Evidence-based medicine. Part 1: The power and pitfalls of meta-analysis

Tim Holt

Article points

1. Meta-analysis allows data from more than one trial to be aggregated, and now sits at the top of the "hierarchy of evidence" as the most reliable means of answering research questions.
2. The process of meta-analysis is greatly facilitated by software products, which automatically calculate overall effect size, confidence intervals, and heterogeneity values, as data from different studies are entered.
3. The technique of meta-analysis equips us to assess evidence in a way that supersedes that of the opinion-based era.

Key words

- Evidence
- Meta-analysis
- Research
- Statistics

Tim Holt is Clinical Lecturer at the Health Sciences Research Institute, University of Warwick, Coventry.

Evidence-based medicine has challenged opinion-based clinical decision making since the 1990s. Meta-analysis improves statistical power by combining data from more than one study. However, although well conducted meta-analyses sit at the top of the "hierarchy of evidence", poor technique can cause confusion and negatively impact on medical practice. The first in a series of articles looking at evidence-based medicine, this article explains the method of meta-analysis and demystifies the concept. A recent diabetes meta-analysis is used as an example.

Starting in the early 1990s, evidence-based medicine (EBM) challenged the opinion-based decision-making of previous decades. Opinions may be personal, resulting from a clinician's own experience, or based on an expert consensus. Recognition of the unreliability of personal and, indeed, expert opinion prompted change towards high-quality research evidence. This change itself required much stricter rules to improve quality in the design and reporting of trials, and also the modern tools of evidence synthesis. These included not only the physical software environment, but also the statistical process of meta-analysis. This became available in the 1980s (DerSimonian and Laird, 1986), and the use of electronic bibliographical databases and the internet soon followed.

Meta-analysis allows data from more than one trial to be aggregated, and now sits at the top of the "hierarchy of evidence" (Guyatt et al, 1995) as the most reliable means of answering research questions. If data from several (or many) studies addressing the same question can be safely brought together, this increases the effective sample size and so the statistical power.

Early success and failure

A famous early example of the success of meta-analysis was the discovery of the benefits of thrombolysis for acute myocardial infarction, a finding that rapidly changed clinical practice in the 1980s, but which could have happened several years earlier had meta-analysis been available (Stampfer et al, 1982). This sparked great interest in the technique.

In 1993, however, a meta-analysis suggested that intravenous magnesium was also beneficial in this situation (Yusef et al, 1993), a conclusion later revoked through the findings of one very large randomised controlled trial (RCT), the Fourth International Study of Infarct Survival (ISIS; ISIS Collaborative Group, 1995). From this episode a better understanding of the potential and the pitfalls of the new technique emerged. While meta-analysis reputedly sits

above single RCTs in the hierarchy, a single RCT may be superior if large enough (such as a "mega-trial").

Type 1 and type 2 errors

By increasing statistical power, meta-analysis aims to overcome two problems affecting our ability to recognise and exclude true treatment effects. These are termed type 1 and type 2 errors.

Type 1 errors occur when an apparently beneficial effect in fact results from chance. At the 5% level of statistical significance, this would be expected to occur in about one in 20 published studies. But in fact it occurs more often than this, through a regrettably common form of bias – publishing bias. Positive studies demonstrating benefits are more likely to be submitted for publication, more welcomed by journals, more likely to be publicised and disseminated, and therefore more likely to influence our practice. Nobody likes negative studies – not the researchers, the participants, the trial sponsor, the journal editor, nor the reader, except when delighting in seeing a bogus therapy exposed! Many treatments have gained acceptance on the basis of type 1 errors, later to be withdrawn when more evidence accumulates.

Type 2 errors result from a failure to demonstrate a true effect due to inadequate sample size. Studies with an insufficient sample should not, in theory, be undertaken, but in practice an estimate of the necessary sample size includes several unknowns including, of course, the effect size of the intervention. A trial may be sufficiently powered to demonstrate an effect provided the intervention changes the outcome by, for example, 10%. An intervention that changes the outcome by 5% will produce a non-significant result, even though this smaller benefit is a true one. A sinking feeling may occur in the researcher's mind as the trial results are analysed: "the best I can hope for here is that this study (that has taken up several years of my life) will contribute to someone else's future meta-analysis"!

Apples, oranges, or fruit?

The process of meta-analysis is greatly facilitated by software products, which automatically calculate overall effect size, confidence intervals, and heterogeneity values, as data from different studies are entered. Heterogeneity is a measure of how dissimilar the studies are, but is based only on the numerical data that are entered. It will be high if the effect sizes of different studies vary, and if confidence intervals overlap poorly. However, a major pitfall is that low heterogeneity values (suggesting highly comparable studies) could occur even though the methods, setting, type of intervention, and outcome measures are different.

Statistical heterogeneity is different from methodological and clinical diversity, but this may not be recognised. Attempts to shoe-horn trials that are actually quite dissimilar into the same meta-analysis may, unfortunately, succeed. The concern is that

apples and oranges are combined in the same analysis. This can sometimes be justified if what we are actually interested in is a wider issue concerning fruit (Higgins and Green, 2009), but is otherwise risky.

Rubbish in, diamonds out

Rather than "rubbish in, rubbish out", the use of meta-analysis software may produce beautiful diagrams that seduce users into believing they have uncovered a diamond. The diamond is the black symbol at the bottom of the forest plot that indicates the overall effect size and confidence interval. No-one should attempt this endeavour without adequate training and statistical support, particularly when combining studies that are clinically or methodologically diverse. In expert hands, a high-quality meta-analysis deserves its place at the top of the hierarchy, but in the wrong hands can simply confuse.

Before data are entered into the software, the literature must be searched systematically for relevant publications. In addition to the publishing bias mentioned earlier, there is also the potential for human factors to influence the choice of titles. Fairly subtle changes in inclusion and exclusion criteria may have a considerable effect on the final sample of included studies. More importantly, studies might be excluded because they report results in a way that is incompatible with other studies in the synthesis. For instance, some may report binary outcomes, others continuous outcomes.

In diabetes, a binary outcome might be the achievement of a target HbA_{1c} level; each participant has, or has not, achieved this at the end of the study. The study will report the odds of achieving the target in the intervention and control arms, and the ratio of these odds as the effect size. A different study might simply measure the change in HbA_{1c} level – a continuous outcome – in each arm, the difference between the arms, and its confidence interval. These trials will be difficult to combine in a meta-analysis (although techniques are available to assist with this), even though they are both measuring something very similar. A further problem may occur where different tools are used to measure the same outcome, such as quality of life. Furthermore, some trials report outcomes as the change from baseline, others simply the outcome value at the end of the study, further complicating data synthesis.

The problem of bias in study selection is offset first of all by publishing a peer-reviewed protocol for the meta-analysis prior to starting it, and second by ensuring that each article identified in the initial searches (apart from obviously irrelevant titles) is assessed by at least two reviewers. Disagreement is resolved either by discussion or by a third reviewer.

An "ideal" meta-analysis involves studies undertaken in the same setting, using the same design, the same outcome (measured in the same way and after the same time interval), has no missing data, and includes high-quality studies whose effect sizes are similar. It is not essential that the study populations are the same – because each arm of the meta-analysis will contain roughly equal numbers of randomised participants – but they all need to be measuring the same effect. As can be imagined, it is rare for all of these ideal conditions to co-exist.

An example

Let's look at an example of a meta-analysis related to diabetes care. Sherifali et al (2010) have meta-analysed the effects of oral antidiabetes agents on HbA_{1c}. The first thing to look at when assessing a meta-analysis is the publication date. Out-of-date meta-analyses may draw different conclusions to more recent reviews. Sherifali et al was published recently, in August 2010.

A strength of this review, in contrast to previous similar studies, is that the methodological criteria were pre-determined, and heterogeneity was reduced by including only high-quality trials with comparable designs. A large number of trials (61) were included, associated with a total of 26 367 participants and 103 different comparisons (one study may, for instance, include data for different doses of a drug, producing more than one comparison). The range of therapies was diverse, including all the major oral antidiabetes drug classes. It was possible to combine different drugs from within the same class (e.g. rosiglitazone and pioglitazone), to derive an estimated effect size for each class.

Page points

1. In addition to publishing bias, there is also the potential for human factors to influence the choice of titles. Fairly subtle changes in inclusion and exclusion criteria may have a considerable effect on the final sample of included studies.
2. The problem of bias in study selection is offset first of all by publishing a peer-reviewed protocol for the meta-analysis prior to starting it, and second by ensuring that each article identified in the initial searches (apart from obviously irrelevant titles) is assessed by at least two reviewers.
3. An "ideal" meta-analysis involves studies undertaken in the same setting, using the same design, the same outcome (measured in the same way and after the same time interval), has no missing data, and includes high-quality studies whose effect sizes are similar.
4. The first thing to look at when assessing a meta-analysis is the publication date. Out-of-date meta-analyses may draw different conclusions to more recent reviews.

This is an example of legitimately comparing apples with oranges (different specific drugs within the class), because what we are interested in is fruit (the blood glucose-lowering effect common to all of them within the class). Combining different classes would not be appropriate. What was more difficult was that different trials measured outcomes after different time intervals – an example of methodological diversity. The authors included in their main forest plot all comparisons reporting outcomes at 13–18 weeks. This was only a subset of all the included comparisons. As a separate figure the authors plotted effect size against time interval, enabling us to see whether longer duration of therapy leads to further reduction in HbA_{1c} level or not.

The authors found that most agents reduced HbA_{1c} levels by 0.5–1.25 percentage points. The thiazolidinediones and sulphonylureas produced changes at the upper end of this range. The benefits were largely evident within the first 6 months. They looked for baseline variables associated with increased likelihood of response to the therapies. They then carried out a meta-regression, which allows a predicted effect on the outcome (HbA_{1c} level) to be estimated for unit change of a predictor variable (in this case, baseline HbA_{1c} level). The participants whose HbA_{1c} level started high tended to have a greater reduction in HbA_{1c} level.

Summary

Clinicians today do what they did before the advent of evidence-based medicine, i.e. the best they can for the individual given the knowledge and evidence available to them. Sometimes this evidence is very adequate in quality and volume and at other times it is poor. The technique of meta-analysis equips us to assess evidence in a way that supersedes that of the opinion-based era. Some may welcome this change, others lament the demotion that "personal experience" has suffered in the process. Meta-analysis is difficult, requires experience, and can sometimes lead us astray, but when conducted properly allows the maximum benefit to be gained from the often piecemeal evidence available. ■

DerSimonian R, Laird N (1986) *Control Clin Trials* **7**: 177–88
Guyatt GH, Sackett DL, Sinclair JC et al (1995) *JAMA* **274**: 1800–4
Higgins JPT, Green S (2009) *Cochrane Handbook for Systematic Reviews of Interventions.* Available at: http://www.cochrane-handbook.org/ (accessed 21.12.10)
International Study of Infarct Survival Collaborative Group (1995) *Lancet* **345**: 669–85
Sherifali D, Nerenberg K, Pullenayegum E et al (2010) *Diabetes Care* **33**: 1859–64
Stampfer MJ, Goldhaber SZ, Yusuf S et al (1982) *N Engl J Med* **307**: 1180–2
Yusuf S, Teo K, Woods K (1993) *Circulation* **87**: 2043–6