**Meta-analysis Duet**

**Summing up evidence: one answer is not always enough**

Joseph Lau, John PA Ioannidis, Christopher H Schmid

Meta-analysis (combining quantitatively evidence from different studies) is increasingly popular in clinical medicine. 16 meta-analyses of randomised trials were published in the 1970s, 279 in the 1980s, and 134 from 1990 to 1992. Over 500 appeared in 1996 alone. More than 2000 meta-analyses have been published. The number is small compared with the estimate of half a million randomised controlled trials, but is rapidly catching up, proportionately, with its prime relative. Compared with traditional reviews and expert opinion, meta-analysis provides a more objective and quantitative summary of the evidence that is amenable to statistical tools. Meta-analyses can enhance precision, provide robust estimates, and answer questions that single trials are underpowered or were not designed to address.

**Heterogeneity**

The growth of meta-analysis has not been unchallenged. Critics have voiced mixed enthusiasm or objections to the principle of combining data. Some worry about combining studies with different degrees of bias: studies of different quality may reach different conclusions and studies with negative results may not be published. However, these biases permeate all clinical research, not just meta-analysis and, in fact, careful meta-analysis can help to identify these problems. One major controversy concerns what it is appropriate to combine, even if biases are recognised and accounted for. How much heterogeneity is acceptable? For example, should studies with different dose schedules, follow-up, types of participants, or modes of treatment (eg, oral vs parenteral) be combined?

Meta-analysts usually try to convince the reader that the data are homogeneous to justify combining them for a focused question. For example, does intravenous magnesium reduce mortality in patients with acute myocardial infarction without thrombolytic treatment? Sometimes, perhaps forced by sparse data, meta-analysts have combined diverse studies across a whole field with broad questions in mind. For example, a meta-analysis combined various endoscopic modalities for haemostasis of bleeding peptic ulcer. In the first case, the answer applies specifically to the treatment of acute myocardial infarction in patients without thrombolytic treatment, but may be useless in other patients. Generalisation beyond the populations and situations studied may be inappropriate. In the second case, the answer may apply to all endoscopic treatments, but does it actually apply to any single treatment?

In either case, critics argue: why perform meta-analysis of small trials instead of a single, large, adequately powered trial with a variety of patients so that results would be generalisable? However, similar concerns about heterogeneity apply to single trials, irrespective of size. Diversity in clinical trials is unavoidable (figure 1). Trials may target different populations of patients, and even populations defined by the same eligibility criteria change over time. In HIV infection, patients with CD4 cell counts below 200/µL had a 1-year mortality of 50% in the first antiretroviral trials in 1987; a decade later, 1-year mortality of such patients is less than 5%. Outcomes may not be immutable, even for seemingly similar patients. Large trials, while more precise than smaller trials, may miss important treatment variation and may not be any more generalisable than smaller studies unless their inclusion criteria and recruitment capture broad populations and different settings. Thus, if we want to evaluate how a treatment works, we must ask first whether the best answer is a single estimate, or separate estimates for individuals or subgroups of patients.

Is meta-analysis simply an extension of clinical trials seeking to confirm a single answer or is it a unique discipline aiming to explore multiple answers?

**In search of a single answer**

**Models of fixed and random effects**

Most meta-analyses have asked how well a treatment works overall. This is also what trials traditionally ask. When the combined trials are a homogeneous set designed to answer the same question in the same population, a fixed-effects model, in which the estimated treatment effects vary across studies only from random error, is appropriate. To assess homogeneity, heterogeneity is often tested, based on the $\chi^2$ distribution, but this method lacks power. When heterogeneity is detected, the traditional approach is to abort the meta-analysis or to use random-effects models. Random-effects models assume that no single true treatment effect exists, but each study has a different true effect, with all treatment effects being jointly drawn from the same distribution.
META-ANALYSIS DUET

derived from a population of such truths, commonly assumed to follow a normal distribution. There are different ways to estimate the mean and variance.14,15

Limitations of fixed and random effects
Neither fixed-effects nor random-effects models are entirely satisfactory. In fixed effects, the premise that all studies replicate the same experiment is often oversimplified. Aborting the meta-analysis because of heterogeneity fails to address the question of how to synthesise the data. Random-effects models are less intuitive and do not explain why heterogeneity exists. The relative validity of the assumptions of the two models have been widely debated, but hopefully, in most circumstances, their estimates of the pooled treatment effect tend to agree. Still, the level of statistical significance may differ (random effects are typically more conservative), posing a problem for those who rely too much on p values.16

Large versus small trials
More perplexing discrepancies may occur when a large study disagrees with smaller studies. Fixed effects are driven by the weight of the large study, while random effects compromise between the small and the large—conclusions may then differ. Some mega-trials obtained different results compared with smaller trials and their meta-analyses. A meta-analysis3 of seven trials of calcium (n=1301) had shown a 59% (95% CI 38 to 73%) risk reduction of mortality in acute myocardial infarction. A subsequent trial17 of 2300 patients found a 23% risk reduction and a mega-trial of 58 050 patients18 found no benefit (5% risk increase [-1 to 12%]). Similarly, a meta-analysis of 11 small trials of nitrates in myocardial infarction (n=2170) showed a 41% (19 to 56%) risk reduction for mortality.7 Two subsequent mega-trials of 18 895 and 58 050 patients showed trivial, non-significant risk reductions (6% [-5 to 15%] and 3% [-3 to 8%], respectively).19,20 13 moderately sized trials of aspirin in pregnancy (n=5579) showed a 38% (16 to 55%) reduction in the risk of pre-eclampsia, whereas a mega-trial with 9309 patients failed to detect benefit (11% [-3 to 23%] risk reduction).21 Finally, a large trial19 (n=4589) of calcium to prevent pre-eclampsia reported different results (relative risk 0·94 [0·76 to 1·16]) from a meta-analysis of 14 small studies (n=2459; relative risk 0·38 [0·22 to 0·65]).22

The frequency of such discrepancies has been systematically studied.23-25 The methods for defining what constitutes a discrepancy differed (overlap in confidence intervals, agreement in p values above or below 0.05, and disagreement beyond statistical chance), but all teams would agree that discrepancies are seen about 10–23% of the time when disagreement beyond chance is considered.26 Thus small and large trials usually agree, but discrepancies, and thus heterogeneity, are not uncommon.

In search of multiple answers
Meta-regressions
Meta-regressions have been used to suggest reasons for observed heterogeneity.27 As in any regression analysis, meta-regressions attempt to identify significant relations between the treatment effect (the dependent variable) and covariates of interest (the independent variables). Whereas in trials the unit of observation is the individual patient, in meta-regressions the unit of observation is the study or subgroup. From that perspective the only major difference between meta-analyses and clinical trials is the unit of measurement. The independent variables may be either constant for the whole study or subgroup (eg, dose schedules, study duration, the use of proper...
randomisation and masking) or may represent a summary measure of individual values (eg, mean or median blood pressure or percentage of male patients) in each study.

As an exploratory tool, meta-regression may provide more insight than traditional pooling and may help formulate new hypotheses: the benefit from reducing cholesterol correlates with the risk of coronary artery disease; the harm of antiarrhythmics depends on the risk of arrhythmia; the relative benefit from changing to new antiretroviral treatments is larger at earlier disease stages of HIV infection; and a lower dose of aspirin may lead to less gastrointestinal toxicity. Several variables may be modelled, describing the multidimensional effects of several factors acting together. For example, a low dose of co-trimoxazole is better tolerated than higher doses for prophylaxis of Pneumocystis carinii infection, and discontinuations because of intolerance also depend on the location of the patients (USA, north or south Europe) and the duration of follow-up. In these examples, reporting an aggregate pooled effect might have been misleading.

Limitations of meta-regressions
A significant association in a regression analysis does not prove causality. Moreover, meta-regressions pose additional challenges. First, ecological fallacies ensue when summary data for a group misrepresent the individual patients. This may be a problem for covariates with different values for each patient within a group. What is true for a study with patients of average age 50 may not apply to a 50-year-old patient. Second, selection of variables is a challenge in any area of research, but is heightened in meta-regression. In addition to the bias from the meta-analyst selecting the variables deemed important, trial investigators target only specific data in each trial. Some trialists advocate randomising very large numbers of patients with minimum data collection. Although this strategy minimises cost and complexity, it impedes the development of informative exploratory regression models when data collection does not include all the known potentially important biological variables. Even when the data are collected with future analyses in mind, meta-regression can suffer if important variables are undiscovered or unavailable when the trial is designed.

Overcoming the limitations
The use of individual patients’ data rather than summary data from each study circumvents ecological fallacies. Such analyses provide maximum information about covariates to which heterogeneity can be ascribed, facilitate collaboration with the original investigators which may improve data quality, encourage further follow-up of patients, and allow for time-to-event analysis. Despite the increased cost, time, and complexity in obtaining these detailed data from all pertinent trials, individual patients’ analyses must be encouraged. The success of the approach, such as the Early Breast Cancer Trialists’ Collaboration, should be better publicised. But those embarking on this course must be wary of introducing retrieval bias when, unable to obtain the data from all studies, they use data from an incomplete set of trials. Until now, the literature has been dominated by meta-analyses of pooled data and the potential advantages of individual patients’ data need to be validated with more experience in the future.

Variable-selection bias can be circumvented, if covariates are available in all trials. One ubiquitous covariate that may serve as a surrogate for others that are missing is the control rate, the proportion of patients who experience the event of interest among those assigned to the control group. It is a global measure reflecting multiple factors that contribute to the occurrence of events in the study cohort. The control rate is related to the severity of illness or risk of disease, as well as to effects from concomitant treatments, the duration of follow-up, and the play of chance (random error) across different trials.

The effect of the play of chance should be eliminated. A weighted least-squares regression of trials with different control rates may erroneously suggest a larger benefit at higher control rates and harm at lower rates, only because of random error. A study having by chance a very high control rate invariably shows a benefit, even if there is none, since it is unlikely that the rate of events would also be as inflated by chance in the experimental arm. Conversely, if a study has by chance a very low control rate, the experimental arm cannot do any better. Statistical methodology has been developed to correct for this problem. With this methodology, it is not uncommon for the control rate to be correlated with the treatment effect. In five of 15 discrepancies between large and smaller trials reported by Cappelleri et al, treatment-effect differences correlated with significant differences in the control rate of small and large trials. In the case of trials of magnesium in

---

**Figure 2: Summing-up evidence in single and multiple dimensions**
acute myocardial infarction, even after correction for chance, there is significant evidence that a higher benefit is seen with increasing control rates, while no benefit would be expected at the control rate of 7% observed in ISIS-4. A meta-regression excluding this mega-trial would still predict no benefit at a control rate of 7%.

Although control-rate meta-regression may suggest that groups of patients at high risk respond differently to treatment than groups at lower risk, one should not directly extrapolate the results to individual patients. For example, for magnesium, it would be erroneous to claim that a benefit is expected for a patient with a higher than 7% risk of mortality. Subgroups of patients with substantially higher mortality rates in ISIS-4 (eg, patients with heart failure [control rate 14·2%]) did not benefit from magnesium. Populations enrolled in clinical trials are frequently largely skewed in terms of the disease risks of individuals. A few high-risk patients contribute most of the observed mortality—eg, in myocardial infarction, half of all deaths occur in patients with cardiogenic shock, who comprise less than 5% of the cohort. Studies with different average control rates may have different percentages of such high-risk patients who dominate the cohort mortality. Thus, showing efficacy for an average control rate higher than 7% may reflect a benefit among patients at very high risk, probably many times higher than 7%, rather than a benefit for all patients with a risk above 7%.

Furthermore, the control rate may not only reflect the individual severity of illness. Even when adjusted for follow-up and corrected for random error, it is a composite of illness severity, concomitant treatments, efficacy of care, and several other factors differing among trials. Moreover, important pathophysiological determinants of the treatment effect may have no relation with the baseline disease-risk. For example, differences in the timing of the administration of magnesium may also be responsible for discrepancies among magnesium trials: the ion is ineffective, if given after reperfusion, as in the majority of the ISIS-4 patients. Nevertheless, control-rate meta-regression can be an important analytic tool to understand between-study heterogeneity.

Conclusions

Given the problems, it is perhaps surprising that meta-analyses have agreed quite well with large trials addressing a similar “homogeneous” question. Clinical trials and meta-analyses mostly have addressed the question of how well a treatment works overall. Both of these approaches, while useful in estimating a population effect, do not show how to treat individuals.

Patients may respond differently to treatment. To address this diversity, meta-analysis needs to evolve from deterministic pooling to multidimensional exploration, creating response-surface models to summarise evidence along multiple covariates of interest (figure 2). Meta-analysis of individual patients’ data would be optimal to create such models with accuracy. In the absence of readily available individual patients’ data, this role may also be assumed by traditional meta-analyses, provided an adequately large number of summary data are available, conclusions are drawn with caution, and it is understood that the more questions that are asked, the less will be the certainty of the answers.

One size may not always fit all. Often, obtaining one answer for the whole population may be appropriate, such as when satisfying regulatory and public-health requirements. However, even in these situations, a single recommendation could be controversial, as in the debate about the use of screening mammography for women 40–49 years of age. On many occasions, meta-analysis may be better suited to explore differences in harm and benefit across subgroups of patients. Having different approaches to meta-analysis is not a drawback, provided their strengths and limitations are recognised (panel) and exploited appropriately. Of course, meta-analyses require technical expertise and should be performed by teams that have knowledge in advanced areas, such as meta-regression. Moreover, the multivariate approach to meta-analysis needs to be validated prospectively as more experience is accumulated on its application. This will help determine the extent to which additional information is gained and whether false-positive associations are claimed through the multiplicity of analyses.
There is much to improve in both clinical trials and meta-analysis, if we expect to obtain useful information more often than biased compilations. The need to improve the quality, and reporting of clinical research cannot be over emphasised, but besides quality we need to look at a broader picture of advance planning. Clinical trials and meta-analyses offer a continuum of evidence. Prospective meta-analyses should be encouraged and trials should be carried out with meta-analyses of similar trials in mind. Bayesian analyses can be used to combine the prior evidence from the meta-analysis and the new evidence from the clinical trial together. Patients' heterogeneity should be anticipated and sought up front, not ignored in pursuit of aggregate statistical significance. Statistical heterogeneity, instead of being thought of as a nuisance, may be a blessing and may help us to understand clinical and biological heterogeneity. Knowing how best to treat the individual should be the ultimate goal of both clinical trials and meta-analyses.

References