Intern Emerg Med (2009) 4:423–427
DOI 10.1007/s11739-009-0296-6

CE - COCHRANE'S CORNER

Heterogeneity and meta-analyses: do study results truly differ?

Gianni Virgili · Andrea A. Conti · Lorenzo Moja · Gian Franco Gensini · Roberto Gusinu

Published online: 28 August 2009

The methodologist's point of view

Gianni Virgili, Andrea A. Conti, Lorenzo Moja

As the Cochrane Corner hosts analytical comments on Cochrane systematic reviews (SRs), it is important that readers are comfortable in understanding the methodology related to SR science. In the first article in this series, we presented an approach to understand all basic information reported in a meta-analysis graph [1]. In this issue, we cover concepts and tips related to heterogeneity, and whether to combine the results of the studies is appropriate. The decision to meta-analyse or not to meta-analyse studies may appear imponderable to many clinicians. How do authors decide whether patients, interventions and outcomes considered in individual studies are sufficiently similar to be pooled in meta-analyses?

G. Virgili (✉)
Department of Oto-Neuro-Ophthalmological Surgical Sciences, Eye Clinic, University of Florence, V.le Morgagni 85, 50134 Florence, Italy
e-mail: gianni.virgili@unifi.it

A. A. Conti
Department of Critical Care Medicine and Surgery, University of Florence, Florence, Italy

A. A. Conti
Don Carlo Gnocchi Foundation, IRCCS Florence, Florence, Italy

L. Moja
Italian Cochrane Centre, Mario Negri Institute for Pharmacological Research, Milan, Italy

G. F. Gensini
Department of Critical Care Medicine and Surgery, University of Florence and Azienda Ospedaliero-Universitaria Careggi, Florence, Italy

R. Gusinu
DAI Cardiologico e dei Vasi Azienda Ospedaliero, Universitaria Careggi, Florence, Italy

Sources of diversity of results across studies in an SR: clinical, methodological and statistical heterogeneity

Systematic reviews synthesise the results of several comparative studies that investigate the same research questions. Such studies often yield diverse estimates of treatment effect, and reviewers need to evaluate whether the variation in the true effects underlying the studies is within the boundaries of chance. As an example, Moncrieff et al. [2] find in randomised controlled trials (RCTs) that antidepressants compared with active placebo are mainly beneficial, although these trials present highly variable treatment effects (Fig. 1a). Is the overall estimate diamond at the bottom of the meta-analysis graph a good descriptor of all study results? There may be reason to be cautiously sceptical about the capacity of the meta-analysis to compact such heterogeneity in a precise estimate. The sceptic should ask first where all this heterogeneity comes from.

Reasons of heterogeneity are usually classified into clinical (participants, interventions and outcomes) and methodological (design and conduct) diversities. For example, interventions can differ because of drug dosage or treatment duration, or, if they are about quality improvement, because they include a number of components that may be only partly similar. Conduct can differ because some researchers have kept study participants and those involved with their management unaware of the assigned treatment (sometimes called blinding or masking), and others have not. Blinding is, particularly, important when the response criteria are subjective, such as an

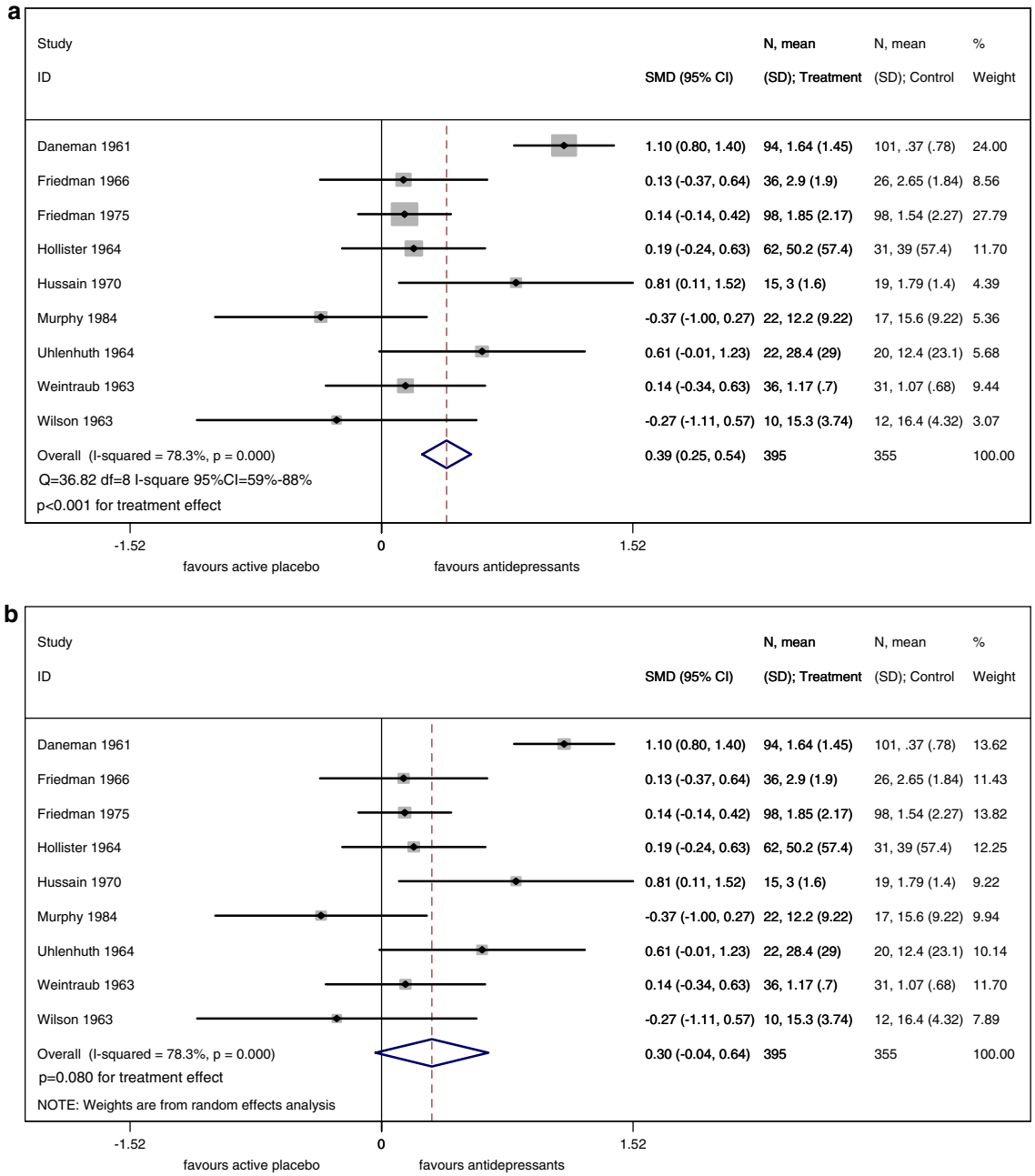

Fig. 1 **a** Fixed-effect meta-analysis of studies comparing antidepressants with active placebo for depression as conducted by Moncrief et al. [2]. **b** The correspondent random-effects meta-analysis. The weights of extreme observations such as Daneman (1991) are reduced using random effects and the resulting 95% CI of the pooled estimates are much wider because of unexplained high heterogeneity ($I^2 = 78\%$, $P < 0.0001$) of treatment effect across studies is accounted

improvement in depression [3]. Moncrieff et al. hypothesise that heterogeneity could arise from the use of different scales to measure the primary outcome: change of mood, inclusion of outpatients versus inpatients, or a variable length of follow-up [2].

The relevance of heterogeneity will vary according to the circumstances. Some reviews are purposely broad, and the authors may need to decide on what is similar and what is not, facing the trade-off between the opportunity to pool 'oversize' accumulated evidence and the risk of pooling apples and oranges, making the meta-analysis less interpretable. Other meta-analyses include studies that deal with similar clinical and methodological aspects, and heterogeneity could be unexpected.

Once scepticism regarding clinical and methodological diversities has been considered, there is a third source to

explore: statistical heterogeneity that can be interpreted as clinical and methodological heterogeneity of unknown source. Statistical heterogeneity manifests itself in the observed intervention effects being more different from each other than one would be expect due to chance (random error) alone. A user-friendly way of describing statistical heterogeneity is the following: does a meta-analysis of similar RCTs suggest a unique true estimate of treatment effect underlying the studies? Or rather, do individual studies' estimates suggest that treatment worked differently in different studies, and we can expect a range of several efficacies?

When there is large heterogeneity of treatment effects across studies, we infer that there is no unique underlying truth to measure, but rather that we are sampling from a range of treatment effects even if a single drug has been used as an intervention. When researchers are unable to explain why this variability occurred, we refer to unknown factors that have modified treatment efficacy across studies.

Methods for measuring statistical heterogeneity

The Cochran's Q and the I^2 statistics are related measures of statistical heterogeneity of treatment effect across studies [4, 5]. All Cochrane meta-analyses report these statistics at the bottom of the meta-analysis graphs.

The Q statistic leads to a P value. The power of this test is often modest, given the small number of studies in typical meta-analyses, and a P value of 0.10 is used as a threshold for significance [4]. The Cochrane Handbook for SRs of interventions recommends the use of the I^2 statistic to measure statistical heterogeneity [5]. The I^2 statistic leads to a percentage value. In effect, the I^2 describes the percentage of the variability in effect estimates that is due to heterogeneity rather than chance (sampling error) [4, 5] and is in fact an estimate of the between-study variance to the total variance (between plus within study). Let the negative values of I^2 be equal to zero so that I^2 lies between 0 and 100%. A value of 0% indicates no observed heterogeneity, and larger values show increasing heterogeneity [4].

At this point, clinicians have to ask themselves whether a meta-analysis is reasonable. The Cochrane Handbook gives the following guidance on this decision based on I^2 values to classify the inconsistency of the effect measures (often relative risks, odds ratios or mean differences) across studies [5]:

- 0–40%: might not be important
- 30–60%: may represent moderate heterogeneity
- 50–90%: may represent substantial heterogeneity
- 75–100%: considerable heterogeneity.

The importance of the observed I^2 values also depends on:

- The magnitude and direction of the effects: if I^2 is >50%, but all studies are in the direction of benefit and a random-effect meta-analysis (see following paragraph) yields highly statistically significant benefit, then we are uncertain about the amount of benefit, but not about its existence. Then, it is safe to conclude that the treatment is beneficial.
- The strength of evidence from heterogeneity, e.g. the P value from the χ^2 test for heterogeneity or a confidence interval for the I^2: as an example, the I^2 statistics of 78% (95% confidence interval from 59 to 88%, $P = 0.000$) obtained from the meta-analysis by Moncrieff et al. [2], provides large evidence of heterogeneity. In this case, three possible solutions are: (1) to avoid a meta-analysis, (2) to explore heterogeneity (i.e. subgroup analysis), or (3) to carry out a random-effect meta-analysis.

Random effects estimate of treatment effect in a meta-analysis

If no clinical, methodological and statistical heterogeneity are found, the studies can be assumed to measure an underlying unique treatment effect. Therefore, we can simply pool individual studies' mean differences (or other effect measures) as a weighted average of the mean difference of each study. Commonly, weights are the inverse of the mean difference variance; the more the study is precise compared with the others in the meta-analysis, the more will be the weight given. This is also called a fixed-effect meta-analysis (Fig. 1a).

When heterogeneity is recognised, we should be able to estimate the variability of the presumed true value of treatment efficacy across studies (between-study variability), taking into account the uncertainty due to the fact that each study is imprecise (within-study variability). It seems straightforward that we place more uncertainty on our mean difference estimate when heterogeneity is found, and this should be done proportionally to how much of statistical heterogeneity exists. A statistic called τ (tau) is the technical estimate of this extra variability of unknown source, and is used to inflate confidence intervals of mean differences in a random-effect meta-analysis (Fig. 1b). This will lead to a more conservative estimate of the amount of benefit, if a benefit is found, as compared to a fixed-effect meta-analysis.

The antidepressant review example, Fig. 1, presents the differences between fixed- and random-effect meta-analyses when large heterogeneity exists [2]. From one approach to the other, there is a change in study weights (last

columns in the meta-analysis graph). The overall estimate of diamond at the bottom of the random-effect meta-analysis graph is inflated incorporating more heterogeneity. Clinicians should consider that incorporating heterogeneity does not eliminate heterogeneity. Some clinicians would still consider insufficient to use a random-effect approach for the different studies considered by Moncrieff et al. and reject the proposed meta-analysis as overly broad.

Heterogeneity could be also explored using other approaches, such as subgroup analyses or the more sophisticated meta-regression, to try to explain diversity of effects, grouping studies by drug type or dose, by patients' age, by high versus low risk of bias according to the study methodological quality, etc. Subgroup analyses and interaction tests will be discussed in a following Cochrane Corner.

Statistical heterogeneity when only one trial is found

Evidence from a single study is believed to be weaker as compared to that from a meta-analysis. When the results of a single trial show that a treatment is superior to control, the limits of the 95% confidence interval of the effect estimate should be far from equivalence to be resistant to assumptions regarding heterogeneity when multiple trials were conducted and included in a meta-analysis. When only one trial is found for a given comparison in an SR, we will be unable to estimate the potential heterogeneity of treatment effects. Borm et al. [6] recently suggest that the confidence intervals of the effect measure from a single study should be inflated to take into account potential heterogeneity amongst multiple trials. Assuming heterogeneity I^2 of 25, 50 or 75% in a meta-analysis including a trial, the 95% confidence interval around the point estimate of this trial would be more consistent with the meta-analysis results if they are inflated by 115, 141 or 200%, respectively. Transferring this information to the P value scale, a P value of 0.001 from a test of statistical significance would become a P value of 0.02 after inflating confidence intervals by 141%, meaning that considerably more uncertainty regarding the amount of treatment benefit would be found assuming moderate heterogeneity of treatment effect in future research.

Conclusion

Heterogeneity of results from RCTs and other studies informs us that a single trial is just a sample drawn from a pool of potentially diverse pieces of research conducted in different settings. As a matter of fact, a single study is not testing the theoretical efficacy of an intervention, as if patients and doctors were acting in a laboratory, but it rather tries to measure the effect of an intervention that may remain a component of a complex and variable clinical pathway, despite the adoption of study protocols. Understanding the concept of heterogeneity is central in clinical research, and possible heterogeneity sources should be targeted and hopefully, resolved on the question on hand.

A clinician's point of view

Gian Franco Gensini, Roberto Gusinu, Andrea A. Conti

Meta-analyses are powerful research tools used to summarise in a quantitative way the results of clinical trials [7]. As for every powerful instrument, their correct use is fundamental, and Gianni Virgili et al. has clearly explained in the first part of this paper the potentialities and boundaries of application of meta-analyses.

Clinicians are, nowadays, more and more familiar with the publication of meta-analyses in prestigious biomedical journals, yet the full elucidation of heterogeneity is appropriate, since heterogeneity has different dimensions, including the clinical, the methodological and the statistical ones, which may relevantly influence the interpretation of scientific literature [8]. Clinical heterogeneity will be briefly discussed here with specific regard to the clinician's point of view.

Even if meta-analyses are today conducted not only on controlled clinical studies, but also on case–control and cohort studies, the number of meta-analyses selectively containing clinical trials is on the increase. The PICO model is a precious guide in this area [9]. The acronym PICO stands for Patients, Interventions, Comparators and Outcomes, and indicates the basic variables that have to be taken into account as a potential generators of clinical heterogeneity.

The variable "Patients" refers to the demographic and clinical characteristics of the people enrolled in controlled trials; the sample included in the clinical study should always be carefully analysed with respect to age and gender distribution and to pathological features (stage, length and seriousness of the single diseases or of the multiple pathologies, i.e. comorbidity, investigated). The variable "Interventions" regards the type, pattern and modality of the health measures implemented, which can include far different interventions, such as, for example, pharmacological, lifestyle and invasive measures. The variable "Comparators" concerns not only the kind of comparator used in controlled trials, but also the way comparisons are performed. Historically, early clinical trials compared one intervention with no intervention or with placebo [10]. Subsequently, established individual health interventions have been head-to-head compared with newly proposed

measures. More recently, the predominant model of clinical trials available effects a comparison, at least in the therapeutic pharmacological area, between a new drug on top of the best therapeutic armamentarium available and the same optimal therapeutic pattern without the new drug. The variable "Outcomes" deals with the evaluation parameters examined in controlled studies. They too are a potential source of heterogeneity, according to whether they are subjective or objective, on the basis of the time period in which they are collected and analysed and with regard to their being simple or composite end points. At present, there is an interesting and ample ongoing international debate on the appropriateness and drawbacks of the use of composite or combined end points.

Although not included in the acronym PICO, the importance of clinical and health-care settings in controlled studies is fundamental, and not by chance does the evidence-based model of synthesis of clinical trials explicitly foresee it. The full consideration of the PICO model, therefore, appears to be of paramount importance in assessing controlled trials and identifying the possible sources of heterogeneity in clinical research. Furthermore, it constitutes a cornerstone of methodological evaluation given that, even when problems regarding clinical diversity are resolved, statistical heterogeneity may still be present and observable.

Conflict of interest statement The authors declare that they have no conflict of interest related to the publication of this manuscript.

References

1. Moja L, Moschetti I, Liberati A, Gensini GF, Gusinu R (2007) Understanding systematic reviews: the meta-analysis graph (also called 'forest plot'). Intern Emerg Med 2:140–142
2. Moncrieff J, Wessely S, Hardy R (2004) Active placebos versus antidepressants for depression. Cochrane Database Syst Rev (1), Art. No.: CD003012. doi:10.1002/14651858.CD003012.pub2
3. Day SJ, Altman DG (2000) Statistics notes: blinding in clinical trials and other studies. BMJ 321:504
4. Higgins JPT, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. BMJ 327:557–560
5. Deeks JJ, Higgins JPT, Altman DG (2008) Chapter 9: analysing data and undertaking meta-analyses. In: Higgins JPT, Green S (eds) Cochrane handbook for systematic reviews of interventions version 5.0.0 (updated February 2008) The Cochrane Collaboration, 2008. Available from http://www.cochrane-handbook.org
6. Borm GF, Lemmers O, Fransen J, Donders R (2009) The evidence provided by a single trial is less reliable than its statistical analysis suggests. J Clin Epidemiol 62:711–715
7. Conti AA, Galanti C, Gensini GF (2000) La Medicina Basata sulle Evidenze è davvero una moda? Sicuramente è di moda criticarla. Un commento metodologico del Centro Italiano per la Medicina Basata sulle Prove. Ital Heart J 1(Suppl):1192–1195
8. Altman DG, Matthews JN (1996) Statistics notes: Interaction 1: heterogeneity of effects. BMJ 313:486
9. Lai NM (2009) Dissecting students' bedside clinical questions using the 'PICO' framework. Med Educ 43:479–480
10. Conti AA, Conti A, Gensini GF (2006) The concept of normality through history: a didactic review of features related to philosophy, statistics and medicine. Panminerva Med 48:203–205