SHIRLEY REYNOLDS, MSc
DAVID STREINER, PhD

1 Hill CE, Corbett MM. A perspective on the history of process and outcome research in counselling psychology. *Journal of Counseling Psychology* 1993;**40**:3–24.

2 Cook TD, Campbell DT. *Quasi-experimentation: design and analysis issues for field settings.* Boston: Houghton Mifflin, 1979.

3 Messick S. Test validity and the ethics of assessment. *Am Psychol* 1980;**35**:1012–27.

4 Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use.* Oxford: Oxford University Press, 1995.

5 Coyne JC. Self-reported distress: analog or ersatz depression? *Psychol Bull* 1994;**116**:29–45.

6 Mintz LB, O'Halloran MS, Mulholland AM, *et al.* Questionnaire for eating disorder diagnoses: reliability and validity of operationalizing DSM-IV criteria into a self-report format. *Journal of Counseling Psychology* 1997;**44**: 63–79.

Understanding and interpreting systematic reviews and meta-analyses. Part 2: meta-analyses

In part 1 (August 1998 issue) we introduced the rationale for the *systematic review* and described the first part of how to appraise critically such articles before using them clinically. The user would want to know: did the review focus on a specific question? Was a comprehensive and clearly described search strategy used? Were the appropriate studies selected? And did the raters agree about which articles should be included?

In part 2, we focus on the statistical combination of the results of a series of studies (meta-analysis). Our objective is to suggest questions that a reader should ask of an article describing a meta-analysis. We will also outline the main methods used in meta-analyses.

Were the results of the individual studies combined and was this appropriate?

Meta-analyses seek to provide the best estimates of treatment effect based upon all the available valid evidence. The simplest type of meta-analysis involves simply counting up the number of statistically significant studies (*vote counting*). Although vote counting is straightforward and superficially easy to interpret, it leads to various potential biases and other problems. Many treatment effects that are of potential clinical importance are only moderately sized. When they have only been investigated in small trials with insufficient statistical power, a simple vote count may fail to identify a true treatment effect that may be important clinically. Of course, such a situation is why many meta-analyses are conducted, and this is therefore a major deficiency of the vote count. Similarly, vote counts treat every study the same, when larger and more powerful studies may appropriately be attributed greater weight than small lower powered ones. A further and important problem is that vote counting does not provide a useful estimate of the magnitude of an effect across a group of studies. For these reasons, more sophisticated methods are needed to synthesise the results of several experimental studies.

APPLYING WEIGHTS TO DIFFERENT STUDIES

A fundamental premise of meta-analysis of randomised controlled trials is that randomisation within individual studies is preserved, and a pooled estimate of the treatment effect is derived from a *weighted average* of study effects. Weights in meta-analyses, like confidence intervals, are based upon the *standard error* (SE) of the study effect. The SE for a study with a dichotomous outcome such as alive or dead will reflect both the number of deaths observed in each group and the total number of patients randomised. Similarly, the SE for a continuous outcome, such as the mean Positive and Negative Symptom Scale score, reflects the observed distribution of the effects (the standard deviation) and the number of patients contributing information. In other words, the SE may be considered to reflect the amount of statistical information available in a study and is therefore an appropriate basis for applying weights to different studies. It follows that studies providing more precise estimates (with *tighter* confidence intervals) receive proportionally more weight than those with less precise estimates (*wider* confidence intervals). Although meta-analyses may be daunting to those without much statistical knowledge or confidence, this basic principle is straightforward in theory and is fundamental to meta-analysis. Different methods for meta-analysis are essentially analogous, and reflect differences in how the weights are calculated for each study.[1]

INTERPRETING FIGURES

Figure 1 describes a meta-analysis of the effects of donepezil hydrochloride on quality of life that uses data from an evidence-based guideline on the treatment of dementia in primary care.[2] This kind of figure is common in reports of meta-analyses and is sometimes called a "blobbogram." Each horizontal line represents a different randomised comparison (often individual trials, but in this case 2 comparisons are made within 2 separate trials). The horizontal bars represent the width of the 95% confidence intervals (which indicate the range of values in which we are 95% certain that the *true* value falls) and the diamond indicates the point estimate of effect. Incidentally, the point estimate is not only the best estimate of the population effect, it is also the most likely value, with the results becoming increasingly less likely as we move towards the 95% confidence limits. In figure 1, the size of the diamond reflects the number of patients randomised.

The pooled effects "*lozenge*" describes a weighted average of the studies, and again provides a point estimate and 95% confidence intervals. This weighted average is little influenced by the comparatively positive result in the trial by Rogers (1996) as the confidence intervals for that study are relatively wide. The pooled effect is influenced more by the results of the 4 other larger comparisons below.

The measure of effectiveness used in meta-analyses depends on the nature of the outcome. When the outcome is event-like or dichotomous (such as mortality or admission to hospital) the measures most commonly used are the odds or risk ratios,

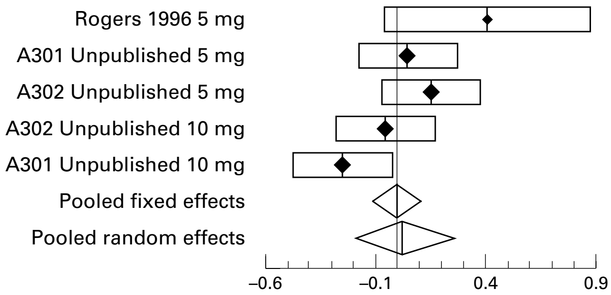

Figure 1 Effects of donepezil on quality of life (standardised effect size).

although risk differences are also used. When the outcome is continuous, such as mean score on the Positive and Negative Symptom Scale, the measure of effectiveness is often expressed as an *effect size* and indeed this is what is used in the example in figure 1, although sometimes the original measure (eg, Positive and Negative Symptom Scale) is used. We discuss the clinical interpretation of different measures below.

TESTS FOR HETEROGENEITY

A careful glance at figure 1 may provide some cause for concern. Although the confidence intervals all overlap, the results of the individual comparisons look quite different. The labels describing each comparison indicate that different doses of drugs are used in the studies. This may mean that we may not be comparing like with like. The studies with higher doses indicate non-significant negative effects upon quality of life, whereas non-significant positive effects are associated with the lower dose studies. Also, the estimate from the Rogers study is much more positive than the others, and nearly significant in its own right (although it is based upon a relatively small number of subjects as reflected by the small diamond representing the point estimate and the wide confidence intervals). These observations might undermine our confidence that there is a single underlying treatment effect and make the fixed effects estimate unhelpful (because this is what it is attempting to estimate).

One way of seeing if the results of the individual studies are similar enough to combine is a statistical test for *heterogeneity*. *Heterogeneity of treatment effect* refers to *systematic differences* between the results of studies that cannot be attributed simply to chance. Standard heterogeneity tests estimate the probability that the observed pattern of results may have occurred simply through the play of chance. In our case, the p value for the heterogeneity test is 0.042. In other words, if the studies really are all estimates of a single underlying treatment effect, we might only expect to see the observed differences between studies (or worse) about 42 times in 1000 simply through the play of chance.

FIXED VERSUS RANDOM EFFECTS

Heterogeneity tests in meta-analysis frequently lack power because they rely upon assumptions about large numbers that may not apply when only a small number of small studies are available for the meta-analyses. This situation occurs frequently because it is one of the main reasons for doing a meta-analysis in the first place. When there is evidence of significant heterogeneity, this is potentially important, although the practical importance of p values alone is hard to interpret. Fortunately, approaches have been devised that can accommodate differences among studies. Fixed effects estimates assume that there is a single underlying population treatment effect, which will be reflected most accurately by larger studies with more statistical power. Random effects models take into account the heterogeneity among studies, both in the point estimate of the treatment effect and in the width of the confidence intervals.

Figure 1 describes a random effects approach to estimate the pooled effect of the donepezil studies on quality of life.[3] The point estimate from the random effects model is very similar to that of the fixed effects approach. It has, however, shifted slightly to the right reflecting an increased relative influence from the smaller study by Rogers with its outlying estimate of effect that is given greater weight than in the fixed effects approach. Perhaps more importantly, the confidence intervals are much wider too, reflecting the variability in the results of the pooled studies.

Standard random effects models are *adaptive*, in that when there is no heterogeneity (p value of 0.5 or greater) they behave as fixed effects models, but as the p value becomes smaller they increasingly take this into account. Fixed effects models may ride roughshod over important differences between study effects. Random effects models may increase the weight of smaller studies that may be more open to systematic bias. The choice between models remains controversial, although reviews that report both may avoid controversy particularly when they give the same answer.

PUBLICATION BIAS AND OTHER RISKS

As we described in part 1, one of the most important biases which occurs in systematic reviews is publication bias. Analyses based only on a small number of small trials are likely to be particularly open to publication bias, and standard statistical methods may also become unstable in this situation, particularly when there are only a few events in the trials. This does not undermine the value of meta-analysis per se, as it remains the best way to summarise the available evidence. A thoughtful and critical approach to interpreting the results of meta-analysis is important. Readers interested to know more about how standard meta-analysis methods work are encouraged to consult the excellent practical review by Fleiss.[4]

What were the results of the meta-analyses and how can they be used clinically?

Once the methods of the review have been critically appraised and appear adequate, the next stage is to consider the results and how they can be used to help your patient. As Peter Szatmari described in his note on using the results of randomised controlled trials (see *Evidence Based Mental Health* 1998;**1**:39–40) there are several measures of clinical effectiveness—one of the most useful for the clinician is the *number needed to treat*. One of the challenges in the clinical interpretation of systematic reviews and meta-analyses is that the most statistically useful approaches may not also be the most clinically meaningful. Here we will outline some of the current approaches to using the results of a meta-analysis.

EVENT–LIKE OUTCOMES

The most commonly used measure of effectiveness in meta-analyses in which the outcome is an event is the odds ratio. The odds ratio is simply the odds of an event occurring in the experimental group divided by the odds of it occurring in the control group. The statistical properties of the odds ratio make it preferable to other metrics in meta-analyses of treatments that attempt to describe the effect of a treatment upon the risk of an event.[5] Although preferable statistically, there are often problems regarding the practical interpretation of odds ratios. For example, on p 115 we abstract a review by Marshall and Lockwood on the effectiveness of active community treatment for people with severe mental disorders. In *Evidence-Based Mental Health,* for reasons of space, we do not provide the blobbogram but figure 2 shows a re-analysis from data in the original article.[3]

Odds ratios are used in epidemiology to approximate risk ratios, which work well when the denominator is large. As an increasing proportion of subjects suffer events, however, odds ratios tend to provide larger estimates compared with the

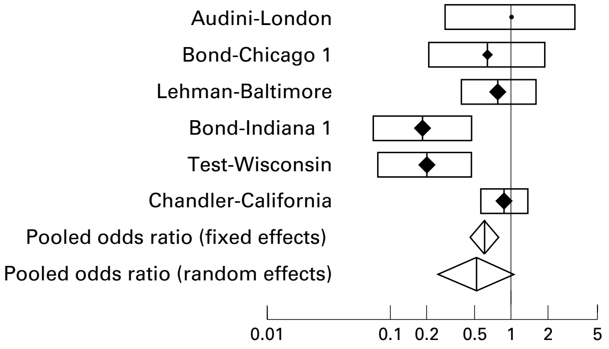

Figure 2 Effect of active community treatment on the odds of admission to hospital.

Percentage of control scores below the average experimental score for various effect sizes

*Effect size**	*Percentage of control scores which would be below the average experimental score*
0	50
0.2	58
0.4	66
0.6	73
0.8	79
1.0	84
1.4	92
1.6	95
2.3	99

*If the effect size is negative, the same percentage of controls do *better* than the average experimental patient.

relative risk. So although active community treatment leads to a 40% reduction in the odds of admission to hospital (odds ratio 0.60; 95% CI 0.45 to 0.79), this is only a 25% reduction in the risk of an event (risk ratio 0.75; CI 0.64 to 0.87). Figure 2 also shows a startling difference in the random effects and fixed effects estimates. Looking at the individual studies it is clear that the majority of the benefits are seen in 2 studies (which are not the largest). The random effects model describes the differences between study effects through wider confidence intervals. Overall, as the point estimate from the random effects model describes, it is likely that active community treatment has a beneficial effect. However, the user of this evidence would also want an explanation for the substantial differences in study effects observed.

Reviewers attempting to describe the practical importance of their results occasionally use simple risk differences, the risk in the intervention group minus the risk in the control group. Risk differences are the inverse of the number needed to treat. There are 2 main problems with this approach. Firstly, risk differences are difficult to interpret without additional information on the length of time in which benefits are accrued. Perhaps more fundamentally, risk differences are easily confounded by study characteristics in ways which odds and risk ratios avoid. For example, if a trial includes a sample of patients at higher underlying risk than those seen by a clinician, she should not expect the same benefits among her patients. In 2 hypothetical studies describing the same treatment in similar sample populations, one of which is twice the length of the other, we might expect twice the risk difference in the longer study (although the ratio of events would be the same). One way of extrapolating from the pooled odds ratio in a meta-analysis is to combine it with the patients' expected event rate (PEER) to produce an estimate of the number needed to treat:

$$\text{NNT} = \frac{1 - [\text{PEER} \times (1 - \text{OR})]}{(1 - \text{PEER}) \times \text{PEER} \times (1 - \text{OR})}$$

In the case of assertive community treatment, if the expected rate of admission was 0.2 or 20%, with a fixed effects odds ratio of about 0.6, the estimate of the number needed to treat is about 15. In our abstract of the Marshall and Lockwood review, we present the findings as numbers needed to treat—these reflect the average effectiveness of assertive community treatment during the course of the studies and will need tailoring to specific clinical circumstances.

CONTINUOUS OUTCOMES

As we mentioned above, the summary measure of effectiveness when continuous measures are used is the *effect size*. The effect size is a measure of the overlap in the distributions of scores on the outcome scale in the control and experimental groups. The effect size is the standardised *z* score which can be thought of as describing outcomes in standard deviation units. Like the odds ratio, the effect size has useful statistical properties, and is more robust than meta-analyses based upon mean scores using the original units. Also, like the odds ratio, the effect size is difficult to interpret practically. One way of interpreting the measure is to estimate the degree of overlap between the 2 populations by obtaining the proportions of the standard Normal distribution above and below the *z* value. This is the proportion of control group scores that are less than the average score in the experimental group. For example, Nowell *et al* (p 117) reviewed the effectiveness of benzodiazepines and zolpidem compared with placebo on total sleep time and found an overall effect size of 0.71 (CI 0.55 to 0.87). This would mean that 76% of control scores would be less than the average score in the experimental group. Unfortunately, the only way to work out the precise percentage is to use statistical tables—unless you can memorise the table for the Normal distribution! As an aide memoire, in the table, we give a summary of the percentages for a range of effect sizes which can be used to give a rough estimate.

Comment

Systematic reviews that include meta-analyses are increasingly common, and when properly conducted may provide the best available indication of the effectiveness of a treatment through summarising the available evidence. However, like all analyses, different assumptions and methods may provide different answers to important questions. In *Evidence-Based Mental Health* we appraise each review using explicit criteria, and aim to ensure the validity of the reviews that we abstract. However, some understanding of the way a systematic review should be put together will be helpful to readers in sorting out what is important and for addressing some of the challenges to the interpretation and implementation of the results.

Nick Freemantle, MA
Medicines Evaluation Group,
Centre for Health Economics,
University of York, UK

John Geddes, MD
Editor, Evidence-Based Mental Health

1 Hedges LV. Meta-analysis. *Journal of Educational Statistics* 1992;**17**:279–96.
2 North of England Evidence Based Guideline Development Project. The primary care management of dementia. *BMJ* 1998; in press.
3 Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* 1995;**14**:2685–99.
4 Fleiss JL. The statistical basis of meta-analysis. *Statistical Methods in Medical Research* 1993;**2**:121–45.
5 Deeks J. Swot's corner: what is an odds ratio? *Bandolier* 1996:**25**:6.