Psychological Medicine, 1997, **27**, 3–7.

EDITORIAL

Systematic reviews and meta-analysis[1]

Conclusions about medical interventions or the causes of disease are based upon reviews of the scientific literature. Single studies usually have limited statistical power or may be difficult to interpret or generalize and so the findings from a single study can rarely justify a change in clinical practice or in an aetiological theory. Even when planning larger studies or mega-trials (Yusuf *et al*. 1984), a thorough review of existing literature is needed and the results of the study need to be placed in that context, though single studies can exert an important and powerful influence. In short, though most scientific effort is expended on conducting individual research investigations, it is the conclusions from reviews upon which most clinical medicine, public health interventions and epidemiological associations are based.

Given the importance of reviews, it is surprising that the methodology of reviewing studies has been relatively neglected. However, over the past 10 or 15 years the need for systematic reviews, in which 'scientific' principles are employed, has gradually been accepted, though the majority of currently published reviews are still expert narrative reviews. Systematic reviews abandon the methodology of reviewing that relies exclusively upon the literature that is remembered or is frequently cited and eschews the reviews that miraculously cite evidence that always supports the author's opinions. As far as is possible, systematic reviews use the methodology of primary research; the aims of the review, the important variables and the criteria for inclusion of relevant literature are determined beforehand and the methodology of searching for both published and unpublished material is specified.

In the course of conducting a systematic review of the qualitative information there will also be opportunities to perform a quantitative summary or meta-analysis; a term probably first coined by Glass (1976). The main advantage with this procedure is that it increases the statistical power and the precision of the estimates of effect. This increase in power may also provide opportunities to study heterogeneity, for example, whether the response to treatment differs between study design or groups of patients (Thompson, 1994). Many of the randomized controlled trials (RCTs) carried out in psychiatry are small and often do not have enough statistical power to detect even quite important clinical differences. For example, to detect a difference in recovery rate of 60% *versus* 70% a clinical trial would need 374 subjects in each group for 80% power at 5% significance.

The principles of systematic reviewing and meta-analysis can be applied to all forms of scientific literature but recently there has been a great deal of interest in applying this method to all RCTs conducted to evaluate medical treatments. The Cochrane Collaboration is an international network committed to organizing systematic reviews of RCTs in all areas of medicine (Chalmers *et al*. 1992). It is named after the epidemiologist Archie Cochrane, one of the early exponents of what is now termed 'evidence based medicine'. The task for obstetrics and gynaecology has largely been completed and the approximately 6000 RCTs are reviewed in *Effective Care of Pregnancy and Childbirth* (Chalmers *et al*. 1989) along with a regularly updated electronic database. The task for mental health professionals is truly Herculean. The RCT in medicine was developed by Bradford Hill in the late 1940s (Hill, 1952) and there are over 600 clinical trials in the *British Journal of Psychiatry* alone and a further 90 or so in *Psychological Medicine* (Clive Adams, personal communication). Since there are many hundreds of journals of relevance to psychiatric disorder, there may well be 10000 or 20000 relevant RCTs that need to be incorporated into a register of RCTs, though once this back log is identified the task of keeping the register up to date will be less

[1] Address for correspondence: Professor Glyn Lewis, University of Wales College of Medicine, Heath Park, Cardiff CF4 4XN.

onerous. The Cochrane Collaboration now includes a review group concerned with schizophrenia and another for depression and neurosis. These will provide a structure that will help to avoid duplication of effort and provide advice and encouragement in identifying relevant literature and conducting systematic reviews.

The arguments to support systematic reviews are based upon the premise that all evidence should be included before coming to conclusions about the effectiveness of an intervention. It might, therefore, seem inconsistent for the Cochrane Collaboration to exclude non-randomized studies from its project. There are also some community wide interventions that are difficult to investigate using randomized designs. However, RCTs are the most robust design for comparing treatments and the majority of non-randomized trials have been conducted without considering the confounding that is inherent in observational studies, or of the appropriate analyses to adjust for confounding. There is also difficulty in deciding upon the methodological criteria for inclusion of non-randomized studies, as these depend upon the satisfactory nature of the adjustment for confounding. For both methodological and practical reasons systematic reviews of RCTs are a reasonable way to assess the evidence for many interventions of interest in psychiatry.

CITATION AND PUBLICATION BIAS

There have been several demonstrations of systematic biases in the submission, publication and citation of studies, particularly for the smaller studies and those with 'negative' results (Gotszche, 1987; Easterbrook *et al.* 1991). This has introduced enough bias to lead to erroneous conclusions. For example, a meta-analysis of smallish trials found that magnesium reduced mortality in acute myocardial infarction but a later mega-trial found no effect. When the results of the individual studies of magnesium were plotted against the number of subjects, (a funnel plot) there seemed to be an under-representation of smallish trials with 'negative' outcomes (Davey Smith & Egger, 1995) so publication bias and perhaps a bias in identification led to this misleading result. This is an example of a single study that had more influence than a meta-analysis. However, it was still important for the mega-trial to be placed in the context of previous research and this indicated why there was an apparent conflict in results.

One of the deficiencies of the psychiatric literature is that RCTs have tended to be too small. We have recently identified 112 reports of 105 RCTs comparing selective serotonin reuptake inhibitors (SSRIs) and the median size was about 30 subjects in each group (Hotopf *et al.* 1996). This could increase the likelihood of publication bias. It has been suggested that publication bias is a problem in the observational studies of the association between obstetric complications and schizophrenia (Geddes & Lawrie, 1995) while there was no evidence of publication bias in the recent reviews of SSRIs (Anderson & Tomenson, 1995).

It is often difficult to identify the relevant literature for a systematic review. Even well conducted literature searches of electronic databases miss many randomized trials. For example the most comprehensive electronic searches of MEDLINE only find about 50% of all RCTs and about 75% of the RCTs listed in that database (Adams *et al.* 1994; Dickersin *et al.* 1994). Hand-searching of all medical journals is the only sure way of identifying all the relevant published material. Reviewers should also consider identifying unpublished results by writing to investigators and pursuing preliminary reports of trials.

HETEROGENEITY: DIVERSE INTERVENTIONS ON DIVERSE GROUPS OF PATIENTS

A meta-analysis produces a summary estimate, a weighted average of the studies, where the weights attach more importance to the larger studies. The simplest procedure, the so-called fixed-effects model, assumes that there is a constant effect size across the studies, and any additional variation results from sampling error. There are many occasions, however, when random error cannot account for the differences between the studies and this will lead to a statistically significant test for heterogeneity (Thompson & Pocock, 1991; Thompson, 1994). The desire to have some overall

average summary has led to the development of methods that allow some variation in effect sizes between studies and these tend to be called 'random-effects models'. These paradoxically give more weight to smaller studies and tend to increase the width of the confidence intervals. However, it seems inconsistent to conclude that there are differences between studies and then produce an average that fails to reflect these differences. It would seem more appropriate to explore and explain the heterogeneity when it is found (Thompson, 1994). The statistical methods for this are at times complex and beyond the scope of this article.

Clinicians are often interested in heterogeneity of this sort as they will want to know if there are particular groups of patients who benefit from a treatment. Many of the interventions in psychiatry are psychotherapeutic in nature, and differences between centres, therapists and therapies may also be of some clinical importance. There are obvious difficulties in standardizing psychotherapeutic interventions both within and between trials and it is more difficult to decide which psychotherapies should be classified together than for pharmacological interventions. This will make the task of investigating heterogeneity in meta-analyses that much more difficult.

There are also the inevitable diagnostic disputes that have plagued psychiatry. Nearly three quarters of the subjects in the cross-national panic disorder trial (Cross-National Collaborative Panic Study, 1992) had phobic symptoms. Those patients would have been entered into a phobia trial in this country, along with those who had phobias without panic. Looking into the more distant past, the criteria for depressive disorder have shown considerable variation over the years and agreed standardized definitions came long after the development of the RCT. There is also the well documented finding that US psychiatrists had a much wider definition of schizophrenia before the advent of DSM-III than was used afterwards (Cooper *et al*. 1972). In contrast, one could argue that the current psychiatric classification is too restrictive, particularly when considering that most neurotic disorders are treated within primary care where the use of pharmacological agents and diagnostic criteria are more fluid (Tyrer *et al*. 1988; Goldberg & Huxley, 1992).When combining trials one must, therefore, be aware of diagnostic differences and it becomes more difficult to investigate whether treatment effects differ according to our current diagnostic criteria if data from the past are used.

POOR METHODOLOGICAL QUALITY OF RCTs

One of the concerns of meta-analysis is that bad studies are grouped together with the good, and both are given equal weight. There are some ways of overcoming this problem including the use of rating scales for methodological quality, which include an assessment of the description of the randomization procedure (Chalmers *et al*. 1981). However, poorly reported studies may also be poor studies, but this is not always the case. Excluding too many studies on the grounds of supposed methodological inadequacy might also lead to a biased sample of studies to review. However, meta-analysis does allow one to test hypotheses about the data, including those about the quality of the studies included. For example, Schulz and colleagues found an association between poor methodological quality and larger treatment effects in their meta-analysis (Schulz *et al*. 1995).

One of the prevalent problems in psychiatry is that studies have not dealt with missing data in a way which maintains the balance of groups at randomization. The commonest way of dealing with this, the so called 'intention to treat' analysis, substitutes previous values for those that are missing. Very few of the older trials in psychiatry (and relatively few of the more modern ones) adopt this procedure, and this reduces the advantage of randomization as the non-random drop out of subjects will allow confounding to be re-introduced. Combining data on RCTs that are incorrectly analysed will be just as misleading as the results of the original trials.

ASSESSMENT OF OUTCOME

Much of the existing work in medical meta-analysis has used either binary outcomes (for example, dead or alive), e.g. (Chalmers *et al*. 1989) or when the outcome is continuous, has used biological meaningful values such as blood pressure or sodium intake (Law *et al*. 1991). Outcomes in

psychiatric research are usually measured with ordinal scales, such as the Hamilton Rating Scale (Hamilton, 1960). A further problem occurs when studies use different scales, though one can calculate the size of effect in terms of the standard deviation of the particular scale used in that study. Deciding upon the standard deviation to use also has its problems and the standard of reporting in many studies in psychiatry is poor. For example, Song *et al*. (1993) found that only 20 of the 58 studies reported both the mean and standard deviation of the Hamilton score and so could be used for their meta-analysis of comparative efficacy between SSRIs and tricyclics.

There are also various ways of analysing continuous data for clinical trials and the lack of a consistent approach makes it difficult to provide a quantitative synthesis. It will, therefore, prove difficult to provide a quantitative estimate for many of the systematic reviews of trials that will be conducted within psychiatry. Much useful information will emerge from qualitative reviews but we should be realistic about the additional information that will emerge from a quantitative approach.

LACK OF ECONOMIC AND QUALITY OF LIFE ASSESSMENT

There is an increasing understanding that decisions about which treatment to use require information about costs in addition to information about clinical effectiveness (Drummond *et al*. 1987). Increasingly, the controversies about new treatments are also concerned with the acceptability of different treatments and the quality of life of patients while on the treatment. However, very few of the existing RCTs in psychiatry will have made either economic or quality of life assessments. Though it is possible to attempt to model the economic effects of treatment using data from outside an RCT, this methodology is crude and cannot help to draw conclusions with any confidence. In order to perform an assessment of cost-effectiveness there is no substitute for performing a randomized clinical trial that includes an assessment of the use of services in the experimental groups. Likewise, the controversy about assessing the relative acceptability of SSRIs and tricyclics has concentrated on the number of drop-outs from clinical trials. This is a poor measure of acceptability of treatment and directly measuring the social functioning and quality of life of subjects in a new clinical trial would be more likely to settle this controversy than analysing the existing data (Simon *et al*. 1996).

If systematic reviews are to be of value it will also be important to disseminate their findings and encourage clinicians to use objective, 'randomized' evidence in coming to decisions about treatment. This is no simple matter and even providing accessible information in both written and electronic forms does not guarantee that the information will be used by clinicians (Paterson-Brown *et al*. 1993).

Finally, it is important to remember to apply our critical faculties towards the systematic reviews themselves, as these can also be of poor quality. Assessing quality depends upon asking questions about the clarity of the aims, the comprehensiveness of the search procedure and the assessment of the scientific validity of the studies. We should also ask whether the results depend upon different methods of analysis or inclusion criteria and pay attention to the authors' interpretation of their findings (Oxman, 1995). The scientific approach towards summarizing evidence is an iterative process that requires criticism and comment. The Cochrane Collaboration, for example, will use electronic forms of publication to aid this process and enable the continuous updating of reviews.

CONCLUSION

There is a wealth of data in the psychiatric literature that have never been systematically drawn together. Like any research endeavour, one can never be sure whether systematic reviews will provide any new and interesting information but it seems sensible to use the data already collected before performing new studies. The poor quality of the literature in psychiatry, the use of a variety of continuous scales to measure outcome and the inconsistent methods of analysis will make quantitative summaries of systematic reviews quite difficult if not impossible to perform. Some

systematic reviews may conclude that new, larger, randomized trials of sound methodology need to be performed, but unless systematic reviews are carried out we will never be sure whether the literature that already exists provides some of the answers.

G. LEWIS, R. CHURCHILL AND M. HOTOPF

REFERENCES

Adams, C. E., Power, A., Frederick, K. & Lefebvre, C. (1994). An investigation of the of the adequacy of MEDLINE searches for trials of the effects of mental health care. *Psychological Medicine* **24**, 741–748.

Anderson, I. M. & Tomenson, B. M. (1995). Treatment discontinuation with selective serotonin reuptake inhibitors compared with tricyclic antidepressants: a meta-analysis. *British Medical Journal* **310**, 1433–1438.

Chalmers, I., Enkin, M. & Keirse, M. J. N. C. (eds.)(1989). *Effective Care in Pregnancy and Childbirth*. Oxford University Press: Oxford.

Chalmers, I., Dickersin, K. & Chalmers, T. C. (1992). Getting to grips with Archie Cochrane's agenda. *British Medical Journal* **305**, 786–788.

Chalmers, T. C., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D. & Ambroz, A. (1981). A method for assessing the quality of a randomized controlled trial. *Controlled Clinical Trials* **2**, 31–49.

Cooper, J. E., Kendell, R. E., Gurland, B. J., Sharpe, L., Copeland, J. R. M. & Simon, R. (1972). *Psychiatric Diagnosis in New York and London: A Comparative Study of Mental Hospital Admissions*. Oxford University Press: London.

Cross-National Collaborative Panic Study (1992). Drug treatment of panic disorder: comparative efficacy of alprazolam, impramine and placebo. *British Journal of Psychiatry* **160**, 191–202.

Davey Smith, G. & Egger, M. (1995). Misleading meta-analysis. *British Medical Journal* **310**, 752–754.

Dickersin, K., Scherer, R. & Lefebvre, C. (1994). Identifying relevant studies for systematic reviews. *British Medical Journal* **309**, 1286–1291.

Drummond, M. F., Stoddard, G. L. & Torrance, G. W. (1987). *Methods for the Economic Evaluation of Health Care*. Oxford University Press: Oxford.

Easterbrook, P. J., Berlin, J. A., Gopalan, R. & Matthews, D. (1991). Publication bias in clinical research. *Lancet* **337**, 867–872.

Geddes, J. R. & Lawrie, S. M. (1995). Obstetric complications and schizophrenia: a meta-analysis. *British Journal of Psychiatry* **167**, 786–793.

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher* **5**, 3–8.

Goldberg, D. & Huxley, P. (1992). *Common Mental Disorders: A Biopsychosocial Approach*. Routledge: London.

Gotszche, P. C. (1987). Reference bias in reports of drug trials. *British Medical Journal* **295**, 654–656.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry* **23**, 56–62.

Hill, A. B. (1952). The clinical trial. *New England Medical Journal* **247**, 113–119.

Hotopf, M., Lewis, G. & Normand, C. (1996). Are SSRIs a cost-effective alternative to tricyclics? *British Journal of Psychiatry* **168**, 404–409.

Law, M. R., Frost, C. D. & Wald, N. J. (1991). By how much does dietary salt reduction lower blood pressure? I–Analysis of observational data among populations. *British Medical Journal* **302**, 811–815.

Oxman, A. D. (1995). Checklists for review articles. In *Systematic Reviews* (ed. I. Chalmers and D. G. Altman and A. D. Oxman), pp. 75–85. BMJ: London.

Paterson-Brown, S., Fisk, N. M. & Wyatt, J. C. (1993). Are clinicians interested in up to date reviews of effective care? *British Medical Journal* **307**, 1464.

Schulz, K. F., Chalmers, I., Hayes, R. J. & Altman, D. G. (1995). Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association* **273**, 408–412.

Simon, G. E., von Korff, M., Helligenstein, J. H., Revicki, D. A., Grothaus, L., Katon, W. & Wagner, E. H. (1996). Initial antidepressant choice in primary care: effectiveness and cost of fluoxetine vs tricyclic antidepressants. *Journal of the American Medical Association* **275**, 1897–1902.

Song, F., Freemantle, N., Sheldon, T. A., House, A., Watson, P., Long, A. & Mason, J. (1993). Selective serotonin reuptake inhibitors: meta-analysis of efficacy and acceptability *British Medical Journal* **306**, 683–687.

Thompson, S. G. (1994). Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal* **309**, 1351–1355.

Thompson, S. G. & Pocock, S. J. (1991). Can meta-analyses be trusted? *Lancet* **338**, 1127–1130.

Tyrer, P., Seivewright, N., Murphy, S., Ferguson, B., Kingdon, D., Barczak, P., Brothwell, J., Darling, C., Gregory, S. & Johnson, A. (1988). The Nottingham study of neurotic disorder: comparison of drug and psychological treatments. *Lancet* **2**, 235–240.

Yusuf, S., Collins, R. & Peto, R. (1984). Why do we need some large, simple randomised trials. *Statistics in Medicine* **3**, 409–420.