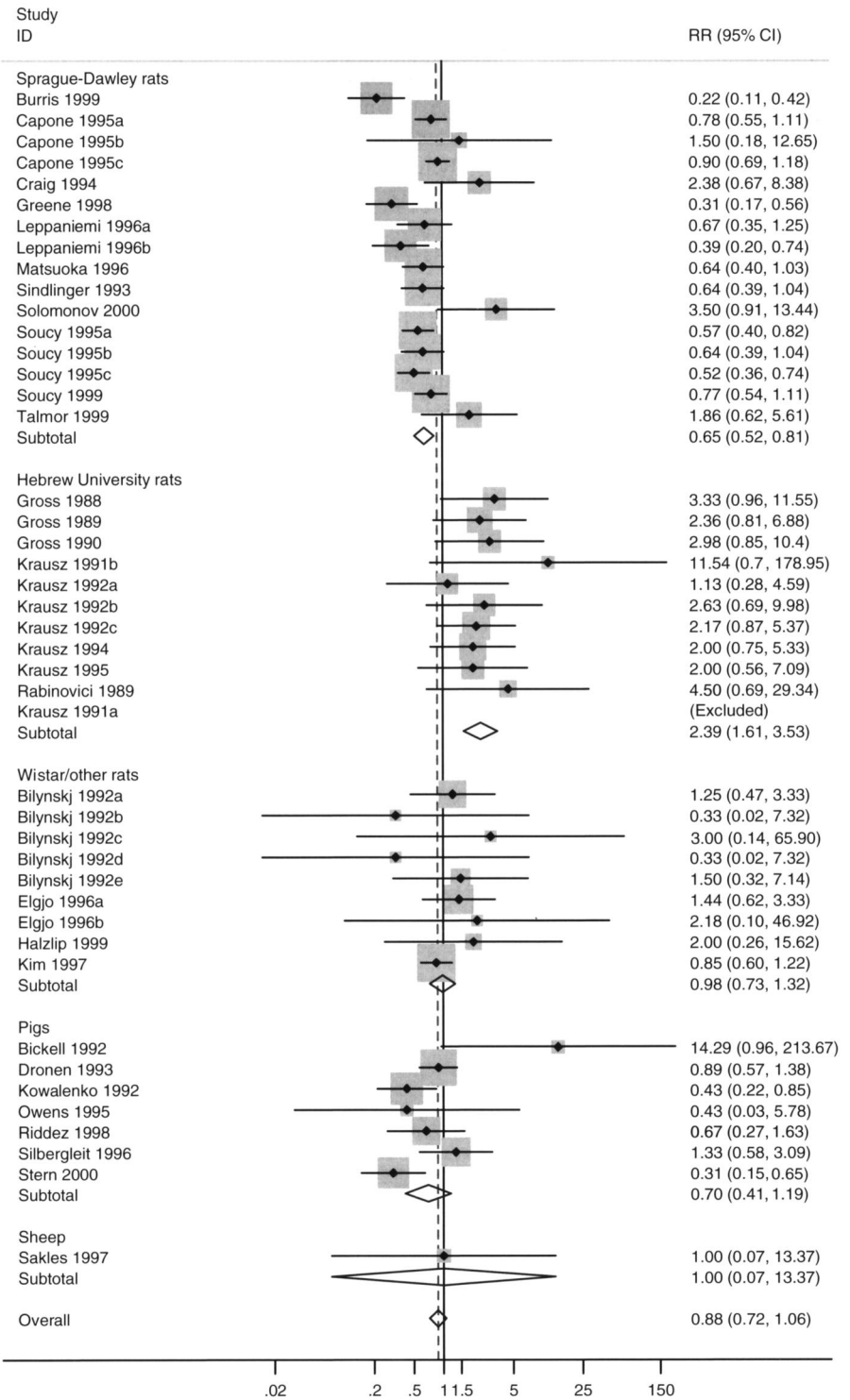

Fig. 1. Forest plot of experiments from Mapstone *et al.* (2003) by species and strain (the weights are from random-effects analysis) used

in the experimental group had an improvement in test scores after they had been identified as likely to do well. Calculation of I^2 suggests evidence of between-study heterogeneity: 50% (95% CI: 15%, 70%), and so we also consider the random-effects estimate of 0.08 (95% CI: −0.02, 0.19). However, as Raudenbush reported, the amount of contact time the teacher has with the experimental group of students (after they have been identified as expected to do well) appears to explain the majority of this between-study heterogeneity. Contact time between student and teacher is dichotomized into *1 week or less* and *more than 1 week*. The fixed effect pooled estimate for studies where the contact time is 1 week or less ($n = 8$) is 0.36 (95% CI: 0.20, 0.51) with an I^2 of 37% (95% CI: 0%, 72%), compared with studies where the contact time is more than 1 week ($n = 11$): the fixed effects pooled estimate is −0.02 (95% CI: −0.10, 0.06) with an I^2 of 0% (95% CI: 0%, 60%), suggesting that teachers who had had more exposure to, and hence longer to form their own expectations of, the students were less amenable to influence by the experimental manipulation. These data, subgrouped by dichotomized contact time, are presented in Fig. 2.

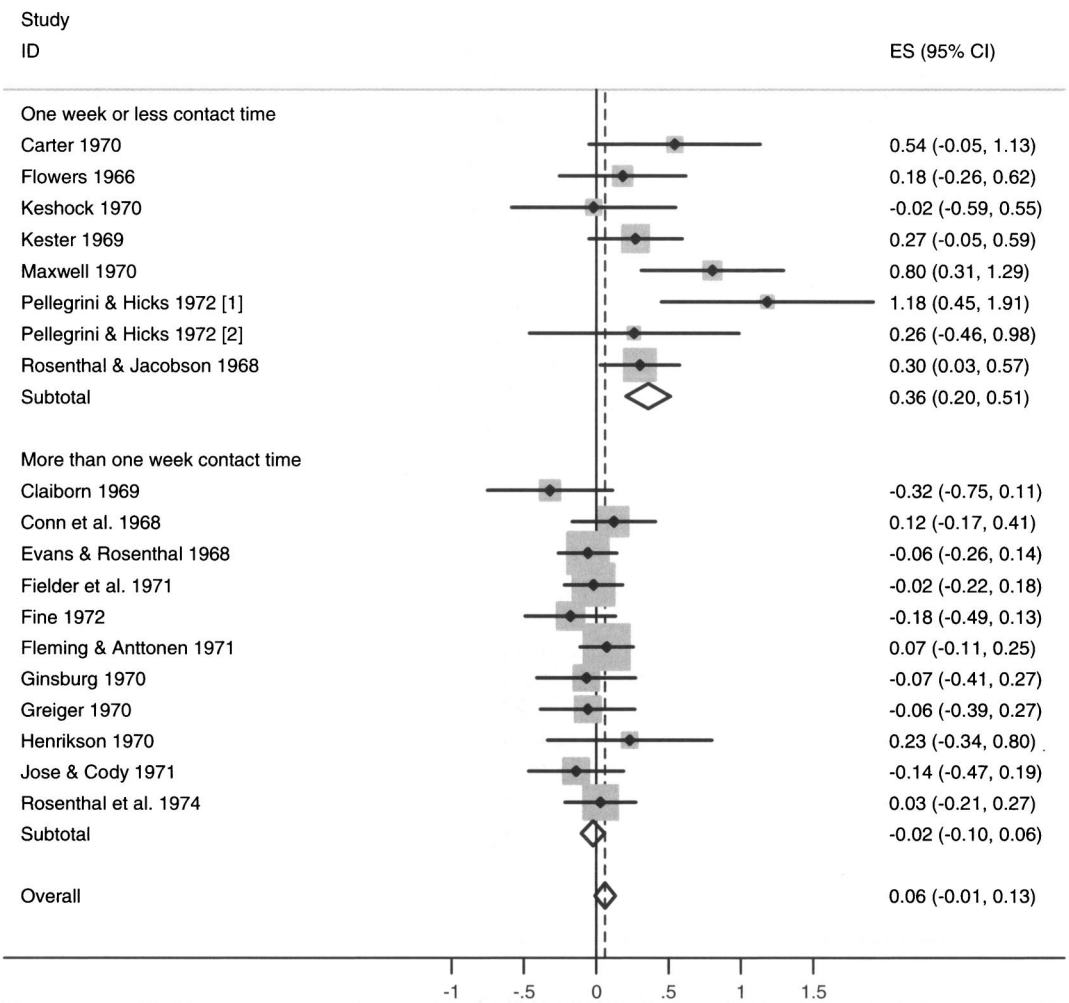

Fig. 2. Forest plot of experiments from Raudenbush (1984) by contact time

4. Testing for publication bias

4.1. Ignoring between-study heterogeneity

The simplest way to assess publication bias when between-study heterogeneity is observed is to ignore the heterogeneity (as in Lewis *et al.* (2005)). We refer to this as the *naive* approach. In Fig. 3 we present a funnel plot of the data from the meta-analysis of Mapstone *et al.* (2003). Visual inspection suggests some evidence of publication bias, with a 'gap' seen in the bottom left-hand corner of the plot suggesting that small studies reporting RR < 1 are missing. Egger's regression test (Egger *et al.*, 1997) suggests some evidence of asymmetry or publication bias ($p = 0.02$) whereas the Peters regression test (Peters *et al.*, 2006) produces a considerably larger p-value (0.65).

A funnel plot of the standardized mean differences in the Raudenbush (1984) meta-analysis suggests evidence of asymmetry or publication bias (Fig. 4), as do the findings from the Egger test ($p = 0.06$). (The Peters test cannot be applied here as it is only for binary outcomes.)

This naive assessment of publication bias in these two meta-analyses is based on the assumption that all studies are estimating the same underlying effect or their underlying effects are sampled from a common, symmetrical, distribution. Since covariate effects are suspected in both example meta-analyses this assumption is brought into question in both instances.

A further consideration (irrespective of whether or not a covariate is an effect size modifier) is that differing levels of publication bias may exist in studies with different covariate values. For example, with respect to the meta-analysis of Mapstone *et al.* (2003), it is conceivable that the propensity for publication bias differs depending on the type of animal that was used in the study. This would be possible because, for ethical reasons, experiments using larger animals (e.g. dogs and monkeys), which are typically smaller studies, may be more likely to be published than studies using smaller animals (e.g. rats and mice) regardless of their size. Differential

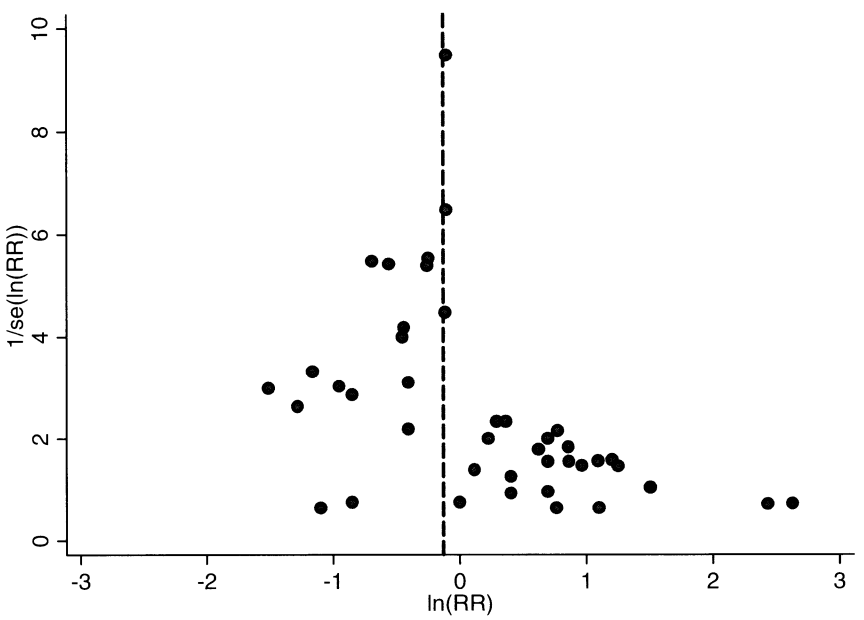

Fig. 3. Funnel plot of experiments in the meta-analysis of Mapstone *et al.* (2003)

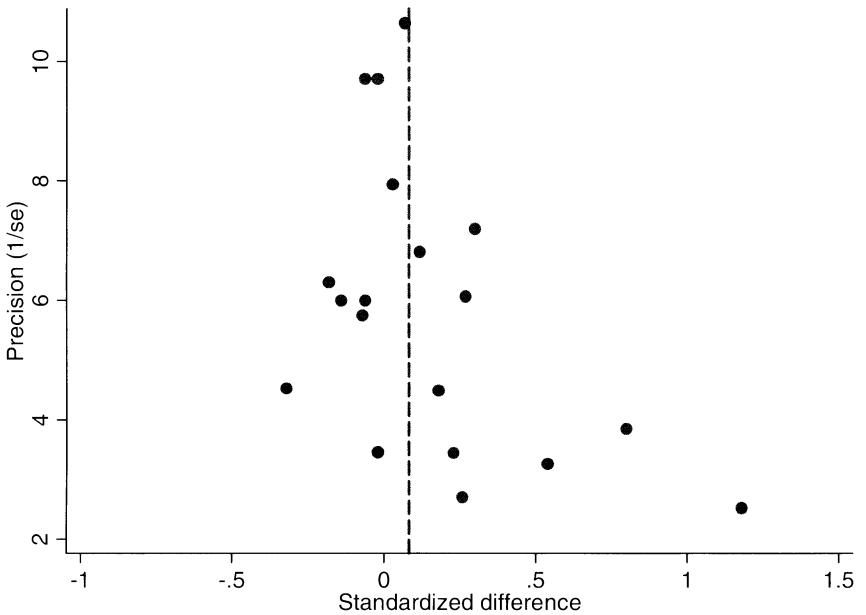

Fig. 4. Funnel plot of studies in the Raudenbush (1984) meta-analysis

Table 1. *p*-values for publication bias: meta-analyses of Mapstone *et al.* (2003) and Raudenbush (1984)

Meta-analysis and subgroups	*Number of experiments*	*p-value*		I^2 *(%)(95% CI)*
		Egger regression test	*Peters regression test*†	
Mapstone et al. (2003)				
Sprague–Dawley rats	16	0.87	0.83	65 (40, 79)
Hebrew University rats	10	0.20	0.93	0 (0, 62)
Wistar rats	9	0.07	0.36	0 (0, 95)
Pigs	7	0.73	0.87	60 (8, 83)
Raudenbush (1984)				
Contact time $\leqslant$ 1 week	8	0.30	—	37 (0, 72)
Contact time $>$ 1 week	11	0.39	—	0 (0, 60)

†Only calculable for binary outcomes, so cannot be used for the Raudenbush meta-analysis.

publication bias is not limited to animal experiments; Sutton *et al.* (2002) have discussed this issue with respect to studies in humans which use different designs, i.e. randomized and observational evidence.

Owing to both issues discussed above, ignoring (explainable) between-study heterogeneity cannot be recommended because it may have serious consequences for the results and conclusions that are drawn from an assessment of publication bias. Alternative approaches are considered below.

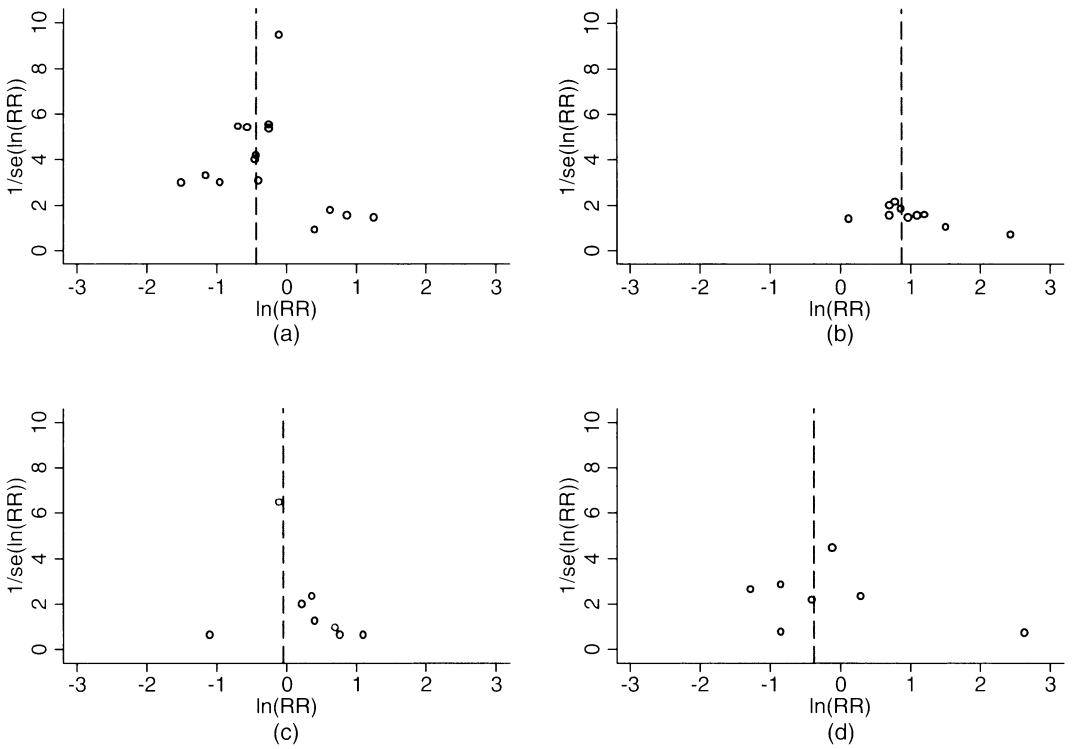

Fig. 5. Funnel plots, with inverse standard error on the *y*-axis, of experiments in the meta-analysis of Mapstone *et al.* (2003) specified by species and rat strain: (a) Sprague–Dawley rats; (b) Hebrew University rats; (c) Wistar rats; (d) pigs

4.2. Assessing publication bias within subgroups

Where covariates are discrete, funnel plots can be plotted for each covariate value. In Fig. 5 separate funnel plots for each species and rat strain subgroup in the data of Mapstone *et al.* (2003) are displayed. Corresponding *p*-values from the two regression tests (Egger *et al.*, 1997; Peters *et al.*, 2006) for publication bias are presented in Table 1 (except for sheep where only one experiment exists). These tests suggest little evidence for publication bias *within* the subgroups. Similarly, funnel plots by contact time are shown in Fig. 6 for the Raudenbush meta-analysis and subgroups, with results for publication bias tests given in Table 1.

This approach allows for the possibility of different effects in the subgroups and differential publication bias, but a large amount of power is lost. For instance, publication bias is assessed within subgroups of 16, 10, nine and seven experiments rather than in a set of 43 experiments in the example from Mapstone *et al.* (2003), whereas recent recommendations suggest that funnel plot tests should only be used when there are at least 10 studies in the meta-analysis (Higgins and Green, 2008). Furthermore, in this meta-analysis, there is still moderate between-study heterogeneity within two of the subgroups (Sprague–Dawley rats and pigs) adding further complexity to the interpretation of the data. Clearly it is possible to assess publication bias in the presence of explainable between-study heterogeneity in this way, but only if the meta-analysis is sufficiently large.

In the next section we consider whether it is possible to distinguish publication bias and between-study heterogeneity through the incorporation of multiple covariates in regression modelling.

4.3. Simultaneous accounting for between-study heterogeneity in publication bias tests

As noted in Section 2.4, random-effects and extended versions of regression tests for publication bias have been evaluated in a simulation study (Peters *et al.*, 2009). In the meta-analysis of Mapstone *et al.* (2003) there is evidence that both residual and explainable heterogeneity exists; a scenario in which the extended tests were not found to be particularly powerful. Moreover, the power of the tests will be decreased further by there being five subgroups in this meta-analysis, when only two subgroups were assessed in the simulation study. For a comparison, an assessment of the meta-analysis of Mapstone *et al.* (2003) gives a p-value of 0.35 from the Egger test with covariates (compared with $p = 0.02$ without covariates as reported earlier in the paper), and a p-value of 0.65 from the Peters test with covariates (compared with $p = 0.70$ without covariates as given earlier). In the Raudenbush meta-analysis only two subgroups are considered and the evidence suggests that the majority of heterogeneity in this meta-analysis is explained by the covariate 'contact time' so there is little residual heterogeneity. In this type of scenario, the Peters extended regression test performed reasonably well, but it cannot be applied here as it is only for binary outcomes. For completeness, however, inclusion of a covariate for contact time in the extended Egger test gives a p-value of 0.7 for the Raudenbush meta-analysis (compared with $p = 0.06$ from the standard Egger test). It is important to note that these non-significant p-values from the extended tests do not necessarily rule out the possibility of publication bias in these two meta-analyses. From the findings of the simulation study, it may be more likely that these models do not have the power to detect publication bias in these meta-analyses.

5. Adjusting for publication bias in the presence of between-study heterogeneity

The previous sections have considered methods to detect and test formally for publication bias. This section considers methods which go further and attempt to adjust meta-analyses for publication bias. We first discuss the non-parametric trim-and-fill method (Duval and Tweedie, 2000a, b) and then consider parametric selection modelling (Hedges and Vevea, 2005).

5.1. Trim-and-fill method

One may wish to apply the trim-and-fill method (Duval and Tweedie, 2000a, b) as a sensitivity analysis to investigate the likely effect that suspected publication bias may have on the overall pooled estimate from a meta-analysis. However, an assessment of the performance of the trim-and-fill method suggests that, in the presence of between-study heterogeneity, it can lead to spurious results (Moreno *et al.*, 2009; Peters *et al.*, 2007; Terrin *et al.*, 2003). Since we have already identified a potential explanation for between-study heterogeneity in the meta-analysis of Mapstone *et al.* (2003) (species and strain) and the Raudenbush meta-analysis (contact time), we would not recommend implementation here unless one was interested in the average effect within the species or strain or contact time groupings. As noted earlier (Section 4.2) assessment within groups reduces the number of studies involved, and the trim-and-fill method has been shown to perform poorly in scenarios where the number of studies is small (Moreno *et al.*, 2009; Peters *et al.*, 2007). Thus with both of the example meta-analyses in this paper we would not advocate use of the trim-and-fill method, even as a form of sensitivity analysis.

5.2. Selection modelling

In this paper we apply a selection model to the Raudenbush meta-analysis, basing selection on the p-value that is associated with a study's outcome. Following Vevea and Woods (2005), we

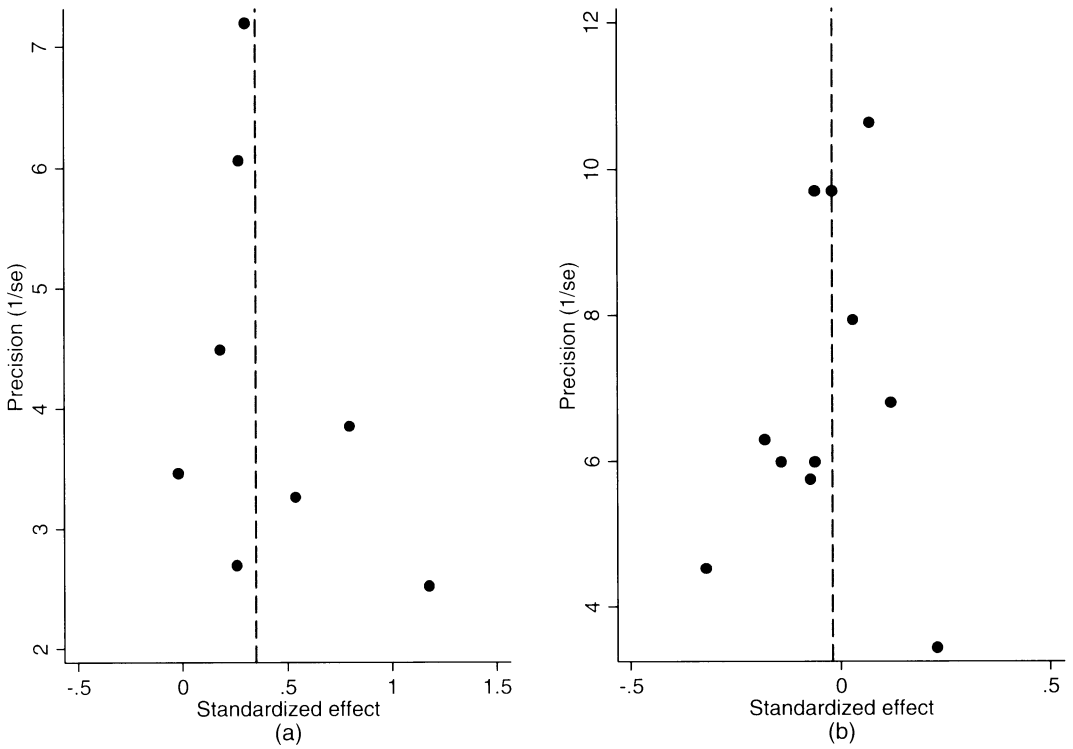

Fig. 6. Funnel plots, with inverse standard error on the y-axis, of studies in the Raudenbush (1984) meta-analysis specified by contact time subgroup: (a) 1 week or less; (b) more than 1 week

specify the weights at each cut point (rather than estimate them as in Hedges and Vevea (1996)). In so doing, we consider the analysis to be a form of sensitivity analysis rather than definitive. In other words, we use specifications of the selection model to investigate how the findings of the meta-analysis may be affected by publication bias, rather than trying to identify an effect size adjusted for publication bias. We assume a fixed effects model and include a binary covariate to account for the prior contact time. The effect size model is

$$y_i = \beta_0 + \beta_1 x_i,$$

where y_i is the standardized mean difference in scores for study i, with x_i indicating whether the contact time was a week or less ($x_i = 0$), or more than a week ($x_i = 1$). The weights to be specified for the selection model are given in Table 2. We specify four selection models: *moderate one-tailed selection, severe one-tailed selection, moderate two-tailed selection* and *severe two-tailed selection* (Vevea and Woods, 2005).

The results in Table 3 show that the estimates do not change appreciably when different selection mechanisms are assumed, so different conclusions on the existence of an effect are unlikely to be drawn, regardless of the model chosen. This therefore suggests that the Raudenbush meta-analysis is robust to the effect of publication bias (on the basis of p-values).

There are many other types of selection model that could be applied. For example, the likelihood of selection could be defined by the size of the estimate, rather than the p-value, or both (Copas, 1999). Furthermore, rather than setting the probabilities of selection as we have done here (Table 2), in a large meta-analysis, these could be estimated by the data and an adjusted effect obtained (Hedges and Vevea, 2005). Random effects could also be included to

Table 2. Selection models assumed

One-sided selection			*Two-sided selection*		
p-value interval	*Probability of selection*		*p-value interval*	*Probability of selection*	
	Moderate	*Severe*		*Moderate*	*Severe*
$0 < p \leqslant 0.01$	1	1	$0 < p \leqslant 0.05$	1	1
$0.01 < p \leqslant 0.05$	0.99	0.99	$0.05 < p \leqslant 0.1$	0.9	0.9
$0.05 < p \leqslant 0.1$	0.95	0.8	$0.1 < p \leqslant 0.4$	0.8	0.4
$0.1 < p \leqslant 0.2$	0.9	0.6	$0.4 < p \leqslant 0.6$	0.6	0.2
$0.2 < p \leqslant 0.5$	0.8	0.4	$0.6 < p \leqslant 0.9$	0.8	0.4
$0.5 < p \leqslant 0.8$	0.75	0.2	$0.9 < p \leqslant 0.95$	0.9	0.9
$p > 0.8$	0.6	0.1	$p > 0.95$	1	1

Table 3. Results of selection models applied to the Raudenbush (1984) meta-analysis

Model assumed		*Intercept* β_0	*Coefficient for covariate* β_1
One or two sided	*Severity of selection*		
No selection assumed		0.35 (0.08)	−0.37 (0.09)
One	Moderate	0.31 (0.09)	−0.36 (0.09)
One	Severe	0.27 (0.09)	−0.33 (0.10)
Two	Moderate	0.34 (0.08)	−0.41 (0.09)
Two	Severe	0.29 (0.09)	−0.37 (0.09)

allow for residual heterogeneity (see Hedges and Vevea (2005)). Further, it would be possible to allow for different levels of publication bias for studies with different covariate values. Using selection models, publication bias can be disentangled from between-study heterogeneity but large meta-analyses and/or strong assumptions are generally needed.

6. Discussion

The aim of this paper was to make a critical assessment of possible approaches to addressing publication bias in the presence of both explainable and residual between-study heterogeneity. This is an important topic since it is not uncommon that both are present in a meta-analysis, although they are usually investigated independently despite the fact that each can have an influence on the assessment and interpretation of the other. As noted, Ioannidis and Trikalinos (2007) observed continuing inappropriate use of publication bias tests. Acknowledgement of the limitations of methods for publication bias needs to be made, particularly in published recommendations, although the *Cochrane Handbook* clearly outlines the shortcomings of publication bias tests in particular (Higgins and Green, 2008).

We have focused here on methods to assess and adjust for publication bias in the presence of explainable and residual between-study heterogeneity, but further work on how publication

bias affects the assessment of heterogeneity would be valuable, e.g. extending Jackson's (2006) investigation of the implications of possible publication bias on estimation of the between-study heterogeneity parameter in random-effects meta-analyses.

Because of this codependence, methods which consider or model both publication bias and heterogeneity simultaneously are clearly desirable from a theoretical point of view. The extended funnel plot tests and selection modelling approaches are both capable of including parameters for publication bias and explainable and unexplainable heterogeneity. However, a disadvantage of such methods is that they lack the visual directness of funnel plots and methods which are based on them (such as trim and fill). A recent simulation study suggests that regression-based models which are commonly used to *test* for publication bias may be useful to *adjust* for publication bias (Moreno *et al.*, 2009). An advantage of this approach, which can account for residual heterogeneity, is that the adjusted estimate can be visualized on a funnel plot (see Fig. 1 of Moreno *et al.* (2009)). However, the performance of these models deteriorates as I^2 exceeds 50%, only residual, not explainable, heterogeneity was simulated and the models have only been assessed with ORs as the effect estimate. Additionally, Moreno *et al.* (2009) did not simulate explainable heterogeneity further than that induced by publication bias in the form of small study effects. Thus, further research is required to assess the performance of the regression-based approach to adjusting for publication bias in such circumstances.

We have only considered categorical covariates to explain between-study heterogeneity in this paper, not continuous study level covariates. For instance, contact time in the Raudenbush meta-analysis was dichotomized into 'a week or less' and 'more than a week', although the actual contact times were available. Assessing publication bias within the subgroup would then not be possible, but selection modelling or the extended funnel plot tests could still be used with continuous covariates.

Furthermore, as seen in Mapstone *et al.* (2003), it is likely that more than one factor may contribute to explaining observed between-study heterogeneity. These other factors could be included in the selection models or extended regression tests but will require a large meta-analysis if there is to be sufficient power to estimate subsequent effects.

In conclusion, even the more promising approaches (selection models and extended funnel plot tests) are limited by the low power that they have in all except the largest meta-analysis. Additionally, there is always the possibility of unknown and unmeasured variables confounding the appearance of the funnel plot. Furthermore, there are issues surrounding the selection of covariates in the extended publication bias tests and the selection modelling.

Thus the message from this paper is that although methods exist for simultaneous assessment of heterogeneity and publication bias, and potential differential publication bias, it is unrealistic to expect reliable disentangling of effects of both in all except the largest meta-analyses. This should be borne in mind when interpreting meta-analyses in which both heterogeneity and publication bias are suspected.

Acknowledgement

JP was funded through a UK Department of Health evidence synthesis award during part of this work. The source of funding had no role in any aspect of this study.

References

Begg, C. B. and Mazumdar, M. (1994) Operating characteristics of a rank correlation test for publication bias. *Biometrics*, **50**, 1088–1101.

Copas, J. (1999) What works?: selectivity models and meta-analysis. *J. R. Statist. Soc.* A, **162**, 95–109.

Deeks, J. J., Macaskill, P. and Irwig, L. (2005) The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J. Clin. Epidem.*, **58**, 882–893.
Duval, S. and Tweedie, R. (2000a) Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, **56**, 455–463.
Duval, S. and Tweedie, R. L. (2000b) A nonparametric "Trim and Fill" method of accounting for publication bias in meta-analysis. *J. Am. Statist. Ass.*, **95**, 89–98.
Egger, M., Davey Smith, G., Schneider, M. and Minder, C. (1997) Bias in meta-analysis detected by a simple, graphical test. *Br. Med. J.*, **315**, 629–634.
Engels, E. A., Schmid, C. H., Terrin, N., Olkin, I. and Lau, J. (2000) Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Statist. Med.*, **19**, 1707–1728.
Harbord, R. M., Egger, M. and Sterne, J. A. C. (2006) A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statist. Med.*, **25**, 3443–3457.
Hedges, L. V. and Vevea, J. L. (1996) Estimating effect size under publication bias: small sample properties and robustness of a random effects selection model. *J. Educ. Behav. Statist.*, **21**, 299–332.
Hedges, L. V. and Vevea, J. L. (2005) Selection method approaches. In *Publication Bias in Meta-analysis: Prevention, Assessment and Adjustments* (eds H. Rothstein, A. J. Sutton and M. Borenstein). Chichester: Wiley.
Higgins, J. P. T. and Green, S. (2008) *Cochrane Handbook for Systematic Reviews of Interventions, Version 5.0.1*. Oxford: Cochrane Collaboration. (Available from `http://www.cochrane-handbook.org`.)
Higgins, J. P. T. and Thompson, S. G. (2002) Quantifying heterogeneity in a meta-analysis. *Statist. Med.*, **21**, 1539–1558.
Higgins, J. P. T., Thompson, S. G., Deeks, J. J. and Altman, D. G. (2003) Measuring inconsistency in meta-analyses. *Br. Med. J.*, **327**, 557–560.
Ioannidis, J. P. A. and Trikalinos, T. A. (2007) The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *Can. Med. Ass. J.*, **176**, 1091–1096.
Jackson, D. (2006) The implications of publication bias for meta-analysis' other parameter. *Statist. Med.*, **25**, 2911–2921.
Lewis, S. J., Ebrahim, S. and Davey Smith, G. (2005) Meta-analysis of MTHFR 677C->T polymorphism and coronary heart disease: does totality of evidence support causal role for homocysteine and preventive potential of folate? *Br. Med. J.*, **331**, 1053.
Light, R. J. and Pillemar, D. B. (1984) *Summing up: the Science of Reviewing Research*. Cambridge: Harvard University Press.
Macaskill, P., Walter, S. D. and Irwig, L. (2001) A comparison of methods to detect publication bias in meta-analysis. *Statist. Med.*, **20**, 641–654.
Mapstone, J., Roberts, I. and Evans, P. (2003) Fluid resuscitation strategies: a systematic review of animal trials. *J. Trauma*, **55**, 571–589.
Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L. and Cooper, N. J. (2009) Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Med. Res. Methodol.*, **9**, article 2.
Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R. and Rushton, L. (2006) Comparison of two methods to detect publication bias in meta-analysis. *J. Am. Med. Ass.*, **295**, 676–680.
Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R. and Rushton, L. (2007) Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statist. Med.*, **26**, 4544–4562.
Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., Rushton, L. and Moreno, S. G. (2009) Performance of regression tests to detect publication bias in meta-analyses in the presence of between-study heterogeneity. *Technical Report*. University of Leicester, Leicester.
Raudenbush, S. W. (1984) Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction. *J. Educ. Psychol.*, **76**, 85–97.
Rücker, G., Schwarzer, G. and Carpenter, J. (2008) Arcsine test for publication test in meta-analyses with binary outcomes. *Statist. Med.*, **27**, 746–763.
Schwarzer, G., Antes, G. and Schumacher, M. (2002) Inflation of type I error rate in two statistical tests for the detection of publication bias in meta-analyses with binary outcomes. *Statist. Med.*, **21**, 2465–2477.
Song, F., Eastwood, A. J., Gilbody, S., Duley, L. and Sutton, A. J. (2000) Publication and related biases. *Hlth Technol. Assessmnt*, **4**, no. 10.
Sterne, J. A. C., Becker, B. J. and Egger, M. (2005) The funnel plot. In *Publication Bias in Meta-analysis: Prevention, Assessment and Adjustments* (eds H. R. Rothstein, A. J. Sutton and M. Borenstein), pp. 75–98. Chichester: Wiley.
Sterne, J. A. C., Gavaghan, D. and Egger, M. (2000) Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J. Clin. Epidem.*, **53**, 1119–1129.
Sutton, A. J., Abrams, K. R. and Jones, D. R. (2002) Generalized synthesis of evidence and the threat of dissemination bias: the example of electronic fetal heart rate monitoring (EFM). *J. Clin. Epidem.*, **55**, 1013–1024.
Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A. and Song, F. (2000a) *Methods for Meta-analysis in Medical Research*. Chichester: Wiley.

Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R. and Jones, D. R. (2000b) Empirical assessment of effect of publication bias on meta-analyses. *Br. Med. J.*, **320**, 1574–1577.
Terrin, N., Schmid, C. H. and Lau, J. (2005) In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *J. Clin. Epidem.*, **58**, 894–901.
Terrin, N., Schmid, C. H., Lau, J. and Olkin, I. (2003) Adjusting for publication bias in the presence of heterogeneity. *Statist. Med.*, **22**, 2113–2126.
Vevea, J. L. and Woods, C. M. (2005) Publication bias in research synthesis: sensitivity analysis using a priori weight functions. *Psychol. Meth.*, **10**, 428–443.