Contemporary Clinical Trials xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Advances in the meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model

Suhail A.R. Doi [a,*], Jan J. Barendregt [b,c], Shahjahan Khan [d], Lukman Thalib [e], Gail M. Williams [c]

[a] Research School of Population Health, Australian National University, Canberra, Australia
[b] Epigear International, Sunrise Beach, Australia
[c] School of Population Health, University of Queensland, Brisbane, Australia
[d] School of Agricultural, Computational and Environmental Sciences, University of Southern Queensland, Toowoomba, Australia
[e] Department of Community Medicine, Kuwait University, Kuwait

ARTICLE INFO

Article history:
Received 5 February 2015
Received in revised form 10 May 2015
Accepted 15 May 2015
Available online xxxx

Keywords:
Fixed effect
Heterogeneity
Meta-analysis
Quasi-likelihood
Random effects

ABSTRACT

This article examines an improved alternative to the random effects (RE) model for meta-analysis of heterogeneous studies. It is shown that the known issues of underestimation of the statistical error and spuriously overconfident estimates with the RE model can be resolved by the use of an estimator under the fixed effect model assumption with a quasi-likelihood based variance structure — the IVhet model. Extensive simulations confirm that this estimator retains a correct coverage probability and a lower observed variance than the RE model estimator, regardless of heterogeneity. When the proposed IVhet method is applied to the controversial meta-analysis of intravenous magnesium for the prevention of mortality after myocardial infarction, the pooled OR is 1.01 (95% CI 0.71–1.46) which not only favors the larger studies but also indicates more uncertainty around the point estimate. In comparison, under the RE model the pooled OR is 0.71 (95% CI 0.57–0.89) which, given the simulation results, reflects underestimation of the statistical error. Given the compelling evidence generated, we recommend that the IVhet model replace both the FE and RE models. To facilitate this, it has been implemented into free meta-analysis software called MetaXL which can be downloaded from www.epigear.com.

1. Introduction

In the era of evidence based medicine, meta-analyses of well-designed and executed randomized controlled trials have the potential to provide high levels of evidence to support therapeutic interventions in all areas of clinical practice. Despite the potential of the outcome of such trials to guide decision making, they may sometimes fail to produce credible conclusive results or may disagree if there were multiple independent trials that investigate the same clinical question. In this situation, a meta-analysis of the trial results has the potential to combine conceptually similar and independent studies with the purpose of deriving more reliable statistical conclusions (based on a much larger sample data) than any of the individual studies [1,2]. Today, clinical decision making relies heavily on this methodology as is evident by the work of the Cochrane collaboration and the high volume of publications for meta-analyses outside the collaboration [3].

Meta-analyses, customarily, are performed using either the fixed effect (FE) or the random effects (RE) models [4,5]. The FE model of meta-analysis is underpinned by the assumption that one identical true treatment effect is common to every study included in the meta-analysis, from which they depart under the influence of the random error only [4]. As such, only within-study variation is considered to be present. In practice, this model ensures that the larger studies (with the lowest probability of random error) have the greatest influence on the pooled estimate. The drawback however is that this model also demonstrates increasing overdispersion as heterogeneity increases. Overdispersion here refers to an estimator that has a greater observed variance (true variance often assessed through simulation) than that theoretically expected which is based on the statistical model (used in the confidence interval computation).

In an attempt to tackle the issue of overdispersion, the RE approach was suggested which attempts to create a more fully specified model [6]. This model makes the additional assumption that the true treatment effects in the individual studies are different from each other and these differences follow a normal distribution with a common variance. The assumption of normally distributed random effects is not justified [7] because the underlying effects included in the meta-analysis do not constitute a random sample from the population. This model nevertheless ignores the need for randomization in statistical inference [8] and the variance of these underlying effects is usually approximated by a moment-based estimator [9]. The application of this common variance to the model has the unintended effect of redistributing the study weights in only one direction: from larger to smaller studies [10]. Thus the studies with the lowest probability of random error are penalized

* Corresponding author at: Chronic Disease Epidemiology Division, Research School of Population Health, Australian National University, Mills Road, Acton ACT 2601, Australia.
E-mail address: sardoi@gmx.net (S.A.R. Doi).

http://dx.doi.org/10.1016/j.cct.2015.05.009
1551-7144/

and do not influence the combined estimates as strongly. The inclusion of this estimated common between-studies variance also seems to be the mechanism that attempts to address overdispersion with increasing heterogeneity, yielding wider confidence intervals and lesser statistical significance than would be attained through the conventional fixed effect model. Yet, given the faulty assumptions, it does not work as expected and as heterogeneity increases, the coverage of the confidence interval drops well below the nominal level [7,11]. Though corrections have been suggested [12,13] they have not been easy to implement and this estimator substantially underestimates the statistical error and remains potentially overconfident in its conclusions [13,14].

A careful look at the inputs to these models demonstrates that both use inverse variance weighting to decrease observed variance of the estimator. However, the approach taken with the RE estimator disadvantages it because as heterogeneity increases, the inverse variance weights are moved towards equality thus increasing estimator variance. This also leads to a failure to specify the theoretical variance correctly so that it now falls short of the observed variance and nominal coverage is not achieved.

While alternative frequentist RE models that attempt to improve on the conventional theoretical model in various ways have been described in the literature [11], they all continue to be based on the assumption of normally distributed random effects which, as mentioned above, leads to several problems. There is therefore the need for a better method and this paper argues that the random effects model should be replaced by a distributional assumption free model. Such a model has been proposed by us as a variant of the quality effects model that sets quality to equal (called the IVhet model) [15]. This paper reviews the model's theoretical construct and presents an evaluation of its performance using standard performance measures [16].

2. Difference between empirically weighted means and the arithmetic mean

Consider a collection of k independent studies, the jth of which has estimated effect size $\hat{\delta}_j$ which varies from its true effect size, δ_j through random error. Also consider that the true effects, δ_js, also vary from an underlying common effect, θ, through bias. There is the possibility of some diversity of true effects (which remain similar) across studies (in which case θ would simply be the mean of the true (unbiased) effects). A greater diversity that leads to dissimilarity of effects would not be meta-analyzed [17]. This underlying common effect, θ, can be estimated through the effect sizes in the k studies using an empirically weighted mean estimator, say $\hat{\theta}_w$. This estimator differs quantitatively from the non-empirically (or naturally) weighted arithmetic mean estimator,

$$\hat{\theta}_{AM} = \frac{1}{k}\sum_{j=1}^{k}\hat{\delta}_j,$$

by the following expression [18]:

$$\hat{\theta}_w = \hat{\theta}_{AM} + \sum_{j=1}^{k}(w_j - 1/k)\left(\hat{\delta}_j - \hat{\theta}_{AM}\right) = \hat{\theta}_{AM} + k\rho_{w\hat{\delta}}\sigma_w\sigma_{\hat{\delta}}, \quad (1)$$

where $\sigma_{\hat{\delta}}$ is the standard deviation of $\hat{\delta}_j$s, σ_w is the standard deviation of the system of weights and $\rho_{w\hat{\delta}}$ is the correlation between the weights and the estimates. Expression (1) serves to demonstrate how and why all empirically weighted estimators are biased [18,19] as defined by deviation from the arithmetic mean estimator (which is unbiased [19]). If $\sigma_{\hat{\delta}}$ was zero (i.e., all effects across all studies are the same), then all methods will default to the arithmetic mean, a similar situation to equal weights where both $\rho_{w\hat{\delta}}$ and σ_w have their minimum value of zero. Therefore, as pointed out by Shuster [19], $\rho_{w\hat{\delta}}$ would determine the extent of bias in an estimator. $\rho_{w\hat{\delta}}$ will only be greater than zero if both the weights and the study effects are correlated and of course, there will be such a correlation if the weights are derived from the data.

For the inverse variance weighted FE model, the weighted estimator,

$$\hat{\theta}_{FE} = \sum_{j=1}^{k} w_j\hat{\delta}_j,$$

has weights that sum to 1 given by:

$$w_j = \frac{1}{v_j}\Big/\sum_{j=1}^{k}\frac{1}{v_j}, \quad (2)$$

where the sampling error variance of the jth study is v_j. Given the previous discussion, this inverse variance estimator is likely to be biased (unless there are equal variances across studies). However, despite the expected bias, the FE estimator does improve over the arithmetic mean estimator because the weights don't just increase the bias, but (by doing so) they also make the variance of the estimator much smaller [20] and trade off this bias through reduction in the mean squared error (MSE). The point of having such increased precision despite, on average, the estimator being biased is that while the estimator is biased "on average", it will still be "more correct" most of the time if the MSE of the biased estimator is smaller than that (MSE) of the unbiased estimator. Since, as researchers, we usually have only one set of sample data for meta-analysis, the bias in the estimator is overshadowed by the decreased MSE.

This works well under the model assumption that studies suffer only from sampling errors. When heterogeneity due to systematic errors is also present, this model suffers from overdispersion. To resolve this, the RE weighted estimator was put forward by DerSimonian and Laird in 1986 [9] where the pooled effect is computed in the same way:

$$\hat{\theta}_{RE} = \sum_{j=1}^{k} w_j^*\hat{\delta}_j,$$

but where the weights that sum to 1 are now given by:

$$w_j^* = \frac{1}{\sigma_j^2}\Big/\sum_{j=1}^{k}\frac{1}{\sigma_j^2} \quad (3)$$

and $\sigma_j^2 = v_j + \tau^2$ in which v_j is the sampling error variance of the jth study and τ^2 is a moment-based estimate of the between-studies variance proposed by DerSimonian and Laird [9], which is applied to all studies within the meta-analysis. It thus becomes clear from expressions (1) and (3) that as $\tau^2 > 0$ increases, σ_w, the standard deviation of the weights, may change unpredictably, but with larger increases ultimately decreases with this system of weights. Consequently, the weights under the random effects model, given large heterogeneity (large τ^2), decrease estimator bias by making the weights more similar and thus the expected value of this estimator comes closer to that of the unbiased arithmetic mean estimator. The problem is that the latter are not optimal weights for variance reduction and so the observed or true variance of the RE estimator continues to increase and exceeds that of the FE estimator as heterogeneity increases. The result is that the decrease in bias is completely overshadowed by a much greater increase in observed variance and such an approach therefore does not make sense [10,21]. The FE estimator therefore is a better performing estimator and can be expected, with increasing heterogeneity, to have a lower variance (ie observed or true variance) and MSE than the RE estimator. It has nevertheless been shunned with heterogeneous studies because of the problem of overdispersion mentioned previously.

Please cite this article as: S.A.R. Doi, et al., Advances in the meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model, Contemp Clin Trials (2015), http://dx.doi.org/10.1016/j.cct.2015.05.009

Table 1
Summary of the three methods[a] of estimation.

	IVhet	RE	AMhet
Weights that sum to 1	$w_j = \frac{1}{v_j} \Big/ \sum_{j=1}^{k} \frac{1}{v_j}$,	$w_j^* = \frac{1}{\sigma_j^2} \Big/ \sum_{j=1}^{k} \frac{1}{\sigma_j^2}$	$\frac{1}{k}$
Pooled effects	$\hat{\theta}_{IVhet} = \sum_{j=1}^{k} w_j \hat{\delta}_j$	$\hat{\theta}_{RE} = \sum_{j=1}^{k} w_j^* \hat{\delta}_j$	$\hat{\theta}_{AM} = \frac{1}{k} \sum_{j=1}^{k} \hat{\delta}_j$
Variance of pooled effect	$var\left(\hat{\theta}_{IVhet}\right) = \sum_{j=1}^{k} \left[\left(\frac{1}{v_j} \Big/ \sum_{j=1}^{k} \frac{1}{v_j} \right)^2 (v_j + \tau^2) \right]$	$var\left(\hat{\theta}_{RE}\right) = 1 \Big/ \sum_{j=1}^{k} \left(1/\sigma_j^2\right)$	$var\left(\hat{\theta}_{AMhet}\right) = \sum_{j=1}^{k} \left[(1/k)^2 (v_j + \tau^2) \right]$.
Comments	Quasi-likelihood model	More "fully" specified model	Quasi-likelihood model

[a] For abbreviations or expansion of the notation please see the text.

3. Variance of the estimator under different models

It is clear from the previous discussion that overdispersion is a problem with both (RE and FE) estimators, more so with the FE estimator. The variance of any weighted estimator $[var\left(\hat{\theta}_w\right)]$ in general is given by:

$$var\left(\hat{\theta}_w\right) = \sum_{j=1}^{k} \omega_j^2 \mathrm{var}\left(\hat{\delta}_j\right), \tag{4}$$

where ω_js are the weights that sum to 1. When there is heterogeneity, the observed variance (or true variance) of the FE model and arithmetic mean (AM) estimator are larger than that computed through the theoretical model, consequently the coverage probability is reduced. However, with the random effects model, the specification of the additional random effects variance expands the computed variance and thus mitigates the reduction in coverage somewhat, but this is still not optimal because the theoretical and observed estimator variances still diverge. Thus, as the heterogeneity increases, the coverage probability of the confidence interval for the RE estimator falls well below the nominal level [7,11]. This is because the optimal weight for each trial is not $1/k$ even though the RE model under increasing heterogeneity weights them more or less equally.

One way to get model based estimator variance closer to the observed variance is to model overdispersion through a quasi-likelihood approach [22,23]. This implies that the meta-analysis is performed under a fixed effect assumption ($\tau^2 = 0$) and the variance of the estimator inflated to account for the heterogeneity, thus preventing a reduction in coverage. This has the advantage of being based purely on the variance-to-mean relationship rather than on distributional assumptions with variance appropriately inflated using a scale parameter, Ψ_j [6]. The latter can be defined by interpreting the multiplicative factor as an intra-class correlation (ICC) as described by Kulinskaya and Olkin [6] where the $ICC_j = \tau^2/(\tau^2 + v_j)$ and the scale parameter is defined as

$$\psi_j = \frac{\sigma_j^2}{v_j} = \frac{1}{1 - ICC_j}. \tag{5}$$

In expression (4), $\omega_j^2 \mathrm{var}\left(\hat{\delta}_j\right)$ is then inflated to $\omega_j^2 \mathrm{var}\left(\hat{\delta}_j\right)\psi_j$ based on expression (5) and this inflation of the random error variance using a quasi-likelihood approach is what we term the inverse variance heterogeneity (IVhet) model of meta-analysis. We point out that if we use these rescaled variances to compute the weights, they would be identical to that in expression (2) for FE weights and thus the weighted FE estimator and weighted IVhet estimator are identical but the model derived variances are different. Thus, incorporating the scale parameter, the variance of the estimator under the IVhet model is given by:

$$var\left(\hat{\theta}_{IVhet}\right) = \sum_{j=1}^{k} \left[\left(\frac{1}{v_j} \Big/ \sum_{j=1}^{k} \frac{1}{v_j} \right)^2 (v_j + \tau^2) \right]. \tag{6}$$

The arithmetic mean estimator can also have a similar correction, $\left(\hat{\theta}_{AMhet}\right)$, but will be expected to have poor coverage with increasing heterogeneity because again the optimal weights are not $1/k$ and would thus mirror the problem seen with the RE estimator. This would take the form:

$$var\left(\hat{\theta}_{AMhet}\right) = \sum_{j=1}^{k} \left[(1/k)^2 (v_j + \tau^2) \right]. \tag{7}$$

4. Examining estimator performance using simulation

We now proceed to examine the performance of the three estimators under varying degrees of heterogeneity. These estimators are what we now call the inverse variance (fixed effect) heterogeneity (IVhet) estimator, the arithmetic mean heterogeneity (AMhet) estimator and the RE estimator (see Table 1 for the mathematical form of the three estimators and their variances). The log odds ratio is used as the effect size (but the models can work with any of the normally distributed effect sizes) and the simulation is modeled around the magnesium meta-analysis [24] data which was previously reviewed by Al Khalaf et al. [10]. This meta-analysis comprises 19 studies, the majority being small studies of under 200 subjects but also has a mega-trial of 58,050 subjects [25] and a smaller mega-trial of 6213 subjects [26]. The latter two studies demonstrated a null effect and the smaller studies a positive effect of magnesium on preventing mortality following myocardial infarction. A key controversy has been that fixed effect meta-analyses (no-effect of magnesium) disagreed with random effects meta-analyses (a strong effect of magnesium) and debates have ensued over the conclusiveness or not of the meta-analytic approach [10]. This sample size discrepancy across the 19 studies was mimicked using a Delaporte distribution with parameters that result in a median study size of 175 and a distribution that resembled the original meta-analysis that included the occasional mega-trial. A simulation study was set-up fixing the true effect size as the OR between 0.4 and 4.0 and allowing the study sample size (N_j) as well as the proportion of events and non-events in the jth study to vary in a similar pattern as in the original studies. The true OR was subjected to randomly generated variance due to bias and chance, the magnitude of the added variance varying over runs to generate different levels of heterogeneity. Description of the distribution parameters used and the simulation protocol are presented in detail elsewhere [15]. Every run generated k studies and the data from 10,000 iterations of these k studies for each model at each heterogeneity level were generated using MetaXL and Monte Carlo simulation software, Ersatz (www.epigear.com). The performance measures were computed from the simulated data as detailed by Burton et al. [16]. The various measures were also plotted as a function of increasing heterogeneity, the latter being indicated by the median τ^2 in a particular simulation run. Each iteration randomly used one of the 3 combinations of sample size (three methods of selection of N_j). The three methods for selection of N_j were from a Delaporte distribution (with parameters 0.1, 8000, 160), a uniform distribution between 50

Please cite this article as: S.A.R. Doi, et al., Advances in the meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model, Contemp Clin Trials (2015), http://dx.doi.org/10.1016/j.cct.2015.05.009

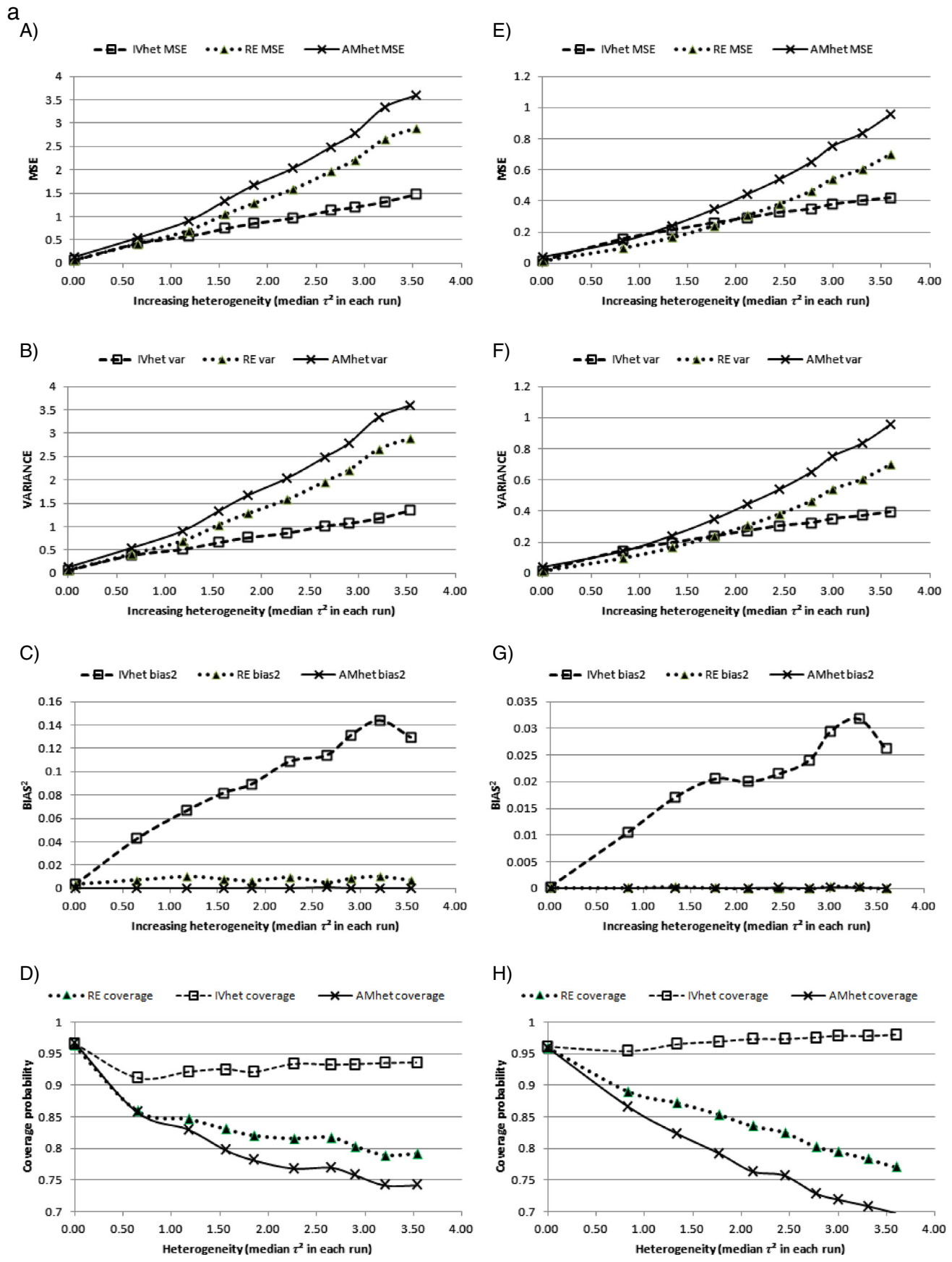

Fig. 1. Simulation of (a) ln OR = ln(0.4) and (b) ln OR = ln(4.0). Left panels restricted to 5 studies while right panels have 19 studies. The panels depict MSE (A & E), variance (B & F), bias squared (C & G) and coverage probability (D & H). The MSE (A & E) is lowest for the IVhet model estimator. The coverage probability (D & H) demonstrates that as heterogeneity increases, the RE model estimator has a somewhat similar coverage to the AMhet model estimator (same weight structure) and drops markedly. The IVhet model estimator clearly has the correct coverage probability (D & H).

Please cite this article as: S.A.R. Doi, et al., Advances in the meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model, Contemp Clin Trials (2015), http://dx.doi.org/10.1016/j.cct.2015.05.009

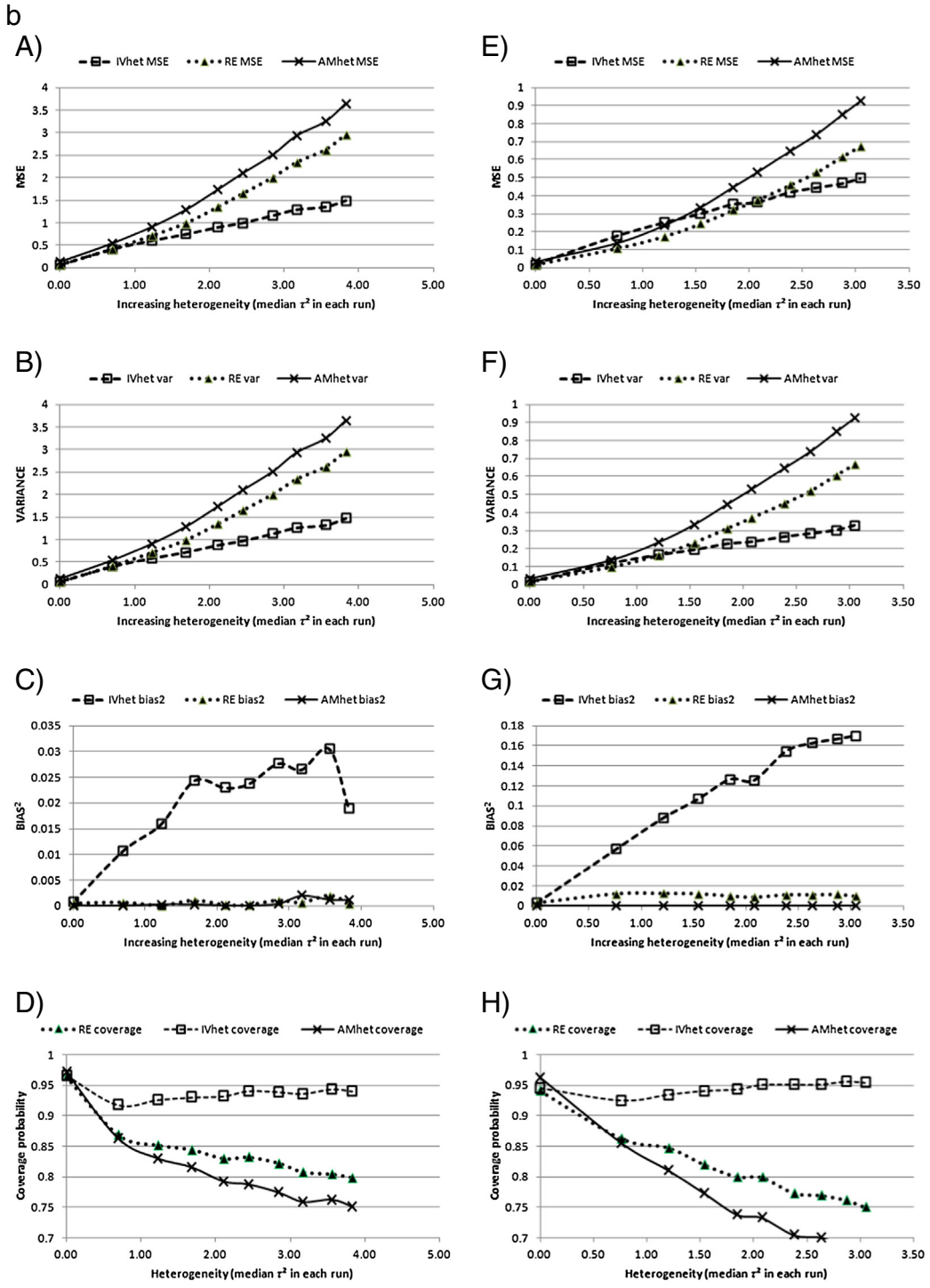

Fig. 1 (*continued*).

and 58,000 in increments of 50 and finally from a uniform distribution between 25 and 200 in increments of 25. This simulation was then run across the different effect size magnitudes (OR 0.4 to 4.0 in increments of 0.4) and for either $k = 5$ or $k = 19$ studies per meta-analysis. A total of 10 separate simulations involving one million separate meta-analyses were therefore performed for each value of k, but only selected results from two effect sizes are reported because they all concurred in terms of IVhet estimator performance.

The first observation from this simulation was to confirm that the RE and AMhet estimators had more or less a similar MSE but the IVhet estimator had a clearly lower MSE (Fig. 1). Additionally, since both empirically weighted models (IVhet and RE) discount studies with larger sampling variability when heterogeneity is low, they did have a similar MSE especially when studies were homogeneous (starting point in Fig. 1). Since the MSE is lower for the IVhet estimator under increasing heterogeneity, the RE model estimator is less efficient (in this respect)

Please cite this article as: S.A.R. Doi, et al., Advances in the meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model, Contemp Clin Trials (2015), http://dx.doi.org/10.1016/j.cct.2015.05.009

than the IVhet estimator. This is because adjustment for the variance due to sampling error deteriorates with increasing heterogeneity and there is therefore no expectation that the RE estimator can produce any additional gains in efficiency. As the number of studies increases, the variance of both estimators declines and thus the gap between the two models' MSE declines since IVhet bias does not change with number of studies. The MSE of the IVhet estimator will exceed that of the RE estimator only when the difference in variance is exceeded by the IVhet estimator bias and this would only be important when the number of studies is large as then estimator variance is already at low levels. Even in such extreme situations, the IVhet coverage is maintained compared to the RE coverage.

A comparison across the three models of the confidence interval width (not shown) and coverage probability (Fig. 1) confirms that with the overdispersion correction (expression (6)) the IVhet estimator keeps coverage at the nominal level because the confidence interval width now keeps pace with the increase in observed variance as heterogeneity increases. The coverage probability of the other two estimators drops well below the nominal level as heterogeneity increases (Fig. 1). Although the results of two simulations are shown in Fig. 1, the results of the remaining 8 simulations were similar when the simulations were run with different magnitudes of effect sizes being simulated (range of OR of 0.8 to 3.6).

Bias was greatest with the IVhet estimator. However, the contribution of bias to the MSE is completely overshadowed by the decrease in variance. Fig. 1B & C indicates magnitudes of variance and bias squared and Fig. 1A then combines this into MSE. It can be seen that there is practically not much change from B to A suggesting that plot C (bias squared) has no practical impact even at these extremes of the ORs.

5. Real data examples from the literature

We also looked at the controversial meta-analysis of intravenous magnesium for prevention of early mortality after myocardial infarction mentioned earlier which consisted of 19 English language randomized trials (published prior to June 2006) [10]. Early mortality was defined as occurring in hospital during the acute admission phase or within 35 days of onset of myocardial infarction. When the meta-analytic estimates were computed using the three methods, they were most extreme with the AMhet estimator (OR 0.44; 95% CI 0.29–0.66), less extreme with the RE estimator (OR 0.71; 95% CI 0.57–0.89) and most conservative with the IVhet estimator (OR 1.01; 95% CI 0.71–1.46). The IVhet estimator has the most bias (towards the null) but the confidence interval, given the performance of this model under simulation, is most likely to reflect the correct coverage probability. Since the confidence interval of the AMhet estimate falls outside this interval, this suggests that the point estimate is too extreme, and occurs simply by chance because of the increased MSE. The RE estimate depicts support for the smaller studies and just falls within the confidence interval based on the IVhet estimate so it may be plausible but the confidence interval around it is probably too narrow (given the simulation performance) and does not extend to cover the OR representing no effect (OR = 1). What the IVhet estimate depicts (Fig. 2) is support for the results of the large studies (pooled estimate) while at the same time support for the smaller but discordant studies by increasing uncertainty around the pooled estimate as evidenced by the expanded (but presumably correct) confidence interval. The RE estimator on the other hand underestimates the statistical error as was expected given its poorer simulation based performance.

Two other examples from the popular meta-analysis literature also reveal a similar problem. An early meta-analysis by Collins et al. [27] demonstrates a significant effect under the RE model but not under the IVhet model (Fig. 3). This meta-analysis was re-analyzed using more conservative approaches by Cornell et al. [11] and the inference was similar to that we demonstrate with the IVhet model. Additionally, a recent meta-analysis by Wang et al. [28] seems to demonstrate that fruit and vegetable intake can protect against all-cause mortality. This again seems to be a consequence of the problems with the random effects model underestimating the statistical error and under the IVhet model there is no significant effect (Fig. 3). This inference is possible since we now know that the IVhet estimator exhibits nominal coverage and has a lower observed variance, thus there is a higher probability that it reflects the true result when compared to the RE estimator result.

6. Discussion

The IVhet model estimate differs from the RE model estimate in three perspectives: Pooled IVhet estimates favor larger trials (as opposed to penalizing larger trials in the RE model), have a more conservative confidence interval with correct coverage probability and exhibit a lesser observed (true) variance irrespective of the degree of heterogeneity. While the RE model represents the conventional method of fitting the overdispersed study data, it is clear from the simulated results that using this more specified probability model with untenable assumptions does not provide better results. The implication based on the IVhet results for the meta-analysis of the magnesium intervention studies in myocardial infarction (Fig. 2) as well as the trials of diuretics in pregnancy (Collins et al.; Fig. 3) is that the evidence for the intervention suggests no benefit, but this remains inconclusive given the relatively wide confidence intervals of 0.71–1.46 and 0.38–1.19 respectively. In terms of the fruit and vegetable intervention for mortality, the IVhet result suggests that there is no evidence at all in support of the latter with a HR of 0.99 (0.93–1.04). This observation, given the comparative RE model results, suggests again that the RE pooled estimate can be less conservative than fixed effect estimates and this has previously been flagged [14].

It should be kept in mind that there are two aspects to conservative results from meta-analysis approaches — conservative in terms of point estimate and conservative in terms of width of the confidence interval. When it comes to the point estimate, Poole and Greenland [14] have highlighted instances where the RE point estimate is indeed less conservative than the FE point estimate (keeping in mind that IVhet and FE point estimates are identical). In addition, the first example (magnesium after myocardial infarction) has point estimates in opposite directions for each model. Thus, depending on the meta-analysis, the IVhet approach does not always produce a conservative point estimate. When it comes to the confidence interval, in many cases the RE confidence interval is too narrow (overdispersion is not adequately addressed) and for all three examples above this has led to spurious significance. These examples therefore highlight that the IVhet results do not exhibit spurious significance because the coverage of the model derived confidence interval remains at the nominal level. All three examples were chosen where spurious significance was present to highlight the latter issue.

The actual or observed variances are the same across the IVhet and RE models only when heterogeneity is low or absent and they diverge as heterogeneity increases. In this situation, the theoretical (model based) variance underestimates the true variance in the RE model and thus the confidence interval has poor coverage. A 95% prediction interval has been suggested as a way to mitigate this for the RE estimate and is defined as the expected effect of a treatment when it is applied within an individual setting and provides its bounds in 95% of the individual study settings. The prediction interval cannot replace the confidence interval and we would not recommend calculating it around the IVhet model estimate because it is based on the specific conceptualization of the underlying model as the RE summary treatment effect and utilizes the between-trial variance [29]. An assumption behind this interval is that trials are considered more or less homogeneous entities and included patient populations and comparator treatments should be considered exchangeable. We therefore agree with Kriston who states that *if this is not the case, then prediction intervals are probably just as useless as random effect estimates* [30].

Please cite this article as: S.A.R. Doi, et al., Advances in the meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model, Contemp Clin Trials (2015), http://dx.doi.org/10.1016/j.cct.2015.05.009

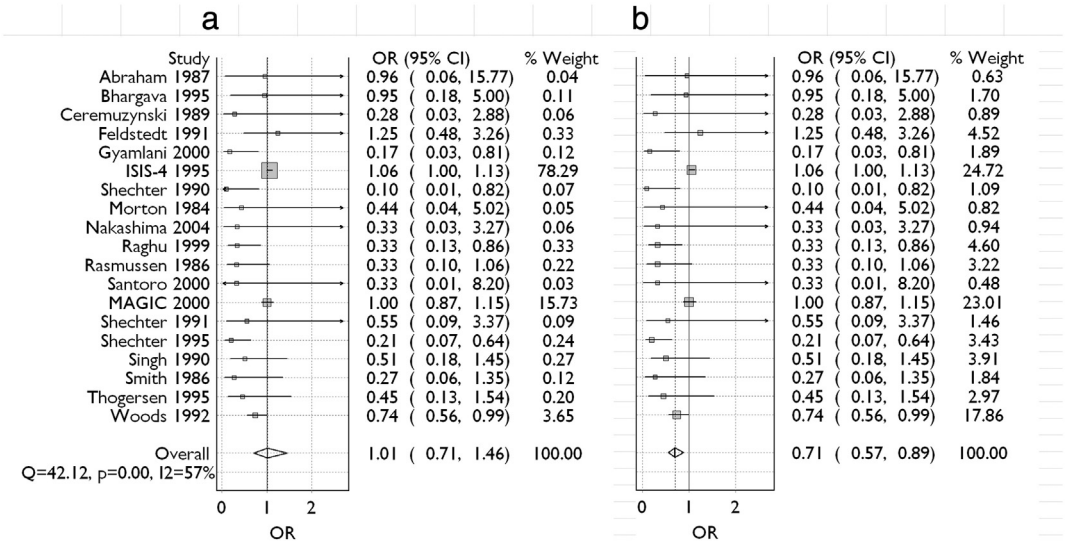

Fig. 2. The meta-analysis results for intravenous magnesium to prevent mortality post-myocardial infarction using the IVhet model (left) and the random effects model (right). The IVhet model (left panel) demonstrates that indeed the statistical error is likely to be greater than what the RE model portrays (right panel). Forest plots created using MetaXL version 2.0 (www.epigear.com).

Senn [31] suggests that the fixed effect approach is an attempt to discover whether the "best use" of the treatment might lead to its being useful. While Senn also suggests that this analysis tests the null hypothesis that the treatment effect is identical in every trial [18], this implicitly assumes exchangeability, i.e., the underlying effects are similar yet non-identical implying ignorance regarding differentiation

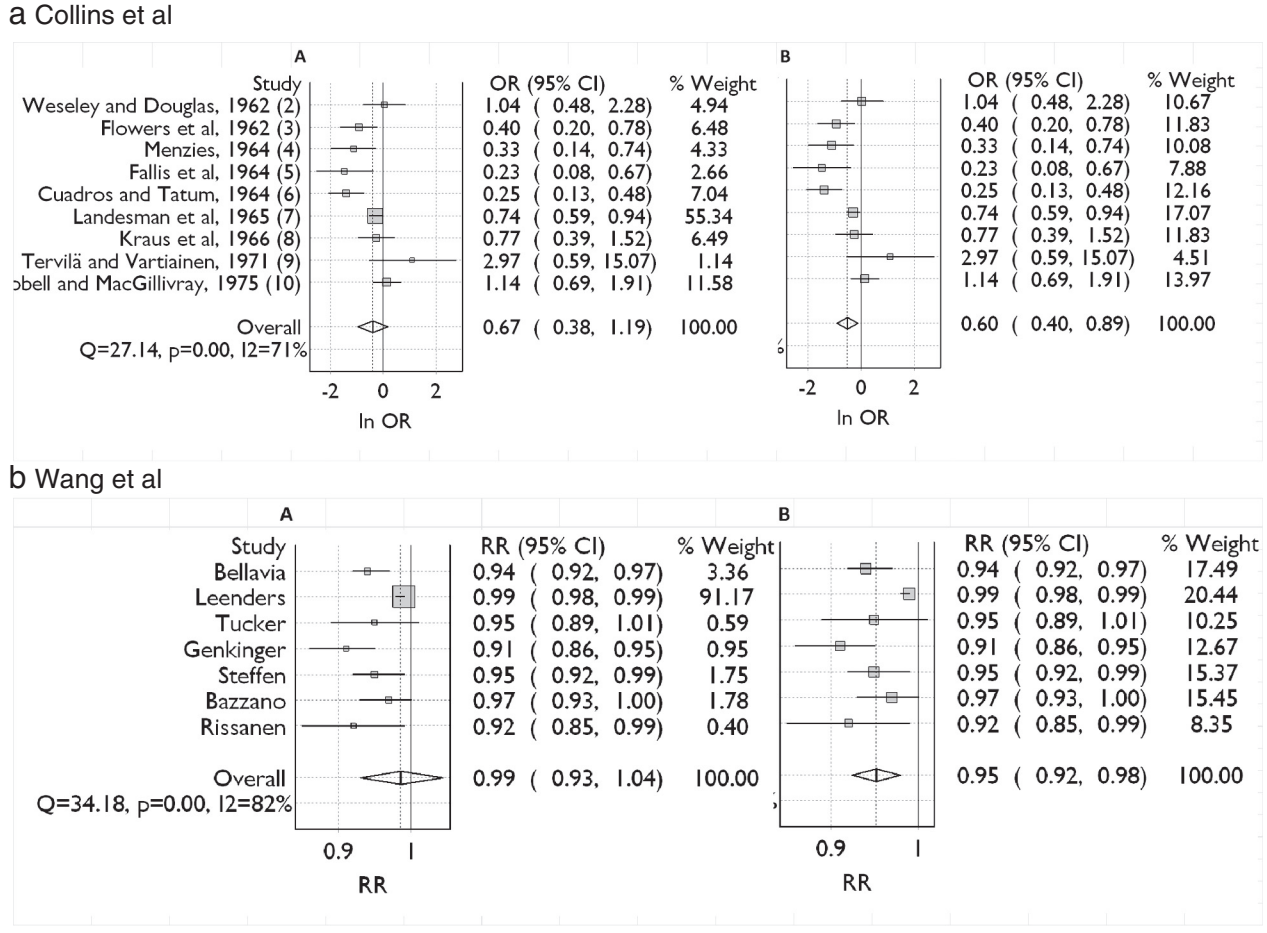

Fig. 3. Examples from two meta-analyses where the RE model probably underestimates the statistical error. The top panel (A) is from a 1985 meta-analysis by Collins and colleagues [27] on the effects of administering diuretics to women with pre-eclampsia and which was discussed by Cornell et al. The second (B) is from a 2014 meta-analysis by Wang et al. [28] on the effects of fruits and vegetables on all cause mortality, the effect size (ES) being the hazard ratio. In both meta-analyses, the IVhet model (left panels) results suggest that the statistical error is greater than what the RE model portrays (right panels). Forest plots created using MetaXL version 2.0 (www.epigear.com).

Please cite this article as: S.A.R. Doi, et al., Advances in the meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model, Contemp Clin Trials (2015), http://dx.doi.org/10.1016/j.cct.2015.05.009

between the magnitudes of the effects. We however believe that when study heterogeneity exists, clinical or methodological covariates are important in the differences between the magnitudes of effects and in this situation, the IVhet analysis should only be viewed as an attempt to portray "best use" of the treatment under the assumption that more precise trials better reflect such "best use". The IVhet pooled estimate could be wrong, (and addressing this requires further input from the studies) but this margin of error is now clearly specified in the confidence interval around the IVhet estimate.

A weighted estimator also requires proper specification of its variance through the theoretical model. Given the demonstrated correct coverage probability of its confidence interval, the IVhet estimator properly specifies this variance and should mitigate the disagreements between meta-analyses and mega-trials [32]. While equal weighting (the arithmetic mean) may seem to avoid the situation where a very few studies play a dominant role when small and large studies disagree, this disagreement is reflected by increasing observed variance and overdispersion that grossly overshadows any benefit from equal weighting. It is true that the dominance of large studies on the IVhet pooled estimate may increase bias if indeed the big studies do not indicate "best use" of the intervention, but this is offset by the variance gains demonstrable with inverse variance weighting and the correct coverage probability of the IVhet confidence interval. As meta-analysts we only do meta-analysis once and therefore what we need to use for this meta-analysis is the estimator with the lowest MSE as then we have a greater probability of being closer to the target we are estimating.

Senn [21] has shown that estimator variance contribution to the MSE tends to go asymptotically to zero as numbers of subjects accrue in a trial and analogously this also applies to meta-analysis as numbers of studies increase. The number of studies in a meta-analysis however has no impact on estimator bias. The only other way of decreasing estimator variance is to use appropriate weights when studies are limited. With the RE model, because heterogeneity essentially reverses weighted averaging and moves the estimator towards the arithmetic mean, true estimator variance is more because there is a less than optimum effect of the weights. Additionally, since the arithmetic mean is unbiased, bias is less for this reason too but the variance increase is much greater thus disadvantaging this estimator. This raises the scenario (we have not mentioned in the paper to avoid too much confusion) of when we have a large number (100 or more) of studies in a meta-analysis. In this situation, we already have the minimum variance of the estimator and addition of any empirical set of weights will simply disadvantage the estimator by increasing bias and MSE. There are thus a threshold number of studies beyond which even the optimal weights are unhelpful and the arithmetic mean suffices. We plan to investigate this threshold in future studies but we estimate that this threshold will be well beyond the numbers seen in most meta-analyses. Additionally, another implication that these results have is for meta-regression. If the RE weights are faulty, that calls into question the rationale behind random effects meta-regression and indeed these results suggest that we should revert back to fixed effects meta-regression. However, the caveat here is that fixed effect meta-regression is only meaningful when there is heterogeneity of studies. We have not studied this per se, but open it up for future investigation.

Finally, when detailed additional information (over and above the study effect and its standard error) becomes available, several options open up. Bias quantification has been proposed as a theoretical way to improve the estimator performance [33] but this remains impractical in meta-analysis because there is no definite relationship between a quality deficiency and the quantitative magnitude or direction of bias in the study effect [34]. A promising development however is the use of the additional information to model the component of variance likely to be contributed by systematic error in individual studies which has been shown to lead to gains in estimator efficiency and is discussed in the next paper in this series [35].

We conclude that the IVhet model of meta-analysis is an improvement over the RE and/or FE models to handle the heterogeneity and performs better than them. This immediately brings into question implementation by the research community of the new and improved method. To facilitate this, our software, MetaXL (available for free download at www.epigear.com), has been updated to version 2.0 to run the IVhet model as well as all other models for comparison.

Conflict of interest

JJB owns *Epigear International Pty Ltd.* which sells the Ersatz Monte-Carlo simulation software used in this study.

References

[1] A. Finckh, M.R. Tramer, Primer: strengths and weaknesses of meta-analysis, Nat. Clin. Pract. Rheumatol. 4 (3) (2008) 146–152.
[2] S.M. Richards, Meta-analyses and overviews of randomised trials, Blood Rev. 9 (2) (1995) 85–91.
[3] I. Bohlin, Formalizing syntheses of medical knowledge: the rise of meta-analyses and systematic reviews, Perspect. Sci. 20 (3) (2012) 273–309.
[4] F.L. Schmidt, I.S. Oh, T.L. Hayes, Fixed- versus random-effects models in meta-analysis: model properties and an empirical comparison of differences in results, Br. J. Math. Stat. Psychol. 62 (Pt 1) (2009) 97–128.
[5] L.V. Hedges, J.L. Vevea, Fixed- and random-effects models in meta-analysis, Psychol. Methods 3 (4) (1998) 486–504.
[6] E. Kulinskaya, I. Olkin, An overdispersion model in meta-analysis, Stat. Model. 14 (1) (2014) 49–76.
[7] S.E. Brockwell, I.R. Gordon, A comparison of statistical methods for meta-analysis, Stat. Med. 20 (6) (2001) 825–840.
[8] R.C. Overton, A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects, Psychol. Methods 3 (3) (1998) 354–379.
[9] R. DerSimonian, N. Laird, Meta-analysis in clinical trials, Control. Clin. Trials 7 (3) (1986) 177–188.
[10] M.M. Al Khalaf, L. Thalib, S.A. Doi, Combining heterogenous studies using the random-effects model is a mistake and leads to inconclusive meta-analyses, J. Clin. Epidemiol. 64 (2) (2011) 119–123.
[11] J.E. Cornell, C.D. Mulrow, R. Localio, et al., Random-effects meta-analysis of inconsistent effects: a time for change, Ann. Intern. Med. 160 (4) (2014) 267–270.
[12] S.E. Brockwell, I.R. Gordon, A simple method for inference on an overall effect in meta-analysis, Stat. Med. 26 (25) (2007) 4531–4543.
[13] H. Noma, Confidence intervals for a random-effects meta-analysis based on Bartlett-type corrections, Stat. Med. 30 (28) (2011) 3304–3312.
[14] C. Poole, S. Greenland, Random-effects meta-analyses are not always conservative, Am. J. Epidemiol. 150 (5) (1999) 469–475.
[15] S.A. Doi, J.J. Barendregt, S. Khan, L. Thalib, G.M. Williams, Simulation comparison of the quality effects and random effects methods of meta-analysis, Epidemiology (2015), http://dx.doi.org/10.1097/EDE.0000000000000289.
[16] A. Burton, D.G. Altman, P. Royston, R.L. Holder, The design of simulation studies in medical statistics, Stat. Med. 25 (24) (2006) 4279–4292.
[17] J.P. Higgins, S.G. Thompson, D.J. Spiegelhalter, A re-evaluation of random-effects meta-analysis, J. R. Stat. Soc. Ser. A Stat. Soc. 172 (1) (2009) 137–159.
[18] S. Senn, Trying to be precise about vagueness, Stat. Med. 26 (7) (2007) 1417–1430.
[19] J.J. Shuster, Empirical vs natural weighting in random effects meta-analysis, Stat. Med. 29 (12) (2010) 1259–1265.
[20] W.L. Cochran, S.P. Carroll, A sampling investigation of the efficiency of weighting inversely as the estimated variance, Biometrics 9 (4) (1953) 447–459.
[21] S. Senn, Lessons from TGN1412 and TARGET: implications for observational studies and meta-analysis, Pharm. Stat. 7 (4) (2008) 294–301.
[22] A.S. Detsky, C.D. Naylor, K. O'Rourke, A.J. McGeer, K.A. L'Abbe, Incorporating variations in the quality of individual randomized trials into meta-analysis, J. Clin. Epidemiol. 45 (3) (1992) 255–265.
[23] R.W. Wedderburn, Quasi-likelihood functions, generalized linear models and the Gauss–Newton method, Biometrika 61 (1974) 439–447.
[24] J. Li, Q. Zhang, M. Zhang, M. Egger, Intravenous magnesium for acute myocardial infarction, Cochrane Database Syst. Rev. 2 (2007) CD002755.
[25] 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. ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group, Lancet 345 (8951) (1995) 669–685.
[26] Early administration of intravenous magnesium to high-risk patients with acute myocardial infarction in the Magnesium in Coronaries (MAGIC) Trial: a randomised controlled trial, Lancet 360 (9341) (2002) 1189–1196.
[27] R. Collins, S. Yusuf, R. Peto, Overview of randomised trials of diuretics in pregnancy, Br. Med. J. (Clin. Res. Ed.) 290 (6461) (1985) 17–23.
[28] X. Wang, Y. Ouyang, J. Liu, et al., Fruit and vegetable consumption and mortality from all causes, cardiovascular disease, and cancer: systematic review and dose–response meta-analysis of prospective cohort studies, BMJ 349 (2014) g4490.
[29] R.D. Riley, J.P. Higgins, J.J. Deeks, Interpretation of random effects meta-analyses, BMJ 342 (2011) d549.

Please cite this article as: S.A.R. Doi, et al., Advances in the meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model, Contemp Clin Trials (2015), http://dx.doi.org/10.1016/j.cct.2015.05.009

[30] L. Kriston, Dealing with clinical heterogeneity in meta-analysis. Assumptions, methods, interpretation, Int. J. Methods Psychiatr. Res. 22 (1) (2013) 1–15.
[31] S. Senn, The many modes of meta, Drug Inf. J. 34 (2000) 535–549.
[32] J. LeLorier, G. Gregoire, A. Benhaddad, J. Lapierre, F. Derderian, Discrepancies between meta-analyses and subsequent large randomized, controlled trials, N. Engl. J. Med. 337 (8) (1997) 536–542.
[33] S. Thompson, U. Ekelund, S. Jebb, et al., A proposed method of bias adjustment for meta-analyses of published observational studies, Int. J. Epidemiol. 40 (3) (2011) 765–777.
[34] S. Greenland, K. O'Rourke, On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions, Biostatistics 2 (4) (2001) 463–471.
[35] S.A. Doi, J.J. Barendregt, S. Khan, L. Thalib, G.M. Williams, Advances in the meta-analysis of heterogenous clinical trials II: the quality effects model, Contemp. Clin. Trials (2015) (In process).

Please cite this article as: S.A.R. Doi, et al., Advances in the meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model, Contemp Clin Trials (2015), http://dx.doi.org/10.1016/j.cct.2015.05.009