ELSEVIER

Journal of Clinical Epidemiology 64 (2011) 119−123

Journal of Clinical Epidemiology

COMMENTARY

Combining heterogenous studies using the random-effects model is a mistake and leads to inconclusive meta-analyses

Mohamad M. Al khalaf[a], Lukman Thalib[b], Suhail A.R. Doi[c,*]

[a]*Department of Medicine, Sabah Hospital, Kuwait*
[b]*Department of Community Medicine (Biostatistics), Faculty of Medicine, Kuwait University, Kuwait*
[c]*Clinical Epidemiology Unit, School of Population Health, University of Queensland, Australia*

Accepted 20 January 2010

There has been a lot of skepticism over the years regarding the value of a meta-analysis. One major area of disagreement has been the divergent findings obtained by pooled meta-analytical estimates and larger or mega-trials. Historically, this argument began when LeLorier demonstrated that estimates from larger or mega-trials differed significantly from pooled estimates of the meta-analyses. They compared the conclusions of 40 meta-analyses with those of subsequent large trials [1] and concluded that meta-analyses are faulty, because they do not always agree with larger trials. After this, research focused on methodological differences between the larger trial and other smaller trials within meta-analyses, but the results were conflicting, because it was found that presumably sound large trials could either agree [2] or disagree [3] with smaller trials of adequate quality. Even when trials within meta-analyses have been grouped by methodological quality measures, there have been no consistent differences in the pooled effect sizes between groups [4−6], and Greenland suggested that this might be because ''quality'' (or whatever leads to more valid results) is of fairly high dimension and possibly nonadditive and nonlinear, highly application specific, and hard to measure from published information [7]. Researchers then began to attribute this disagreement (between large trials and meta-analyses) to bias because of ''faulty'' smaller trials, which were termed ''small study effects'' [8], However, small studies may not necessarily be at fault here, because the evidence provided by a single (large) trial is probably less reliable than its statistical analysis suggests [9], and indeed, there is evidence that a treatment could possibly be better evaluated by a series of small trials [10]. It, therefore, becomes plausible that this lack of agreement could be the product of the methodological process used to arrive at the pooled point estimate in meta-analyses [11,12], an issue that has consistently been overlooked whenever this sort of discussion came up.

The discussion took a new turn when Nuesch and Juni [13] asked a different but crucial question: ''Which meta-analyses are conclusive?'' using the magnesium meta-analysis as an example [14]. This time, these authors take us through funnel plots, tests of interaction, and trial sequential analysis (TSA) in an attempt to help us determine whether a meta-analysis should be trusted or could be considered conclusive. Although the analyses performed by the authors do provide some understanding of how meta-analytic models behave, we realized that the root cause (the models themselves) of faulty meta-analyses have again been overlooked. We will, therefore, attempt to address some of these issues using the same example given by Nuesch and Juni [13]. We should point out at the outset that it has conventionally been thought that a fixed-effects approach estimates the pooled effect under the assumption that the effect is the same in all individual studies, whereas a random-effects approach is used to describe the distribution of the effects in the individual studies (which are allowed to vary). We have, therefore, been advised in the process of writing this commentary that a fixed-effects approach should not be used in the presence of heterogeneous effects, and (conversely) the mean estimate from a random-effects approach should not be used as a benchmark for the effects in each individual study. In this discussion, however, we take the stand of Senn [15] that a fixed-effects meta-analysis tests the null hypothesis that treatment effects are identical in all trials. When, and if, this is rejected, then the alternative hypothesis that may be asserted is, ''there is at least one trial in which the treatment effects differed.'' Indeed, we agree with Senn, who says that ''the position of simultaneously holding that fixed effects meta-analyses are inappropriate but that 'uninformative' prior distributions for random effect variances must be used in connection with random effects meta-analyses seems untenable'' [15]. We,

* Corresponding author. Clinical Epidemiology Unit, School of Population Health, University of Queensland, Herston, Brisbane, QLD 4006, Australia. Fax: +617-3365-5599.

E-mail address: sardoi@gmx.net (S.A.R. Doi).

0895-4356/$ - see front matter
doi: 10.1016/j.jclinepi.2010.01.009

therefore, proceed to use these models mathematically based on how they actually alter the estimates of treatment effects rather than base their use on what we believe that these models are estimating.

We will begin in the same way as Nuesch and Juni, by dividing the magnesium trials into four stages based on trial date: (1) trials available until 1991, before the Second Leicester Intravenous Magnesium Intervention Trial (LIMIT-2) trial; (2) trials until 1995, before the Fourth International Study of Infarct Survival (ISIS-4) mega-trial became available; (3) all trials until 1995, including ISIS-4; and (4) all trials available to date [13]. They first point out that the funnel plots of the meta-analysis at various stages reveal that there is no significant asymmetry in the first stage (stage A), but there is significant asymmetry in all subsequent stages (B to D). They then suggest that there is bias, because there is funnel plot asymmetry, and they imply that this bias is because of ''inadequately sized'' smaller studies. We disagree. Funnel plot asymmetry simply means discordance between the results of a few more-precise (larger) vs. many less-precise (smaller) studies. Conventionally, this bias has been attributed to smaller (less-precise) studies, although the funnel plot does not tell us which group (larger or smaller) has been affected by bias. We do not think that the possibility of a biased large trial can be completely discarded, and there is ample evidence to argue this point of view [9,10,16–18]. There could also be bias toward the null as a consequence of the recruitment drive in mega-trials, and thus, the contrast between treatment and no treatment or between subgroups could have been blunted either by nonprotocol therapy or by inaccuracy of data [18]. These errors are then randomized between comparison groups [19,20], leading ultimately to a reduction in the level of experimental control. This makes it very difficult for a mega-trial to reveal an effect despite the trial size, unless that effect is considerable. It is, therefore, theoretically possible that funnel plots may be asymmetric because of bias created by a large study.

Second, the authors show that before ISIS-4, studies were insufficiently powered to detect a true effect (using TSA methods and depicted in the second figure of their commentary). Unfortunately, when the sample size is large enough to achieve sufficient power to detect an effect (after inclusion of ISIS-4), the meta-analysis models diverge: a random-effects meta-analysis now favors magnesium, whereas fixed-effects meta-analysis now shows no effect of magnesium. Although they conclude that TSA might help by telling us if a meta-analysis has sufficient power to be trusted, this divergence of results clearly demonstrates that sufficient power is not a satisfactory indicator of conclusive findings from meta-analyses. The authors correctly point out that when the smaller studies are removed, the results of the random-effects meta-analysis concur with the results of the fixed-effects meta-analysis (namely, that magnesium has no effect). Mainly on this basis, they conclude that this discrepancy in the meta-analyses is a result of small trials that should be distrusted. What they do not realize is that the fixed-effects meta-analyses differ simply because, after the inclusion of ISIS-4, less-precise (higher variance) trials have their inverse variance weights markedly decreased. In essence, the pooled effect size is simply the most precise trial's effect size—in this case, the effect size of ISIS-4. This is why a fixed-effects analysis cannot, in this instance, be considered a meta-analysis, and in our opinion, results rather in a *meta-reduction*, where the meta-analysis is reduced to the effect of a single study. Quantitative analysis of the robustness of such a trial suggests that the best evidence of a treatment's clinical performance is probably better reflected by multiple smaller trials that are independent and differ in their protocols [9], precisely the type of trials that the fixed-effect model would minimize.

In this sort of situation, are we implying that we should focus on the random analysis results as being conclusive? Again, if we look critically at computational aspects of the random model, we immediately realize that the random-effect model meta-analysis is simply a process whereby the inverse variance weighting of the fixed model is reversed (to a variable extent), thus, moving the weighted mean effect size back toward an unweighted mean. The extent of this reversal is solely dependent on two factors (Fig. 1):

1. Heterogeneity of precision (study size): The extent of the spread of precision of the studies involved as indicated by the maximum minus minimum inverse variance weights;
2. Heterogeneity of effect size (adjusted τ^2): How many times bigger than the average variance of studies within the meta-analysis is the value of τ^2 (τ^2 alone is not comparable across different meta-analyses).

This model does away with the *meta-reduction* of the fixed analysis by reducing the extreme diminution of the effect size of the smaller studies so that the pooled effect size

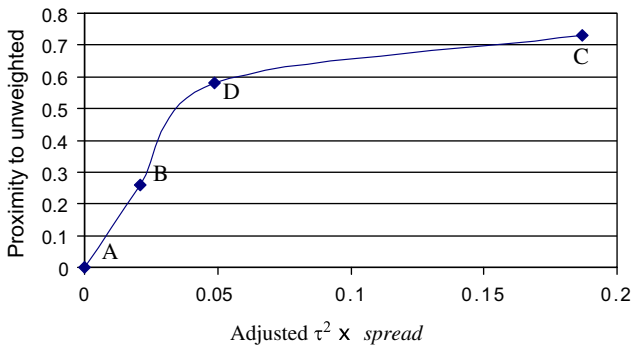

Fig. 1. Relationship of study weights with study size and effect size heterogeneity. The y-axis plots unweightedness that increases from 0 (inverse variance weighting) to 1 (completely unweighted and, thus, equal study weights). This scale was computed as follows: unweightedness $= \frac{(SD_{Fw} - SD_{Rw})}{SD_{Fw}}$, where SD is standard deviation, Fw is fixed-model weights, and Rw is random-model weights. Spread of precision and adjusted τ^2 are defined in text. It is clear that unweightedness increases with increase in spread or τ^2 and is irrespective of the stage of the meta-analysis (letters A to D). The line connecting the points is a free-form line but suggests a logarithmic relationship.

now moves toward an intermediate value between trials with extremes of precision [21]. This movement backward toward an unweighted mean, however, is based on penalizing larger studies based on sample size and effect size (τ^2) heterogeneity. Is the random-model analysis now valid, given our rejection of the fixed-effects analysis? Unfortunately, the answer is *no*, because there is no reason to automatically assume that a larger variability in study sizes or effect sizes (meta-analysis τ^2) automatically indicates a faulty larger study or more reliable smaller studies. Indeed, there is no reason why the conclusiveness of a meta-analysis should be associated with this method of reversal of the inverse variance weighting process of the included studies. As such, the changes in weight introduced by this model to each study have no statistical or probabilistic interpretation and, thus, bear no relationship with what the studies actually have to offer.

Third, the authors argue that the tests of interaction imply that the increase in weight of smaller studies in a random-effects meta-analysis is not because of chance alone. It is, indeed, because of chance, because it is attributed to the timing of the introduction of ISIS-4. Stage C has both the highest spread of inverse variance weights and the biggest τ^2 (relative to the average variance), and after more studies are added, the meta-analysis becomes *less* unweighted in stage D. In this instance, one would expect a test of interaction to become positive simply because the fundamental methodology of a random-effects model is redistribution of study weights based on differences in study size. Although it is the heterogeneity of effect size that generates a constant (τ^2), its ability to redistribute study weights depends on the extent of the variability of study sizes. The random-effects model operates on the premise that, with increasing heterogeneity, larger studies should have weight subtracted, whereas smaller studies should be given additional weight, and indeed, when study sizes are equal, the random-effects model gives the same pooled effect sizes as the fixed-effects model (i.e., when there is no heterogeneity in study sizes). Is there any researcher who would agree that, given heterogeneity of outcome, bigger studies should have lesser say, simply because they are big and have no other reason? The answer is *no*.

We firmly believe that getting rid of this misconception requires an understanding that redistribution of study weights can only be done using tangible information from the studies themselves and not on the concept of smaller studies being good and larger studies being bad or vice versa. Hence, is the solution indeed the removal of smaller studies, as the authors have suggested, or is it looking up the list of studies and addressing their differences? We believe the former is irrelevant in determining which results to trust. The reason why removal of smaller studies allows the random-effects and fixed-effects meta-analyses results to concur is not that small studies cannot be trusted. In reality, by removing earlier smaller studies, the authors are actually removing heterogeneity of effect sizes from the analysis. By doing so, the random model defaults to a fixed model (a random model in the absence of heterogeneity equals a fixed model).

The removal of smaller studies by the authors assumes, without proper reason, that the earlier smaller studies were not credible. The way the authors address study credibility in their commentary is not really comprehensive. Although it turns out in the practical example chosen (the magnesium studies) that the earlier smaller studies were the ones without proper allocation concealment, how about other parameters of quality? We have examined all the studies in detail (Appendix [available on the journal's Web site at www.elsevier.com]), and it turns out that allocation concealment cannot explain all differences. Using a more comprehensive quality score [22] to incorporate all components of quality into this meta-analytic model, we find that size and rigorousness do not correlate (Fig. 2). It seems clear that smaller studies are not less credible, and therefore, the only valid option available to judge the conclusiveness of a meta-analysis is to address the source of heterogeneity, if at all possible.

The question is how do we address this source of heterogeneity? It can be done if we use a statistical model that assumes that, if heterogeneity can be quantified, then the main source would be the differences in quality between the studies [22]. This has been attempted previously [5], but models have been conceptually flawed by attempts to relate quality to the model weights directly [23]. Others that ran quality-based sensitivity analyses reported no benefit with or without quality groupings but did not realize that even though τ^2 is generated from effect size heterogeneity, the random model only changes the pooled effect size when there is additional heterogeneity of precision. In the unusual situation where all trials are of similar precision, the τ^2 only affects the width of the confidence interval (CI), but the overall random-model effect size remains unchanged from the overall fixed-model effect size. As such, even when we stratify studies in the random model by quality groups into high and low, for example, the difference in pooled effect size depends not only on the relationship between effect size and quality but also on how the precision in both groups gets redistributed, a parameter that was completely overlooked, thus, nullifying the conclusion that

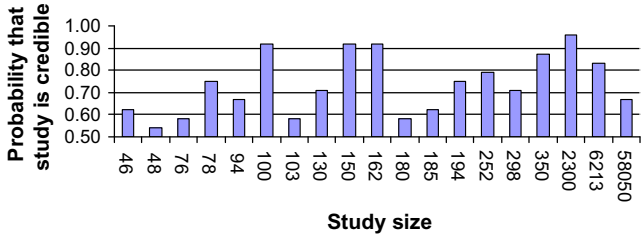

Fig. 2. Bar chart showing quality of each study in the magnesium meta-analysis listed in increasing order of study size. There is clearly no relationship. Quality was graded as the probability (0–1) that each study was credible. See Appendix (available on the journal's Web site at www.elsevier.com).

there is no effect of quality scores applied to meta-analysis models [4,5].

We take the position that a correction for the quality-adjusted weight of the *i*th study must be visualized as a composite based on the quality of all other studies except the study under consideration, and this composite should be used to redistribute weights. This means that, as studies increase in quality, redistribution should become progressively less and should cease when all studies are judged to be of maximum credibility. We proposed such a model in 2008 and called this the quality-effect model [22,24]. To implement this model, we transform Q_i (the probability [0−1] that study i is credible) to a new variable $\widehat{Q}_i$ by means of a quality adjustor ($\hat{\tau}_i$) for the ith study. If N is the number of studies in the analysis and w_i is the inverse variance weight of the ith study, then $\widehat{Q}_i$ is given by [22,24]:

$$\widehat{Q}_i = Q_i + \left(\frac{\hat{\tau}_i}{w_i}\right) \quad \text{where } \hat{\tau}_i = \left(\sum_{i=1}^{N} \tau_i \times N \times \frac{Q_i}{\sum_{i=1}^{N} Q_i}\right) - \tau_i$$

$$\text{and } \tau_i = \frac{w_i - (w_i \times Q_i)}{N-1}$$

The final summary estimate for the meta-analysis is then given by:

$$\overline{\text{ES}}_{(\text{QE})} = \frac{\sum(\widehat{Q}_i \times w_i \times \text{ES}_i)}{\sum(\widehat{Q}_i \times w_i)}$$

whereas the variance of this weighted average is:

$$v_{\overline{\text{ES}}(\text{QE})} = \frac{\sum\left(\widehat{Q}_i^2 \times w_i\right)}{\left(\sum(\widehat{Q}_i \times w_i)\right)^2}$$

This adjustor redistributes the sum of weight removed from each study to each of the remaining studies proportionate to their quality. In this case, the total sum of the redistributed weight is the same as that with inverse variance weighting, but the individual studies receive a slightly different amount based on their quality. Although, there is no gold standard and we still do not know the best way to measure quality, this is not an obstacle to this quality model, because it works with any quality score [22].

If we apply this concept to the magnesium meta-analysis, then the ISIS-4 trial's weight should change proportionate to its assumed decrease in credibility for reasons outlined in the quality assessment given in the Appendix (available on the journal's Web site at www.elsevier.com). Furthermore, the lower-precision trials only get upgraded if this high-precision trial is deemed to have flaws in quality. As such, if the ISIS-4 trial were to be judged to be of perfect quality, then there will be no trade-off, and the quality-based model defaults to a fixed-effects model. Based on our (subjective) assessment of ISIS-4, we deemed it *not* to be of perfect quality. When we then ran our model on the magnesium data, the quality-effect meta-analysis had an overall effect size (odds ratio [OR]) of 0.78 (95% CI: 0.68−0.90) in favor of the use of magnesium (Appendix [available on the journal's Web site at www.elsevier.com]). This is in contrast to the results of both the fixed-effects and random-effects analysis. The fixed-effects model showed no effect of magnesium, whereas the random-effects model showed a much more favorable effect. The study that had the maximum weight in all three analyses was the ISIS-4 study [25] because of its large sample size (58,050). It accounted for 78% of the total weight in the fixed-effects model but only 25% in the random-effects model, the latter being a result of the large discordances in results among this group of studies induced by the mega-trials. The quality-effect weight for the ISIS-4 study was, however, in between these two extremes at 52%. The random-effects approach weighted the ISIS-4 trial much less than the other models did while weighting the rest of the studies higher. The gain in weight, however, was not consistent across studies, and although some studies gained up to 14%, others gained less than 1%. The two multicenter trials (ISIS-4 [25] and Magnesium in Coronaries Trial [26]) were not in favor of the use of magnesium, with ORs of 1.059 and 1.002, respectively. All other studies had an OR that favored the use of magnesium, except that by Feldstedt et al. [27]. Details of these studies can be found in the Appendix (available on the journal's Web site at www.elsevier.com).

The incorporation of quality scores in summary estimates from meta-analyses has been criticized by Greenland and O'Rourke on the grounds that it fails to account for the magnitude or direction of bias induced by a quality deficiency and the fact that we are mixing "objective" measures with "arbitrary judgements" [7]. Although the magnitude/direction problem has been overcome by our model [22], we are certainly making arbitrary (albeit guided) judgments about quality. But then, is there not a different kind of arbitrariness with the random model when we inflate the variance of more precise studies in the model with a constant that bears no relationship with heterogeneity arising from differences in the individual studies themselves? This would imply that the random-model point estimates are not going to be invariably closer to the null value nor are their P values going to be invariably larger than those of fixed-effects summaries. Indeed, Poole and Greenland provide evidence that this is true, and they even give an example in which the random-effects summaries are less conservative in both of these alternative senses and possibly more biased than the fixed-effects summaries [28]. This again is simply because the random model redistributes study weights based on sample size or effect size variability alone, disregarding all other differences that constitute something tangible, such as randomization, blinding, losses to follow-up, and others. In essence, therefore, even though a random model may seem more precise and less arbitrary than a quality-adjusted model, unfortunately, it turns out to be more precise about nothing tangible.

We conclude that when heterogeneity is present, a fixed-effects analysis results in a *meta-reduction* and, therefore, cannot be used, because it throws away all other studies in favor of ISIS-4. Unless, for some reason, all other studies cannot be trusted (or the big trial is for some reason a gold standard), this meta-analytic result cannot be considered conclusive. On the other hand, we are also forced to refute the very concept of a random-effects meta-analysis. It is basically a computational method of moving an inverse variance−weighted meta-analysis back toward an unweighted mean estimate based on the spread of study sizes and the value of τ^2 relative to the average variance in the group of studies making up the meta-analysis. Because neither of these two numbers have any relationship with the content of the studies in a meta-analysis, we have a precise yet uninterpretable redistribution of weights in the magnesium meta-analysis. This certainly cannot be conclusive in any meaning of the word. So what then is the effect of magnesium? We have two options. Individual centers follow the results of their studies and ignore other centers or we implement a more tangible model as an alternative meaningful way of combining heterogeneous studies that would lead to a conclusive result [12].

Supplementary information

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jclinepi.2010.01.009

References

[1] LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. N Engl J Med 1997;337:536−42.

[2] Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med 2001;135:982−9.

[3] Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned. J Clin Epidemiol 2006;59:1249−56.

[4] Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 2005;5:19.

[5] Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002;287:2973−82.

[6] Verhagen AP, de Vet HC, Vermeer F, Widdershoven JW, de Bie RA, Kessels AG, et al. The influence of methodologic quality on the conclusion of a landmark meta-analysis on thrombolytic therapy. Int J Technol Assess Health Care 2002;18:11−23.

[7] Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics 2001;2:463−71.

[8] Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol 2000;53:1119−29.

[9] Borm GF, Lemmers O, Fransen J, Donders R. The evidence provided by a single trial is less reliable than its statistical analysis suggests. J Clin Epidemiol 2009;62:711−5.

[10] Borm GF, Donders R. A treatment should be evaluated by small trials. J Clin Epidemiol 2009;62:887−9.

[11] Doi SA. The influence of methodologic quality on the conclusion of a landmark meta-analysis on thrombolytic therapy. Int J Technol Assess Health Care 2009;25:107−9.

[12] Doi SA. Modelling methodologic quality into meta-analyses and pitfalls of not doing this. Ann Thorac Surg 2009;87:985−6. author reply 986.

[13] Nuesch E, Juni P. Commentary: which meta-analyses are conclusive? Int J Epidemiol 2009;38:298−303.

[14] Li J, Zhang Q, Zhang M, Egger M. Intravenous magnesium for acute myocardial infarction. Cochrane Database Syst Rev 2007;(2). CD002755.

[15] Senn S. Trying to be precise about vagueness. Stat Med 2007;26:1417−30.

[16] Lin Z. An issue of statistical analysis in controlled multi-centre studies: how shall we weight the centres? Stat Med 1999;18:365−73.

[17] Senn S. Some controversies in planning and analysing multi-centre trials. Stat Med 1998;17:1753−65. discussion 1799−8000.

[18] Woods KL. Mega-trials and management of acute myocardial infarction. Lancet 1995;346:611−4.

[19] Charlton BG. Fundamental deficiencies in the megatrial methodology. Curr Control Trials Cardiovasc Med 2001;2:2−7.

[20] Charlton BG. Megatrials are based on a methodological mistake. Br J Gen Pract 1996;46:429−31.

[21] Helfenstein U. Data and models determine treatment proposals—an illustration from meta-analysis. Postgrad Med J 2002;78:131−4.

[22] Doi SA, Thalib L. A quality-effects model for meta-analysis. Epidemiology 2008;19:94−100.

[23] Berard A, Bravo G. Combining studies using effect sizes and quality scores: application to bone loss in postmenopausal women. J Clin Epidemiol 1998;51:801−7.

[24] Doi SA, Thalib L. An alternative quality adjustor for the quality effects model for meta-analysis. Epidemiology 2009;20:314.

[25] ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group. 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. Lancet 1995;345:669−85.

[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 2002;360:1189−96. Notes: Corporate name: Magnesium in Coronaries (MAGIC) Trial Investigators.

[27] Feldstedt M, Boesgaard S, Bouchelouche P, Svenningsen A, Brooks L, Lech Y, et al. Magnesium substitution in acute ischaemic heart syndromes. Eur Heart J 1991;12:1215−8.

[28] Poole C, Greenland S. Random-effects meta-analyses are not always conservative. Am J Epidemiol 1999;150:469−75.