Mawdsley, D., Higgins, J., Sutton, A. J., & Abrams, K. (2017). Accounting for heterogeneity in meta-analysis using a multiplicative model—an empirical study. *Research Synthesis Methods*, *8*(1), 43–52. https://doi.org/10.1002/jrsm.1216

Accounting for heterogeneity in meta-analysis using a multiplicative model–an empirical study

David Mawdsley[*,†], Julian P. T. Higgins[*], Alex J. Sutton[†], and Keith R. Abrams[†]

[*]School of Social and Community Medicine, University of Bristol
[†]Department of Health Sciences, University of Leicester

Abstract

In meta-analysis the random-effects model is often used to account for heterogeneity. The model assumes that heterogeneity has an additive effect on the variance of effect sizes. An alternative model which assumes multiplicative heterogeneity has been little used in the medical-statistics community, but is widely used by particle physicists. In this paper we compare the two models using a random sample of 448 meta-analyses drawn from the Cochrane Database of Systematic Reviews. In general, differences in goodness of fit are modest. The multiplicative model tends to give results that are closer to the null, with a narrower confidence interval. Both approaches make different assumptions about the outcome of the meta-analysis. In our opinion, the selection of the more appropriate model will often be guided by whether the multiplicative model's assumption of a single effect size is plausible.

Introduction

A meta-analysis (MA) provides a quantitative summary of a measure taken from a collection of studies usually obtained via a systematic review process. MAs are widely used in economics, the social sciences and in medicine (see, e.g. Whitehead (2002); Borenstein et al. (2009)). Although the studies included in a MA should be comparable, it is not uncommon for there to be heterogeneity between the effect sizes of the studies.

Two main models are commonly applied in MA; the fixed-effect model, and the random-effects model. The former assumes that there is a single true effect size, and which each study included in the MA is measuring, and so any variation between studies is assumed to be due to random error. In contrast, the random-effects model allows for heterogeneity and assumes that the effect sizes of the studies included in the MA are a random sample of all effect sizes from studies addressing the question that met the criteria for inclusion in the underlying systematic review. The distribution of these effect sizes, with a focus on its mean, is usually the outcome of the random-effects MA (Borenstein et al., 2009).

Under the random-effects model (see, e.g. Whitehead (2002)):

$$\hat{\theta}_i \sim \mathrm{N}(\theta, v_i^2 + \tau^2)$$

where $\hat{\theta}_i$ and v_i^2 are the observed estimates of the effect size and sample variance of the ith study, and τ^2 is the between-studies variance. A variety of estimators have been proposed for τ^2 (DerSimonian and Kacker, 2007; Veroniki et al., 2015), the most frequently used of which is probably the DerSimonian and Laird estimator (DerSimonian and Laird, 1986).

Under the random-effects model, the weight assigned to each study in the MA is given by:

$$w_i = \frac{1}{v_i^2 + \tau^2}$$

where v_i is the variance of the ith study. Both the study variances and τ^2 are generally treated as known, and hence their uncertainties ignored, although profile likelihood (Hardy and Thompson, 1996) and Bayesian methods (Smith et al., 1995) can relax these assumptions. Thus, in the presence of heterogeneity, as τ^2 increases relative to the v_is, the weights assigned to each study will become relatively more similar. This has the effect of reducing the influence of studies with a small variance (usually the larger studies in the MA).

Thompson and Sharp (1999) considered an alternative method of accounting for heterogeneity by the use of a multiplicative parameter:

$$\hat{\theta}_i \sim \mathrm{N}(\theta, v_i^2 \phi)$$

with ϕ, which is estimated from the data, set to 1 if the estimate is < 1. This model gives the same effect size as a fixed-effect analysis, but with standard errors inflated by $\sqrt{\phi}$. The appropriate value of ϕ can be established by running a linear regression of the observed effect sizes against a constant, with weights $w_i = 1/v_i$ and extracting the mean squared error (Thompson and Sharp, 1999). We note that Higgins's $H^2 = Q/k - 1$ (Higgins and Thompson, 2002) is equivalent to this

quantity (Stanley and Doucouliagos, 2015), which is also referred to as Birge's ratio (see, e.g. Cooper et al. (2009)).

In contrast to the random-effects model, which accounts for heterogeneity by adding a constant, τ^2 to each study's variance, the variance of each study is multiplied by a constant, ϕ. Smaller studies therefore have a smaller absolute increase in variance under the multiplicative model than the additive random-effects model.

This model has, until recently, received little attention in the medical statistics literature, although as Baker and Jackson (2013) observe, its use is well established in the synthesis of the results of particle physics experiments (Rosenfeld, 1975). In contrast to the random-effects model, the relative weights of each study are independent of the level of observed heterogeneity. Interestingly, the particle physicists exclude studies with a very low precision, relative to the rest of the MA, using an "arbitrarily chosen" threshold (Beringer et al., 2012). They reason that, although these studies can only make a very modest contribution to the overall effect size, they can make significant contributions to the amount of heterogeneity observed.

Thompson and Sharp did not recommend the multiplicative model, noting that:

> "The rationale for using a multiplicative factor for variance inflation is weak. The idea that the variance of the estimated effect within each study should be multiplied by some constant has little intuitive appeal, and leads to the same dominance of large studies over small studies that has been criticised in the context of fixed-effect meta-analysis."

A multiplicative model has recently been re-investigated by Stanley and Doucouliagos (2015). They show, via a simulation study, that under large levels of additive heterogeneity the coverage of the model outperforms the additive random-effects model (although they do not impose the requirement $\phi \geq 1$). They also note the multiplicative model has better coverage and resistance to bias in the presence of publication bias. As they observe "the advantage of the weighted least squares meta-regression [i.e. the multiplicative model] is largest in those exact conditions for which random-effects are designed—large additive heterogeneity".

Implications of the fixed-effect, additive random-effects and multiplicative models are illustrated for two example data-sets (discussed in more detail later in the paper) using the funnel plots (Light and Pillemer, 1984) in Figure 1. In addition to plotting the observed studies on the plots, we include 95% pseudo confidence regions, in which we expect 95% of the observed study estimates to lie under each of the models. They serve to illustrate how the variation in heterogeneity with respect to the precision of studies may be better modelled by the either the additive or multiplicative term.

This paper uses a large collection of MAs from the January 2008 issue of the Cochrane Database of Systematic Reviews (CDSR) to perform an empirical study of the effects of using the multiplicative model instead of the random-effects model. The database contains 2,321 reviews, which contain 22,453 meta-analyses, classified by medical speciality, type of intervention in the comparison (e.g. pharmacological, surgical, etc.) and type of outcome (e.g. mortality, pain, quality of life, etc.). The data cover a wide range of subject areas, interventions and outcome types and are described more fully by Davey et al. (2011).

Methods

The fixed-effect, random-effects and multiplicative models were fitted to a random sample of binary outcome data MAs taken from the CDSR. A systematic review could use the same study in more than one meta-analysis (for example in meta-analyses looking at different outcomes, such as an efficacy outcome and a safety outcome). In order to ensure independence between analyses a single meta analysis from each systematic review was selected at random. Studies were combined across subgroups within the same CDSR forest plot. As subgroup analyses were not necessarily mutually exclusive it would have been possible for a study to be included more than once in a meta-analysis; these duplicates had previously been removed from the data as described by Davey et al. (2011).

It would be difficult to compare the appropriateness of the two models in meta-analyses with very few studies, owing to the difficulties in estimating τ^2 and ϕ, therefore if a meta-analysis had fewer than 10 studies it was excluded from the database.

The odds-ratio was used as the primary outcome measure, although the analysis was repeated using the risk-ratio as a sensitivity analysis. The metafor package (version 1.9-3) (Viechtbauer, 2010) for R (version 3.1.2) (R Core Team, 2014) was used to calculate (logged) effect sizes and to fit the random-effects models. τ^2 was calculated using the Paule and Mandel estimator (Paule and Mandel, 1982), as this has been shown to be statistically optimal (DerSimonian and Kacker, 2007). We extended the metafor package's code to allow the fitting of the multiplicative model. This was done by inflating the variances of the studies contributing to a fixed-effect meta-analysis, calculating ϕ as described above. Following Thompson and Sharp (1999), where $\phi < 1$ we set $\phi = 1$.

A continuity correction of 0.5 was added to all cells of the 2x2 table for trials with at least one zero in any of the cells. Trials with no events in either arm were removed, as such trials provide no information on which group has the higher risk when using the odds ratio as an outcome measure (Higgins and Green, 2011).

The goodness of fit of the models was compared by taking differences in AIC

between the additive and multiplicative model. As smaller values of the AIC indicate a better fitting model, positive values of this quantity indicate that the multiplicative model is favoured over the additive model.

The magnitudes of the effect sizes and widths of their associated confidence intervals were compared under each model. An exploratory regression analysis was performed to investigate characteristics of the MAs that were better supported by each model. Covariates (described below) were regressed in univariate analyses against the difference in AIC. As positive values of this quantity indicate that the multiplicative model is favoured over the additive model, positive values of the regression coefficient, β, suggest that larger values of the covariate favour the multiplicative model.

We considered plausible covariates by considering the differences in the assumptions underlying the models, and examination of the forest plots of MAs that were particularly well fitted by one or other of the models. These showed that the multiplicative model tended to favour MAs where the largest studies had similar effect sizes. In the presence of heterogeneity these studies will have larger weights in the multiplicative model than under the additive model. In order to quantify this difference, we calculated each study's difference in effect size from the effect size of the most precise study in the MA. As the sign of this difference is unimportant, and because it seems reasonable to penalise larger differences more, this quantity was squared. The squared difference was divided by the study's variance, reflecting the weight it has in the fixed-effect MA. This gives:

$$D^2 = \sum_{i=1}^{k} \frac{(y_i - y_{\text{max prec}})^2}{v_i}$$

Where y_i and v_i are the effect size and variance of the ith study, and $y_{\text{max prec}}$ is the effect size of the study with the smallest variance in the MA. The quantity is similar to Cochran's Q statistic, which we also considered. Logs of these quantities were taken to reduce skew.

As the multiplicative model assumes a single effect size, it may be more difficult to satisfy this requirement when there are many studies included in the MA. Consequently, we considered number of studies in the MA as a covariate.

When there is little variation in the precisions of studies contributing to the MA there can be less difference between the additive and multiplicative models, as the relative weights of the studies will be quite similar. To quantify the level of variation in study precisions, the variance of the relative variances (i.e. $v_i / \sum v_i$) was calculated for each MA. Logs of this quantity were taken to reduce skewness.

We considered an alternative method of exploring the heterogeneity within a MA by splitting each MA into tertiles by study size. Additive and multiplicative models were then separately fitted to the largest third of studies and the smallest

third of studies. As the median number of studies per MA included in the regression analysis was 12, we note that these estimates will, in general, be associated with considerable uncertainty.

Results

Of the 22,453 MAs in the database, 14,886 were for binary outcomes (excluding MAs that had reported binary results as part of a MA of a survival analysis). There were 7,265 studies with no events in both treatment and control arms, which were removed from the MA they were included in. After these studies were removed, MAs with fewer than 10 studies remaining were removed giving 1,402 potential meta-analyses. One MA from each systematic review was randomly selected, to ensure independence, giving a final sample of 448 MAs. Table 1 provides summary statistics of the MAs that were included.

Of the 448 meta-analyses, 160 (35.7%) had $\hat{\phi} < 1$; $\hat{\phi}$ was set to 1 for these studies. 145 MAs (32.4%) had $\hat{\tau}^2 = 0$, all of which also had $\hat{\phi} = 1$.

Figure 2 shows the relationship between the variance of the log odds ratios, τ^2, and the variance scaling factor, ϕ, for each of the meta-analyses. There is a fairly strong correlation between the two measures of heterogeneity ($\rho_{\text{pearson}} = 0.536$, excluding studies with $\tau^2 = 0$ and $\phi = 1$, $\rho'_{\text{pearson}} = 0.478$), although some MAs have a relatively large value of τ^2 in the additive analysis, and a small value of ϕ in the multiplicative analysis (or vice versa).

Excluding the 145 MAs for which heterogeneity was estimated to be zero, and thus equivalent to the fixed-effect model, the median difference in AIC was 0.028 (Inter-quartile range: −0.5 to 0.884). A histogram of AIC differences is shown in Figure 3. A difference in AIC of at least two is often taken as showing a substantial improvement in a model. By this criterion 9.38% of the studies were substantially better fitted by the multiplicative model and 6.92% of MAs are better fitted by the random-effects model. Considering a more stringent requirement of taking differences of at least five as being important, 2.01% of MAs are better fitted by the additive model. The same percentage (2.01%) are better fitted by the multiplicative model.

Effect sizes and confidence intervals

Two hundered and twenty MAs recorded an absolute effect size closer to the null under the multiplicative model than the additive model. Eighty three had a larger absolute effect size. The remaining MAs were those that were equivalent to the fixed-effect model. MAs' effect sizes were slightly smaller under the multiplicative model (paired t-test, mean difference in LOR: 0.021, p=0.002).

Two hundred and eighty six MAs had a narrower 95% confidence interval under the multiplicative model, 17 MAs had a wider interval (the remaining MAs had $\phi = 1$ and $\tau^2 = 0$ and so had identical CIs). There was strong evidence for a difference in CI widths between the two models (paired t-test, mean difference: 0.076, p<0.001).

Table 2 shows the number of MAs that were statistically significant at the 5% level under each of the three models. As the multiplicative model has the same point estimate as the fixed-effect model, but with a wider CI when $\phi > 1$, none of the MAs which were not statistically significant under fixed-effect can be statistically significant under the multiplicative model. Twenty three MAs are significant under the fixed-effect model, but not under either the additive or multiplicative models. Twenty six MAs were statistically significant under only one of the additive or multiplicative models.

Predictors of the more appropriate model

We performed an exploratory linear regression analysis to investigate the factors that influenced whether a particular MA was likely to be better fitted by the additive or multiplicative model. Owing to the large amount of unexplained residual variation we verified that the univariate linear regressions explained the broad trends in the data by visually comparing the results of the linear regression to a LOWESS smoothed curve (Cleveland, 1981). The 145 MAs which had $\tau^2 = 0$ under the additive model and $\phi = 1$ under the multiplicative model were excluded from this analysis, as they would be best fitted by the fixed-effect model. A sensitivity analysis showed that their inclusion had little influence on the model.

There was a statistically significant association between MAs with large values of $\log(D^2)$ and their being better fitted by the additive model ($\beta = -0.697$, $p < 0.001$), although this explained relatively little of the observed variation ($R^2 = 0.046$).

We found that the log of Cochran's Q statistic explained less of the observed variation than $\log(D^2)$ ($\beta = -0.417$, $p = 0.039$, $R^2 = 0.014$). No significant association between number of studies and model appropriateness was found ($\beta = -0.003$, $p = 0.742$, $R^2 = 0$). MAs with a large amount of variation in variances tended to be better fitted by the multiplicative model ($\beta = 0.283$, $p = 0.029$), although this explained little of the variance in the regression($R^2 = 0.016$).

When large amounts of heterogeneity were estimated in the smaller studies (using either additive or multiplicative models), this was strongly associated with the multiplicative model being favoured (Additive: $\beta = 1.029$, $p < 0.001$, $R^2 = 0.189$, multiplicative: $\beta = 0.651$, $p < 0.001$, $R^2 = 0.091$). Large amounts of heterogeneity in the larger studies were associated with the additive model being favoured, although this result was not statistically significant using the additive

model (additive: $\beta = -0.086$, $p = 0.6$, $R^2 = 0.001$, multiplicative: $\beta = -0.179$, $p < 0.001$, $R^2 = 0.078$.).

The relative levels of heterogeneity between the largest and smallest thirds of studies was also considered. The ratio $\phi_{\text{large}}/\phi_{\text{small}}$ was calculated for each MA. Large amounts of heterogeneity in the small studies, relative to the large studies, led to the multiplicative model being favoured ($\beta = -0.76$, $p < 0.001$, $R^2 = 0.289$).

Multivariate models were explored, considering the number of studies, $\log(D^2)$, the logged variance of relative study variances, and $\phi_{\text{large}}/\phi_{\text{small}}$ as covariates. These covariates were selected as they represent different aspects of the model, and are relatively uncorrelated (all pairwise correlations have $|\rho| < 0.5$, with the exception of number of studies and logged variance of relative study variances ($\rho = -0.677$).)

The glmulti package (Calcagno and de Mazancourt, 2010) for R was then used to find the best fitting main-effects model, using these covariates, taking the AIC as the model selection criterion. This did not result in the removal of any covariates. All two level interactions were then included in the potential model, and the selection process repeated. Three way or higher level interactions were not considered. This gave the model shown in Table 3, which contained all two-way interactions with the exception of the $k \times D^2$ interaction.

In order to assist interpretation of this model, Table 4 shows the predicted differences in AIC for hypothetical MAs with the minimum, maximum and median value of each covariate when the remaining covariates take their median values. For typical values of the other parameters, the ratio of heterogeneity between the largest and smallest studies appears to be the most significant predictor of whether the additive or multiplicative model is preferred. MAs with highly heterogeneous large studies, relative to the heterogeneity of the smaller studies tend to be favoured by the additive model.

Alternative outcome measures

The use of the log risk-ratio was considered as an alternative outcome measure. The difference in AIC between the additive and multiplicative models was calculated, and compared to the differences in AIC when the OR was used as the outcome measure. Taking differences in AIC of at least two as significant, 323 of the 448 MAs in the sample were not significantly better fitted by either model under either outcome measure. There were 78 MAs that were significantly better fitted by the same model under both outcome measures, and 5 MAs whose significantly preferred model depended on the outcome measure used. The remaining 42 MAs were significantly better fitted by one or other of the models under only one of the outcome measures.

Illustration with example data-sets

Thompson and Sharp (1999) criticise the "dominance of large studies over small studies" under the multiplicative model, compared with the additive model. In some circumstances this could be considered an advantage. Consider the well known and much studied example of magnesium to treat acute MI (see, e.g. Li et al. (2007); Higgins (2002)). Here, small, early trials and meta-analyses showed substantial reductions in mortality for patients treated with intravenous magnesium. In contrast, the 1995 ISIS-4 "mega trial" of 58,050 participants, showed "no significant reduction in 5-week mortality, either overall. . . or in any subgroup examined". The substantial heterogeneity in the MA has the effect of down-weighting the ISIS-4 trial to the extent that it is impossible for it to supersede the earlier results; a random-effects analysis including the trial suggests that magnesium is an effective intervention (OR: 0.55, 95% CI: 0.39 to 0.78). In contrast, the multiplicative or fixed-effect models suggests there is little evidence for the intervention's efficacy (OR: 1.02, 95% CI (multiplicative model): 0.92 to 1.13, 95% CI (fixed-effect model): 0.96 to 1.08). Given the very different effect sizes reported in the studies underlying the analysis, the more liberal assumptions of the random-effects model appear to be more appropriate, even though this model gives a more favourable result than a recent Cochrane review, which concluded magnesium was unlikely to be effective (Li et al., 2007). We note, however, that as with any meta-analysis, (possibly unmeasured) study level covariates could explain some of this heterogeneity, and that any residual heterogeneity would still need to be modelled appropriately having modelled these.

Figure 1 shows funnel plots for the datasets used in our empirical investigation in which the difference in fit between additive and multiplicative models was the greatest. The additive model fitted dataset 1 the best while the second dataset plotted was fitted best by the multiplicative model. The 95% pseudo confidence intervals included on the plots indicate the assumptions about the nature of the heterogeneity, as described in the introduction. For neither dataset does a fixed effect model look reasonable (Figures 1a and d) as several studies exist outside the confidence region. For dataset 1 (Figures 1a, b, c) we see that the heterogeneity is fairly constant over the whole range of standard errors—the assumption that is consistent with the additive random effects model. In contrast, dataset 2 (Figure 1d, e, f) would appear to have heterogeneity which increases with standard error and such an appearance is best accommodated by the multiplicative error term (considerably inflated beyond that allowed under the fixed effect assumption).

Discussion

This work has shown that, according to goodness of fit statistics, there are rarely significant differences in model fit observed when comparing the use of an additive or multiplicative meta-analysis model. There are, however, some MAs that are clearly better fitted by one or other of the models. The reasons behind this were examined using an exploratory regression analysis, which showed that large amounts of heterogeneity in the larger studies (particularly when this was large relative to the amount of heterogeneity in the smaller studies) led to the additive model being favoured. This is due to the relatively lower weights that will be assigned to the large studies under the additive model.

The multiplicative model tended to give effect sizes that were closer to the null and narrower confidence intervals than the additive model. We note that smaller (and potentially more biased) studies tend to show larger effect sizes (Nüesch et al., 2010; Sterne et al., 2000, 2001; Turner et al., 2012), and that these will have less weight under the multiplicative model, in the presence of heterogeneity. As this is an empirical study, the true effect size and confidence interval are unknown.

Although it is useful to think about the additive and multiplicative models in terms of comparing goodness of fit statistics, the models have quite different assumptions. In common with the fixed-effect model, the multiplicative approach assumes a single overall effect size. In contrast, the outcome of the random-effects model is the mean of the effect sizes for all the studies that could have contributed to the MA. Although the model's goodness of fit is clearly an important factor, we would argue that the appropriateness of the underlying the model is also important. For the multiplicative model, the assumption of a single effect size will often be unrealistic; the characteristics of participants and the precise nature of the intervention in the trials contributing to the meta-analysis will often vary.

A single effect size is likely to be a more tenable assumption in some circumstances. The particle physics meta-analyses considered by Baker and Jackson (2013) should be measuring a single effect (for example, the mass of a fundamental physical particle). In medical statistics there are few situations where we can be confident that a single effect size is the most appropriate outcome of a meta-analysis. In the presence of very similar study protocols, as might be observed if the results of a multi-centre trial were combined using a meta-analysis, it may well be appropriate to assume a single effect size. We cannot, however, rule out that the true effect measured by each centre is different; for example, different experience levels of staff performing the intervention or differences in the baseline characteristics of participants could lead to the efficacy of the intervention varying between sites.

The models used in this paper were applied to binary outcomes, using the log odds-ratio as the outcome measure. A sensitivity analysis showed that similar

results were obtained using the log risk-ratio. The models used in this paper could, in principle, be applied to other types of outcome measure, provided the normality assumptions that underlie the models can be satisfied.

This work used on a frequentist approach, although all of the models discussed here can be readily implemented using MCMC methods, in WinBUGS (Lunn et al., 2000), for example. Implementing the models using an MCMC approach provides an estimate of the uncertainties in τ^2 and ϕ and incorporates this uncertainty into the models' estimated effect sizes.

In this work, we set $\phi = 1$ if its estimate was < 1. Others have considered similar MA models with an over-dispersion parameter and allowed it to be estimated as < 1 (i.e. implying under-dispersion) (Stanley and Doucouliagos, 2015; Moreno et al., 2009; Kulinskaya and Olkin, 2014). $\phi < 1$ implies there is less between study variation in the MA than we would expect by chance; this may occur if the same study is included multiple times in the MA or in the presence of publication bias. We would argue against allowing $\phi < 1$, but would suggest that it could be useful as a diagnostic tool.

It is known that the random-effects model is sensitive to publication bias (Brockwell and Gordon, 2001). Henmi and Copas (2010) proposed a model that uses the point estimate obtained from a fixed-effect analysis, but incorporates the uncertainty due to τ^2 in the confidence interval. As the fixed-effect model down-weights the smaller studies (which are more prone to bias), relative to the random-effects model, they show that their approach is less prone to bias. The multiplicative model discussed in this paper shares many of the appealing features of Henmi and Copas's model; under both models the relative weights of the studies will be equal, and identical to the fixed-effect weights. Both may return a wider confidence interval than the fixed-effect model, reflecting any observed heterogeneity.

Baker and Jackson (2013) explored the goodness of fit of the additive and multiplicative models by maximising the likelihood of a model that allows one to interpolate between the additive and multiplicative models. They fitted this model to (separate) collections of medical and physics MAs, and obtained results consistent with the additive model, on average, better fitting both groups of MAs. Their model raises the intriguing possibility that a 'hybrid' model that combines elements of the multiplicative and additive models may be more appropriate for some MAs. This would avoid the rather strict assumption of a single effect size, while mitigating the equalisation of study weights that is characteristic of the additive model in the presence of heterogeneity. We note that some of the MAs examined in this paper had rather large estimates of τ^2 and rather small estimates of ϕ, and vice versa (Figure 2), suggesting they may well be measuring different aspects of heterogeneity. The outcomes of the Baker and Jackson model for intermediate values of the parameter that controls the 'additiveness' of the MA are,

unfortunately, rather difficult to interpret. We plan to explore an alternative formulation of their model, whose parameters are more closely analogous to ϕ and τ^2, in a forthcoming paper.

Cochran's Q statistic (Cochran, 1954) is often used to test for heterogeneity in a MA, although it is known that its power is poor where there is limited information (Hardy and Thompson, 1998). In the circumstances where the outcome of the MA must be a single true effect (for example the physics experiments referred to previously), then the multiplicative model would appear to be the more appropriate model to use. Since the multiplicative model takes account of heterogeneity, the CI returned will be wider if it is present. In the absence of heterogeneity, the model reduces to fixed effect model. As such, the multiplicative model may be considered to be a more general model, although the value of ϕ will clearly be of interest in interpreting the results of the MA. In the medical area the assumption of a single true effect is much less likely to be tenable, so the random-effects model would usually appear more appropriate.

In this paper we showed that the statistical significance of both models tended to agree. In general, if visual inspection of the MA using a funnel plot suggests that the chosen model is inappropriate we would suggest running the other model as a sensitivity analysis. If this leads to a different conclusion we would suggest reporting this and applying appropriate caution when drawing conclusions from the analysis.

Acknowledgements

Thanks to Rebecca Turner of the MRC biostatistics unit, who provided the data used in this paper, and to Dan Jackson and John Thompson for useful discussions.

DM was supported by a National Institute of Health Research (NIHR) Research Methods Fellowship, and is currently supported by an MRC industry collaboration agreement, part funded by Pfizer. KRA is partially supported as a National Institute for Health Research (NIHR) Senior Investigator (NF-SI-0512-10159), and has acted as a paid consultant on methodological and strategic aspects of HTA for Janssen, GSK, Merck, Novartis and Roche, as well as for a number of HTA consultancy companies (Amaris, Creativ-ceutical, OptumInsight and RTI). The remaining authors have no interests to declare. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

References

Baker RD and Jackson D. Meta-analysis inside and outside particle physics: two traditions that should converge? *Research Synthesis Methods*, 4(2):109–124, June 2013. ISSN 17592879.

Beringer J et al. Review of Particle Physics. *Physical Review D*, 86(1):010001, July 2012. ISSN 1550-7998.

Borenstein M, Hedges LV, Higgins JPT, and Rothstein H. *Introduction to Meta Analysis*. Wiley-Blackwell, 2009. ISBN 0470057246.

Brockwell SE and Gordon IR. A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, 20(6):825–40, Mar. 2001. ISSN 0277-6715.

Calcagno V and de Mazancourt C. glmulti: An R package for easy automated model selection with (generalized) linear models. *Journal of Statistical Software*, 34, 2010. R package version 1.0.7 `http://CRAN.R-project.org/package=glmulti`.

Cleveland WS. LOWESS: A Program for Smoothing Scatterplots by Robust Locally Weighted Regression. *The American Statistician*, 35(1):54, feb 1981. ISSN 00031305.

Cochran WG. The Combination of Estimates from Different Experiments. *Biometrics*, 10(1):101, Mar. 1954. ISSN 0006341X.

Cooper H, Hedges LV, and Valentine JC, editors. *The Handbook of Research Synthesis and Meta-Analysis*, chapter 15. Russell Sage Foundation, second edition, 2009.

Davey J, Turner RM, Clarke MJ, and Higgins JPT. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. *BMC Medical Research Methodology*, 11(1):160, Jan. 2011. ISSN 1471-2288.

DerSimonian R and Kacker R. Random-effects model for meta-analysis of clinical trials: an update. *Contemporary Clinical Trials*, 28(2):105–14, Feb. 2007. ISSN 1551-7144.

DerSimonian R and Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, Sept. 1986. ISSN 01972456.

Hardy RJ and Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine*, 15(6):619–29, Mar. 1996. ISSN 0277-6715.

Hardy RJ and Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*, 17(8):841–56, Apr. 1998. ISSN 0277-6715.

Henmi M and Copas JB. Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in Medicine*, 29(29):2969–83, Dec. 2010. ISSN 1097-0258.

Higgins JPT. Being sceptical about meta-analyses: a Bayesian perspective on magnesium trials in myocardial infarction. *International Journal of Epidemiology*, 31(1):96–104, Feb. 2002. ISSN 14643685.

Higgins JPT and Green S, editors. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*. Wiley, 2011. `www.cochrane-handbook.org`, Accessed 19 June 2014.

Higgins JPT and Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539–58, June 2002. ISSN 0277-6715.

Hlavac M. *stargazer: LaTeX/HTML code and ASCII text for well-formatted regression and summary statistics tables*. Harvard University, Cambridge, USA, 2014. R package version 5.1.

ISIS-4. ISIS-4: A randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58,050 patients with suspected acute myocardial infarction. *The Lancet*, 345(8951):669–685, Mar. 1995. ISSN 01406736.

Kulinskaya E and Olkin I. An overdispersion model in meta-analysis. *Statistical Modelling*, 14(1):49–76, Feb. 2014. ISSN 1471-082X.

Li J, Zhang Q, Zhang M, and Egger M. Intravenous magnesium for acute myocardial infarction. *The Cochrane Library*, (2):CD002755, Jan. 2007. ISSN 1469-493X.

Light R and Pillemer D. *Summing up. The science of reviewing research*. Harvard University Press, Cambridge, MA, 1984.

Lunn DJ, Thomas A, Best N, and Spiegelhalter D. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337, Oct. 2000. ISSN 1573-1375.

Moreno SG, Sutton AJ, Ades AE, Stanley TD, Abrams KR, Peters JL, and Cooper NJ. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology*, 9(1):2, Jan. 2009. ISSN 1471-2288.

Nüesch E, Trelle S, Reichenbach S, Rutjes AWS, Tschannen B, Altman DG, Egger M, and Jüni P. Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. *BMJ (Clinical research ed.)*, 341(Jul 16):c3515, Jan. 2010. ISSN 1756-1833.

Paule R and Mandel J. Consensus Values and Weighting Factors. *Journal of Research of the National Bureau of Standards*, 87(5):377, Sept. 1982. ISSN 0160-1741.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

Rosenfeld AH. The Particle Data Group: Growth and Operations-Eighteen Years of Particle Physics. *Annual Review of Nuclear Science*, 25:555–598, 1975.

Smith TC, Spiegelhalter DJ, and Thomas A. Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine*, 14(24): 2685–2699, Dec. 1995. ISSN 02776715.

Stanley TD and Doucouliagos H. Neither fixed nor random: Weighted least squares Meta-Analysis. *Statistics in Medicine*, 34:2116–2127, 2015. ISSN 02776715.

Sterne JAC, Gavaghan D, and Egger M. Publication and related bias in meta-analysis. *Journal of Clinical Epidemiology*, 53(11):1119–1129, Nov. 2000. ISSN 08954356.

Sterne JAC, Egger M, and Smith GD. Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis. *BMJ*, 323(7304):101–105, July 2001. ISSN 0959-8138.

Thompson SG and Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine*, 18(20):2693–708, Oct. 1999. ISSN 0277-6715.

Turner RM, Davey J, Clarke MJ, Thompson SG, and Higgins JPT. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *International journal of epidemiology*, 41(3):818–27, June 2012. ISSN 1464-3685.

Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, Kuss O, Higgins JP, Langan D, and Salanti G. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research synthesis methods*, Sep 2015. ISSN 1759-2887.

Viechtbauer W. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48, 2010.

Whitehead A. *Meta-Analysis of Controlled Clinical Trials*. Wiley, 2002. ISBN 0471983705.

Tables

	Min	Q1	Median	Q3	Max
Trials per meta-analysis	10	10	12	16	158
Participants in treatment groups	38	567.5	1018	2259.75	243928
Events in treatment groups	5	89	210	477	29866
Participants in control groups	74	523.5	964	2138.25	211559
Events in control groups	0	92	218	449.5	8216

Table 1: Summary information for the 448 meta analyses used. Participant and event totals are taken across all studies included in the meta-analyis.

Fixed-effect		SS		NSS	
		Multiplicative			
		SS	NSS	SS	NSS
Additive	SS	235	10	0	1
	NSS	15	23	0	164

Table 2: Numbers of MAs that were statistically significant (SS) or not statistically significant (NSS) at the 5% level under each of the models, split by whether the fixed-effect MA was statistically significant.

	Difference in AIC
	Interactions model
Number of studies	0.019 (−0.103, 0.141)
$\log(D^2)$	2.860 (0.696, 5.030)
$\log(v_i/\sum v)$	−0.304 (−1.580, 0.970)
$\phi_{\text{large}}/\phi_{\text{small}}$	−0.189 (−1.370, 0.992)
Number of studies $\times \log(v_i/\sum v)$	−0.012 (−0.023, −0.0001)
$\log(D^2) \times \log(v_i/\sum v)$	0.377 (−0.024, 0.778)
Number of studies $\times \phi_{\text{large}}/\phi_{\text{small}}$	−0.058 (−0.085, −0.032)
$\log(D^2) \times \phi_{\text{large}}/\phi_{\text{small}}$	−0.197 (−0.374, −0.020)
$\log(v_i/\sum v) \times \phi_{\text{large}}/\phi_{\text{small}}$	−0.183 (−0.407, 0.041)
Constant	−3.960 (−10.900, 2.960)
Observations	303
R^2	0.424
Adjusted R^2	0.406
Residual Std. Error	1.840 (df = 293)
F Statistic	23.900*** (df = 9; 293)

Table 3:[1] The model for predictors of whether the additive or multiplicative model is preferred. Coefficient estimates and 95% confidence intervals are given. *** $p < 0.01$.

[1]Table created using Stargazer (Hlavac, 2014).

	Estimate	Lower bound	Upper bound
$\log(D^2)$ 0% (2)	-0.07	-3.72	3.59
$\log(D^2)$ 50% (3.31)	0.75	-2.88	4.38
$\log(D^2)$ 100% (5.67)	2.23	-1.55	6.02
$\log(v_i/\sum v)$ 0% (-10.63)	-2.26	-6.24	1.71
$\log(v_i/\sum v)$ 50% (-5.24)	0.75	-2.88	4.38
$\log(v_i/\sum v)$ 100% (-3.11)	1.94	-1.73	5.62
Number of studies 0% (10)	0.75	-2.89	4.38
Number of studies 50% (12)	0.75	-2.88	4.38
Number of studies 100% (158)	1.05	-8.54	10.64
$\phi_{\text{large}}/\phi_{\text{small}}$ 0% (0.14)	1.45	-2.19	5.09
$\phi_{\text{large}}/\phi_{\text{small}}$ 50% (1.33)	0.75	-2.88	4.38
$\phi_{\text{large}}/\phi_{\text{small}}$ 100% (13.06)	-6.09	-10.44	-1.74

Table 4: Predicted values of the difference in AIC from the model in Table 3, including 95% prediction intervals. The covariate listed is set to the appropriate percentile (values given in brackets), while the other three variables are set to their median value.